计算机科学
COMPUTER SCIENCE

大语言模型与谣言:生成与检测的综述

潘杰, 王娟, 王楠

引用本文

潘杰, 王娟, 王楠. 大语言模型与谣言:生成与检测的综述[J]. 计算机科学, 2025, 52(11): 1-12.

PAN Jie, WANG Juan, WANG Nan. Large Language Models and Rumors:A Survey on Generation and Detection [J]. Computer Science, 2025, 52(11): 1-12.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.250700034

大语言模型与谣言:生成与检测的综述

潘 杰 王 娟 王 楠
中国人民警察大学智慧警务与大数据技术研究中心 河北 廊坊 065000
(2023903005@cppu.edu.cn)

摘 要 谣言检测自20世纪中期起便是跨学科研究议题,微博、Twitter等社交媒体的迅速普及让该任务受到持续关注,并在2016年美国总统选举期间因谣言泛滥而进入更广泛的公共视野。随着大语言模型的发展,其在自然语言理解与生成方面实现了突破,并推动了谣言检测领域发生深刻变革。文中系统综述了大语言模型在谣言生成与检测领域的最新研究:首先回顾社交媒体谣言的概念,概述了当前用于谣言检测的多种基准数据集以及传统机器学习、深度学习与图神经网络等检测框架的演进历程;继而重点分析大语言模型(Large Language Models,LLMs)在谣言检测中的4类核心角色,即参数微调、零/少样本提示、知识增强、多模态融合,梳理了由LLM生成谣言的数据集,以及水印、语言指纹、语义熵等针对AI生成内容的检测技术;最后展望了未来研究的方向以及面临的挑战。

关键词:谣言;大语言模型(LLMs);深度学习;检测技术
中图分类号 TP311

Large Language Models and Rumors:A Survey on Generation and Detection

PAN Jie,WANG Juan and WANG Nan
Smart Policing and Big Data Technology Research Center,China People's Police University,Langfang,Hebei 065000,China

Abstract Rumor detection has been an interdisciplinary research topic since the mid-20th century. The rapid rise of social-media platforms such as Weibo and Twitter has kept the task in the spotlight,and the surge of rumors during the 2016 U.S. presidential election brought it to wider public attention. Breakthroughs in LLMs have dramatically advanced natural-language understanding and generation,catalyzing profound changes in the field of rumor detection. This paper presents a systematic survey of the latest studies on rumor generation and detection in the LLM era. It firstly revisits the concept of social-media rumors and summarizes widely used benchmark datasets,tracing the evolution of detection frameworks from traditional machine learning to deep learning and graph neural networks. It then analyzes in depth the four core roles that LLMs play in rumor detection,parameter fine-tuning,zero/few-shot prompting,knowledge augmentation and multimodal fusion. In addition,it catalogs datasets containing LLM-generated rumors and examines emerging detection techniques for AI-generated content,such as watermarking,linguistic fingerprints,and semantic-entropy-based methods. This paper concludes by outlining future research directions and the key challenges that remain.

Keywords Rumor,Large Language Models(LLMs),Deep learning,Detection techniques

1 引言

当前,随着互联网与社交媒体的迅猛发展,信息传播的速度与范围达到了前所未有的水平。微信、微博、Twitter和Facebook等日渐成为人类获取信息、传递看法、开展交流的核心手段[1]。然而,网络环境也给谣言的产生与传播提供了便利[2-3]。社交媒体中的谣言具有相对的煽动性、扩散速度更快、受众范围更广等鲜明特征,不仅会产生误导性效应,损害人们对真相的正确把握,还会引发民众心理与行为的混乱[4],甚至对经济、政治和国家安全造成一定的损害[5],尤其在公共卫生事件(如COVID-19大流行[2,6-7])、自然灾害或社会动荡时期,谣言的负面影响更为凸显[8]。如何识别与制止谣言在社交媒体中扩散的问题,也逐渐成为学术界、工业界乃至全球各国政府高度关注的问题[9-10]。

随着深度学习、自然语言处理(Natural Language Processing,NLP)等技术的发展,人工智能为自动化检测谣言提供了技术支持[1,10]。传统的谣言检测方法大多基于人工特征工程和传统机器学习模型[8],而海量、异构、动态的社交网络数据处理给这种方法带来了极大的困难[11]。近年来,深度学习模型,如CNN、RNN及其衍生模型LSTM(Long Short-

到稿日期:2025-07-07 返修日期:2025-08-29
基金项目:河北省社会科学基金(HB22SH011)
This work was supported by the Social Science Foundation of Hebei(HB22SH011).
通信作者:王娟(wangjuan@cppu.edu.cn)

Term Memory)、GRU(Gated Recurrent Unit)和 GNN(Graph Neural Network)等,因其强大的学习特征能力和建模能力在谣言检测任务上取得了重大突破[1,7,11]。这些模型可以从文本、用户以及传播结构等多个方面发现谣言的细微特征[9,12]。

值得关注的是大语言模型,例如 BERT(Bidirectional Encoder Representations from Transformers)[13]及其后继者 GPT(Generative Pre-trained Transformer)等,通过在大规模文本数据集上的预训练获得了丰富的语言知识和推理表达能力,这为谣言检测带来了新的机遇与挑战。一方面,借助 LLMs 优秀的表示学习能力和丰富的知识储备,研究者开始尝试利用 LLMs 检测谣言的方法,以提高谣言检测的准确性和鲁棒性[5,13-14]。有学者利用 LLMs 捕捉传统方法未能观察到的、隐藏在上下文中的谣言线索,如不一致、荒谬的推理、错误的情感分析等[15]。另一方面,利用 LLMs 自身强大的文本生成能力,可以通过一定手段生成和发布更多难以识别、具有迷惑性的虚假新闻和谣言[16-17]。与传统由人工撰写并传播的谣言相比,其文本语义高度连贯、内容高度逼真且缺乏明显风格特征,其来源既包括模型幻觉也包括恶意指令驱动的内容生成。同时,低成本生成与多账号投放,使谣言借助平台推荐迅速形成"信息级联",扩散速度和规模明显放大。实证研究表明[18],在语义一致的对照下,人类仅能识别约 10%的幻觉新闻,Llama 等主流检测器的识别准确率也普遍下降 10%～30%,显示出更高的隐蔽性和检测难度。因此,识别由大语言模型生成的谣言,成为了谣言研究的新方向。

因此,本文系统性地梳理和综述了基于大语言模型的社交媒体谣言检测相关技术及发展历程,提出了两个核心问题。第一,如何运用 LLMs 提高对于社交媒体中的各种谣言信息的甄别能力。第二,如何甄别出 LLMs 所产生的信息,探究其中的技术难点与现有解决方法。同时,对于现有研究中运用到的谣言数据集进行梳理和总结[3,9,19],为后续研究提供借鉴。最后,总结现有研究的挑战和未来发展趋势,从而推动这一领域研究不断向更高、更深发展的目标努力。

2 理论基础与相关工作

2.1 谣言的定义与特征

目前,学术界尚未对谣言形成统一明确的定义,有研究将谣言定义为被视为虚假的信息,而另一些研究则强调其未经证实的特性[20],还有研究将谣言视为一种非正式的信息噪声,具有生命周期特征:在传播过程中迅速出现、短暂存在,最终如其产生般迅速消退[21]。心理学家则将谣言定义为一种与新闻相关的陈述,这类陈述通常具有表面可信性并被广泛传播,但并未经过验证[22]。表 1 列出了不同权威机构对谣言的阐述。

表 1 不同来源对谣言的定义

Table 1 Definition of rumor from different sources

来源	定义
柯林斯词典	谣言指未经官方证实的,混合真相与谎言并通过口头流传的信息。其核心特征包括:非官方性、传播性、不确定性
剑桥词典	一种未经官方确认的有趣故事或消息,可能是真实的,也可能是虚构的,并且迅速在人与人之间传播
现代汉语词典	作为中国最具影响力的现代汉语词典,以规范现代汉语为目标,其对谣言的定义为没有事实根据的消息
维基百科	谣言可能是针对公众所关心的事物,所提出的一种未经证实的解释或理由。进一步来说,谣言牵涉到的是未经可靠来源证实的消息,换言之,谣言是一种人与人之间,口耳相传,但缺乏可靠证据支持的陈述或信念[23]
信息科学	谣言是未经权威验证的信息单元,通过非正式人际网络或数字平台快速扩散,其内容通常涉及公众关注的事件(如公共安全、政治争议),并在传播过程中因群体认知偏差产生语义变异
牛津词典	世界上最权威的英语词典之一。其中"rumor"(谣言)相关解释为"a piece of information, or a story, that people talk about, but that may not be true.",即一条人们谈论但可能不实的信息或故事
韦氏大词典	美国具有较高权威的词典,对谣言的解释分为两层含义:1)广泛传播且无明显来源的说法或观点;2)在当前没有已知权威来源来证明其真实性的陈述或报告

结合这些描述,可以提炼出谣言的两个核心特征:1)广泛传播性;2)可验证的错误性。因此,本文将社交媒体谣言界定为:在社交媒体平台被广泛传播,并最终通过可靠渠道证实与客观事实不符的信息。

2.2 常用谣言检测数据集

用于社交媒体谣言检测的数据集主要来源于三大社交媒体平台:Twitter、Facebook 和微博。它们因用户基数大、数据开放性强而成为研究的主要来源。其中,超过一半的数据集包含 3 种真实性标签:真实、错误以及未经验证。其余数据集则仅提供两种标签:真实与虚假[24]。

表 2 详细列出了谣言检测研究中常用的数据集,提供了对数据集特征的全面理解。其中,"数据集详细信息"详细说明了数据集的数据量、具体内容,以帮助研究人员了解数据集的全面结构;"内容类型"包括文本、图片、视频等不同类型的数据;"平台"指数据来源平台;最后"语言"指定了数据集内容可用的语言。

表 2 谣言数据集特征比较

Table 2 Comparison of rumor dataset feature

数据集	数据集详细信息	内容类型	平台	语言
PHEME[25]	包含 6425 条关于 9 个重大事件下的推文的数据集,数据来自 Twitter。根据推文对数据集进行分类,并将其验证为真实、虚假或未经验证	文本	Twitter	英文
PHEME Updated Version[26]	它是一个扩展版本,将线程分类为谣言或非谣言,该数据集有 2402 条谣言,其中 1067 条为真,638 条为假,697 条未经验证	文本	Twitter	英文
Twitter15/Twitter16[27]	分别包含了 1490 和 818 条推文,提取高度转发或回复的热门源推文构建了传播树,标注了源推文的真实性,并使用辟谣网站的标签进行标注	文本	Twitter	英文

（续表）

数据集	数据集详细信息	内容类型	平台	语言
FakeNewsNet[28]	从两个事实核查网站 GossipCop 和 PolitiFact 收集，包含新闻内容及由专业记者和专家标注的标签，以及社交环境信息	文本图像	Twitter	英文
Weibo[29]	从 2012 年 5 月到 2016 年 1 月期间，从微博的官方辟谣系统收集了所有已确认的假谣言帖子，共包含 2 万条推文	文本	新浪微博	中文
Weibo21[30]	涵盖政治、娱乐、体育等 9 个领域的 9128 条数据，涉及公众的认知和情感	文本图像	新浪微博	中文
BuzzFeedNews[31]	收集了 2016 年美国大选期间 9 家新闻机构发布的 2282 篇帖子，每个帖子中都由 BuzzFeed 记者进行事实核查，只包含了每篇新闻的标题和文本	文本	Facebook	英文

2.3　基于大语言模型生成的谣言数据集整理

随着大语言模型生成虚假信息能力的提高，学术界已构建多个此类数据集支持谣言检测研究。表 3 详细列出了常用的由大语言模型生成谣言的数据集。其中，LLMFake[18] 通过多种提示策略实现同义改写生成、重写生成与开放式扩写，由 ChatGPT 等模型生成不同类型的谣言文本。CoSMis[32] 则聚焦公共卫生事件相关的谣言，包含等量的 AI 生成与人工撰写的新闻。VLPFN[33] 借助对抗式 VLPrompt 对事实报道进行重写生成，产出高度迷惑性的谣言。在医疗健康领域，MM-Health[34] 和 Med-MMHL[35] 均为多模态医疗谣言数据集，纳入了大语言模型生成的虚假内容，并支持跨疾病的检测场景。此外，MegaFake[36] 基于社会心理理论与大规模新闻语料，采用重写生成与开放式扩写生成多种欺骗性风格的假新闻。

表 3　大语言模型生成谣言数据集的比较

Table 3　Comparison of rumor datasets generated by large language models

数据集	数据集详细信息	内容类型	平台	语言
LLMFake	LLMFake 数据集包含逾 1 万条由大型语言模型生成的虚假或误导性文本，涵盖医疗、科学、政治等八大领域，使用多种虚假信息生成策略（如幻觉生成、随机生成、改写生成、信息操控等）	文本	GPT-3.5-turbo，Llama2(7B/13B/70B) Vicuna(7B/13B/33B)	英文
CoSMis	该混合数据集共包含 2400 条新闻报道，均来自 LLM 与人工撰写，聚焦公共卫生专题，并配对科学摘要，主要用于科学领域误导信息的检测评估	文本	GPT-3.5，LLama2(7B)	英文
VLPFN(VLPrompt fake news)	共计 3174 条文章，涵盖真实文本、人工虚假新闻与 LLM 生成的虚假新闻，用于评估 LLM 在 VLPrompt 类攻击下的生成行为及检测模型的鲁棒性	文本	GPT-3.5-turbo，GPT-4，Vicuna(7B/13B)	英文
MM-Health	共收集 34746 篇多模态新闻文章，其中 5776 条为人工生成、28880 条为 AI 生成，涵盖文本与视觉数据，用于健康领域虚假信息的检测与溯源分析	文本图像	Llama-3.1-8B，Qwen-2.5-7B，ChatGLM-4-9B，Gemma-2-9B	英文
Med-MMHL	专门面向医疗领域谣言检测的数据集，旨在同时覆盖多疾病、多模态场景，首次纳入 LLM 自动生成的虚假信息。相比以往只含文本、且主题单一的数据集，Med-MMHL 在内容广度与研究难度上都做了显著扩展	文本图像	GPT-3.5-turbo	英文
MegaFake	基于 LLM-Fake Theory 理论构建，总计包含 46096 条假新闻和 17871 条真实新闻，构成一个大规模共 63967 条的样本集，并包含 4 种生成假新闻的方法和 2 种生成真实新闻的策略，可用于研究 LLM 生成假新闻的特征及其检测	文本	ChatGLM3-6B，GLM-4	英文

这些数据集在生成式谣言检测研究中具有重要意义，它们为算法提供了标准化的评测基准和数据支撑，为文本、多模态及跨领域检测模型建立了可比性强的对照环境，促进了零样本、少样本与增量学习场景下的公平评测，加快了今后“生成式”内容安全治理在未来的落地。

3　传统谣言检测方法

近年来，面对谣言传播的新态势，谣言检测技术发展迅速。从方法论上看，谣言检测大致经历了从传统机器学习特征工程到深度神经网络的建模，再到大语言模型辅助检测的发展过程。图 1 给出了一个通用谣言检测流程的示意图（采用二分类方式），第一阶段为数据采集与数据集构建，大多数学者通过社交媒体平台、事实核查网站等平台收集数据或者使用基准数据集，并运用预处理技术构建输入数据集。数据采集后，各种词嵌入技术被用于生成特征向量，随后将这些特征向量输入检测模型中进行训练并实现最终预测，即判断输入消息为谣言或非谣言。

早期的谣言检测研究大量依赖于手工设计的特征，Castillo等[37] 开创性地研究了社交媒体消息的可信度评估，使用了包含用户信誉、帖文内容、传播动态等在内的 53 项特征训练分类模型。这些特征通常可以从内容、用户和传播 3 个维度进行提取[9]，如图 2 所示。

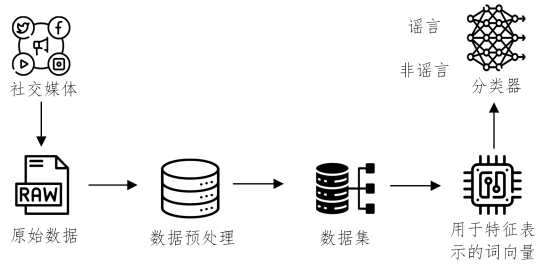

图 1　通用谣言检测流程示意图

Fig. 1　Schematic diagram of the general rumor detection process

内容特征关注信息本身的文本和视觉元素，例如文本的语言风格（如词汇使用、句子结构、情感倾向[15]）、信息熵、是否存在疑问或否定词语、文本中包含的实体或 URL 等。Zoleikha等[38] 借鉴 Allport 和 Postman 的理论，提出了基于内容的重要性和模糊性两大类，共 42 个特征来计算谣言的传播

力，并发现虚假谣言和真实谣言在传播力上存在显著差异，这表明内容特征对于区分谣言类型具有指示作用。用户特征描述的是传播用户自身属性，如用户的注册时间、用户粉丝数、关注数、历史发帖行为、用户是否认证以及社交网络中的关系和影响力等[4]，这有助于分析信息的来源可靠性；传播特征是描述信息在社交网络中传播的模式，包括转发/评论的数量、速度、深度、广度、参与转发/评论的群体用户特征、信息传播形成的级联结构或网络拓扑结构[11-12]。Ma 等[12]指出在谣言生命周期中捕获社交上下文特征的变化十分重要，他们构建的基于时序的方法比只使用静态特征的方法有效得多，同时在谣言早期检测方面也表现出很大的潜力[39]。

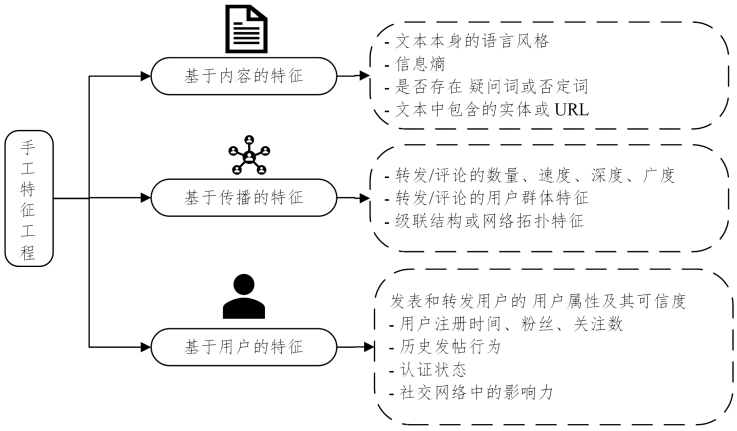

图 2 早期谣言检测特征提取示意图

Fig. 2 Schematic diagram of feature extraction for early rumor detection

3.1 基于机器学习的方法

在提取了手工特征之后，研究者通常采用各种传统的机器学习算法来构建谣言分类器。常用的算法包括逻辑回归(Logistic Regression)、支持向量机(SVM)、决策树(Decision Tree)、随机森林(Random Forest)、梯度提升机(Gradient Boosting)等。Taha 等[8]的研究比较了多种机器学习算法在基于 TF-IDF(Term Frequency-Inverse Document Frequency)文本特征的假新闻检测任务上的表现，结果显示决策树和梯度提升算法取得了接近 99.5%的准确率，略优于逻辑回归(98.5%)和随机森林(98.9%)。这表明在特征表示得当的情况下，传统机器学习模型也能达到较高的检测精度。然而，传统机器学习方法严重依赖输入特征的质量，一旦手工提取的特征未能抓住谣言的关键差异，即使分类器再强也难以奏效。此外，这些方法往往将特征视为相互独立，难以捕捉复杂的交互行为和深层语义。随着数据量和特征维度的增加，其训练效率和效果可能下降，因此从 2016 年后研究重心向自动学习特征的深度学习方法转移。

3.2 基于深度学习的方法

深度学习技术的兴起为谣言检测带来了突破。与传统方法不同，深度学习模型能够端到端地从原始数据中自动学习层次化的特征表示，避免了繁琐且主观性强的手工特征工程[1]。近年来，多种深度学习架构被应用于谣言检测任务，并取得了优于传统方法的性能。

3.2.1 基于循环神经网络(RNN)及其变体的方法

考虑到社交媒体帖文及其传播过程天然具有时序性，循环神经网络及其变体(如 LSTM 和 GRU)被广泛应用于谣言检测。Ma 等[12]的早期工作虽未显式使用 RNN，但通过时间序列模型揭示了时序特征在谣言检测中的重要性。Al-Sarem 等[7]提出了一个融合 LSTM 与并行 CNN 的混合模型(LSTM-PCNN)，用于社交媒体中公共卫生相关的谣言检测，在 ArCOV-19 数据集上获得了很好的效果。LSTM 可以有效地建模文本序列中的长范围依赖关系，这对于理解上下文以及判断信息的一致性非常重要。Rani 等[40]利用双向的 LSTM(Bi-LSTM)模型持续监测社交媒体信息流以识别新出现的谣言，并取得了高达 99.63%的准确率，从而实现早期预警。

RNN 及其变体在处理文本序列和传播时序方面具有优势，能够捕捉动态演化特征。但其也存在一些潜在问题，例如对于非常长的序列可能出现梯度消失或爆炸问题，模型训练难以并行化，计算开销相对较大，以及单纯依赖 RNN 无法充分捕捉文本中的局部关键信息。

3.2.2 基于卷积神经网络(CNN)的方法

CNN 最初应用于计算机视觉，后来被广泛引入自然语言处理任务，如谣言检测。该模型通过在文本序列上滑动不同窗口大小的卷积核，可学习到局部的 n-gram 特征，挖掘文本中可能出现的煽动词、矛盾语句或规律句式，从而为谣言识别提供关键信息。Huang 等[13]提出了一种结合 BERT 和 CNN 的方法(BERT-CNN)，利用 BERT 强大的上下文理解能力生成词嵌入，再使用 CNN 提取局部特征，用于最终的谣言分类。他们在中国互联网用户数据集上进行实验，结果表明，该方法相比传统方法在准确率、召回率和 F1 分数等指标上均有提升。Khan 等[41]的 HCovBi-Caps 模型也包含了卷积层，用于在 Bi-GRU 和 Capsule 网络之前提取初始特征。

CNN 擅长捕捉局部模式，并且可以并行化计算，效率较高。其缺点在于卷积核的大小限制了其感知范围，对于长距离依赖关系的捕捉不如 RNN 直接。另外，CNN 本身不直接考虑词序信息，通常需要结合位置编码或其他序列模型来使用。

3.2.3 基于图神经网络(GNN)的方法

谣言在社交媒体上的传播过程天然形成一种图结构，其

中帖子、用户可以被视为节点，转发、评论、关注等关系可以被视为边。图神经网络可以直接学习和利用图数据上的拓扑结构以及节点之间的相互联系，在很大程度上有助于谣言检测模型对信息传播的建模。Chen 等[42]提出了一种基于时序感知的异质图神经网络谣言检测模型 SHGN。该模型将事件与用户构建为异质图，显式融合事件内响应的时序依赖与事件间由共同用户诱导的全局结构。模型以位置编码与自注意力获取响应序列，并与源贴经图注意力聚合形成局部表征，再以元素级注意力学习全局表征，拼接后进行谣言识别，使得模型具有良好的谣言检测能力。Jiang 等[43]注意到基于 GNN 的现有谣言检测模型中，对于传播树的深度变化敏感：浅层树提供的信息不够，而深层树则容易包含噪声。因此他们把流行病传播模型中的知识融入到 GNN 框架中，用于增强模型针对不同深度传播树的鲁棒性。实验证明，EIN 在多个数据集中和不同传播树深度下具有良好的性能和鲁棒性。

GNN 聚合邻居节点的信息进行学习，从而能够建模节点中的依赖关系，获得更丰富的节点表征，但其也面临着挑战。

1）传播图的构建需要完整的交互数据，在现实情况中很难收集到完整的传播数据，数据集的质量也不高。

2）社交网络的图往往很大且动态变化，对 GNN 的计算效率和可扩展性要求高。

3）Jiang 等[43]指出，GNN 模型对图的结构可能敏感，基于 GNN 的谣言检测模型也可能受到精心设计的恶意攻击（注入恶意的谣言消息改变传播图），影响其性能[17]。

3.2.4　注意力机制和多模态融合的方法

注意力机制能够使模型动态关注更为重要的信息，在谣言检测任务中具有重要作用。例如在处理一个贴中包含的多项评论时，并非所有评论都同等重要，其中立场鲜明或包含关键信息的评论对谣言真伪判断更为重要[44-45]。Ge 等[15]构建了一种基于双重情感感知的可解释性谣言检测模型，通过协同注意力机制分别学习了谣言语义与用户评论情感、谣言情感与用户评论情感之间的关系，该方法不仅提升了检测准确率，还能从注意力权重的角度给出基于情感方面的解释。Tao 等[46]提出的方法是在前后期融合中都引入注意力机制，实现特征与决策的自适应加权，并且利用注意力机制对词和视觉信息进行双路融合，赋予对谣言检测贡献更大的词和视觉神经元更大的权重。Yang 等[47]在其弱监督联合学习架构中，提出了一种层次化注意力机制，把“真实性－立场”的多类任务拆成多组二元子模型，树注意力层在传播树上将帖子级立场证据聚合成二元真实性，再用判别性注意力在多组二元分类器输出间加权，得到各类别的概率分布，并以其中概率最大的类别作为预测值。

随着社交媒体内容的多元化，谣言常常以图文并茂或短视频的形式流传[14]。单模态信息不足以判别真假，例如真实图片上叠加虚假的文字描述，因此多模态谣言检测成为研究热点。Zhou[4]针对多模态早期谣言检测存在的预训练模型和目标领域差异性，提出了基于领域自适应的多模态方法，对文本和视觉特征分别进行领域自适应优化，基于特征和决策级进行特征融合。Jing 等[48]提出了 TRANSFAKE 模型——基于 Transformer 的多任务学习框架，能够联合建模新闻主体内容和用户评论，并处理多模态输入。该模型通过谣言得分预测和事件分类两个辅助任务来提取跨模态的隐藏关系，在两个真实数据集上的检测精度和 F1 值均提升了 10%以上。

基于注意力机制，模型能更好地提升对关键信息的关注，提高模型的可解释性，而多模态方法能更好地利用多源数据，做出更可靠的判断[49]。然而，多模态方法也面临挑战，如模态间的对齐问题、特征融合策略的选择以及处理多模态数据带来的更高计算复杂度。特别是对于视频等多媒体内容，其信息密度高、结构复杂，给特征提取和融合带来了更大难度[14]。

3.2.5　深度学习方法在谣言检测中的总结

为系统对比各类深度学习框架在谣言检测任务中的适用性，在之前讨论 RNN，CNN，GNN 与 Transformer 的基础上，进一步凝练它们在输入类型、建模对象及性能表现上的共性与差异。表 4 列出了 4 类主流方法在 4 个维度上的对比结果。

表 4　深度学习方法在谣言检测中的共性和差异

Table 4　Commonalities and differences of deep learning methods in rumor detection

模型框架	输入类型	建模对象	优势	局限
RNN	文本序列	谣言内容及用户互动的时间演化	可按时间增量更新隐藏状态，能够实时或早期检测	训练不可并行；长序列梯度消失；对早期信息稀缺敏感
CNN	文本内容	谣言文本的局部信号	易于捕获关键信息和局部特征	长距离建模能力有限；未显式利用传播结构与时序信息
GNN	社交传播图	谣言传播结构	利用拓扑信息提升准确率；可同时编码文本信息实现多源信息融合	对图完整性要求高，大规模动态图训练开销大
Transformer（自注意力）	文本序列	谣言文本的全局语义与上下文关系	避免 RNN 的长距梯度衰减问题；注意力分数提供一定的可解释性；所有位置可并行计算	参数规模大、训练成本高

4　大语言模型在谣言检测中的应用

4.1　Transformer 架构

Transformer 模型是由 Vaswani 等[50]于 2017 年提出的一种革命性模型架构，已成为自然语言处理的基石。该架构采用了最初为机器翻译任务而设计的编码器-解码器结构，其中编码器负责处理输入序列，解码器则生成输出序列。两个组件均由堆叠的相同层构成，这些层集成了注意力机制与前馈网络，其结构如图 3 所示。这种全新的架构不仅显著提升了模型并行计算的效率，还极大地改善了语义信息的捕捉能力。

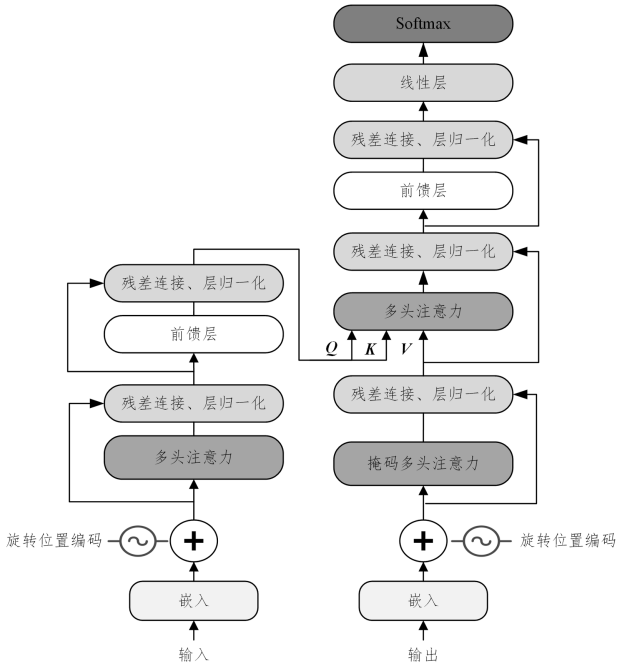

图 3 Transformer 模型架构

Fig. 3 Architecture of Transformer model

Transformer 的核心创新是提出自注意力机制，使得模型可以一次考虑输入序列的不同部分。相较于传统模型采取的顺序处理，Transformer 可以并行处理序列中的所有位置，从而更方便寻找句子中词语之间的依赖和关联，避免了传统顺序建模带来的长依赖建模困难。自注意力机制的运作过程如下：首先，将输入序列转为词嵌入以获取每一个词的语义表征，这些嵌入通过训练得到的权重矩阵分别转为查询向量(Query)、键向量(Key)和值向量(Value)；然后，对序列中每个位置的查询向量与所有键向量计算相似度分数，再经 softmax 函数归一化后得到注意力权重，这些权重作为系数对各位置的值向量进行线性组合，从而获得每一个词的最终表达。其计算过程可形式化为：

$$Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{1}$$

其中，d_k 为键向量的维度。通过允许序列中每个位置关注所有其他位置，自注意力机制使得模型能够捕捉长距离依赖，并生成具有上下文信息的表示，从而增强对词语关系和上下文的理解能力。

4.2 基于预训练模型的方法

Transformer 模型的成功为大规模预训练模型(如 BERT、GPT 系列)奠定了坚实的基础。通过在海量语料上进行预训练，这些模型不仅学习了语言的句法与语义规律，还隐式积累了丰富的世界知识。而预训练模型通过微调，在各种文本分类、问答及谣言检测等任务中取得了出色的成绩。

早期 LLM 的使用就是将 BERT 等预训练模型作为编码器来获取高维文本表示。在大量的未标注语料中进行预训练(如 MLM(Masked Language Model)任务)，预训练模型能够学习大量的句法、语义，甚至是世界知识。将训练好的模型在下游的谣言检测任务上进行微调，通常能取得比从零开始训练的模型或基于传统词嵌入的模型更好的效果。Jin 等[51]在其少样本假新闻检测框架 DetectYSF 中，使用基于 Transformer 的 PLM(Pretrained Language Models)作为骨干网络，并采用基于 MLM 的伪提示学习范式进行模型调优(prompt-tuning)。这表明 PLM 已成为当前谣言检测研究中获取文本表示的主流方法之一。

4.3 基于大语言模型的方法

基于 PLM 的微调方法能够有效捕捉文本的深层语义特征，在谣言检测任务中表现出了显著的性能优势。近年来，随着模型规模的进一步增大和能力的提升，GPT 系列、LLama 系列等大语言模型迅速崛起，研究者逐渐将注意力从传统语言模型转向了大语言模型的方法。如图 4 所示，大语言模型在谣言检测流程中可以承担不同角色，主要包括数据增强、特征生成、知识增强、提示学习、参数微调、多模态融合和可解释性输出等，几乎覆盖了谣言检测从数据预处理、模型训练到结果解释的整个环节，为提升谣言检测性能提供了多种途径。

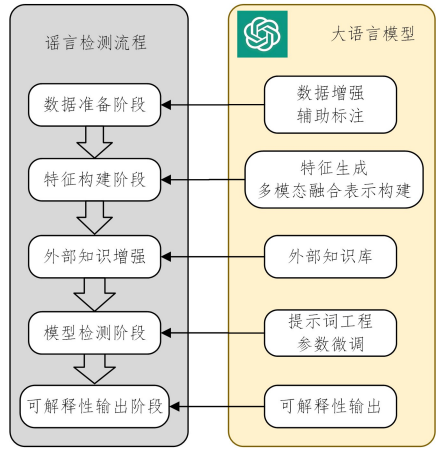

图 4 LLM 赋能谣言检测全流程的职能示意图

Fig. 4 Illustration of LLM-enabled functional roles across the full rumor-detection pipeline

4.3.1 基于参数微调的方法

对大语言模型的微调方法取得了快速发展，并逐渐成为提升模型性能的重要途径。其中，全参数微调[52](Full Fine-tuning)尽管效果显著，但往往对计算资源要求极高。为降低微调成本并提高效率，研究者提出了参数高效微调(Parameter-Efficient Fine-tuning)方法，如前缀微调[53](Prefix-Tuning)和低秩适配[54](LoRA)等策略。这些方法通过冻结原模型的大部分参数，只需训练少量的附加或适配参数，即可实现接近全参数微调的效果。Pavlyshenko[55]尝试对 Llama2 进行 LoRA 微调，使模型能够分析文本中的宣传性叙事、进行事实核查和假新闻判别，微调后模型可按预设 JSON 格式输出链式推理结果，既便于专家质检，也能直接作为下游预测特征。实验结果显示，微调后的 Llama2 具备生成结构化深度分析的能力，扩大数据与模型规模并结合 RLHF 可进一步提升精度，而 LoRA 显著降低了计算与存储开销。Liu 等[56]围绕金融领域的虚假信息检测展开，首次构建了金融虚假信息检测的指令微调数据集，并将 Llama 微调为“FMDLlama3”模型，FMDLlama3 在 FMD-B 上的分类准确率和解释质量均超越了所有开源模型及闭源 ChatGPT。与仅针对单一任务微

调的 BERT，RoBERTa 相比，FMDLlama3 在复杂长文本场景下优势更明显。

4.3.2　零样本/少样本提示(Prompting)

LLMs 有着强大的零样本(zero-shot)和少样本(few-shot)学习能力。通过精心设计提示(prompt)，可以直接引导 LLM 对给定信息进行真伪判断，而无需或只需极少量的标注样本进行微调。Lin 等[57]提出基于 RPL 框架的零样本谣言检测，首次将结构信息融入零样本提示检测框架。他们先将响应帖按照“时间＋树搜索”的多策略响应排序，构造多条传播线程增强上下文，将谣言传播线程编码进 prompt。在 Weibo，Twitter 和 CatAr-COVID19 的完全零样本场景下，RPL-Bre 取得了最高 74.5% 的准确率和 71.9% 的 F1 值，与多种微调、适配器和提示基线相比，均有显著提升。Xu 等[5]提出了一种综合性的辟谣流程，其中就利用提示工程技术将检索到的相关知识和待检测信息输入 LLM，实现了有效的谣言识别，同时避免了微调带来的计算成本。这种方法对于处理标注数据稀缺的新发事件谣言(如文献[6]中提到公共卫生事件的谣言)具有重要意义。而 Pendyala 等[58]假设 LLM 内部丰富的预训练知识可用于虚假信息检测，以 Llama2，Orca，Falcon 和 Mistral 为代表，检验零样本和少样本提示下 LLM 直接判别谣言的可行性，并辅以 LIME，SHAP，Integrated Gradients 等可解释 AI 方法剖析模型关注的关键词。实验结果显示，在 COVID-19 数据集上，Mistral 和 Orca 零样本准确率可达约 60% 以上，而 Falcon 最差，仅略高于随机。在颗粒度更细的 LIAR 数据集上，4 个模型的最高准确率仅为 23%，并在文中指出 LLM 的谣言检测能力主要受训练数据的时效性与覆盖度限制，单靠提示无法支撑高可靠度，需要结合 RAG、微调或更丰富的语料来弥补知识缺口。

4.3.3　辅助标注与特征生成、提取

LLMs 还可以用于辅助人工标注或自动生成一些有用的中间特征。Jiang 等[43]利用 LLM 生成用户对源信息的立场标签，并将其作为其流行病学启发网络的优化目标。Yang 等[19]提出的数据集中也包含了由 LLM 生成的证据。这些方法可以大大降低数据标注的成本和时间。部分研究把 LLM 视为强大的特征提取器，用其输出的文本嵌入来提升检测模型的表现。Pattanaik 等[59]提出了一种结合大语言模型嵌入与 TextGCN 的谣言检测方法，具体做法是在图神经网络模型中同时使用 BERT 和 GPT 来提取词级与句级的向量表示，将两种 768 维向量线性对齐后拼接成节点特征，在 PHEME，Twitter15 与 Twitter16 这 3 个公开谣言数据集上均取得了 80% 以上的准确率。他们还指出 GPT 的单向编码擅长捕捉词语的上下文关系，能补充 BERT 的能力并显著提升模型性能。Chen 等[60]提出了“LLM-based detector”和“LLM-enhanced detector”两条技术路线，并系统比较了 5 种主流 LLM 与传统模型的配合效果，最新的实证研究表明，在特征增强场景下，由 Qwen，NV-Embed 等 LLM 生成的文本嵌入能够在大多数基准数据集上显著提升 MLP，EANN，GCN，Bi-GCN 的表现，为后续将 LLM 深度融入谣言检测流程提供了清晰的思路与实证依据。

4.3.4　数据增强

在谣言检测任务中，标注数据往往稀缺且类别不平衡[16,61]。LLMs 可以用来生成与现有谣言样本风格、内容相似的合成数据，用于扩充训练集，缓解数据不足和类别不平衡问题。Lai 等[16]提出了一种名为 RumorLLM 的模型，这是一个使用谣言写作风格和内容进行微调的大语言模型，专门用于为假新闻检测任务生成增强数据，特别是针对小类别进行增强。实验表明，使用 RumorLLM 增强数据后，模型在 BuzzFeed 和 PolitiFact 数据集上的性能(尤其在 F1 分数和 AUC-ROC 上)优于基线方法。此外，Askarizade[62]也使用 GPT-2 生成与真实谣言相似的文本，在 3 个不平衡数据集 PHEME，Twitter15 和 Twitter16 上进行扩充，随后将 GPT-2 细调为二分类器，只取最后 token 的嵌入表示作为分类特征，并通过 Softmax 输出类别，实现“零特征工程、端到端”的检测流程。Dai 等[63]提出了利用 ChatGPT 进行文本数据增强的方法 AugGPT，首先将训练集中每个句子改写为多个意义相同但表述不同的变体，以扩充少样本场景下的训练数据，以该数据微调 Bert 完成分类任务，在少样本文本分类实验结果中，AugGPT 生成的数据显著提升了模型性能，各项指标均优于其他增强方法。这类方法将 LLM 用作“谣言生成器”，帮助现有模型获得更多训练样本。

4.3.5　可解释性

除了进行两分类，LLMs 的强大生成能力同样可用于判断依据或者生成详实的辟谣说明。Wang 等[64]提出了 FND-LLM(Fake News Detection with Large Language Models and Multimodal Fusion)框架，有效融合了 SLM(小语言模型)和 LLM 的互补优势。FND-LLM 框架采用了多种分支，文本特征分支和视觉语义分支分别用于提取新闻文本内容和视觉信息，共注意网络提取文本和视觉信息的相关性，视觉篡改分支提取新闻图像篡改特征，跨模态特征分支通过 CLIP 模型强化了模态间的互补性，而大型语言模型分支利用 LLM 的推理能力为检测过程提供辅助解释。Xu 等[5]提出的流程不仅可以检测谣言，还能生成解释性内容来反驳信息的真实性，对于提升用户对检测结果的信任度至关重要。这与 Ge 等[15]追求的可解释性目标一致，但 LLM 能够生成更自然、更丰富的解释文本。

4.3.6　多模态信息处理与融合

对于包含图像、视频等多模态信息的谣言，具有多模态能力的 LLMs 可以扮演更核心的角色。Zhong 等[14]针对短视频谣言检测提出了多模态融合 LLM 框架 VMID(Visual-Textual-Metadata Integrated Detection)，其通过视频内容多维解析获取不同模态表示生成文本描述，然后输入 LLM 进行综合评估。这种方法利用 LLM 强大的文本理解和推理能力来处理融合后的信息，在准确性、鲁棒性和多模态信息的融合利用上均优于基线模型，准确率达到 90.93%。Zeng 等[65]则构建了大规模合成图文谣言数据集，并设计了基于语义和分布相似度的选样机制，从中筛选最能代表真实数据的子集用于微调多模态 LLM。实验表明，只用精选的合成样本微调一个参数大小为 1.3×10^{10} 的 LLM，也能在真实图文谣言检测任务上超越 GPT-4V。因此，多模态信息＋LLM 的结合为

谣言检测提供了新的思路。

4.3.7 外部知识库

LLMs本身蕴含大量知识，但可能存在知识过时或不准确的问题。将LLM与实时更新的外部知识库或事实核查数据库相结合，可以提升检测的准确性和时效性。Xu等[5]设计的ECCW(Expert-Citizen Collective Wisdom)模块中包含一个检索模块，负责从实时更新的辟谣数据库中检索相关知识，再提供给LLM进行判断。Hang等[66]提出的TrumorGPT模型采用了GraphRAG(基于知识图谱的检索增强生成)策略，将LLM与定期更新的领域知识图谱相结合用于事实核查，并确保模型基于最新的外部信息进行判断，在PolitiFact数据集上取得了88.5%的准确率和88.1%的F1分数，明显优于不引入外部知识的大语言模型。相较于HybridRAG等现有检索增强方案，GraphRAG的结构化检索有着更高的精度。此外，Shao等[67]设计的LaReF框架通过权威新闻检索模块将大模型与最新的权威新闻数据相融合，并采用主动学习机制在检测时不断利用新辟谣样本更新模型，提升泛化性能。该方法充分利用了LLM在自然语言理解上的优势，同时引入权威新闻等可信知识来规避大模型潜在的幻觉。实验证明，与不利用权威新闻或LLM的模型相比，该框架的谣言检测准确率有明显提高，且能随着迭代学习持续优化检测效果。Zhu等[68]提出的方法也通过外部知识图谱获取帖子中实体和概念的解释，以提供更多上下文信息，增强语义理解。虽然其未使用LLM，但体现了结合外部知识的重要性，可与LLM方法进一步融合。Yang等[19]构建了一个包含多粒度证据的谣言检测数据集RD-E，旨在支持模型利用外部证据来验证社交媒体上的声明，这表明利用LLM检索证据并进行推理判断，是未来的一个重要方向。

总而言之，对于谣言检测来说，LLMs为其提供了强大的语义理解能力、零/少样本学习能力、基于知识的推理和建模、可解释地生成、多模态应用能力以及数据增强能力。与其他传统的深度学习方法或者直接将PLM作为编码器的做法相比，基于LLM的提升方法则更加灵活。这对于推动谣言检测技术在准确率、鲁棒性、适配性、可解释性等方面的发展是十分有益的。与此同时，其带来的挑战也是十分显著的，如高昂的计算成本、模型的“幻觉”问题，以及如何更好地控制LLM实现可靠推理。

4.3.8 基于大语言模型方法的总结

为了更系统地梳理和比较已有工作，本文将前述主流方法进行归纳性分析，可大致分为4类：提示学习方法、参数微调方法、知识增强方法和多模态融合方法。各类方法在适用场景、建模成本、效果表现和发展趋势等方面存在显著差异。表5列出了各类方法的特点，为后续在选择与设计模型时提供了参考。

表5 基于大语言模型的4种技术路径对比

Table 5 Comparison of four technical paradigms based on large language models

技术路径	典型方法	适用场景	优势	局限性	发展趋势
提示学习	Few-Shot提示；Zero-Shot提示	缺少标注数据的任务场景下直接使用预训练LLM	无需额外训练，直接调用LLM推理，能够利用模型知识快速适应谣言任务	对提示工程依赖严重，不同提示对结果影响大；模型对长上下文处理有限	自动化提示优化，减少对人工经验的依赖
参数微调	LoRa微调；Adapter适配器微调	有充足领域数据需定制LLM至特定任务或领域的场景	通过更新模型参数获得针对谣言检测任务的最优表现，通常微调后性能优于直接提示	训练开销高；可能遗忘原有知识	持续学习和多任务微调，避免遗忘，实现一模多能
知识增强	RAG，GraphRag	任务需要模型获取外部知识、提供事实支撑的场景	将外部知识引入LLM，缓解模型封闭语料局限，提高准确性和知识覆盖面；利用知识图谱等结构化知识可提升复杂推理能力	检索不相关或错误信息会误导模型；模型可能对检索到的知识依赖过强而缺乏自主推理	引入动态知识更新机制，保持模型知识的时效性
多模态融合	多模态LLM如GPT-4	带有文本、图像、视频等混合场景	融合图文等多模态线索，信息更全面	模型结构复杂，需解决跨模态对齐和噪声问题；通常依赖大量多模态标注数据，训练难度高	提升跨模态对齐技术，引入OCR等模型处理特殊模态任务

5 大模型生成内容的检测技术

目前已有多种指标用于区分机器生成文本。大型语言模型有时会出现细微的逻辑漏洞或事实偏差，因此文本中的不一致性和错误可以被视为潜在的机器生成信号[69]。由于这些模型在大规模语料上进行训练，其输出往往反映出训练数据固有的模式和偏见，文本中若出现不寻常或倾向性的观点，则可能暗示其生成来源。此外，对文本上下文相关性与连贯性的考察亦具有一定鉴别价值。尽管在构建符合语境且连贯的叙述方面，大型语言模型表现出了较高的能力，但在确保信息一致性和捕捉细微语境方面仍存在不足，尤其是在涉及复杂或多阶段推理的任务中，偶尔出现的语调或主题偏移均可能成为判别依据。

水印技术作为识别机器生成内容的另一手段，其核心在于在生成的文本中嵌入特定的可检测模式，同时尽量保持输出的质量和多样性。Atallah等首次提出了针对自然语言的水印概念[70]，该方法后续逐步演进为适用于神经语言模型输出的技术。近期，文献[71]针对Transformer模型展开研究，提出了AWT(Adversarial Watermarking Transformer)，而Kirchenbaue等[72]则提出了一种模型无关的方法，使得水印技术能够在当下主流的自回归语言模型中得到应用。但在实际应用中，该方法仍面临着大规模数据存储和高效检索机制的挑战。

随着机器生成技术的不断进步，人机文本间的界限逐渐模糊，严格的来源验证变得尤为必要。目前，识别机器生成文本的研究主要集中在以下几个方面：语言特征分析、风格与语调评估、基于知识的分析以及对抗性测试。

在语言特征分析中，研究者关注文本中n-gram频率、句

法结构和语义一致性等语言属性，这些指标能够反映出文本是否符合人类撰写的特征。An 等[73]的实验研究利用来自 2012 年 William and Flora Hewlett 基金会 Kaggle 竞赛的数据集，对 7 至 10 年级学生撰写的 12 978 篇作文进行分析，探讨了人类文本与机器生成文本之间的差异。

1）风格与语调评估则侧重于文本在情感表达、语气和风格细节上的差异。机器生成文本往往缺乏人类文本中所具有的微妙情感和个性化表达，通过情感分析、风格计量学方法以及对特定词汇和短语使用情况的检测，可以为区分两者提供参考依据[74]。

2）基于知识的分析关注文本中信息的时效性和准确性。通过对文本内容进行事实核查、与权威信息源的比对以及寻找潜在逻辑矛盾，可以判断文本是否存在知识缺陷，这也是识别机器生成文本的重要手段[74-75]。

3）对抗性测试方法则通过构造专门的挑战任务，考察模型对精心设计的输入的反应情况，从而评估其在基本逻辑推理、幽默识别以及非结构化对话等方面的表现。这一方法在假新闻检测等领域显示出了较高的实用价值[76]。

除了以上较为常见的技术，还有基于模型内部统计特性的新的检测方法。Farquhar 等[77]针对大模型（如 ChatGPT 和 Gemini）在生成回答时容易产生的“幻觉”现象，尤其是论文中所称的“虚构”——即模型在回答过程中生成的既不真实也不一致的内容，提出了一种基于语义熵的无监督检测方法，该方法通过对同一问题进行多次采样，并利用双向蕴涵来对生成答案进行语义聚类，进而计算聚类内的熵值，以度量模型对其回答含义的不确定性。与传统的仅基于词级熵的检测方法相比，这种方法能够更准确地捕捉同一语义表达的不同表述，从而有效识别那些因随机性导致的虚构信息。在多个数据集上的实验结果表明，该方法在提高问答准确性和拒答可能产生虚构回答的问题上均有显著优势，为大模型在开放域生成任务中的安全性和可靠性提供了新的评估思路和技术支持[77]。

6　挑战与展望

基于大语言模型的社交媒体谣言检测技术尽管取得了显著进展，但仍然面临诸多挑战，同时也展现出广阔的研究前景。

6.1　挑战

1）早期检测问题：谣言出现之初，可供利用的传播特征（如转发、评论、用户反应等）非常有限且稀疏，判断依据不足。与此同时，谣言检测具有严格的实时性要求，模型需要在极短时间窗口内做出决策，以尽早阻止虚假信息扩散。这一挑战实质上是对检测时机与准确率的权衡：过早判断可能因证据不足影响准确率，但过晚又削弱了干预价值。因此，想要在传播初期信息极度匮乏的情况下保持高准确率较为困难。

2）多模态信息的有效融合与理解：大语言模型需要在多模态空间中学习到一个共享的语义表示，使得文本、图像或视频等模态的语义信息能够相互理解和融合，如何处理模态间的歧义和冗余信息成为一大挑战。此外，在融合多模态信息后如何有效地推理不同模态间的语义一致性，并识别潜在的虚假或不一致的信息，这种语义理解能力对于识别利用图像篡改、断章取义等方式制造的谣言至关重要。

3）可解释性：目前很多深度学习模型，尤其是 LLMs，在输出判定结果时缺乏透明的推理过程，降低了用户对检测结果的可信度。如何使模型具有可解释性，能够给出较为可信的证据和逻辑，是提高谣言检测系统可实用化和应用化的重要途径。

4）模型的鲁棒性与对抗性攻击：谣言制造者可能会故意设计一些内容来欺骗或攻击谣言检测系统。Luo 等[17]的研究表明，现有的基于 GNN 的检测器容易受到利用 LLM 生成的恶意消息注入攻击。提升模型对噪声与对抗攻击的鲁棒性，是确保其在真实环境中可靠运行的关键。Li 等[78]的最新研究分析了不同大语言模型在“越狱”提示下的脆弱性，指出很多大模型经过了安全对齐训练，仍可能被精心设计的输入诱导生成违禁内容。这一发现揭示了现有安全机制的局限，也对未来大模型开发者提出了更高的要求。

5）检测 AI 生成谣言的新挑战：综合前面所探讨的内容，LLMs 生成的谣言具有高仿真、迷惑性强的特点，对检测形成了全新的挑战，如何辨别人工编写的谣言和 AI 生成的谣言，并研发有效的检测技术，是未来研究的重要且迫切的方向。

6）动态性与时效性：谣言传播的媒介环境以及传播模式不断变化，检测模型必须跟上这种变化，同时保持检测的时效性。

6.2　展望

1）发展更强大的多模态谣言检测模型。随着多模态 LLM 的发展，未来将出现更强大的能够联合处理文本、图像、视频、音频等多模态信息的谣言检测模型。研究重点可聚焦于高效的跨模态表示学习、信息融合机制及内容一致性验证方法。

2）针对 AI 生成谣言的检测技术。随着 AIGR 威胁的加剧，探索语言学“指纹”、内容溯源技术，以及利用模型对自身生成内容的识别技术将成为研究热点。

3）融合 LLM 与外部知识。通过引入实时更新的知识图谱、事实核查数据库和多源证据，结合 LLM 的推理能力，可缓解其“幻觉”问题，提升检测的准确性与可靠性。

4）人机协同的谣言治理体系。完全自动化的谣言检测可能难以实现百分之百的准确性。未来的发展方向是构建人机协同的系统，由模型负责大规模筛选与分析，再由人类专家提供深度判断与决策，实现更有效、更可靠的谣言治理。

结束语　本文围绕“大语言模型与谣言”的主题，系统地综述了社交媒体中谣言检测的发展路径与技术方法。从传统机器学习与深度学习模型出发，逐步引入注意力机制、多模态融合以及近年来兴起的大语言模型技术，全面回顾了谣言检测在方法、数据集和实践应用等方面的研究现状。同时，也对基于大模型生成内容的检测技术进行了梳理，并总结了当前在语义一致性、风格特征、水印识别等方面的代表性方法。

参 考 文 献

[1]　WAN Q B，HU F，ZHOU M T，et al. A survey of rumor detection oriented to Twitter platform [J]. Information & Communi-

cations, 2019(12): 137-139
[2] TASNIM S, HOSSAIN M M, MAZUMDER H. Impact of rumors and misinformation on COVID-19 in social media[J]. Journal of preventive medicine and public health, 2020, 53(3): 171-174.
[3] LIU X N, HONG X Y, CAO Z Y, et al. Rumor detection on social media: methods, challenges, and trends [J]. Computer Engineering and Applications, 2025, 61(11): 31-50.
[4] ZHOU H H. Research on rumor detection and harmfulness prediction in social networks [D]. Nanjing: Nanjing University of Information Science and Technology, 2022.
[5] XU J, XIAN L, LIU Z, et al. The future of combating rumors? Retrieval, discrimination, and generation [J]. arXiv: 2403.20204, 2024.
[6] LU H Y, FAN C Y, WU X J. Few-shot COVID-19 rumor detection for online social media [J]. Journal of Chinese Information Processing, 2022, 36(1): 135-144, 172.
[7] AL-SAREM M, ALSAEEDI A, SAEED F, et al. A novel hybrid deep learning model for detecting COVID-19-related rumors on social media based on LSTM and concatenated parallel CNNs [J]. Applied Sciences, 2021, 11(17): 7940.
[8] TAHA M A, JABAR H D A, MOHAMMED W K. Fake news detection model basing on machine learning algorithms [J]. Baghdad Science Journal, 2024, 21(8): 2771-2781.
[9] WU S Y, DONG Q X, SONG Z J, et al. A review of misinformation detection methods on social media [J]. Journal of the China Society for Scientific and Technical Information, 2022, 41(6): 651-661.
[10] LIU Z Y, SONG C H, YANG C. Early automatic detection of rumors on social media platforms [J]. Global Media Journal, 2018, 5(4): 65-80.
[11] LIN H, ZHANG X, FU X. A graph convolutional encoder and decoder model for rumor detection [C]// 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA). IEEE, 2020: 300-306.
[12] MA J, GAO W, WEI Z, et al. Detect rumors using time series of social context information on microblogging websites[C]// Proceedings of the 24th ACM International Conference on Information and Knowledge Management. ACM, 2015: 1751-1754.
[13] HUANG D, YANG S, GAO S. Rumor detection in networks based on the BERT-CNN approach[J]. Applied and Computational Engineering, 2024, 77(1): 1-6.
[14] ZHONG W, XIAO Y, XU M, et al. VMID: A Multimodal Fusion LLM Framework for Detecting and Identifying Misinformation of Short Videos [J]. arXiv: 2411.10032, 2024.
[15] GE X Y, ZHANG M S, WEI B, et al. Explainable rumor detection based on dual sentiment perception [J]. Journal of Chinese Information Processing, 2022, 36(9): 129-138.
[16] LAI J, YANG X, LUO W, et al. Rumorllm: A rumor large language model-based fake-news-detection data-augmentation approach [J]. Applied Sciences, 2024, 14(8): 3532.
[17] LUO Y, LI Y, WEN D, et al. Message Injection Attack on Rumor Detection under the Black-Box Evasion Setting Using Large Language Model [C]// Proceedings of the ACM Web Conference. ACM, 2024: 4512-4522.
[18] CHEN C Y, SHU K. CAN LLM-Generated Misinformation Be Detected? [EB/OL]. (2023-09-25) [2025-07-06]. https://arxiv.org/abs/2309.13788.
[19] YANG Z, LIN J, GUO Z, et al. Towards rumor detection with multi-granularity evidences: A dataset and benchmark [J]. IEEE Transactions on Knowledge and Data Engineering, 2024, 36(11): 7188-7200.
[20] ZUBIAGA A, AKER A, BONTCHEVA K, et al. Detection and resolution of rumours in social media: A survey [J]. ACM Computing Surveys, 2018, 51(2): 1-36.
[21] GAILDRAUD L, SAMIER H, BRUNEAU J M. The generation of a rumour: from emergence to percolation [C]// Proceedings of the European Symposium of Competitive Intelligence. 2009.
[22] ALLPORT G W, POSTMAN L. The psychology of rumor [M]. New York: Henry Holt and Company, 1947.
[23] BONSTEEL S. APA PsycNET [J]. The Charleston Advisor, 2012, 14(1): 16-19.
[24] LI Q, ZHANG Q, SI L, et al. Rumor detection on social media: Datasets, methods and opportunities [J]. arXiv: 1911.07199, 2019.
[25] DERCZYNSKI L, BONTCHEVA K, LUKASIK M, et al. PHEME: computing veracity—the fourth challenge of big social data[C]// European Semantic Web Conference(ESWC). 2014.
[26] KOCHKINA E, LIAKATA M, ZUBIAGA A. All-in-one: Multi-task learning for rumour verification [J]. arXiv: 1806.03713, 2018.
[27] MA J, GAO W, WONG K F. Detect rumors in microblog posts using propagation structure via kernel learning [C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2017: 708-717.
[28] SHU K, MAHUDESWARAN D, WANG S, et al. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media [J]. Big Data, 2020, 8(3): 171-188.
[29] JIN Z, CAO J, GUO H, et al. Multimodal fusion with recurrent neural networks for rumor detection on microblogs [C]// Proceedings of the 25th ACM International Conference on Multimedia. ACM, 2017: 795-816.
[30] NAN Q, CAO J, ZHU Y, et al. MDFEND: Multi-domain fake news detection [C]// Proceedings of the 30th ACM International Conference on Information and Knowledge Management. ACM, 2021: 3343-3347.
[31] SILVERMAN C. This analysis shows how viral fake election news stories outperformed real news on Facebook[EB/OL]. (2016-11-16)[2025-06-28]. https://www.buzzfeednews.com/article/craigsilverman/viral-fake-election-news-outperformed-real-news-on-facebook.
[32] CAO Y P, NAIR A M, SOOF N J, et al. CoSMis: A Hybrid Human-LLM COVID Related Scientific Misinformation Dataset and LLM pipelines for Detecting Scientific Misinformation in the Wild [C]// Proceedings of the AAAI Conference on Artificial Intelligence. AAAI, 2025.

[33] SUN Y S, HE J F, CUI L M, et al. Exploring the Deceptive Power of LLM-Generated Fake News: A Study of Real-World Detection Challenges [J]. arXiv: 2403.18249, 2024.

[34] ZHANG Z, ZHANG Y, ZHOU X, et al. From Generation to Detection: A Multimodal Multi-Task Dataset for Benchmarking Health Misinformation [J]. arXiv: 2505.18685, 2025.

[35] SUN Y, HE J, LEI S, et al. Med-MMHL: A Multi-Modal Dataset for Detecting Human- and LLM-Generated Misinformation in the Medical Domain [J]. arXiv: 2306.08871, 2023.

[36] WANG L Z, MA Y, GAO R, et al. MegaFake: A Theory-Driven Dataset of Fake News Generated by Large Language Models [J]. arXiv: 2408.11871, 2024.

[37] CASTILLO C, MENDOZA M, POBLETE B. Information credibility on Twitter [C] // Proceedings of the 20th International Conference on World Wide Web. ACM, 2011: 675-684.

[38] JAHANBAKHSH-NAGADEH Z, FEIZI-DERAKHSHI M R, RAMEZANI M, et al. A model to measure the spread power of rumors [J]. Journal of Ambient Intelligence and Humanized Computing, 2023, 14: 13787-13811.

[39] PI D C, WU Z Y, CAO J J. Early rumor detection method based on knowledge graph representation learning [J]. Acta Electronica Sinica, 2023, 51(2): 385-395.

[40] RANI P, JAIN V, SHOKEEN J, et al. Blockchain-based rumor detection approach for COVID-19 [J]. Journal of Ambient Intelligence and Humanized Computing, 2024, 15(1): 435-449.

[41] KHAN S, KAMAL A, FAZIL M, et al. HCovBi-caps: hate speech detection using convolutional and Bi-directional gated recurrent unit with Capsule network [J]. IEEE Access, 2022, 10: 7881-7894.

[42] CHEN L W, SONG Y R, SONG B. Temporal-aware heterogeneous graph neural network for rumor detection [J]. Mini-micro Systems, 2024, 45(1): 45-51.

[43] JIANG W, CHEN T, GAO X, et al. Epidemiology-informed network for robust rumor detection [C] // Proceedings of the ACM on Web Conference 2025. ACM, 2025: 3618-3627.

[44] LI H. Research on key issues of concept-aware rumor detection in English texts on social media [D]. Guilin: Guilin University of Electronic Technology, 2023.

[45] HAN X M, JIA C Y, LI X Y, et al. Rumor detection model based on dual attention to nodes and paths in propagation tree structures [J]. Computer Science, 2023, 50(4): 22-31.

[46] TAO X, ZHU Y, LI C P. Rumor detection method based on attention and multimodal hybrid fusion [J]. Computer Engineering, 2021, 47(12): 71-77.

[47] YANG R, MA J, LIN H, et al. A weakly supervised propagation model for rumor verification and stance detection with multiple instance learning [C] // Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2022: 1761-1772.

[48] JING Q, YAO D, FAN X, et al. TRANSFAKE: multi-task transformer for multimodal enhanced fake news detection [C] // 2021 International Joint Conference on Neural Networks (IJCNN). IEEE, 2021: 1-8.

[49] HUANG Y, GAO H, GAO D, et al. Research on Tibetan Rumor Detection Method Based on Word2Vec-BiLSTM-Att [C] // 2024 5th International Symposium on Computer Engineering and Intelligent Communications (ISCEIC). IEEE, 2024: 331-337.

[50] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [C] // Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017: 6000-6010.

[51] JIN W, WANG N, TAO T, et al. A veracity dissemination consistency-based few-shot fake news detection framework by synergizing adversarial and contrastive self-supervised learning [J]. Scientific Reports, 2024, 14(1): 19470.

[52] HOWARD J, RUDER S. Universal language model fine-tuning for text classification [J]. arXiv: 1801.06146, 2018.

[53] LI X L, LIANG P. Prefix-tuning: Optimizing continuous prompts for generation [EB/OL]. (2021-01-01) [2025-07-06]. https://arxiv.org/abs/2101.00190.

[54] HU E J, SHEN Y, WALLIS P, et al. Lora: Low-rank adaptation of large language models [C] // Proceedings of the 10th International Conference on Learning Representations. ICLR, 2022.

[55] PAVLYSHENKO B M. Analysis of disinformation and fake news detection using fine-tuned large language model [EB/OL]. (2023-09-08) [2025-07-06]. https://arxiv.org/abs/2309.04704.

[56] LIU Z, ZHANG X, YANG K, et al. Fmdllama: Financial misinformation detection based on large language models [C] // Companion Proceedings of the ACM on Web Conference 2025. ACM, 2025: 1153-1157.

[57] LIN H, YI P, MA J, et al. Zero-shot rumor detection with propagation structure via prompt learning [C] // Proceedings of the AAAI Conference on Artificial Intelligence. 2023: 5213-5221.

[58] PENDYALA V S, HALL C E. Explaining Misinformation Detection Using Large Language Models [J]. Electronics, 2024, 13(9): 1673.

[59] PATTANAIK B, MANDAL S, TRIPATHY R M, et al. Rumor detection using dual embeddings and text-based graph convolutional network [J]. Discover Artificial Intelligence, 2024, 4(1): 86.

[60] CHEN M, WEI L, CAO H, et al. Explore the potential of LLMs in misinformation detection: An empirical study [C] // AAAI 2025 Workshop on Preventing and Detecting LLM Misinformation (PDLM). AAAI, 2025.

[61] HAN S, GAO J, CIRAVEGNA F. Neural language model based training data augmentation for weakly supervised early rumor detection [C] // Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). IEEE, 2019: 105-112.

[62] ASKARIZADE M. Enhancing rumor detection with data augmentation and generative pre-trained transformer [J]. Expert Systems with Applications, 2025, 262: 125649.

[63] DAI H, LIU Z, LIAO W, et al. AugGPT: Leveraging ChatGPT for Text Data Augmentation [J]. IEEE Transactions on Big Data, 2025, 11(3): 907-918.

[64] WANG J, ZHU Z, LIU C, et al. LLM-enhanced multimodal detection of fake news [J]. PLoS ONE, 2024, 19(10): e0312240.

[65] ZENG F, LI W, GAO W, et al. Multimodal misinformation detection by learning from synthetic data with multimodal LLMs [EB/OL]. (2024-09-30)[2025-07-06]. https://arxiv.org/abs/2409.19656.

[66] HANG C N, YU P D, TAN C W. TrumorGPT: Graph-Based Retrieval-Augmented Large Language Model for Fact-Checking [J]. IEEE Transactions on Artificial Intelligence, 2025, 1: 1-15.

[67] SHAO Z J, CAI G Y, LIU Q H, et al. Iterative rumor detection method enhanced by large language models and authoritative news [J]. Journal of Nanjing University (Natural Sciences), 2024, 60(6): 970-980.

[68] ZHU Y, WANG G S, JIN W W, et al. Online rumor detection based on text semantic enhancement and comment stance weighting [J]. Journal of Frontiers of Computer Science and Technology, 2024, 18(12): 3311-3323.

[69] GALLEGOS I O, ROSSI R A, BARROW J, et al. Bias and fairness in large language models: A survey[J]. Computational Linguistics, 2024, 50(3): 1097-1179.

[70] ATALLAH M J, RASKIN V, CROGAN M, et al. Natural language watermarking: Design, analysis, and a proof-of-concept implementation [C] // Information Hiding: 4th International Workshop (IH 2001). Pittsburgh, PA, USA, Berlin: Springer, 2001: 185-200.

[71] ABDELNABI S, FRITZ M. Adversarial watermarking transformer: Towards tracing text provenance with data hiding [C]// 2021 IEEE Symposium on Security and Privacy (SP). IEEE, 2021: 121-140.

[72] KIRCHENBAUER J, GEIPING J, WEN Y, et al. A watermark for large language models [C] // International Conference on Machine Learning. PMLR, 2023: 17061-17084.

[73] AN R, YANG Y, YANG F, et al. Use prompt to differentiate text generated by ChatGPT and humans [J]. Machine Learning with Applications, 2023, 14: 100497.

[74] RAY P P. ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope [J]. Internet of Things and Cyber-Physical Systems, 2023, 3: 121-154.

[75] YUAN A, COENEN A, REIF E, et al. Wordcraft: story writing with large language models [C]// Proceedings of the 27th International Conference on Intelligent User Interfaces (IUI '22). New York, NY, USA: ACM, 2022: 841-852.

[76] WANG H, DOU Y, CHEN C, et al. Attacking fake news detectors via manipulating news social engagement[C]// Proceedings of the ACM Web Conference. ACM, 2023: 3978-3986.

[77] FARQUHAR S, KOSSEN J, KUHN L, et al. Detecting hallucinations in large language models using semantic entropy [J]. Nature, 2024, 630(8017): 625-630.

[78] LI T, WANG Z, LIU W, et al. Revisiting Jailbreaking for Large Language Models: A Representation Engineering Perspective [C]// Proceedings of the 31st International Conference on Computational Linguistics (COLING 2025). ACL, 2025: 3158-3178.

PAN Jie, born in 1998, postgraduate. His main research interests include natural language processing and rumor detection.

WANG Juan, born in 1979, Ph.D, associate professor. Her main research interests include crime prediction and online public opinion analysis.

(责任编辑:喻藜)