

XFacta：基于多模态语言模型的多模态虚假信息检测之当代、真实世界数据集与评估

Yuzhuo Xiao^{1,*} Zeyu Han^{2,*} Yuhan Wang³ Huaizu Jiang²

¹Guizhou University ²Northeastern University ³UC Santa Cruz

* Equal contribution

Abstract

社交媒体上多模态错误信息的快速传播需要更有效和更强大的检测方法。最近利用多模态大语言模型 (MLLMs) 的进展显示了在应对这一挑战中的潜力。然而，现有方法的瓶颈（证据检索对比推理）究竟在哪里仍不清楚，这阻碍了该领域的进一步进展。在数据集方面，现有的基准要么包含过时事件，导致由于与当代社交媒体场景的不一致而产生的评估偏差，因为 MLLMs 可以简单地记住这些事件，要么是人为合成的，无法反映真实世界的错误信息模式。此外，缺乏对基于 MLLM 的模型设计策略的全面分析。为了解决这些问题，我们引入了 XFacta，这是一个当代的、真实世界的数据集，更适合评估基于 MLLM 的检测器。我们系统地评估了各种基于 MLLM 的错误信息检测策略，评估了不同架构和规模的模型，并与现有的检测方法进行了基准测试。在这些分析的基础上，我们进一步启用了——一个半自动检测循环框架，该框架不断更新 XFacta 以包含新内容，以保持其当代相关性。我们的分析为推动多模态错误信息检测领域的发展提供了宝贵的见解和实践。我们的数据集和代码可在 <https://github.com/neu-vi/XFacta> 获取。

1 引言

谎言能够在真相穿上鞋子之前就走遍半个地球——这一说法在社交媒体时代显得尤为真实。随着平台信息传播的加速，人们在识别网上假新闻时面临越来越大的挑战。现代假新闻通常是多模态的，结合了文本和看似支持虚假或不相关事件的图像，这使得检测更加困难。深度伪造技术的兴起进一步降低了创作此类欺骗性内容的门槛。这些发展突显了开发更先进和更强健的方法以自动检测多模态虚假信息的必要性。

多模态大型语言模型 (MLLMs) 的出现具备在文本和图像中强大的推理能力，为检测多模态虚假信息提供了一个有前途的方向。最近的研究已经开始探索这种潜力。一些方法 (Qi et al., 2024; Liu et al., 2024a; Zeng et al., 2024; Shalabi et al., 2024) 在特定的虚假信息数据集上对通用目的的 MLLM 进行微调，以创建特定任务的模型。其他方法 (Khaliq et al., 2024; Xuan et al., 2024; Liu et al., 2024b; Geng et al., 2024) 采用零样本设置并依赖于更强大的模型，如 GPT-4 或 Gemini，这些模型在现有的虚假信息数据集上表现更好。一般来说，现有的基于 MLLM 的虚假信息检测器模仿人类的验证过程，涉及两个主要步骤：证据检索，即从互联网检索外部信息以作为证据，然后进行推理，即系统地分析和结合新闻发布和检索到的证据以做出最终判断。

尽管这些研究报告的结果令人鼓舞，但现有基于 MLLM（多模态大语言模型）的误导信息检测方法的瓶颈究竟在哪里（证据检索与推理）仍不清楚，阻碍了该领域的进一步发展。从数据集的角度来看，真实世界社交媒体上的误导信息通常涉及 MLLM 训练数据中缺乏的新颖和及时事件。检测这些事件需要模型积极检索证据并基于证据进行彻底的推理。相比之下，现有的误导信息基准 (Vlachos & Riedel, 2014; Wang, 2017; Thorne et al., 2018; Hanselowski et al., 2019; Khanam et al., 2021) 大多包含过时数据，其中的事件可能已经存在于 MLLM 的训练数据中，使得模型能够简单依赖记忆而不是基于证据的推理。这引入了显著的评估偏差，如图 1 所示的例子所证明。此外，一些数据集 (Luo et al., 2021; Chakraborty et al., 2023; Liu et al., 2024b; Shao et al., 2023; Aneja et al., 2021) 是合成的，意味着误导信息样本是利用 AI 模型人工构建的，而不是从现实世界来源中收集的。这限制了它们反映真实误导信息创建者所使用的复杂性和策略的能力。关于技术方法，尽管现有

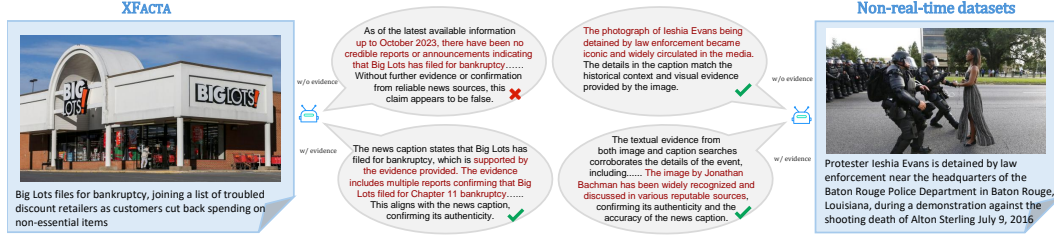


Figure 1: 左图：来自我们数据集的一个例子，其中 MLLM (GPT-4o) 必须依靠证据来判断真假。右图：来自非实时数据集的一个例子，其中证据的重要性较低。在 XFacta 上评估基于 MLLM 的虚假信息检测器引入的评估偏见较少。

Dataset	Multimodal	Contemporary	Real-world	Evidence-based annotations	Real Num	Fake Num
FEVER (Thorne et al., 2018)	✗	✗	✗	✓	93,367	43,107
LIAR (Wang, 2017)	✗	✗	✗	✓	7,085	5,751
NewsCLIPpings (Luo et al., 2021)	✓	✗	✗	✗	816,922	816,922
Fakeddit (Nakamura et al., 2019)	✓	✗	✗	✗	527,049	628,501
Snopes+Reuters Zlatkova et al. (2019)	✓	✗	✗	✓	592	641
DGM ⁴ (Shao et al., 2023)	✓	✗	✗	✗	77,426	152,574
FACTIFY 3M (Chakraborty et al., 2023)	✓	✗	✗	✗	406,000	316,000
MMFakeBench (Liu et al., 2024b)	✓	✗	✗	✗	3,300	7,700
COSMOS (Aneja et al., 2021)	✓	✗	✗	✓	1,700	1,700
Mocheg (Yao et al., 2023)	✓	✗	✗	✓	5,144	5,855
MediaEval (Boididou et al., 2016)	✓	✗	✓	✓	292	410
VERITE (Papadopoulos et al., 2023)	✓	✗	✓	✓	338	662
Post-4V (Geng et al., 2024)	✓	✓	✓	✓	81	105
XFacta (Ours)	✓	✓	✓	✓	1,200	1,200

Table 1: 不同错误信息数据集的比较。Contemporary 指的是 2024 年 1 月 1 日之后发布的数据；Real-world 意味着由实际用户创建的假帖子，而不是使用 AI 模型人工生成的；Evidence-based annotations 意味着有足够证据支持的注释以验证数据。

研究通常关注提出新的模型或方法并在特定数据集上展示其有效性，但它缺乏对 MLLM 检测不同设计选择的系统分析和严格比较。因此，至今仍难以识别构建可靠的多模态误导信息检测器的最佳实践或可推广的见解。

在本文中，我们通过整理一个新的错误信息数据集来解决这些限制，该数据集名为 XFacta (从 X (Twitter) 收集，并用于 Fact 的审核)。我们数据集中的所有数据点都在 2024 年 1 月之后，确保其具有现代相关性 (例如，比 GPT-4o 的 2023 年 10 月截止时间更为近期)。此外，它们来自社交媒体上的谣言传播者，反映了现实世界中观察到的模式。基于该数据集，我们分别从证据检索和推理的角度系统探索如何构建基于 MLLM 的错误信息检测器。此外，我们评估了不同架构和规模的各种 MLLM，以及现有的错误信息检测方法。通过这些实验和分析，我们提供了有关基于 MLLM 错误信息检测的宝贵见解。基于这些见解，我们应用得到的检测器标记新帖子，提供初步评估供人类审核者核实并添加到 XFacta。这种半自动化的检测循环保持数据集的时效性，防止其随着时间的推移而过时。我们相信，XFacta 数据集以及我们的研究结果将成为未来多模态错误信息检测研究的有用基准。

总之，我们的贡献是：

- 我们策划了一个用于多模态错误信息检测的当代现实世界数据集，并整合了一个半自动检测循环过程，以使其不断更新，这将进一步推动基于 MLLM 的检测研究。
- 通过 XFacta，我们从证据检索和推理两个角度，对开发一个好的基于 MLLM 的虚假信息检测模型进行了全面和深入的分析，为该领域提供了宝贵的见解。
- 我们对各种基于 MLLM 的错误信息检测策略进行了全面评估，评估了跨不同架构和规模的模型，并与现有的检测方法进行了基准测试。

2 相关工作

Datasets: 以往的研究引入了各种基于文本的单模态虚假信息数据集 (Vlachos & Riedel, 2014; Wang, 2017; Thorne et al., 2018; Hanselowski et al., 2019; Khanam et al., 2021)。社交媒体的兴起导致了对多模态虚假信息检测的关注增加，并发布了各种数据集 (Nakamura et al., 2019; Zlatkova et al., 2019; Yao et al., 2023; Boididou et al., 2016; Papadopoulos et al., 2023)。然而，现实世界的虚假信息数据集通常要么规模较小，要么面临噪声标注的问题。



Figure 2: XFacta 中错误信息类型、主题和发布时间的例子与分布。

因此，一些其他的工作 (Luo et al., 2021; Chakraborty et al., 2023; Liu et al., 2024b; Shao et al., 2023; Aneja et al., 2021) 利用启发式规则或 AI 模型来合成虚假信息检测数据集。由于这种合成，这些数据集未能捕获现实世界中虚假信息创作者的复杂模式和策略。此外，以上所有数据集都并非当代的，且通常与 MLLMs 的训练数据重叠，这阻碍了基于 MLLM 的虚假信息检测器的公平和稳健评估。Post-4V (Geng et al., 2024) 通过收集更近期的例子来应对这一问题，但其规模非常有限，且数据收集和处理的细节记录不足，这使得它不适合作为广泛接受的基准。相比之下，我们的 XFacta 数据集确保了同时具有当代性和现实世界特征，同时保持了足够评估 MLLMs 零次设置所需的中等规模。此外，我们的数据集为假新闻提供了详细的记者证据，可以帮助验证检测模型的推理路径。针对不同数据集的多维度比较可以在表格 1 中找到。

一些传统的多模态错误信息检测器 Abdelnabi et al. (2022); Yuan et al. (2023); Brahma et al. (2023a); Aneja et al. (2023); Mu et al. (2023); Zhang et al. (2023); Brahma et al. (2023b); Yang et al. (2024) 是在特定的数据集上进行了训练和评估的，例如常用的 NewsCLIPPings 数据集 (Luo et al., 2021)。随着开源 MLLM 的出现，最近有一些工作 (Qi et al., 2024; Liu et al., 2024a; Zeng et al., 2024; Shalabi et al., 2024) 采用了一种不同的方法，即在错误信息数据集上微调预训练的 MLLM，从而获得更好的性能。然而，这些方法通常带有其训练数据特有的偏见，无法应对每天在社交平台上出现的新、更复杂的错误信息。因此，一些研究探索了更强大的闭源 MLLM，并取得了更好的结果。然而，这些模型要么声称无证据 (Geng et al., 2024)，要么在更新不足或现实世界的数据集上进行评估 (Khaliq et al., 2024; Xuan et al., 2024; Liu et al., 2024b; Jin et al., 2024)，这对其在不断发展的社交媒体部署时的有效性提出了质疑。

3 我们的 XFacta 数据集

多模态虚假信息检测是指评估包含支持图片和文本的新闻帖子的真实性。正式来说，给定一组支持图片 $I = \{I_1, \dots, I_n\}$ 和一个文本声明 T ，此任务是判断帖子 $\mathcal{P} = (I, T)$ 是真实的还是假的。

支持图像 I 可以让文本声明 T 看起来更可信，即使它们没有关联或具有误导性，这使得检测比在单模态环境中更困难。因此，大多数方法将检索到的证据 $\mathcal{E} = (E_i, E_t)$ 纳入它们的检测流程，其中 E_i and E_t 分别是图像类型和文本类型的证据。

3.1 数据来源 & 集合

我们的数据集来源于 X/Twitter。真实新闻帖子来自权威新闻机构，包括 CNN、Fox News、The Guardian 和 BBC。假新闻帖子则是由 BBC 认证记者标记为虚假的内容和 X Community Notes 精选而来。

我们首先收集虚假新闻帖子，因为它们较为少见且需要仔细识别，之后我们收集大约五倍的真实帖子，以确保选择的多样性。这使我们能够采样一个子集的真实帖子，使其在数量和

分布上与虚假帖子相匹配，从而减少潜在的评估偏差。我们在两个方面确保分布的一致性：(1) Topic-aligned selection，在这个步骤中，我们为真实和虚假的帖子标记主题。然后，我们确保每个主题的真实和虚假帖子数量相同，这有助于通过保持内容的语义对齐来减少语义差异。关于帖子的主题的详细描述将在第 3.2 节中提供。(2) Image similarity selection：前一步主要关注文本陈述 T 的对齐，这里我们处理图像 I 的对齐。我们使用 SigLip (Zhai et al., 2023) 来提取图像特征，并应用最优传输算法 (Genevay et al., 2016) 来选择图像特征分布最接近虚假帖子的真实帖子。这种对齐有助于最小化由视觉差异引起的偏差，从而确保评估准确反映模型从文本和视觉角度检测虚假信息的能力。

此外，为了确保新闻帖子的标签具有可靠性，除了帖子内容 \mathcal{P} 之外，还提供了对应的元数据，包括帖子作者 ID、帖子 URL、日期、主题等。对于假新闻帖子，我们还收集标识这些帖子为假的标注，包括理由和证据，并基于这些标记帖子，我们还注释了错误信息的类型，更多细节在第 3.2 节中提供。我们对每个条目进行人工审核，只有那些具有明确虚假信息证据的才被包含在数据集中。

3.2 数据统计 & 分析

我们的 XFacta 数据集共有 2400 个数据点，包括 1200 个真实帖子和 1200 个假帖子。为了方便模型开发，我们随机选择了 120 个真实帖子和 120 个假帖子作为 Dev 集，而其余的 2160 个帖子用作 Test 集。如图 2 的右下角所示，所有数据收集于 2024 年 1 月至 2025 年 4 月，其中大部分数据是在 2024 年 9 月之后收集的。这种当代性质确保了数据集反映出真实和虚假新闻的最新模式和发展特征。

为了更好地理解数据集，我们根据主题和假新闻中的错误信息类型对每个新闻帖进行标注。对于主题分类，每个帖子的主题 \mathcal{P} 被分类为以下类别之一：政治、社会、娱乐、科学、历史、自然和体育，如图 2 左下角所示。值得注意的是，与政治和冲突相关的错误信息占主导地位，但也伴随着其他领域的信息，这与当前的全球趋势紧密一致。

对于假帖子，如图 2 顶部所示，我们根据收集到的标记帖子中提供的明确证据，从三种预定义的错误信息类型中分配一个或多个标签。我们不基于提供证据之外的推测或假设来分配标签。这三种错误类型定义如下：

- 深度伪造：图像根据标记帖子被识别为生成或数字操控的。
- 图像上下文不符 (OOC)：图像是真实的，但根据标记的帖子，图像来源于与附带文本中描述事件不同的事件。这并不表明文本的真实性。
- 文本误导：文本内容传达了一个已经被标识帖子明确指出为虚假的声明。这并不表明图像是否真实或相关。

通过用这些更细粒度的虚假信息标签注释每个虚假帖子，我们实现了对多模态虚假信息特征的更细致理解，并使我们能够对虚假信息检测器在不同虚假信息类型上的性能进行更详细的分析。

4 如何构建一个基于多模态大模型的有效错误信息检测器？

在本节中，我们探索了基于 MLLM 的错误信息检测在 XFacta 数据集上的不同设计策略。我们主要探讨两个问题：(1) 不同类型的证据如何对错误信息检测产生影响，以及如何更好地利用它们；(2) 不同的 LLM 推理方法如何影响模型的预测。

对于要验证的给定帖子 \mathcal{P} ，我们假设检索到的证据可以在两个主要方面帮助检测模型：(1) 验证帖子中描述的事件的真实性，以及 (2) 验证附带的图像是否在脱离背景的情况下使用。基于这些假设，我们引入了八种旨在支持这些目标的证据检索策略：

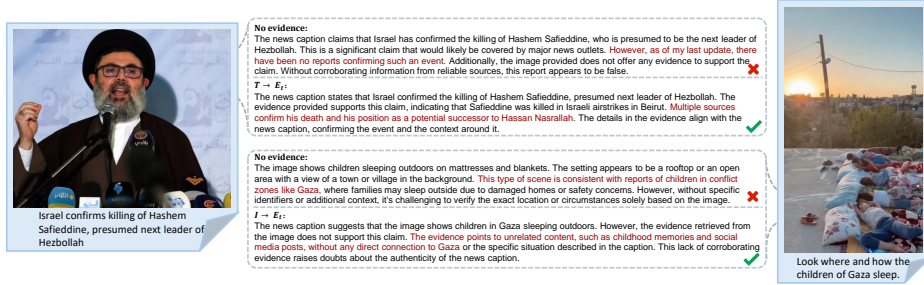
- ① 单模态证据：使用帖子文本 T 检索文本证据 E_t 以支持方面 (1)。这模仿了人类通过在线搜索相关信息来验证新闻的方法。
- ② - ③ 跨模态证据：分别使用帖子文本 T 和图像 I 来检索图像类型证据 E_i (策略 ②) 和文本类型证据 E_t (策略 ③)，遵循 Abdelnabi et al. (2022) 中的跨模态检索方法以支持方面 (2)。
- ④ - ⑤ LLM 查询：使用 LLM 生成关于帖子中不确定或可疑细节的问题，然后形成搜索查询以检索图像型证据 E_i (策略 ④) 和文本型证据 E_t (策略 ⑤)。这模拟了人类通过提出有针对性的问题来调查不清楚的说法的方式。

Evidence Type	GPT-4o			Gemini-2.0-flash			Qwen-vl-7b		
	Acc.	R. Acc.	F. Acc.	Acc.	R. Acc.	F. Acc.	Acc.	R. Acc.	F. Acc.
no evidence	70.8	50.8	90.8	71.7	78.3	65	60.8	76.7	44.4
Google Search									
① $T \rightarrow E_t$	87.1	97.5	76.7	81.3	98.3	64.2	59.7	82.7	38.1
② $T \rightarrow E_i$	81.7	75.8	87.5	77.9	90.8	65	62.2	84.2	40.2
③ $I \rightarrow E_t$	77.9	70	85.8	78.8	83.3	74.1	55.7	71.7	39.3
④ Query $\rightarrow E_i$	69.2	51.7	86.7	71.9	76.5	67.2	55.8	91.2	20
⑤ Query $\rightarrow E_t$	77.5	80	75	77.7	83.9	71.7	56.1	63.9	48.1
DuckDuckGo Search									
⑥ $T \rightarrow E_t$	84.2	94.2	74.2	79.2	97.5	60.8	64	91	37
⑦ $T \rightarrow E_i$	76.3	64.2	88.3	76.7	87.5	65.8	53	79	26.5
⑧ $T \rightarrow E_{news}$	84.2	80	88.3	75.3	92.4	58.3	68.4	84.4	52.9

Table 2: 不同证据检索方法下 MLLM 性能比较。

Method	Acc.	R. Acc.	F. Acc.
$T \rightarrow E_t$	87.1	97.5	76.5
Domain Filter	88.3	98.3	78.3
Evidence Extraction	87.5	95.8	79.2
$T \rightarrow E_i$	81.7	75.8	87.5
Domain Filter	83.8	80.8	86.7
Evidence Extraction	82	77	87
$I \rightarrow E_t$	77.9	70	85.8
Domain Filter	79.6	72.5	86.7
Evidence Extraction	81.3	75	87.5

Table 3: 不同证据后处理方法与 GPT-4o 的比较。

Figure 3: $T \rightarrow E_t$ 和 $I \rightarrow E_t$ 如何纠正无证据检测错误的例子。

- ⑥ - ⑧ DuckDuckGo 变体：为了探索不同搜索引擎如何影响检索结果，我们在策略 ① 和 ② 中分别用 DuckDuckGo 替换策略 ⑥ 和 ⑦ 中使用的搜索引擎。我们还使用 DuckDuckGo 的“搜索新闻”功能进行新闻证据 E_{news} （策略 ⑧），调查其是否能检索到更具权威性的证据。

此外，我们相信后期处理可以帮助清理证据以减少其噪声。在此，我们提出了两种受 Xuan et al. (2024) 启发的证据后期处理方法：

- 领域过滤器：滤除来自不可信域的证据。¹
- 证据提取：使用多模态语言模型（本论文中为 GPT-4o）选择与新闻帖高度相关的证据部分，并去除不相关的部分。

为了评估每种证据类型的影响，我们首先运行模型，不使用任何证据，仅依赖于 MLLM 的内部知识。然后，我们分别添加八种证据类型，并与无证据基线和彼此进行比较。我们使用链式思维 (CoT) (Wei et al., 2022) 提示以获得可解释的推理输出，而非简单的二元决策。实验在 Dev 集上进行，使用三个不同规模的模型：GPT-4o (Yang et al., 2023)、Gemini-2.0-Flash (Team et al., 2023) 和 Qwen-VL-7B (Wang et al., 2024)，以减少模型特定的偏差。在后处理阶段，我们分别测试 GPT-4o 在每种策略下的表现。我们报告三个指标：总体准确度 (Acc.)、真实帖子上的准确度 (R. Acc.) 和虚假帖子上的准确度 (F. Acc.)。模型还输出置信度评分 (0-100)，我们在某些表中报告平均置信度 (Avg. Conf.) 以反映预测的确定性。

4.0.1 结果与分析

表 2 和 3 分别展示了不同证据检索和后处理策略的表现。表 4 对假贴文的错误信息类型进行了更详细的比较。我们总结了几个关键观察如下。

1. 所有类型的证据都能在无证据的基准线之上持续提高准确性。没有证据时，模型在行为上表现出显著的差异：GPT 更为保守，而 Gemini 和 Qwen 则更倾向于将帖子标记为真实。有了证据后，它们的分类变得更加平衡，显示出外部证据在虚假信息检测中的重要性。

¹默认情况下，从数据集整理中使用的领域证据被排除在外以避免泄露。

Evidence Type	Deepfakes		Image OOC		Text Misleading	
	Acc.	Avg. Conf.	Acc.	Avg. Conf.	Acc.	Avg. Conf.
no evidence	89.7	87.4	93.8	77	91.1	82.6
Google Search						
① $T \rightarrow E_t$	79.3	87.8	77.1	84	80.4	87
② $T \rightarrow E_i$	93.1	85.9	85.4	81	87.5	83
③ $I \rightarrow E_t$	100	88.5	83.3	85.2	80.4	88.2
④ Query $\rightarrow E_i$	89.7	88.8	79.2	82.9	89.3	85.4
⑤ Query $\rightarrow E_t$	86.2	91.2	60.4	88.7	82.1	89.6
DuckDuckGo Search						
⑥ $T \rightarrow E_t$	79.3	90	64.6	85.5	82.1	86.1
⑦ $T \rightarrow E_i$	89.7	85.9	83.3	79.6	91.1	80.9
⑧ $T \rightarrow E_{news}$	93.1	84.8	81.3	80.6	92.9	80.8

Table 4: 使用 GPT-4o 比较不同证据检索策略在各类虚假信息之间的效果。

Reasoning Method	GPT-4o			Gemini-2.0-flash			Qwen-vl-7b		
	Acc.	R. Acc.	F. Acc.	Acc.	R. Acc.	F. Acc.	Acc.	R. Acc.	F. Acc.
Chain of Thought	88.3	98.3	78.3	83.8	98.3	69.2	54.8	84.2	24.1
Prompt Ensembles	90	100	80	85.4	98.3	72.5	67.1	90	44
Self Consistency	88.3	97.5	79.2	86.7	98.3	75	61	64	58.2
Multi-step Reasoning	91.3	91.7	90.8	81.3	90	72.5	62.1	78.4	45.9

Table 5: 使用各种推理方法在 Dev 集合上比较 MLLM 性能。

2. $T \rightarrow E_t$ (策略 ①) 显著提升了性能, 特别是在真实帖子的情况下。这是可以预期的——即使对人类来说, 真实新闻更有可能得到在线证据的支持, 因此更容易验证。参见图 3 左侧的例子。然而, 假帖子的准确性并没有显著提高, 对于 GPT-4o 而言甚至略有下降。我们将这归因于 OOC 虚假信息, 其中 $T \rightarrow E_t$ 没有提供关于图像 I 的信息, 而 E_t 中对 T 的强支持误导模型, 颠覆其对假帖子的原本正确预测。

3. 针对情境外的错误信息, $I \rightarrow E_t$ (策略 ③) 比 $T \rightarrow E_i$ (策略 ②) 更有效。尽管表格 2 中的 $T \rightarrow E_i$ 显示了更高的整体准确性, 但手动检查表明 $I \rightarrow E_t$ 对情境外的情况检测更佳。这是因为 $I \rightarrow E_t$ 直接检索包含查询图像的网页并提取高度相关的文本, 而 $T \rightarrow E_i$ 基于标题进行模糊搜索, 常常检索到关系松散的图像。此外, 在这些情况下, 文本证据通常信息量更大, 因为基于图像的对比通常仅限于一般场景或人物等粗略特征。这些表面相似性通常在情境外错误信息中被保留, 从而使得通过图像类型的证据检测操控变得困难。第 5 点还进一步分析了策略 ③ 在虚假帖子上的表现。

3. LLM 生成的查询 (策略 ④ 和 ⑤) 不如直接使用字幕搜索有效。在大多数情况下, LLM 无法生成高度针对性的查询; 大多数查询只是原始字幕的改写。因此, 使用这些改写版本进行搜索不如直接使用字幕 T 本身检索证据来得准确。在某些假帖子中, 如果 LLM 提出的问题或疑虑未能针对帖子为什么是假的实际原因, 检索到的证据甚至可能导致模型做出错误判断, 如下一个要点中进一步分析的那样。

. 在不同的虚假信息类型中, $I \rightarrow E_t$ (策略 ③) 提供了一贯具有信息量的证据, 特别是在识别假帖子时。不像早先基于总体准确性的分析, 对假帖子的细粒度虚假信息类型的分析需要超越仅仅考虑准确性。GPT-4o 倾向于保守地将模棱两可的帖子分类为假, 甚至没有额外的证据, 这可能会提高假帖子的准确性。因此, 平均置信度分数变得至关重要, 因为它们表明所检索的证据是否提供了真正有助于模型判断的明确和信息丰富的知识。如表格 4 所示, 策略 ③ 不仅取得了高准确率, 而且在 Deepfakes 和 Image OOC 类别中始终保持高置信度。虽然查询 $\rightarrow E_t$ (策略 ⑤) 在文本误导类别中表现出稍优的综合表现, 但它导致模型在 Image OOC 情况下做出了置信度很高但错误的预测, 从而显著降低了其整体效用。因此, $I \rightarrow E_t$ 在各种虚假信息类型中仍然是最佳的证据检索策略。

6. DuckDuckGo 提供的证据质量比 Google 低 (strategies ⑥ - ⑧)。实验表明, 与 Google 搜索相比, 使用 DuckDuckGo 检索的证据往往表现较差。此外, $T \rightarrow E_{news}$ 并未产生预期的改进。

7. 域过滤器可以减少证据噪声。如表 3 所示, 域过滤器可以总体上提高准确性, 这表明来自低可信度网站的证据确实是噪声且可能具有误导性。

8. 基于 LLM 的证据提取可以减轻证据噪声。我们检查了提取结果, 发现 LLM 可以成功保留检测虚假信息所需的关键信息, 并过滤掉一些不相关的证据, 这导致了检测准确性的提

Model	Scale	Acc.	R. Acc.	F. Acc.
GPT	GPT-4o-mini	83	84.6	81.3
	GPT-4o	88.6	87.6	89.6
Gemini	Gemini-2.0-lite	76.2	77.2	75.2
	Gemini-2.0-flash	78.9	83.6	74.2
Qwen	Qwen-vl-7b	65	80.9	48.5
	Qwen-vl-72b	81	82.3	79.6

Table 6: 不同 LLM 在 Test 集合上的检测性能比较。

Methods	Training Set / LLM	Acc.	R. Acc.	F. Acc.
SENS	NewsCLIPPings	52.6	58.4	46.7
Mocheg	Mocheg	51.7	57.6	45.5
HAMMER	DGM [†]	57.2	78.2	36.1
Sniffer	NewsCLIPPings	56.1	71	41.1
MMFakeBench *	GPT-4o	68.2	61.5	75.6
LEMMA	GPT-4o	77.3	63.9	90.8
Ours	GPT-4o	88.8	87.2	90.4

Table 7: 对 Test 集的不同虚假信息检测方法的比较。*：由于输出模糊，排除 284 个样本。

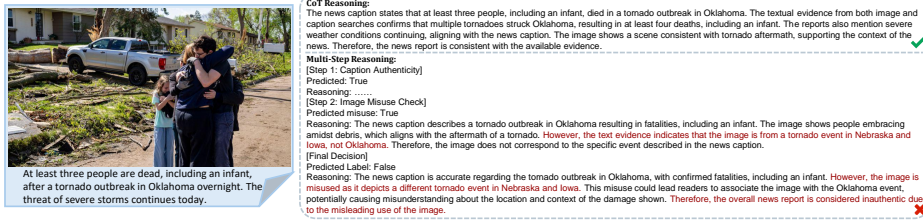


Figure 4: 多步骤推理可以检测 CNN 新闻帖子中的图片“误用”。我们认为这种“过于严格”的行为实际上有助于可靠的虚假信息检测。

升，如表 3 特别是对于 $I \rightarrow E_t$ 。然而，需要注意的是，证据提取会引入巨大的 token 开销。

4.1 推理分析

4.1.1 实验设置

基于对证据检索策略的分析，我们在推理阶段使用 $T \rightarrow E_t$ （策略 ①）和 $I \rightarrow E_t$ （策略 ③），因为它们可以很好地互相补充。我们还应用领域过滤器以减少证据噪声，但跳过证据提取，以便更好地评估模型在噪声证据上的推理能力。我们测试了四种推理策略，包括 CoT（Wei et al., 2022）、提示集合（Geng et al., 2024）、自一致性（Wang et al., 2022）和多步推理。更多详细信息请参阅附录。

根据表格 5 中报告的结果，我们在下面总结了几个关键观察。

1. MLLM 越强，它受到不同推理方法的影响就越小。像 GPT-4o 这样的较强模型通常默认具有良好的推理能力，并且在不同推理技术上有类似的准确性。
2. 不同的模型架构对不同的推理方法表现出不同的偏好。因此，在实践中，部署基于 MLLM 的错误信息检测器应包括测试各种推理方法，特别是对于较小的 MLLM，以获得更好的性能。
3. 对于 GPT-4o，多步推理在整体上具有最佳平衡准确度。对于表现最佳的模型 GPT-4o，通过对各种策略的推理路径进行人工检查，我们发现多步推理始终提供了最清晰且结构化的推理。尤其是在检测假新闻方面，其准确度优于其他方法。然而，在真实新闻上的准确度却不太理想。有趣的是，我们发现一些来自信誉良好新闻来源的真实帖子可能会使用来源于不相关事件的图片（我们不认为这是误导信息，因为没有误导的意图）。多步推理可以识别并标记这些由于图片不匹配而导致的假案例。我们认为这种“过于严格”的行为对于可靠的误导信息检测实际上是有益的。图 4 中可以找到一个例子。

在本节中，我们对 XFacta 进行进一步评估，以全面探讨 MLLM 性能、不同的误信息检测方法和数据集特定的特性。

不同 MLLM 的比较。我们在 Test 集上评估各种 MLLM 的性能。具体来说，我们分别分析了闭源模型（GPT 和 Gemini）以及开源模型（Qwen）在不同模型规模下的性能差异。我们遵循第 4.1 节的证据类型并使用多步骤推理策略。结果如表 6 所示。对于相同的模型架构，较大的模型总是能够获得更高的准确率。

现有多模态虚假信息检测方法的比较。我们使用我们的 Test 集对现有多模态虚假信息方法进行横向比较,包括从头开始训练的模型:SEnS (Yuan et al., 2023)、Mocheg (Yao et al., 2023) 和 HAMMER (Shao et al., 2023); 微调 MLLMs 的方法: Sniffer (Qi et al., 2024), 以及使用闭源 MLLMs 的零样本方法: MMFakeBench (Liu et al., 2024b)、LEMMMA (Xuan et al., 2024)。表格 7 中的结果显示,针对特定数据集训练的专业方法存在严重的泛化问题。相反,使用 GPT-4o 的模型表现出相对良好的性能。基于对证据检索和推理策略的系统分析,我们的方法优于这些模型,在 XFacta 上建立了 SOTA 准确度。

相似数据集的比较。我们研究了相似数据集中对证据的依赖性。为此,我们分别从 (Papadopoulos et al., 2023)、Snopes+Reuters Zlatkova et al. (2019) 和 NewsCLIPPings (Luo et al., 2021) 数据集中选择一个子集。我们按照第 4.1 节中的证据类型,并应用带有 GPT-4o 的 CoT 推理。结果在附录的表 8 中报告。值得注意的是,对于我们的 XFacta 数据集,没有任何证据的情况下, GPT-4o 运作不佳,这验证了我们建立一个需要证据检索来捕获及时事件的当代现实世界基准的动机。

检测器在较新的和分布外数据上的有效性。我们评估训练在原始 XFacta 数据集上的错误信息检测器在应用于较新的和分布外社交媒体内容时是否仍然有效。首先,我们选择 Snopes 作为测试平台,因为与 X 相比,它是分布外的数据。此外,该网站提供由专业记者提供的真实/虚假注释,因此可以作为评估检测器性能的参考。我们从 Snopes 收集了 2024 年 7 月至 2025 年 7 月间的 1200 条事实核查的新闻(600 条真实新闻和 600 条虚假新闻),这些新闻比原始 XFacta 数据集更新。我们的检测器在该数据集上实现了 89.2 % 的整体准确率。具体来说,它在真实声明上达到了 85.5 % 的准确率,在虚假声明上达到了 93.0 % 的准确率,显示出模型在最新和分布外数据上表现良好。

在本节中,我们展示了经过原始 XFacta 数据集验证的虚假信息检测器如何能有效地用于支持数据集扩展。以往的实验表明,该检测器在较新的和分布外的数据上都能保持稳定的性能,这表明它能够很好地推广到不断出现的、先前未见过的内容。因此,它可以集成到数据集收集管道中,帮助人工审核人员验证虚假信息,从而实现一个加速并扩展数据整理流程的检测循环框架。

为了验证这一想法,我们进行了一个案例研究以证明概念的可行性。这一次,我们不依赖记者标注的帖子或官方新闻账号的帖子进行真实/虚假的参考。相反,我们选择了一些定期发布关于热门或有争议话题并拥有大量粉丝的账号。我们抓取并识别了这些账号在 2025 年 6 月至 2025 年 7 月之间的 500 条帖子。其中,我们的检测器将 265 条帖子识别为虚假,将 235 条识别为真实。对于每个预测,检测器还生成了一个解释来支持其决策。该解释可以帮助人类审核员更有效地验证预测,决定是否将这些帖子纳入 XFacta。示例可在附录中找到。作为我们公开发布的一部分,附加数据集已与主要的 XFacta 数据集一起发布。

在本文中,我们介绍了一种用于多模态虚假信息检测的现代真实世界数据集 XFacta。使用该数据集,我们从两个角度分析如何构建一个有效的基于 MLLM 的虚假信息检测器:证据检索和推理。我们的实验提供了开发强大检测系统的实用见解。此外,我们实施了一个半自动检测循环以便不断用新标记的内容更新 XFacta。我们还在更现实的设置中使用我们的数据集对现有 SOTA MLLMs 和检测方法进行了基准测试。我们相信,XFacta 和我们的研究结果将促进多模态虚假信息检测的未来研究。

5

局限性

为了实现现实世界的特征,我们的数据集中所有假数据均来自社交媒体上的实际帖子,其作为假的标签是通过人工确定的。因此,与通过自动化方法收集的虚假信息数据集相比,该数据集在规模上不可避免地较小,也不太适合支持模型训练。然而,我们的数据集规模已经足够,可以作为一个测试集用于评估 MLLM 在零样本设置中的检测性能。

随着 MLLMs 的不断演进,该数据集可能会最终变得过时或与未来 MLLMs 的训练数据重叠。尽管如此,我们的初步基准测试和分析已经为构建多模态虚假信息检测器提供了宝贵的指导和实用的标准。我们也希望未来的工作可以利用开发的检测模型从网上半自动化地识别和收集假新闻,以进一步扩展数据集并保持其现当代性。

由于有限的令牌使用和高成本的闭源 MLLMs,我们无法对所有不同策略组合进行基准测试,因此我们的评估可能不会得出最佳解决方案,因为不同策略之间可能存在相互作用。我们期待未来的工作能够解决这些限制。

伦理声明 我们的研究遵循 Twitter 开发者条款²中规定的指南。我们确保我们的数据收集和使用符合这些条款，包括对 Twitter API 的合理使用。虽然我们计划出于研究目的发布我们的数据集，但我们将以遵循所有适用规则和指南的方式进行发布。

我们的研究集中在检测多模态错误信息，这是数字时代的一个重大问题。通过识别和减少错误信息的传播，我们的工作为维护网络信息的完整性作出了积极贡献。由于我们的数据集由互联网假新闻帖子组成，某些帖子可能包含冒犯性内容。然而，我们的研究在减少错误信息方面的积极贡献远远超过了潜在的负面影响。

References

- Sahar Abdelnabi, Rakibul Hasan, and Mario Fritz. Open-domain, content-based, multi-modal fact-checking of out-of-context images via online resources. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 14940–14949, 2022.
- Shivangi Aneja, Chris Bregler, and Matthias Nießner. Cosmos: Catching out-of-context misinformation with self-supervised learning, 2021. URL <https://arxiv.org/abs/2101.06278>.
- Shivangi Aneja, Chris Bregler, and Matthias Nießner. Cosmos: catching out-of-context image misuse using self-supervised learning. In Proceedings of the AAAI conference on artificial intelligence, volume 37, pp. 14084–14092, 2023.
- 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.
- Debarshi Brahma, Amartya Bhattacharya, Suraj Nagaje Mahadev, Anmol Asati, Vikas Verma, and Soma Biswas. Dpod: Domain-specific prompt tuning for multimodal fake news detection. arXiv preprint arXiv:2311.16496, 2023a.
- Debarshi Brahma, Amartya Bhattacharya, Suraj Nagaje Mahadev, Anmol Asati, Vikas Verma, and Soma Biswas. Leveraging out-of-domain data for domain-specific prompt tuning in multi-modal fake news detection. arXiv preprint arXiv:2311.16496, 2023b.
- Megha Chakraborty, Khusbu Pahwa, Anku Rani, Adarsh Mahor, Aditya Pakala, Arghya Sarkar, Harshit Dave, Ishan Paul, Janvita Reddy, Preethi Gurumurthy, et al. Factify3m: A benchmark for multimodal fact verification with explainability through 5w question-answering. arXiv preprint arXiv:2306.05523, 2023.
- Aude Genevay, Marco Cuturi, Gabriel Peyré, and Francis Bach. Stochastic optimization for large-scale optimal transport. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (eds.), Advances in Neural Information Processing Systems, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/2a27b8144ac02f67687f76782a3b5d8f-Paper.pdf.
- Jiahui Geng, Yova Kementchedjieva, Preslav Nakov, and Iryna Gurevych. Multimodal large language models to support real-world fact-checking, 2024. URL <https://arxiv.org/abs/2403.03627>.
- Andreas Hanselowski, Christian Stab, Claudia Schulz, Zile Li, and Iryna Gurevych. A richly annotated corpus for different tasks in automated fact-checking. arXiv preprint arXiv:1911.01214, 2019.
- Yiqiao Jin, Minje Choi, Gaurav Verma, Jindong Wang, and Srijan Kumar. Mm-soc: Benchmarking multimodal large language models in social media platforms. arXiv preprint arXiv:2402.14154, 2024.

²<https://developer.x.com/en/developer-terms/more-on-restricted-use-cases>

- M Abdul Khaliq, Paul Chang, Mingyang Ma, Bernhard Pflugfelder, and Filip Miletić. Ragar, your falsehood radar: Rag-augmented reasoning for political fact-checking using multimodal large language models. arXiv preprint arXiv:2404.12065, 2024.
- Zeba Khanam, BN Alwasel, H Sirafi, and Mamoon Rashid. Fake news detection using machine learning approaches. In IOP conference series: materials science and engineering, volume 1099, pp. 012040. IOP Publishing, 2021.
- Xuannan Liu, Peipei Li, Huaibo Huang, Zekun Li, Xing Cui, Jiahao Liang, Lixiong Qin, Weihong Deng, and Zhaofeng He. Fka-owl: Advancing multimodal fake news detection through knowledge-augmented lvlms. In Proceedings of the 32nd ACM International Conference on Multimedia, pp. 10154–10163, 2024a.
- Xuannan Liu, Zekun Li, Peipei Li, Shuhan Xia, Xing Cui, Linzhi Huang, Huaibo Huang, Weihong Deng, and Zhaofeng He. Mmfakebench: A mixed-source multimodal misinformation detection benchmark for lvlms, 2024b. URL <https://arxiv.org/abs/2406.08772>.
- Grace Luo, Trevor Darrell, and Anna Rohrbach. Newscippings: Automatic generation of out-of-context multimodal media. arXiv preprint arXiv:2104.05893, 2021.
- Michael Mu, Sreyasee Das Bhattacharjee, and Junsong Yuan. Self-supervised distilled learning for multi-modal misinformation identification. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 2819–2828, 2023.
- Kai Nakamura, Sharon Levy, and William Yang Wang. r/fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. arXiv preprint arXiv:1911.03854, 2019.
- Stefanos-Iordanis Papadopoulos, Christos Koutlis, Symeon Papadopoulos, and Panagiotis C. Petrantonakis. Verite: A robust benchmark for multimodal misinformation detection accounting for unimodal bias, 2023. URL <https://arxiv.org/abs/2304.14133>.
- Peng Qi, Zehong Yan, Wynne Hsu, and Mong Li Lee. Sniffer: Multimodal large language model for explainable out-of-context misinformation detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13052–13062, 2024.
- Fatma Shalabi, Hichem Felouat, Huy H Nguyen, and Isao Echizen. Leveraging chat-based large vision language models for multimodal out-of-context detection. In International Conference on Advanced Information Networking and Applications, pp. 86–98. Springer, 2024.
- Rui Shao, Tianxing Wu, and Ziwei Liu. Detecting and grounding multi-modal media manipulation, 2023. URL <https://arxiv.org/abs/2304.02556>.
- Snopes. URL <https://www.snopes.com/>.
- Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805, 2023.
- James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a large-scale dataset for fact extraction and VERification. In Marilyn Walker, Heng Ji, and Amanda Stent (eds.), Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pp. 809–819, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1074. URL <https://aclanthology.org/N18-1074/>.
- Andreas Vlachos and Sebastian Riedel. Fact checking: Task definition and dataset construction. In Proceedings of the ACL 2014 workshop on language technologies and computational social science, pp. 18–22, 2014.

- Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191, 2024.
- William Yang Wang. “liar, liar pants on fire”: A new benchmark dataset for fake news detection. In Regina Barzilay and Min-Yen Kan (eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 422–426, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-2067. URL <https://aclanthology.org/P17-2067/>.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. arXiv preprint arXiv:2203.11171, 2022.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.
- X Community Notes. URL <https://communitynotes.x.com/guide/en/about/introduction>.
- X/Twitter. URL <https://x.com/>.
- Keyang Xuan, Li Yi, Fan Yang, Ruochen Wu, Yi R Fung, and Heng Ji. Lemma: Towards lvlm-enhanced multimodal misinformation detection with external knowledge augmentation. arXiv preprint arXiv:2402.11943, 2024.
- Yuzhou Yang, Yangming Zhou, Qichao Ying, Zhenxing Qian, and Xinpeng Zhang. Search, examine and early-termination: Fake news detection with annotation-free evidences. arXiv preprint arXiv:2407.07931, 2024.
- Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of llms: Preliminary explorations with gpt-4v (ision). arXiv preprint arXiv:2309.17421, 9(1):1, 2023.
- Barry Menglong Yao, Aditya Shah, Lichao Sun, Jin-Hee Cho, and Lifu Huang. End-to-end multimodal fact-checking and explanation generation: A challenging dataset and models. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '23*, pp. 2733–2743. ACM, July 2023. doi: 10.1145/3539618.3591879. URL <http://dx.doi.org/10.1145/3539618.3591879>.
- Xin Yuan, Jie Guo, Weidong Qiu, Zheng Huang, and Shujun Li. Support or refute: Analyzing the stance of evidence to detect out-of-context mis- and disinformation. arXiv preprint arXiv:2311.01766, 2023.
- Fengzhu Zeng, Wenqian Li, Wei Gao, and Yan Pang. Multimodal misinformation detection by learning from synthetic data with multimodal llms. arXiv preprint arXiv:2409.19656, 2024.
- Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training, 2023. URL <https://arxiv.org/abs/2303.15343>.
- Yizhou Zhang, Loc Trinh, Defu Cao, Zijun Cui, and Yan Liu. Detecting out-of-context multimodal misinformation with interpretable neural-symbolic model. arXiv preprint arXiv:2304.07633, 2023.
- Dimitrina Zlatkova, Preslav Nakov, and Ivan Koychev. Fact-checking meets fauxtography: Verifying claims about images. arXiv preprint arXiv:1908.11722, 2019.

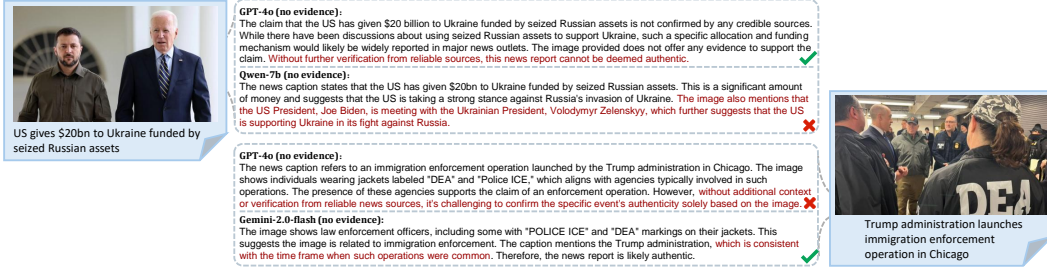


Figure 5: 不同 MLLMs 在没有证据的情况下表现比较。不同 MLLMs 在没有支持性证据的情况下展现出不同的预测倾向。

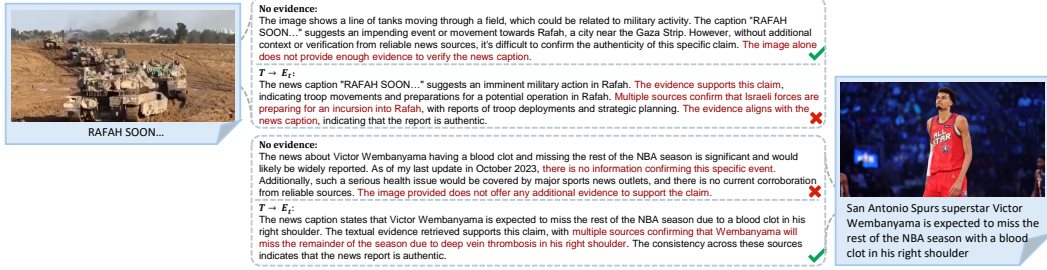


Figure 6: $T \rightarrow E_t$ 策略在识别真实和虚假帖子中的有效性。 $T \rightarrow E_t$ 擅长寻找真实帖子的证据。然而，它也可能导致模型翻转其最初正确的预测，尤其是在图像超出语境的错误信息中。

A 更多的证据检索分析

MLLMs 在没有证据的情况下表现出显著的行为差异。如图 5 所示，GPT 在多模态误导信息检测中表现出保守的倾向。无论帖子是假的（如上图所示）还是真实的（如下图所示），当没有支持证据时，GPT 倾向于将其分类为假。然而，Gemini 和 Qwen 表现出相反的行为：如果没有观察到图像和标题之间的明显不一致，它们更可能将新闻分类为真实。这进一步强调了仅依赖模型的内部知识而没有外部证据在误导信息检测中是不可靠的。

$T \rightarrow E_t$ （策略①）显著提升了性能，特别是对于真实帖子。在图 6 的底部和图 7 的顶部显示了另外两个例子。然后我们讨论为什么对虚假帖子的准确度没有显著提高，甚至对于 GPT-4o 略有下降。我们将此归因于图像 OOC 虚假信息，其中 $T \rightarrow E_t$ 未提供关于图像 I 的信息，而 E_t 中对 T 的强有力支持误导了模型，使其翻转其原本正确的虚假预测。作为说明性的例子，图 8 和图 6 的顶部展示了两个案例。没有证据时，模型给出了谨慎且正确的答案，而当 $T \rightarrow E_t$ 支持帖子声明 T 时，它变得更加自信但做出了错误预测。因此，直接针对图像 OOC 虚假信息的证据是此类证据的重要补充，例如图 3 右侧所示的例子。

$I \rightarrow E_t$ （策略③）比 $T \rightarrow E_i$ （策略②）对于上下文之外的虚假信息更有效。虽然在表格 2 中， $T \rightarrow E_i$ 显示出更高的整体准确性，但人工检查发现 $I \rightarrow E_t$ 能更好地检测出上下文之外的情况。这是因为 $I \rightarrow E_t$ 直接检索包含查询图像的网页并提取高度相关的文本，而 $T \rightarrow E_i$ 进行基于标题的模糊搜索，经常检索到松散相关的图像。在这些情况下，文本证据也更加具有信息性，因为基于图像的比较通常仅限于粗略特征，如一般场景或人物。这些表面级别的相似性通常在上下文之外的虚假信息中被保留下来，使得仅通过图像证据难以检测操控。如图 8 所示， $I \rightarrow E_t$ 正确地追溯到一个 Facebook 用户的图像并将其识别为上下文之外的虚假信息。相比之下， $T \rightarrow E_i$ 仅提供粗略的背景比较，未能精确检测虚假信息。

DuckDuckGo 提供的证据质量低于 Google。我们在图 10 中提供了一个例子，其中 DuckDuckGo 新闻搜索（策略⑧）未能检索到任何相关证据，而 Google 搜索（策略①）准确找到了能够帮助模型做出正确推断的“直接匹配”新闻事件。

域过滤器可以减轻证据噪音。域过滤器可以通过过滤掉来自低可信度网站的误导性证据来提高准确性。这里我们给出一个例子。如图 9 所示，没有域过滤器时，证据中有支持图注中所做声明的相关内容。因此，即使新闻帖图注本身提到这可能是一个“未经证实的传闻”，

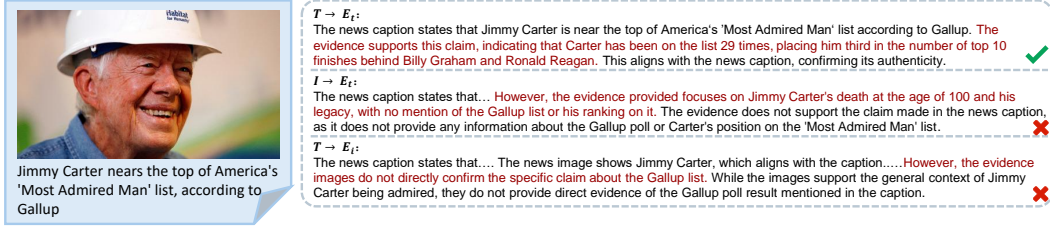


Figure 7: 不同类型证据在真实帖子上的有效性。 $T \rightarrow E_t$ 可以有效地检索真实帖子相关的证据，但在这种情况下，跨模态证据的作用较小。

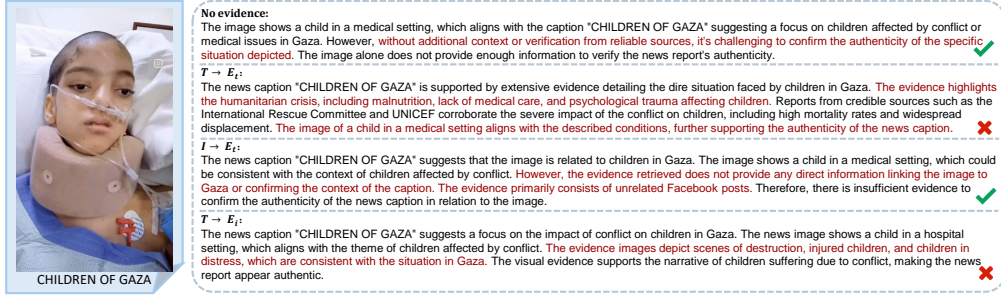


Figure 8: 不同证据类型对脱离上下文的虚假信息帖子的有效性

模型仍然根据支持证据将其分类为真实。然而，域过滤器排除了这条错误的证据，使得模型更容易做出正确的判断。这表明，即使在正确分类应是显而易见的情况下，错误的证据仍能对检测产生显著的负面影响。

基于 LLM 的证据提取可以减轻证据噪声。我们在图 11 中展示了一个例子。提取的证据更加简洁，使模型更容易进行推理。

B 推理的更多分析

推理方法的详细信息：

- 思维链：模型除了输出二值标签外，还输出额外的推理依据。
- 提示集合：受 Geng et al. (2024) 启发，我们使用多种提示来生成多个响应，然后要求模型聚合这些响应以获得更稳健的结果。
- 自我一致性：执行多轮推理并使用多数投票来获得最终结果。
- 多步推理：当提供多个证据来源时，模型可能会陷入困惑。因此，我们对每种类型的证据分别调用 LLM，然后总结所有中间推理过程以得出最终的聚合答案。

CoT 和多步骤推理之间的比较。图 12 说明了 CoT 推理和多步骤推理之间的不同路径。在此例中，多步骤推理通过分析 $I \rightarrow E_t$ ，准确识别出图像来源于另一个事件，从而正确地将其分类为图像 OOC 误导信息。然而，CoT 推理未能充分利用每一条证据，导致其忽视了 $I \rightarrow E_t$ 证据，并产生了错误的推论。

不同模型规模的推理性能比较。图 13 展示了 GPT-4o 和 GPT-4o-mini 的推理路径。GPT-4o 拥有比 GPT-4o-mini 更强的推理能力，这使得它能够更精确地识别证据中的短语“起初提出无罪抗辩”，因此作出正确的判断。

C 不同虚假信息检测数据集的比较

结果如表 8 所示。GPT-4o 在不使用任何其他数据集上的证据情况下，能够实现 0.8 甚至 0.9 的准确率，表明它可以仅通过记忆有效地进行虚假信息检测。此外，我们观察到证据带

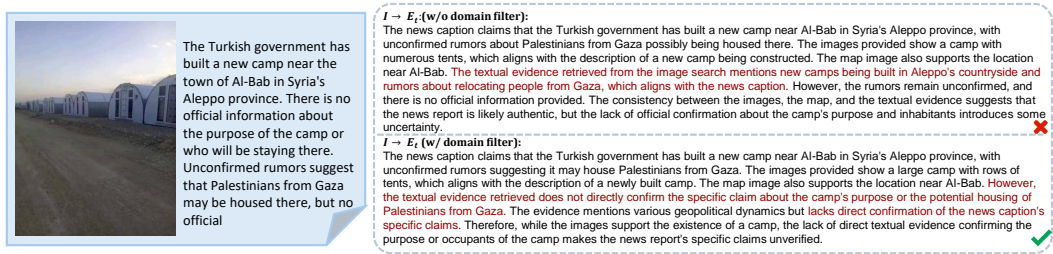


Figure 9: 证据领域过滤器的有效性。

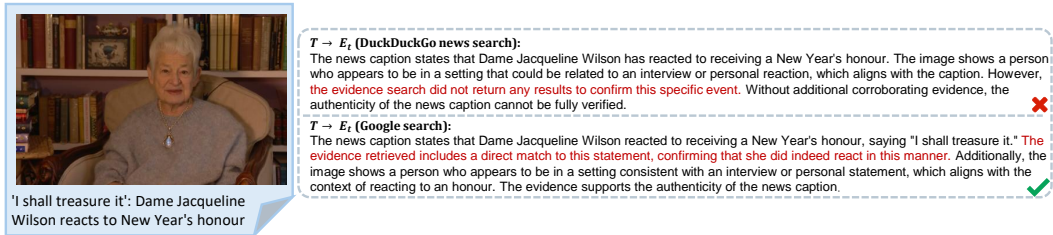


Figure 10: 谷歌搜索与 DuckDuckGo 新闻搜索的比较。

来的改进在我们的数据集上最为显著。因此，我们的数据集更适合用于评估基于检索的虚假信息检测器，并且与现实世界的虚假信息场景相比，评价偏差更小。

D 基于检测器辅助的数据集扩充示例

图 14 显示了使用我们的检测器在 X 的最新帖子上进行错误信息检测的示例，其中包括为每个预测提供的支持性解释，以帮助人工审核人员更有效地验证结果。

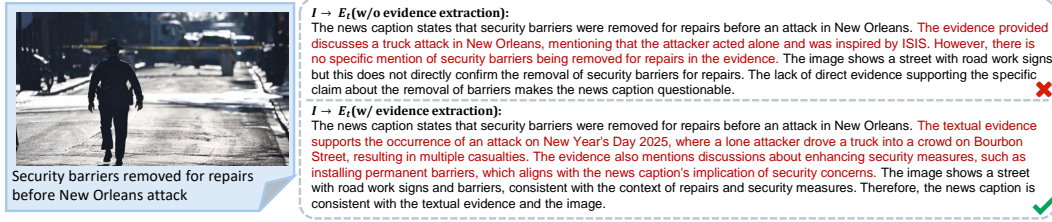


Figure 11: 证据提取的有效性。

Dataset	Evidence	Acc.	R. Acc.	F. Acc.
VERITR	✗	80.1	78	82
	✓	91.9	88.5	95.2
Snopes+Reuters	✗	91.9	93.5	90.3
	✓	96.7	94.6	99
Newsclipping	✗	80.7	88.2	73.2
	✓	89.5	91.2	87.7
Xfacta(Ours)	✗	70.5	51	90
	✓	89.5	98	81

Table 8: GPT-4o 在不同数据集上的性能比较。

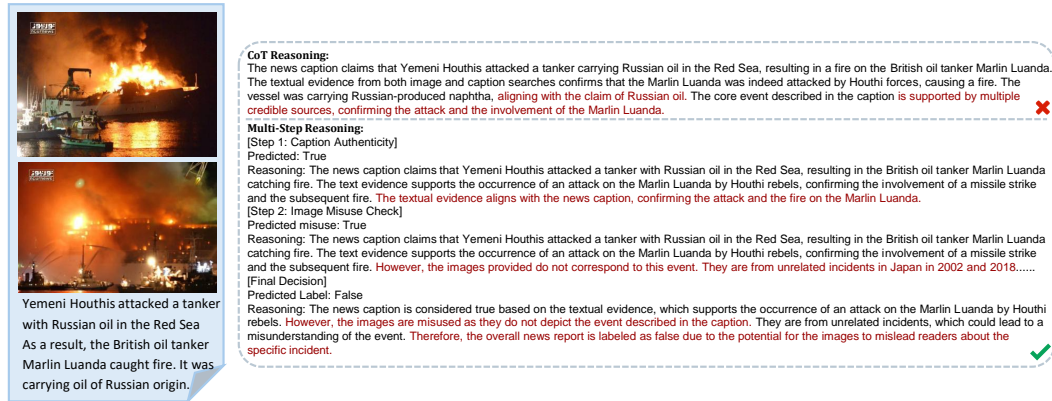


Figure 12: 链式推理与多步推理的比较。



Figure 13: GPT-4o 与 GPT-4o-mini 之间的比较。

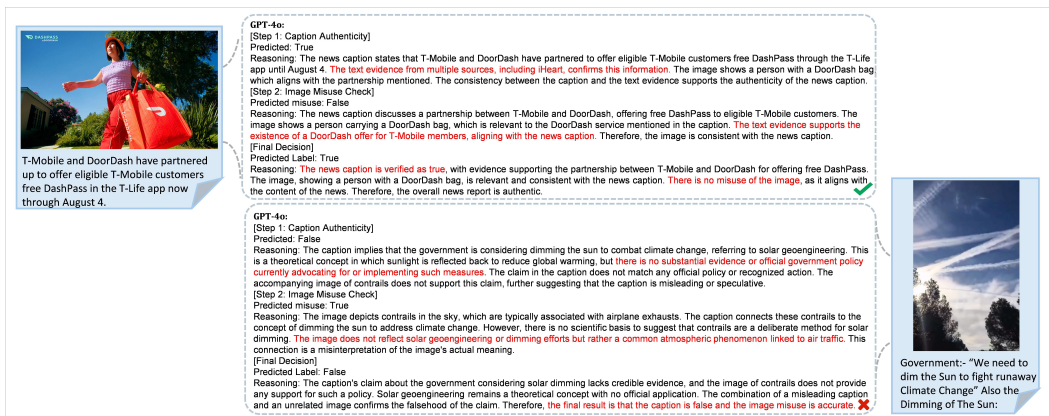


Figure 14: 使用检测器的推断对 X 最新帖子进行错误信息检测的示例