Bad Actor, Good Advisor: Exploring the Role of Large Language Models in Fake News Detection 复现报告

摘要

随着大型语言模型（LLMs）在自然语言处理任务中的广泛应用，它们在假新闻检测中的角色和效能也引起了广泛关注。本文《Bad Actor, Good Advisor: Exploring the Role of Large Language Models in Fake News Detection》[5] 旨在探讨 LLMs 作为假新闻鉴别者（Bad Actor）与检测顾问（Good Advisor）的双重角色，并分析其在这一领域内的潜力和局限性。

复现研究中，首先回顾了原始论文的方法论，包括数据集的选择、预处理步骤、实验设置以及评估指标。然后，基于这些信息，在相似条件下进行了实验，以验证 LLMs 是否能够有效地辨别真实新闻与虚假新闻。结果表明，尽管 LLMs 具备一定的假新闻检测能力，但它们也可能被用来制造更加难以识别的假新闻内容。此外，还发现模型的表现很大程度上依赖于训练数据的质量和多样性。

通过对比原作与复现结果，讨论了影响 LLM 性能的关键因素，并提出了改进模型鲁棒性的建议。最后，本报告强调了开发更先进、可靠的假新闻检测工具的重要性，同时呼吁制定严格的伦理指导方针来规范 LLMs 的应用，确保技术进步服务于社会公共利益。

关键词：大型语言模型；假新闻检测；复现研究；机器学习

1 引言

在信息传播日益加速的当今社会，互联网和社交媒体已经成为人们获取新闻和信息的主要渠道。然而，这种快速的信息流通也带来了假新闻（fake news）泛滥的问题，假新闻不仅误导公众，还可能对社会稳定、政治决策和个人生活造成严重影响。随着人工智能技术的进步，大型语言模型（LLMs）作为自然语言处理领域的新兴力量，在文本生成、理解和分类等任务上展现出了卓越的能力。因此，探讨 LLMs 在假新闻检测中的角色及其潜在影响，成为了学术界和业界共同关注的重要课题。

1.1 选题背景

原始论文《Bad Actor, Good Advisor: Exploring the Role of Large Language Models in Fake News Detection》聚焦于 LLMs 在假新闻检测中的双重角色：一方面，根据原始论文的

研究，尽管 LLMs 展示了强大的文本处理能力，但在假新闻检测任务上的表现却不如经过特定任务微调的小型语言模型（SLMs）。具体来说，LLMs 能够从多个角度提供合理的解释性理由，但其选择和整合这些理由的能力尚显不足，导致整体性能不尽人意，所以称其为 Bad Actor；
另一方面，它们也可以作为“好顾问”（Good Advisor），研究还发现，通过引入视角特定的提示并执行基于规则的判断集成，可以显著提升 LLMs 的表现，表明 LLMs 可以作为有效的辅助工具而非替代方案。通过提供有指导性的推理来辅助检测假新闻。该研究旨在深入分析 LLMs 在这一领域内的潜力和局限性。

1.2 选题依据

原始论文研究基于以下几点考虑：

1. **社会需求**：近年来，假新闻问题日益严重，对社会舆论、政治稳定乃至个人生活造成了诸多负面影响。社交媒体平台成为假新闻传播的主要渠道，使得虚假信息能够迅速扩散并产生广泛的社会效应。社会各界对于高效准确的假新闻检测工具的需求愈发迫切。因此，探索如何借助先进的 AI 技术如 LLMs 来应对这一挑战，不仅具有重要的学术价值，也符合社会发展的实际需要。

2. **现有研究空白**：随着人工智能技术的迅猛发展，大型语言模型（LLMs）在自然语言处理（NLP）领域取得了显著成就。这些模型通过海量数据训练，具备了强大的文本生成、理解和分类能力。特别是在信息检索和问答系统中，LLMs 已经展示了其卓越的应用潜力。然而，它们在假新闻检测这一特定任务上的表现及其潜在影响尚未得到充分探讨。本研究旨在填补这一研究空白，为更广泛地利用 LLMs 提供理论和技术支持。

1.3 选题意义

本研究的目标是复现原始论文中的实验结果，进一步验证 LLMs 在假新闻检测中的实际效能。通过对不同提示策略的实证分析，我们希望揭示 LLMs 在假新闻检测方面的优势与挑战，并为开发更先进、可靠的假新闻检测系统提供理论支持和技术路径。同时，我们也强调了在 AI 技术发展中融入伦理考量的重要性，确保技术创新符合社会公共利益。

此外，本研究还致力于：

- **丰富理论框架**：通过实证研究，补充和完善关于 LLMs 在假新闻检测中作用的相关理论。
- **提供实践指导**：为构建更加智能且可靠的假新闻检测系统提供建议，提高公众获取真实信息的质量。
- **促进政策制定**：为政府机构及相关组织制定相关政策提供科学依据，推动健康网络环境的建设。

综上所述，《Bad Actor, Good Advisor: Exploring the Role of Large Language Models in Fake News Detection》这一选题紧密结合当前科技发展趋势和社会热点问题，旨在从多个角度深入探讨 LLMs 在假新闻检测中的角色，对于推动学术研究和实际应用均有着深远的影响。

以上引言部分结合了原始论文的内容，概述了研究的背景、目的以及它所具有的理论和实践意义。请注意，具体的实验细节和数据结果将在后续章节中详细展开。

2 相关工作

在探讨大型语言模型（LLMs）于假新闻检测中的角色时，本研究建立在现有文献的基础上，综合了多个领域的研究成果。以下部分将详细介绍与本课题相关的先前研究，涵盖假新闻检测方法、LLMs 的应用现状以及伦理考量。

2.1 假新闻检测方法

早期的假新闻检测研究主要依赖于传统的机器学习算法和特征工程。例如，基于文本内容的分类器可以通过提取关键词、主题建模或情感分析等手段来识别虚假信息。随着深度学习的发展，卷积神经网络（CNN）、循环神经网络（RNN）及其变体长短期记忆网络（LSTM）被广泛应用于自然语言处理任务中，并取得了显著进展。此外，预训练的语言模型如 BERT [3]、RoBERTa [7] 等通过大规模无监督数据训练，能够更好地捕捉语义信息，从而提高了假新闻检测的效果。

近年来，研究者们开始关注多模态融合的方法，结合文本、图像甚至视频等多种媒体形式进行综合判断。这些方法不仅考虑了单一模态的信息，还试图利用不同模态之间的关联性来增强检测准确性。例如，一些研究表明，通过联合分析社交媒体帖子的文字描述与其配图，可以有效减少误报率并提高整体性能。

2.1.1 基于社交上下文的方法

社交上下文为基础的方法通过观察假新闻在传播过程中的扩散模式、用户反馈和社会网络结构来进行区分。Zhou 和 Zafarani (2019) [19] 研究了假新闻在社交媒体上的传播路径；Min 等人 (2022) [8] 探讨了用户对假新闻的反应；Nguyen 等人 (2020) [10] 分析了社会网络对于假新闻传播的影响。

2.1.2 基于内容的方法

基于内容的方法则侧重于从给定的内容中寻找线索，包括文本和图像。Przybyla (2020) [13] 提出了针对文本内容的假新闻检测技术；Qi 等人 (2021) [14] 则专注于图像内容的分析。此外，还有些方法借助知识库 (Popat et al., 2018) [12] 和新闻环境 (Sheng et al., 2022) [15] 来辅助检测。

2.2 大型语言模型的应用现状

LLMs 作为新兴的 NLP 技术，因其强大的生成能力和广泛的上下文理解能力而受到广泛关注。它们不仅可以用于文本生成、翻译、问答等多个应用场景，还在对话系统和个性化推荐等领域展现了巨大潜力。然而，在假新闻检测方面，LLMs 的表现却呈现出复杂的一面。

一方面，LLMs 具备从多种角度提供合理解释性理由的能力，这使得它们能够在一定程度上辅助人类专家进行假新闻的甄别。另一方面，由于 LLMs 是基于大量互联网文本训练而

成，其中可能包含了不准确或有偏见的内容，导致其生成的结果存在“幻觉”现象，即产生不符合事实的陈述。因此，如何有效利用 LLMs 的优势同时规避其潜在风险，成为了一个亟待解决的问题。

具体而言：

- LLMs 在假新闻检测任务上的表现不如经过特定任务微调的小型语言模型（SLMs），但能提供合理的解释性理由。
- 通过引入视角特定的提示并执行基于规则的判断集成，可以显著提升 LLMs 的表现，表明 LLMs 可以作为有效的辅助工具而非替代方案。

2.3 小型语言模型与大型语言模型的比较

原始论文首先提出了以下两个问题。

- 能否将大型语言模型应用于假新闻检测任务？
- 如何将大型语言模型应用于假新闻检测任务和如何将大型语言模型应用于假新闻检测任务？

为此原始论文研究人员在两个数据集上进行了一系列实验，以下为实验设置。

2.3.1 数据集

原始论文使用了两个真实世界的数据集，分别涵盖中文和英文新闻样本。中文数据选用了 weibo21 数据集 [9], 英文数据选用了 gossipcop 数据集 [16] 具体统计信息如表1所示：

语言	中文			英文		
	训练集	验证集	测试集	训练集	验证集	测试集
真实新闻数量	2,331	1,172	1,137	2,878	1,030	1,024
假新闻数量	2,873	779	814	1,006	244	234
总计	5,204	1,951	1,951	3,884	1,274	1,258

表 1. 假新闻检测数据集统计

2.3.2 模型选择

为了全面评估 LLMs 在假新闻检测中的表现，作者选择了 GPT-3.5-turbo [11] 作为代表性的大型语言模型，并将其与几种经过特定任务微调的小型语言模型（SLMs, 如 Bert）进行了比较。此外，还引入了基线模型（Baseline），包括未附加理由的 LLM-only 方法以及结合了理由的组合模型。

2.3.3 评估指标

性能评估采用了多种常用的分类评价指标，包括宏平均 F1 分数（macro F1）、准确率（Accuracy）、真实新闻 F1 分数（F1real）和假新闻 F1 分数（F1fake）。这些指标能够从不同角度衡量模型的分类效果，从而更全面地反映其优劣。

2.3.4 实验设计

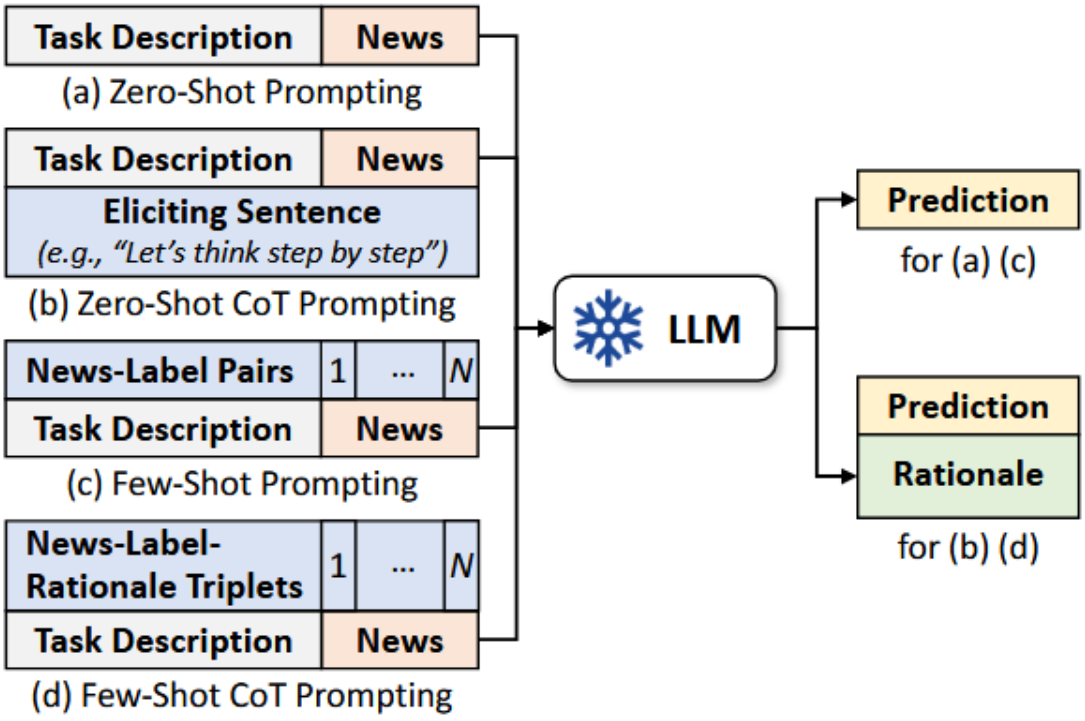

图 1. 大型语言模型的四种 Prompt 调用示例图

大型语言模型（LLM）和小型语言模型（SLM）在中文数据集上和英文数据集上通过新闻文本直接对新闻真假性进行预测，其中大型语言模型（LLM）使用四种不同的 Prompt 调用方式，小型语言模型（SLM）则是通过微调进行预测。

1. **零样本提示（Zero-shot Prompting）：**直接要求 LLM 对给定新闻条目进行真实性判断。
2. **少量样本提示（Few-shot Prompting）：**提供几个相关案例后，再让 LLM 做出判断。
3. **链式思考提示（Chain-of-Thought Prompting, CoT）：**引导 LLM 从多个视角分析新闻内容并提供解释性理由。
4. **规则集成（Rule-based Ensemble）：**将来自不同视角的理由整合起来，通过基于规则的方法得出最终判断。

2.3.5 结果分析

最终的实验结果如上表2 通过对不同提示策略的实证分析，发现尽管 LLMs 能够生成合理的解释性理由，但在选择和整合这些理由方面仍存在不足。尤其在零样本和少量样本提示条件下，LLMs 的表现不如经过特定任务微调的小型语言模型。然而，当引入链式思考提示时，LLMs 的表现得到了显著提升，但是相比于 SLM 仍然有一定差距。

虽然 LLMs 本身直接对新闻真假性进行判断能力较差但是在推理过程中 LLMs 能够生成丰富且复杂的理由，随之作者通过 Prompt 设置尝试让 LLMs 从不同的角度进行分析，实验

Model	Usage	Chinese	English
GPT-3.5-turbo	Zero-Shot	0.676	0.568
	Zero-Shot CoT	0.677	0.666
	Few-Shot	0.725	0.697
	Few-Shot CoT	0.681	0.702
BERT	Fine-tuning	**0.753** (+**3.8%**)	**0.765** (+**9.0%**)

表 2. 实验结果

结果如下图3 从表3可以看出 LLMs 从单一角度出发对新闻文本进行分析能够更好的判断新闻

表 3. LLMs 多角度分析结果

Perspective	Chinese		English	
	Prop.	**macF1**	**Prop.**	**macF1**
Textual Description	65%	0.706	71%	0.653
Commonsense	71%	0.698	60%	0.680
Factuality	17%	0.629	24%	0.626
Others	4%	0.649	8%	0.704

的真实性，其中从文字描述的角度和从社会常识的角度出发得到结果表现较为突出，但是从事实性分析的角度出发反而得到了较差的结果，作者推测有可能是由于 LLMs 自身的幻觉问题导致。

通过以上实验，作者得出本文的核心观点，LLMs 本身具有极为丰富的内部记忆以及多角度分析的能力，但由于其本身结合多角度的能力较差以及幻觉问题，导致其本身难以直接应用于假新闻检测任务中，即”Bad Actor”，SLM 具有灵活且强大的多角度结合的能力却不具有丰富的内部记忆，所以利用 LLMs 生成多角度分析的原理（Rationale）弥补 SLM 的缺陷，起到两者互补的作用，即”Good Advisor” 为了验证以上结论作者统计了 SLM 和 LLMs 在两个数据集上的预测分布，使用 Oracle Vote 的方式给出两者结合的最理想情况，具体如下表4

通过表4可以观察到最理想的情况下（Oracle Vote）最终的实验结果远超于单 LLMs 和 SLM 微调的结果，而最简单的结合方式（Majority Voting）同样能够得到不错的结果。这一结果验证了作者的猜想。

表 4. LLMs 多角度分析和 SLM 微调的实验结果汇总

Model	Usage	Chinese	English
GPT-3.5-turbo	Zero-Shot CoT	0.677	0.666
	from Perspective TD	0.667	0.611
	from Perspective CS	0.678	0.698
BERT	Fine-tuning	0.753	0.765
Ensemble	Majority Voting	0.735	0.724
	Oracle Voting	0.908	0.878

3 本文方法

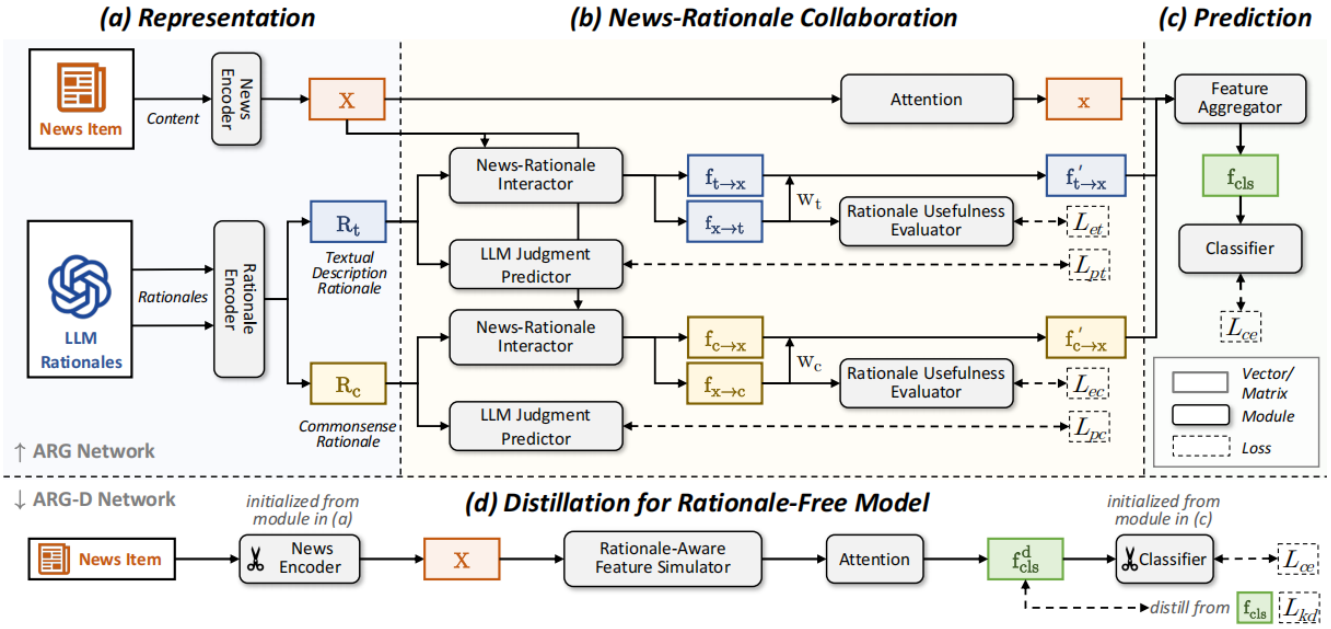

图 2. ARG 模型和 ARG-D 模型

通过上文中作者针对 LLMs 和 SLM 进行的一系列实验，之后作者的工作则围绕着如何结合 SLM 和 LLMs 展开。作者在文中提出了 ARG 模型和 ARG-D 模型，其模型定义如图2

3.1 本文方法概述

ARG 模型的输入有两种，一种为原始数据集中的新闻文本（News Item），一种为通过 LLMs 生成的判定原理（LLM Rationales），这里 ARG 模型中作者采用了表现良好的文本描述角度生成的原理和社会常识角度生成的原理。

两种输入通过 Bert 特征提取后得到特征向量，随后通过三个组件 News-Rationale Interactor、LLM Judgment Predictor、Rationale Usefulness 进行融合，最终再将经过特征融合过后的输出进行聚合得到最终的预测结果。

考虑到 LLMs 本身推理速度较慢，在某些对延迟较为敏感的生产环境中，LLMs 带来的延迟是无法忍受的，于是作者提出了 ARG-D 模型，该模型由 ARG 模型蒸馏 [4] 得到，通过模型蒸馏，ARG-D 得以摆脱对于 LLMs 输入的依赖，ARG-D 提升了推理速度的同时一定程度上保留了原始 ARG 模型的性能。

3.2 特征提取与融合模块

新闻文本和 LLMs 生成的原理通过两个不同的 Bert 进行特征提取，然后通过 News-Rationale Interactor 进行交互，如文字描述角度原理与原始新闻文本交互得到得到 $f_{t\to x}$ 和 $f_{x\to t}$，其计算公式如公式123，其中1为点积注意力

$$CA(Q,K,V) = softmax(Q' \cdot K'/\sqrt{d})V' \tag{1}$$

$$f_{t\to x} = AvgPool(CA(R_t, X, X)) \tag{2}$$

$$f_{x\to t} = AvgPool(CA(X, R_t, T_t)) \tag{3}$$

3.3 损失函数定义

ARG 模型使用了三种损失函数，其中两种与 ARG 中的关键组件 LLM Judgment Predictor 和 Rationale Usefulness 有关，下面分别先介绍这两种组件

3.4 LLM Judgment Predictor 组件

$$\hat{m}_t = sigmoid(MLP(R_t)) \tag{4}$$

$$L_{pt} = CE(\hat{m}, m) \tag{5}$$

LLM Judgment Predictor 组件的输入为 Rationale 的特征向量，LLM Judgment Predictor 通过 LLM Rationale 预测 LLM 对该新闻的预测，其目的是为了让提取 Rationale 特征的 Bert 能够更好的”理解 “Rationale 的重点，以此能够更好的提取 LLM Rationale 中与假新闻预测相关的特征。LLM Judgment Predictor 内部比较简单，仅包含一个多层的 MLP，将原先的高维特征向量降维至一维，公式4为 LLM Judgment Predictor 组件的计算逻辑，5为 LLM Judgment Predictor 组件所使用的损失函数，CE 为二元交叉熵损失函数，其定义为 $CE(\hat{y}, y) = -y\log\hat{y} - (1-y)log(1-\hat{y})$。

3.5 Rationale Usefulness Evaluation 组件

$$\hat{u}_t = sigmoid(MLP(f_{x\to t})) \tag{6}$$

$$L_{et} = CE(\hat{u}_t, u_t) \tag{7}$$

$$w_t = sigmoid(MLP(f_{x\to t})) \tag{8}$$

$$f'_{t\to x} = w_t \cdot f_{t\to x} \tag{9}$$

Rationale Usefulness Evaluation 组件的输入为通过 News-Rationale Interactor 特征交互得到的 $f_{x\to t}$, 其内部计算公式如6，Rationale Usefulness Evaluation 通过 $f_{x\to t}$ 预测 LLM 的判断是否正确 (损失函数定义如公式7)，其目的是为了限制错误的 Rationale 对 ARG 模型的影响，此处作者假定如果 LLM 对该新闻样本预测正确则 LLM 提供的 Rationale 则是" 有用的 "，否则为" 无用的 "，以此作为归纳偏置限制 Rationale 的影响，但是 News-Rationale Interactor 并不直接影响 Rationale 在最终分类上的权重，而是影响特征融合模块的 News-Rationale Interactor 组件中的参数来影响最终特征融合的结果（即 $f_{x\to t}$)，随后以 $f_{x\to t}$ 作为输入单独计算一个权重 w_t8, 使用权重 w_t 限制 Rationale 的影响 (如公式9)。虽然原理十分复杂但是最终呈现得到的效果却并不明显，这一点可以通过后文中作者针对 Rationale Usefulness Evaluation 组件的消融实验可以看出。

3.5.1 Feature Aggregator 组件

$$V = \text{concat}(x, f_{t\to x}, f_{c\to x}; \text{dim} = 1) \tag{10}$$

$$MaskAttention = CA(w, V, V) \tag{11}$$

$$Lce = CE(MLP(f_{cls}), y) \tag{12}$$

经过 Rationale Usefulness 组件为 Rationale 与新闻文本的融合特征 $f_{t\to x}$ 重新分配了权重得到了 $f'_{t\to x}$，随后新闻文本特征进行自注意力层得到 x, 最终将 $x, f_{t\to x}, f_{c\to x}$ 三者通过 Feature Aggregator 组件进行聚合得到 f_{cls}，Feature Aggregator 组件使用了 MaskAttention 对三个输入进行加权聚合, 具体计算如公式11, 其中 V 由 $x, f_{t\to x}, f_{c\to x}$ 三者在第一维上的拼接（此处定义第 0 维为样本批量维度，第一维为序列长度维度，第三维为特征维度)，除此之外公式11中 w 为可学习张量 (假设 $V \in \mathbb{R}^{3\times n}$, 其中 3 为 $x, f_{t\to x}, f_{c\to x}$ 三者，n 为特征维度，则 $w \in \mathbb{R}^{1\times n}$, 所以最终 $f_{cls} \in \mathbb{R}^{1\times n}$), 最终 f_{cls} 通过分类头（Classifier）得到最终的预测结果，其损失函数定义如公式12.

$$L = L_{ce} + \beta_1(L_{et} + L_{ec}) + \beta_2(L_{pt} + L_{pc}) \tag{13}$$

最终总的损失函数定义为公式13, 其中 β_1, β_2 为超参数，用于控制两种损失函数5和7的权重

4 复现细节

最终复现在相关工作中提及的 Weibo21 数据集和 Gossipcop 数据集上进行，以下是复现细节

4.1 与已有开源代码对比

首先作者在原文中公开了源代码，但是对于 LLMs 生成的数据集虽然可以通过向作者申请获得但并没有公布实现代码，为了后续的创新工作，本次复现首先使用原作者提供的数据集对论文中展示的实验结果进行复现，其次采用 Qwen [18] [17] 替代 GPT-3.5-turbo 作为 LLMs 实现，在原始 Weibo21 数据集和 Gossipcop 数据集上实验之外，额外扩展 Twitter 数据集 [1] 补充，这是考虑到中文数据集 Weibo21 选自社交媒体，而 Gossipcop 数据集选自娱乐新闻平台 Gossipcop，为了进一步验证 ARG 模型针对现如今谣言泛滥的社交媒体平台的性能，补充英文社交媒体平台数据集。

以上提及的三种数据集中均包含图片模态，而 ARG 模型仅使用文字模态，所以后续改进工作将围绕多模态方向作为切入点。

4.2 实验环境搭建

4.2.1 LLMs Rationale 生成

	Weibo	Gossipcop 和 Twitter
Python 版本	3.10	3.10
CUDA 版本	12.1	12.1
LLM	Qwen2.5	Qwen2 VL
GPU	NVIDIA A800 80GB PCIe * 2	NVIDIA GeForce RTX 4090

表 5. LLMs Rationale 生成实验环境配置

基本实验环境如表5，由于时间和硬件设备关系，早期工作关于 Gossipcop 和 Twitter 数据集相关的 Rationale 生成选用的 LLMs 为 Qwen2-VL-7B-Instruct，随着后续工作的开展，由于早期 Gossipcop 和 Twitter 数据集的 Rationale 质量不佳，近期将使用 Qwen2.5-72B-Instruct-GPTQ-Int8 重新生成相关 Rationale。

Weibo21 数据集 Rationale 则是通过 Qwen2.5-72B-Instruct-GPTQ-Int8 生成，实验中使用 vLLM [6] 配合 flash attention2 技术 [2] 加速推理过程。

表 6. 各数据集的总量、真数据量和假数据量

数据集	总量 (sum)	真数据量 (real data)	假数据量 (fake data)
GPT gossipcop	6416	4932	1484
GPT weibo	9106	4640	4466
Qwen gossipcop	12332	9616	2716
Qwen twitter	14195	6432	7763
Qwen weibo	7961	3642	4319

原始数据集相关指标如表6，其中 Gossipcop 数据和原作者提供的数据集数量差距较大，Weibo 数据集部分数据存在图片缺失的情况，遂过滤部分图片缺失的数据。

4.2.2 ARG 模型复现

ARG 模型相关代码作者已在 GitHub 开源，此次复现工作除了在作者使用的 GPT gossipcop 和 GPT weibo 上进行之外，也在通过 Qwen 生成的 Rationale 上进行实验。所以此次复现工作的主要集中与 LLMs 数据集的生成与模型改进。

4.3 创新点

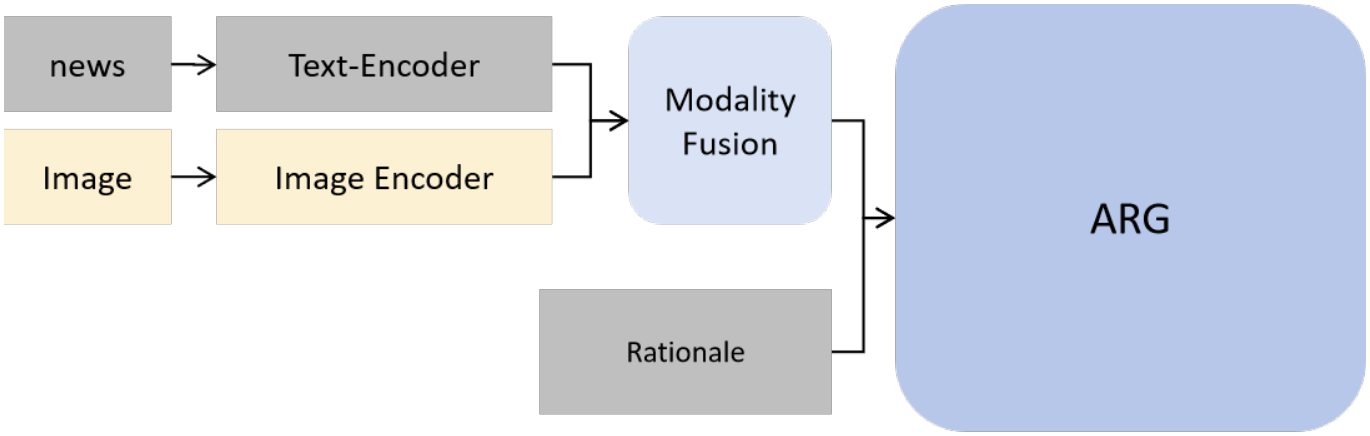

图 3. ARGVL 模型

本次复现工作主要的创新点集中在对 ARG 模型的改进，改进的整体方向为多模态数据的引入与融合，由于 ARG 模型作者本身使用了文字模态，所以此次复现工作的主要改进方向集中在图片模态的融入。然而由于作者提供的数据集中文字模态并未与图片模态相对应因此如果需要加入图片模态则必须从头生成对应的数据集。经过上文中提及的使用 Qwen 大模型生成三个数据集的 Rationale 之后，本文在此提出 ARG VL 模型，该模型在 ARG 模型的基础上接受图像模态作为另一种输入，其模型结果如图3，其基本思路目前比较简单，仅仅通过模态融合模块将原始新闻文本模态特征和图像模态特征融合替代原先 ARG 模型中文本模态特征，其中模态融合模态采用了多层双重交叉注意力，其模型结构如图4，其数学表达如公式1415，其中 I_t, I_v 分别为文字模态特征和图像模态特征，经过多层的融合后将两个特征张量在第 1 维上进行拼接。

$$I_t = MLP(CA(I_t, I_v, I_v) + I_t) + I_t \tag{14}$$

$$I_v = MLP(CA(I_v, I_t, I_t) + I_v) + I_v \tag{15}$$

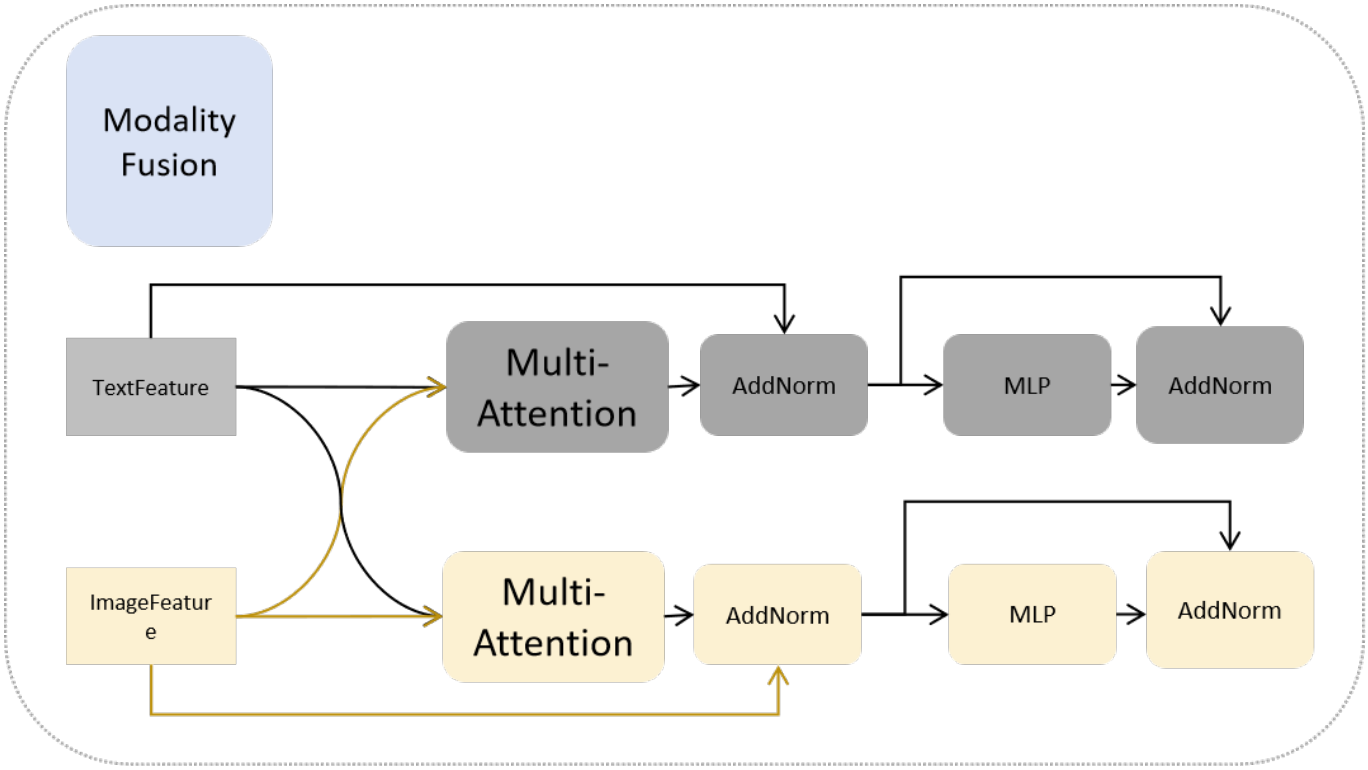

图 4. 双重交叉注意力

5 实验结果分析

2*Model	Chinese				English			
	macF1	Acc.	$F1_{real}$	$F1_{fake}$	macF1	Acc.	$F1_{real}$	$F1_{fake}$
G1: LLM-Only GPT-3.5-turbo	0.725	0.734	0.774	0.676	0.702	0.813	0.884	0.519
G2: SLM-Only								
Baseline	0.753	0.754	0.769	0.737	0.765	0.862	0.916	0.615
$EANN_T$	0.754	0.756	0.773	0.736	0.763	0.864	0.918	0.608
Publisher-Emo	0.761	0.763	0.784	0.738	0.766	0.868	0.920	0.611
ENDEF	0.765	0.766	0.779	0.751	0.768	0.865	0.918	0.618
Baseline + Rationale	0.767	0.769	0.787	0.748	0.770	0.870	0.921	0.633
SuperICL	0.757	0.759	0.779	0.734	0.736	0.864	0.920	0.551
G3: LLM+SLM								
ARG (原始论文结果)	**0.784**	**0.786**	**0.804**	**0.764**	**0.790**	**0.878**	**0.926**	**0.653**
	(+4.2%)	**(+4.3%)**	**(+4.6%)**	**(+3.8%)**	**(+3.2%)**	**(+1.8%)**	**(+1.1%)**	**(+6.3%)**
ARG (复现结果)	0.780	0.782	0.799	0.761	0.795	0.881	0.928	0.662
	(-0.57%)	**(-0.62%)**	**(-0.65%)**	**(-0.50%)**	**(+0.63%)**	**(+0.41%)**	**(+0.02%)**	**(+1.38%)**
w/o LLM Judgment Predictor	0.773	0.773	0.789	0.756	0.786	0.880	0.928	0.645
w/o Rationale Usefulness Evaluator	0.781	0.783	0.801	0.761	0.782	0.873	0.923	0.641
w/o Predictor & Evaluator	0.769	0.770	0.782	0.756	0.780	0.874	0.923	0.637
ARG-D (原始论文结果)	0.771	0.772	0.785	0.756	0.778	0.870	0.921	0.634
	(+2.4%)	**(+2.3%)**	**(+2.1%)**	**(+2.6%)**	**(+1.6%)**	**(+0.9%)**	**(+0.6%)**	**(+3.2%)**
ARG-D (复现结果)	0.759	0.760	0.773	0.746	0.779	0.866	0.9188	0.625
	(-1.43%)	**(-1.47%)**	**(-1.44%)**	**(-1.32%)**	**(+0.78%)**	**(-0.39%)**	**(-0.24%)**	**(-1.42%)**

表 7. 原文结果以及复现结果

首先在原作者提供的数据集上进行实验，最终结果如表7, 基本上与原始论文给出的结果相似。从原始论文的结果中同样可以看出 ARG 模型的 Rationale Usefulness Evaluator 组件

Model	Weibo				Gossipcop				Twitter			
	macF1	Acc.	$F1_{real}$	$F1_{fake}$	macF1	Acc.	$F1_{real}$	$F1_{fake}$	macF1	Acc.	$F1_{real}$	$F1_{fake}$
G1: LLM-Only Qwen2/Qwen2.5												
Textual Description	0.820	0.823	0.788	0.851	0.571	0.781	0.871	0.276	0.554	0.578	0.451	0.657
Commonsense	0.816	0.817	0.788	0.844	0.565	0.785	0.874	0.256	0.558	0.569	0.488	0.627
G3: LLM+SLM												
ARG	0.880	0.880	0.876	0.884	0.796	0.870	0.919	0.673	0.356	0.543	0.703	0.009
ARG VL	0.881	0.881	0.882	0.880	0.788	0.859	0.910	0.666	0.525	0.565	0.387	0.663
	(**+0.114%**)	(**+0.114%**)	(**+0.685%**)	(**-0.453%**)	(**-1.005%**)	(**-1.26%**)	(**-0.979%**)	(**-1.040%**)	(**+47.4%**)	(**+4.05%**)	(**+-45.0%**)	(**+7266.67%**)

表 8. Qwen Rationale 数据集实验结果

并没有很好的过滤较差的 Rationale。

其次在 Qwen 生成的 Rationale 数据集上进行测试最终实验结果如表8，通过该结果可以分析得出：

1. **Weibo 数据集：**
   - Textual Description 和 Commonsense 模型的表现非常接近。
   - ARG 和 ARG VL 模型显著优于前两者。
   - ARG VL 在所有指标上都略高于 ARG。
2. **Gossipcop 数据集：**
   - Textual Description 和 Commonsense 模型的表现也非常接近。
   - ARG 模型在所有指标上都显著优于其他模型。
   - ARG VL 模型在某些指标上略低于 ARG，但在 $F1_{fake}$ 上表现更好。
3. **Twitter 数据集：**
   - Textual Description 和 Commonsense 模型的表现非常接近，且整体较低。
   - ARG 模型在 $F1_{real}$ 上表现较差，但在 $F1_{fake}$ 上表现较好。
   - ARG VL 模型在 $F1_{real}$ 上有所提升，但在 $F1_{fake}$ 上略有下降。
4. **总结：**
   - ARG 和 ARG VL 模型在 Weibo 和 Gossipcop 数据集上表现最好，尤其是在 $F1_{real}$ 和 $F1_{fake}$ 上。
   - ARG VL 模型在某些情况下略微优于 ARG，特别是在 $F1_{fake}$ 上。
   - Twitter 数据集上的模型表现普遍较低，尤其是 $F1_{real}$ 指标，这可能表明该数据集的难度较高或模型对这类数据的适应性较差。
   - 这些数据特征显示了不同模型在不同类型数据集上的表现差异，有助于进一步优化模型和选择最适合特定任务的模型配置。

6 总结与展望

6.1 研究总结

本研究报告探讨了大型语言模型（LLMs）在假新闻检测中的双重角色，即作为“坏演员”（Bad Actor）和“好顾问”（Good Advisor）。通过复现原始论文《Bad Actor, Good Advisor: Exploring the Role of Large Language Models in Fake News Detection》的研究成果，我们验证了以下几点：

1. **LLMs 的局限性**：尽管 LLMs 展示了强大的文本处理能力，但在直接用于假新闻检测任务时表现不如经过特定任务微调的小型语言模型（SLMs）。尤其在零样本和少量样本提示条件下，LLMs 的表现存在不足。

2. **多角度分析的优势**：当引入链式思考提示（Chain-of-Thought Prompting）时，LLMs 能够从多个视角提供解释性理由，显著提升了其在假新闻检测任务上的表现。然而，选择和整合这些理由的能力仍然有限。

3. **幻觉问题的影响**：LLMs 在生成内容时可能出现“幻觉”，即产生不符合事实的陈述，这对其在假新闻检测任务中的可靠性提出了挑战。

4. **辅助工具的角色**：通过引入视角特定的提示并执行基于规则的判断集成，可以显著提升 LLMs 的表现，表明 LLMs 可以作为有效的辅助工具而非替代方案。它们提供了有指导性的推理来辅助检测假新闻。

5. **ARG 模型的有效性**：ARG 模型结合了 LLMs 生成的多角度分析原理（Rationale）和 SLMs 的灵活且强大的多角度结合能力，达到了比单个模型更好的性能。特别是 ARG VL 模型通过引入图像模态进一步增强了检测效果。

6.2 未来展望

基于上述研究成果，提出以下未来研究方向：

1. **改进模型鲁棒性**：开发更先进、可靠的假新闻检测工具，尤其是针对 LLMs 的幻觉问题，提高模型的鲁棒性和准确性。

2. **多模态融合**：探索更多类型的模态（如音频、视频等）与文本信息的融合，以增强假新闻检测系统的综合判断力。

综上所述，《Bad Actor, Good Advisor: Exploring the Role of Large Language Models in Fake News Detection》这一选题不仅揭示了 LLMs 在假新闻检测领域的潜力和局限性，还为未来的研究和发展提供了宝贵的理论支持和技术路径。此次复现工作中 ARG 模型[1]和 Qwen Rationale 数据集生成[2]相关代码已在 Github 上开源。

[1]https://github.com/LYQ1-ai/ARGVL

[2]https://github.com/LYQ1-ai/QwenVLRationaleGenerate

参考文献

[1] Christina Boididou, Symeon Papadopoulos, Duc Tien Dang Nguyen, G. Boato, Michael Riegler, Andreas Petlund, and Ioannis Kompatsiaris. Verifying multimedia use at mediaeval 2016. 10 2016.

[2] Tri Dao. FlashAttention-2: Faster attention with better parallelism and work partitioning. In *International Conference on Learning Representations (ICLR)*, 2024.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[4] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *ArXiv*, abs/1503.02531, 2015.

[5] Beizhe Hu, Qiang Sheng, Juan Cao, Yuhui Shi, Yang Li, Danding Wang, and Peng Qi. Bad Actor, Good Advisor: Exploring the Role of Large Language Models in Fake News Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 22105–22113, 2024.

[6] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

[7] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[8] Erxue Min, Yu Rong, Yatao Bian, Tingyang Xu, Peilin Zhao, Junzhou Huang, and Sophia Ananiadou. Divide-and-conquer: Post-user interaction network for fake news detection on social media. In *Proceedings of the ACM Web Conference 2022*, WWW '22, page 1148–1158, New York, NY, USA, 2022. Association for Computing Machinery.

[9] Qiong Nan, Juan Cao, Yongchun Zhu, Yanyan Wang, and Jintao Li. Mdfend: Multi-domain fake news detection. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 3343–3347, 2021.

[10] Van-Hoang Nguyen, Kazunari Sugiyama, Preslav Nakov, and Min-Yen Kan. Fang: Leveraging social context for fake news detection using graph representation. In *Proceedings of*

the 29th ACM International Conference on Information & Knowledge Management, CIKM '20, page 1165–1174, New York, NY, USA, 2020. Association for Computing Machinery.

[11] OpenAI. ChatGPT: Optimizing language models for dialogue. https://openai.com/blog/chatgpt/, 2022. Accessed: 2023-08-13.

[12] Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. DeClarE: Debunking fake news and false claims using evidence-aware deep learning. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 22–32, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[13] Piotr Przybyla. Capturing the style of fake news. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01):490–497, Apr. 2020.

[14] Peng Qi, Juan Cao, Xirong Li, Huan Liu, Qiang Sheng, Xiaoyue Mi, Qin He, Yongbiao Lv, Chenyang Guo, and Yingchao Yu. Improving fake news detection by using an entity-enhanced framework to fuse diverse multimodal clues. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21, page 1212–1220, New York, NY, USA, 2021. Association for Computing Machinery.

[15] Qiang Sheng, Juan Cao, Xueyao Zhang, Rundong Li, Danding Wang, and Yongchun Zhu. Zoom out and observe: News environment perception for fake news detection. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4543–4556, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[16] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big data*, 8(3):171–188, 2020.

[17] Qwen Team. Qwen2.5: A party of foundation models, September 2024.

[18] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

[19] Xinyi Zhou and Reza Zafarani. Network-based fake news detection: A pattern-driven approach. *SIGKDD Explor. Newsl.*, 21(2):48–60, November 2019.