传播之前：智能体驱动的虚拟传播用于早期假新闻检测

BINCHENG GU,重庆大学，中国MIN GAO*，网络物理社会可信服务计算重点实验室（重庆大学），中国教育部JUNLIANG YU，澳大利亚昆士兰大学ZONGWEI WANG，重庆大学，中国ZHIYI LIU，重庆大学，中国KAI SHU，美国埃默里大学HONGYU ZHANG，网络物理社会可信服务计算重点实验室（重庆大学），中国教育部

arXiv:2601.02750v1 [cs.IR] 6 Jan 2026

早期检测假新闻对于遏制其在社交媒体上的快速传播至关重要，因为假新闻可能严重损害公众信任和社会稳定。近期研究表明，与以往仅基于内容的方法相比，纳入传播动态能够显著提升检测性能。然而，在早期阶段，由于缺乏可观测的传播信号，这仍然颇具挑战。为解决这一局限，我们提出AVOID，一种用于早期假新闻检测的智能体驱动虚拟传播方法。AVOID 将早期检测重新定义为一种全新的证据生成范式，在这一范式中，传播信号是主动模拟生成的，而非被动观测得到的。借助具有差异化角色和数据驱动人设的大语言模型驱动的智能体，AVOID 无需真实传播数据即可逼真地构建早期扩散行为。生成的虚拟轨迹提供了互补的社会证据，丰富了基于内容的检测；同时，去噪引导融合策略使模拟传播与内容语义相匹配。在基准数据集上进行的大量实验表明，AVOID 的表现始终优于最先进的基线方法，凸显了虚拟传播增强方法在早期假新闻检测中的有效性和实用价值。代码和数据见 https://github.com/Ironychen/AVOID。

CCS 概念：•**信息系统** → **数据挖掘**；•**计算方法** → **自然语言处理**。

Ad附加关键词与短语：假新闻检测、社会模拟、基于大语言模型的智能体 s

* 通讯作者

作者联系信息：Bincheng Gu，gubincheng@stu.cqu.edu.cn，重庆大学，中国重庆；Min Gao，gaomin@cqu.edu.cn，网络物理社会可信服务计算重点实验室（重庆大学），教育部，中国重庆；Junliang Yu，昆士兰大学，澳大利亚布里斯班，jl.yu@uq.edu.au；Zongwei Wang，zongwei@cqu.edu.cn，重庆大学，中国重庆；Zhiyi Liu，liuzhiyi@ stu.cqu.edu.cn，重庆大学，中国重庆；Kai Shu，kai.shu@emory.edu，埃默里大学，美国亚特兰大；Hongyu Zhang，hyzhang@cqu.edu.cn，网络物理社会可信服务计算重点实验室（重庆大学），教育部，中国重庆。

允许免费制作本作品全部或部分内容的数字或纸质副本，供个人或课堂使用，但不得出于营利或商业目的制作或分发副本，且副本须附有本声明及首页上的完整引文。作者以外的其他权利人拥有的本作品组成部分，其版权必须得到尊重。注明出处后，允许摘录。若要以其他方式复制、再版、发布到服务器或分发至列表，则须事先获得特别许可和/或支付费用。请通过 permissions@acm.org 申请许可。会议简称 *'XX*，纽约州伍德斯托克© 2018 版权归所有者/作者所有。出版权已授权给 ACM。美国计算机协会国际标准书号 978-1-4503-XXXX-X/2018/06 https://doi.org/XXXXXXX.XXXXXXX

ACM 参考文献格式：
Bincheng Gu、Min Gao、Junliang Yu、Zongwei Wang、Zhiyi Liu、Kai Shu 和 Hongyu Zhang。2018。传播之前：智能体驱动的虚拟传播用于早期假新闻检测。载于请根据版权确认邮件中的信息填写正确的会议名称（会议简称 *'XX*）。ACM，纽约，纽约州，美国，29 页。https://doi.org/XXXXXXX.XXXXXXX

1 引言

Facebook 和 X 等社交媒体平台已成为新闻消费的主要渠道，使信息能够快速、大规模地传播。尽管这提高了信息的可获取性，却也加速了假新闻的传播，侵蚀公众信任，并扭曲社会认知 [15, 56]。错误信息的快速传播，尤其是在政治和公共卫生等敏感领域 [3, 52],，可能在纠正干预得以实施之前造成严重的社会后果 [61]。这些挑战凸显了迫切需要有效的方法，以尽早检测假新闻。

早期假新闻检测的主流方法主要依赖基于内容的分析 [7, 8, 79]。如图 1(a) 所示，这些方法利用直接从新闻内容中提取的文本、视觉和情感特征来评估可信度。然而，纯粹基于内容的方法存在固有的缺陷：它们往往无法捕捉新闻传播所涉及的更广泛的语境和社会动态，从而降低了区分精心伪装的假新闻与真实信息的能力，尤其是在欺骗性新闻高度模仿可信写作风格时 [78]。为弥补纯内容方法缺乏上下文线索的问题，近期研究探索了将社交传播模式用作检测的补充信号 [5, 58]。针对实时数据稀缺的早期检测场景，现有研究进一步采用深度图生成模型合成虚拟传播轨迹 [74, 76, 77], , 如图 1(b) 所示，这些轨迹作为辅助信息，可丰富检测过程中新闻文章的表征。图 1(d) 所示的实证结果直接比较了纯内容模型与整合传播信息的模型的性能，证实纳入传播信号能够显著提高检测准确率。

尽管生成式方法前景广阔，但仍面临一个关键局限：大多数现有图生成模型侧重于拟合历史图结构的统计分布，往往无法捕捉用户与新闻文章之间动态的、由内容驱动的互动。因此，对于语境新颖的全新文章，它们难以生成可靠、可信的轨迹 [74, 76]。在实践中，理想的情况是在任何真实传播发生之前检测出假新闻，此时基于内容的方法便成为唯一可行的选择。为弥合这一差距，我们的研究提出了一种新范式：通过基于智能体的模拟，在社交网络环境中对逼真的用户与内容互动进行建模，从而构建虚拟传播轨迹，为难以判定、置信度较低的新闻文章生成必要证据，以作出明确判断。

大语言模型（LLM）和基于智能体的建模技术[1, 33, 59] 的最新进展，使复杂社交互动和信息传播过程得以更加真实地模拟。新兴研究表明，智能体驱动的模拟能够有效地模拟现实环境中观察到的个体用户行为和特定社会现象，为研究社会动态提供了强有力的工具 [6, 25, 45, 64]。基于这些进展，我们尝试从零开始模拟用户行为和互动，即使在缺乏观测数据的情况下，也能生成合理的虚拟传播路径。尽管社会模拟已经取得进展，要实现真实可靠的虚拟传播仍面临重大障碍。当前的基于智能体的模拟方法通常根据基本的社会人口统计特征初始化智能体，并为其分配统一的行为规则 [38, 43]。这类智能体初始化方法基于

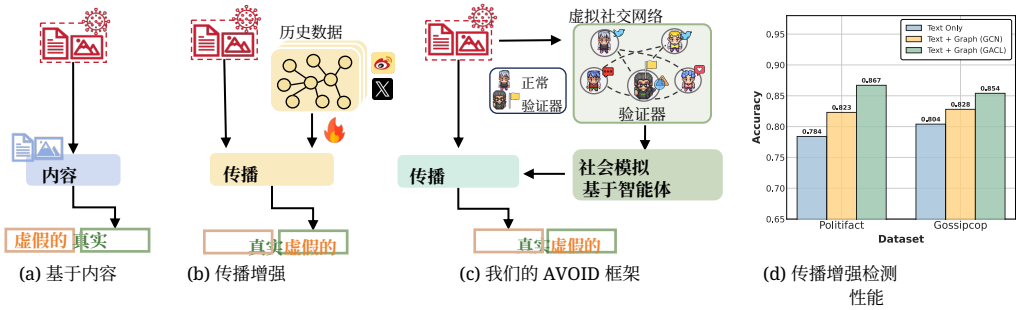

图 1. 图 (a)–(c) 对比了不同假新闻检测方法的架构，图 (d) 则展示了两种数据集上的性能对比，证实传播增强方法相较于纯内容基线方法带来了显著提升。

纯粹的统计特征，忽视了社交互动中现实用户行为的细微复杂性。此外，模拟传播轨迹生成过程可能引入噪声，从而削弱早期检测的效果。

为克服这些局限性，我们提出 AVOID，这是一种由大语言模型赋能、由智能体驱动的虚拟传播框架，用于早期且鲁棒的假新闻检测，无需依赖任何观测到的传播数据，如图 1(c) 所示。我们将参与的智能体划分为两种不同类型：扩散智能体，代表通过被动浏览和转发推动信息传播的普通用户群体；以及验证智能体，具备批判性评估内容真实性并主动介入的能力。通过明确区分并建模这些角色，AVOID 忠实地模拟了真实社交平台中常见的交互异质性。此外，正如 [27, 37, 53, 75], 所指出的，用户画像是交互行为的关键决定因素，而使画像分布与实证观察相吻合，对于实现逼真的模拟行为和传播动态至关重要。因此，我们使智能体画像与真实用户数据相匹配。此外，为减轻虚拟模拟引入的噪声，AVOID 采用去噪引导的跨模态融合策略，从而提高早期假新闻检测的可靠性和准确率。

总体而言，我们的贡献可概括如下：

- 我们提出了一种新颖的早期假新闻检测范式，即使没有观测到任何扩散路径，也能融入虚拟传播增强，从而解决现有方法的重大局限。
- 我们提出 AVOID，这是一种大语言模型驱动的智能体框架，通过基于逼真用户画像生成由智能体驱动的虚拟传播网络，并借助去噪引导的跨模态融合减轻这些路径中固有的噪声，从而增强早期假新闻检测。
- 在真实世界数据集上开展的大量实验表明，AVOID 的表现始终优于最先进的基线方法，验证了我们基于虚拟传播的早期假新闻检测方法的有效性与实用性。

本文结构如下：第 2 节介绍预备知识，包括基于智能体的社交网络设定、传播增强型假新闻检测以及基于置信度的过滤策略。第 3 节介绍传播器和核查代理的设计。第 4 节介绍所提出的 AVOID 框架。第 5 节报告实验设置与结果。第 6 节回顾相关工作，第 7 节总结本文并展望未来方向。

2 预备知识

本节首先概述基于智能体的社交网络仿真设定，随后形式化定义传播增强型假新闻检测，最后定义硬新闻，并提出一种过滤策略：仅将置信度较低的内容用于生成虚拟传播，从而确保计算效率。

2.1 基于智能体的社交网络模拟

借助大语言模型的生成和推理能力，自主智能体能够在复杂的社会语境中模拟人类行为 [2, 45, 69]。在我们的研究中，我们采用了一种典型的智能体架构，由档案模块、记忆模块和行为模块组成 [65]。档案模块定义智能体的身份和行为，记忆模块捕获交互历史，而行为模块则将推理转化为可观察的行为，例如发表评论或转发。

我们将社会环境建模为由相互互动的智能体组成的网络 V。每个节点 $v_i \in \mathrm{V}$ 对应一个自主智能体，其特征包括独特的个人资料、记忆和行动策略。新闻通过相邻智能体之间的互动进行传播。具体而言，当一个智能体 v_i 分享某条新闻并成功影响邻居 $v_j \in \mathrm{N}(v_i)$ 与该内容进行互动时，就会建立一条有向边 $(v_i \rightarrow v_j)$。对于每条新闻条目 i，由此产生的互动会形成一个动态传播图 $\mathrm{G}_i = (\mathrm{V}_i, \mathrm{E}_i)$，其中 $\mathrm{V}_i \subseteq \mathrm{V}$ 表示参与智能体的集合，E_i 则由相应的影响边组成。每个传播图 G_i 刻画了该新闻条目的时间扩散轨迹。

2.2 传播增强假新闻检测

假新闻检测传统上依赖于从新闻内容中多模态信息 X_C 提取的基于内容的特征。这些方法通常采用编码器获取 X_C的语义表示，随后利用这些表示学习条件概率分布$P_\theta(y \mid \mathrm{X}_C)$，以预测新闻的真实性。

为进一步增强检测能力，近期研究 [13, 82] 将社交网络中的传播动态纳入考量。通过将新闻的扩散模式建模为从传播图 G 中提取的结构特征，这些传播增强方法扩展了检测的输入空间。因此，检测目标扩展为 $P_\theta(y \mid \mathrm{X}_C, \mathrm{X}_\mathcal{G})$，其中 $\mathrm{X}_\mathcal{G}$ 捕捉其在社交网络中的扩散拓扑特征。

以内容为中心的方法与传播增强方法都采用相同的训练范式：最小化二元交叉熵损失。形式上，检测模型优化：

$$\mathcal{L}_{cls} = -\sum_{i=1}^{Z} \left[y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i) \right], \tag{1}$$

ˆ其中 $\hat{}y_i \in Y$表示第 i篇新闻文章为虚假的预测概率， $y_i \in Y$表示相应的真实标签， Z 表示新闻样本总数。

2.3 置信度引导的选择性虚拟传播

尽管虚拟传播能为假新闻检测提供有价值的额外证据，但其计算开销较大。为避免对容易判定的新闻条目进行传播模拟，我们引入一种过滤策略，用于检测“难例”新闻。具体而言，我们从一个预训练的 BERT编码器出发，仅在目标训练数据集上微调其最后两个 Transformer层，从而得到一个轻量级的内容分类器 f_θ。给定输入新闻条目 x，分类器会输出后验概率 $\hat{}p(x) = P_\theta(y \mid \mathrm{X}_C)$，其中 X_C表示由 BERT 编码的文本表示。我们将置信度定义为 $conf(x) = \max\{\hat{}p(x), 1 - \hat{}p(x)\}$，用于衡量

表 1. 评论过滤统计。我们使用基于规则的过滤器，将低信息量反应与包含解读或核验线索的信息性评论区分开来。过滤后的数量表示保留的信息性评论数，百分比（%）表示保留评论占所有评论的比例。

数据集	评论总数	过滤后	百分比（%）
PolitiFact	229,370	8,778	4.83
GossipCop	180,902	4,098	2.26
微博	367,070	20,475	5.58

分类器对任一类别的倾向程度。如果 $conf(x)$低于阈值，则将该样本视为低置信度（难例）。我们仅对这些样本调用虚拟传播模块，以生成基于图的表示 $\mathbf{X}_G$。随后，按照第 2.2 节所述，使用联合表示 $(\mathbf{X}_C, \mathbf{X}_G)$ 得出最终预测。对于高置信度（易例）样本，我们完全跳过传播，仅依赖 f_θ，从而在准确率和效率之间实现较好的平衡。

3 个传播模拟智能体

为模拟真实的信息传播，我们引入两种智能体角色，以体现社交网络中的异质性参与。先前研究 [4] 和我们的分析（表 1）表明，绝大多数回应都是信息量较低的反应，而只有一小部分回应包含能够引导后续讨论的有意义线索。基于这一观察，我们设计了两类智能体：传播智能体模拟轻量级、由反应驱动的参与者，而验证智能体则生成信息量更丰富的回应，并增强真实性评估和主观表达。这种基于角色的设计能够更真实地模拟谣言传播中的异质性用户行为。

3.1 扩散者智能体设计

传播智能体代表社交网络中的大多数用户，他们依据自身的经历、兴趣和信念与信息互动。在我们的框架中，它们构成模拟网络的基础，以体现多样化的用户行为。

3.1.1 多样化用户画像建模。 大多数现有模拟依赖于根据统计特征随机初始化的智能体档案，无法反映特定类型新闻 [43, 45]的真实人物画像分布。为克服这一局限，我们从真实世界数据集中提取细粒度人物画像表征，使其更符合用户特征的实际分布。详细方法见第 4.1 节。我们还为这些细粒度人物画像增添了社交情境元数据，从而提升真实感并支持更细致的模拟。

3.1.2 时序情境记忆建模。人类的社会行为不仅取决于当前语境，也取决于过往经验。为捕捉这种时间依赖性，我们引入短期和长期记忆，以模拟社交接触。

模拟短期社交互动。为反映近期的社交互动，该模块会将最新的 k 社交互动（例如新闻摘要和朋友的评论）记录为向量嵌入。遇到新内容时，它会使用 FAISS [14]，根据余弦相似度高效检索最相似的记录，为决策提供即时语境。

模拟长期知识积累。长期记忆通过定期将短期内容总结为紧凑的记录来保留提炼后的知识，在留存重要经验的同时，由遗忘机制丢弃过时的信息。检索时，它会使用一个

大语言模型将新条目分解为语义子查询（实体、事件、主题）。与短期记忆采用的直接向量相似度搜索不同，这种方式能够更精确地上下文感知检索相关的过往知识。指导提示如下：

长期记忆检索提示模板

给定 <新闻内容>，将其分解为 <实体>、 <事件>和 <主题>，然后从 <长期记忆>中检索相关片段。

3.1.3 用户参与行为建模。在社交网络上模拟用户行为，需要智能体主动参与新闻内容互动。为此，我们设计了一个行为模块，支持以下行为。对于扩散智能体，可执行的行为包括：(1) 评论：撰写文本回复以表达立场并提供简要解释；(2) 转发：分享现有帖子，可以直接分享，也可以附加评论；(3) 点赞：通过点赞帖子表达赞同；(4)查看：仅查看新闻，不采取任何进一步行动。这些行为会影响智能体的曝光，并间接塑造传播过程。

3.2 VerifierAgent 的设计

扩散智能体代表普通用户群体，而验证智能体则模拟一小部分具有影响力的用户，他们会积极主动且有策略地影响新闻传播，从而关键性地塑造信息流和舆论动态。

3.2.1 特定核验者画像建模。与扩散智能体不同，核验者角色设定明确取自具有影响力的用户。我们首先根据转发次数、平均点赞数和粉丝数量计算影响力评分，并将排名前 5% 的用户标记为具有影响力的用户。从这些用户的历史评论中，我们仅保留参与度信号强、立场明确且有充分依据的评论，例如引用官方来源、科学出版物或政府统计数据的评论。然后，我们将第 4.1 节的方法应用于经过过滤的评论，以提取核验者特定的人物画像档案。

3.2.2 增强型策略记忆建模。为了实现高级可信度评估推理，验证智能体会维护策略记忆，用于存储过去真实性判断的结构化记录，包括内容、立场、推理依据和社会语境。记忆采用层级结构，分为三个层级：实体层级 π_e 用于追踪关键参与者，事件层级 π_v 用于捕捉因果和时间模式，元层级 π_m 用于概括通用推理策略。这种设计增强了验证器根据语境作出判断的能力，并使其能够在网络中产生可信的影响。

3.2.3 真实性驱动的行为建模。除了所有用户都具备的基本操作外，验证智能体还被赋予了两项额外操作，以体现其在现实世界网络中的权威角色：(5) 事实核查：利用外部知识来源或事实核查工具验证新闻条目；以及 (6)警告：发布公开警告，以遏制错误信息的传播。这些特权操作使验证智能体能够主动介入并重塑信息传播的轨迹。

4 AVOID：智能体驱动的新闻虚拟传播

本节介绍 AVOID，这是一个通过生成和利用虚拟传播轨迹来实现早期假新闻检测的仿真框架。需要注意的是，虚拟传播仅针对第2.3节中介绍的“低置信度”样本进行，以平衡检测效果与计算效率。

4.1 真实人物画像提取

为了让传播器和验证器智能体具备真实的用户行为，我们从真实世界的社交媒体评论中提取智能体人物画像。我们采用多阶段人物画像提取流程，逐步整理、抽样并提炼用户评论，从而构建连贯且有行为依据的个人资料。为确保流程不存在数据泄漏，人物画像提取和智能体初始化仅限于训练数据，且不使用新闻真实性标签。

4.1.1 面向人物画像的层次聚类。 为了从原始新闻评论内容中提炼出具有代表性的用户原型，我们提出了一个两阶段层次聚类过程，从主题发现过渡到观点提取。

该过程始于上下文级分组：将新闻内容嵌入语义空间，并划分为 m 主题簇 $\mathrm{T} = \{t_1, \ldots, t_m\}$。在这些语境的基础上，我们在每个主题簇内进行观点级提炼。对于主题 t_k，令 $\mathrm{R}_k = \{r_{k,1}, \ldots, r_{k,n_k}\}$表示与其相关的用户评论。我们对评论嵌入进行聚类，以捕捉不同的反应模式，并将每个聚类映射回其成员评论，从而生成评论组集合 $\mathrm{P}_k = \{p_{k,1}, \ldots, p_{k,\ell_k}\}$。其中，每个组 $p_{k,j} \subseteq \mathrm{R}_k$聚合了观点相似的评论，对应于一个特定主题的人物画像。全局人物画像池定义为 $\mathrm{P} = \bigcup_{k=1}^{m} \mathrm{P}_k$。这种层次化方法确保提取的人物画像具有多样化的观点，并且扎根于其所涉及的具体新闻主题。

4.1.2 均衡人物画像采样。尽管层次聚类能够生成多样的人物画像组，但使用特定组内的所有评论$p_{k,j}$可能会引入冗余。基于质心的选择 [57] 可能会过度简化用户画像，而边界聚焦方法 [47] 则有过拟合离群值的风险。为缓解这些问题，我们采用均衡人物画像采样策略。对于每个人物画像组$p_{k,j}$我们将其包含的评论嵌入表示为$\mathbf{E}_{k,j} = \{\mathbf{e}_i \mid r_i \in p_{k,j}\}$。随后，我们通过同时优化典型性和多样性，选出一个精简且具有代表性的评论子集 $\mathrm{R}^*_{k,j} \subseteq p_{k,j}$。由此得到以下优化问题：

$$\max_{\mathcal{R}^*_{k,j}} \left(w_p \sum_{r_i \in \mathcal{R}^*_{k,j}} \frac{1}{1+D(\mathbf{e}_i, \boldsymbol{\mu}_{k,j})} + w_d \cdot \frac{2}{|\mathcal{R}^*_{k,j}|} \sum_{r_a, r_b \in \mathcal{R}^*_{k,j}} D(\mathbf{e}_a, \mathbf{e}_b) \right), \tag{2}$$

其中$\boldsymbol{\mu}_{k,j}$是嵌入空间中$\mathbf{E}_{k,j}$的质心。项 D 表示欧几里得距离，而权重w_p 和 $w_d = 1 - w_p$用于平衡典型性与多样性。所选子集 $\mathrm{R}^*_{k,j}$能够体现底层人物画像的核心语义及其变化。我们使用贪心算法求解这一离散优化问题，通过迭代选择边际增益最大的候选项。尽管该方法无法保证全局最优性，但在我们的实验中，它始终能够得到精简且具有代表性的子集。

4.1.3 现实对齐的人物画像提炼。 对于每个采样的评论子集 $\mathrm{R}^*_{k,j}$，我们为相应的用户群组提取细粒度的人物画像档案。由于直接使用原始评论提示大语言模型通常会生成泛化或肤浅的人物画像，我们采用大语言模型引导的迭代反思与精炼流程，以提升社会真实性和行为合理性。在每次迭代中 t，我们从 $\mathrm{R}^*_{k,j}$ 中采样一小批评论作为证据，逐步精炼人物画像档案：

$$\mathcal{P}_{t+1} \leftarrow \mathrm{Reflection}(\mathcal{P}_t, \hat{r}_t, \boldsymbol{\mu}_t, \lambda), \tag{3}$$

其中，$\hat{r}_t$是大语言模型基于当前档案 π_t生成的评论， $\boldsymbol{\mu}_t$表示样本批次的语义中心。当生成的评论嵌入 $\mathrm{Embed}(\hat{r}_t)$ 与中心 $\boldsymbol{\mu}_t$ 之间的余弦相似度低于

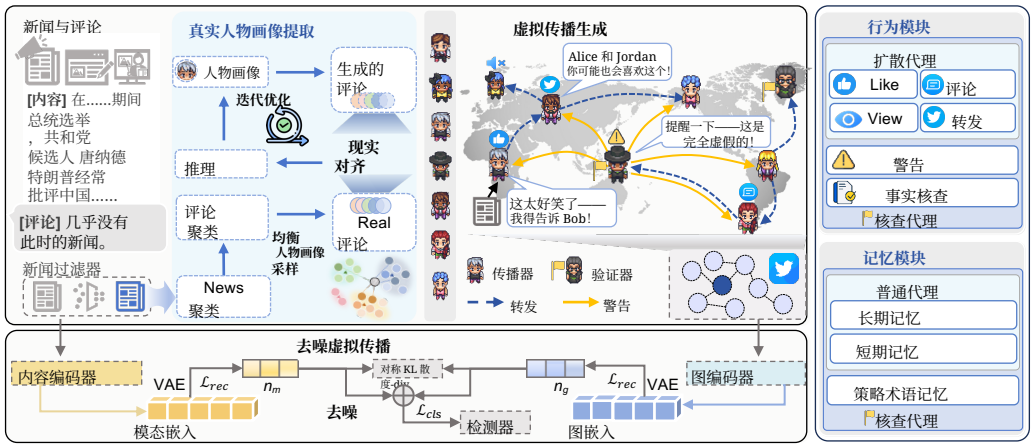

图 2. AVOID 概览。现实世界评论经过提炼形成细粒度人物画像，并据此初始化模拟智能体。这些智能体参与虚拟传播，以促进早期假新闻检测。

阈值 $\lambda = 0.7$时，将触发反思机制。这种差异会促使大语言模型修订 P_t，使其更好地与观测证据保持一致。所使用的系统提示词如下：

人物画像提炼提示模板

根据当前的 <人物画像>，你生成了一条具有代表性的 <评论>，但它与 <下一条评论>相矛盾。请思考 <原因> 和改进人物画像的方法，并输出更新后的 <人物画像>。

通过这一过程，我们从真实世界数据集中提取逼真的细粒度人物画像表示，使模拟智能体的分布与实际社交环境中观察到的分布相一致。请注意，人物画像提取是一次性离线过程，计算开销可以忽略不计。

4.2 虚拟传播生成

在为每个智能体分配所提取的人物画像的基础上，本节将详细介绍如何模拟社交网络中的新闻传播，并生成用于下游分析的结构化传播图数据。

4.2.1 基于真实感角色的初始化。 我们从真实世界网络中提取一个闭合子图来构建社交图，以保留真实的用户关系。每个智能体都以采样得到的人物画像及社交属性构成的个人资料进行初始化。初始时，所有智能体的短期记忆和长期记忆均为空。此外，还为验证智能体分配策略记忆，并赋予其使用事实核查工具进行真实性评估的权限。

4.2.2 基于记忆更新的社交动态建模。记忆更新使智能体能够通过整合社交反馈和过往经验来适应环境；对于核查代理，这一过程还进一步纳入了基于判断错误的反思性自我纠正。在我们的社会模拟中，每个智能体的短期记忆会持续记录其好友对某条新闻的实时行动和评论。这一机制为智能体提供了即时且不断变化的快照，展现其社交圈对新闻的看法。为进一步模拟社交热点话题的长期演变，

智能体会定期将短期记录整合到长期知识库中。在此过程中，重大事件和互动会被提炼为代表当前趋势的高层次见解。将这些见解纳入长期记忆时，我们采用指数时间衰减方案，降低较旧信息的权重并优先考虑近期证据，从而能够持续适应不断演变的社会动态。

借助上述记忆动态，验证智能体还会通过策略更新过程进一步调整自身，该过程由错误的真实性判断触发。当判断有误时，智能体会运用思维链推理 [67] 重新审视其原始推理轨迹 Φ_t。通过将该轨迹与真实标签进行比较，智能体能够准确找出其初步评估中的逻辑错误；该评估考量了新闻内容 x_t、策略记忆 Π_t和社会语境 K_t。这种有针对性的错误分析会直接指导策略记忆的优化，从而改进未来的策略。形式化地说，更新定义如下：

$$\Pi_{t+1} \leftarrow \text{Reflection}(x_t, \Pi_t, K_t, \Phi_t), \tag{4}$$

反思函数会优化策略的多层级组成部分（实体级、事件级和元层级）。检测到错误时，该过程通过以下特定提示实现：

策略记忆优化提示

给定 <新闻内容>、 <策略记忆>、 <社会语境>，以及 <真实标签>，请先给出你的评估并 <说明理由>。如果评估与 <真实标签>不一致，请使用思维链推理进行分析，然后根据分析优化你的 <策略> (π_e， π_v， π_m)。

这一反思性更新使验证智能体能够动态演进其决策策略，从而更准确地检测错误信息，并在社交网络中发挥更大的影响力。

4.2.3 行动驱动传播。 完成记忆初始化后，智能体根据记忆和社会语境逐步做出决策，以此传播新闻并塑造传播轨迹。对于每条新闻条目，传播从选定的种子智能体开始，并以离散步骤展开。在每一步中，智能体在短期记忆和长期记忆的引导下，从预定义的集合中选择一项行动。用于决策的提示如下：

传播智能体行动提示模板

你是 <人物画像>。你刚刚看到了 <新闻内容> ，由你的朋友分享。结合你的短期记忆 <$\mathcal{M}^S$> 和长期记忆 <$\mathcal{M}^L$>，从 <行动列表>中选择最合适的行动。

除了通用的决策机制外，指定的验证智能体还会遵循专门流程：它们依据已学习的策略和事实核查工具评估转发的新闻。如果检测到错误信息，它们会向网络广播警告。具体提示如下：

验证智能体行动提示模板

您是 <已认证的智能体人物画像> ，负责在社交网络上进行真实性评估。使用您的 <政策> 和 <事实核查> 工具来验证 <新闻内容>。然后从 <行动列表>中选择一项行动。行动：警告——这则新闻可能是虚假的。

智能体根据动态更新的记忆传播新闻，通过模拟新闻在固定深度内从随机种子智能体开始的传播，为每条新闻生成一个图 G_i。

4.3 去噪虚拟传播检测

为硬新闻生成虚拟传播路径后，下一步是有效利用这些增强证据进行真实性预测。我们首先对新闻内容和模拟图中的特征进行编码。然后，为了最大限度地发挥模拟信号的效用，并弥合内容与图模态之间的差异，我们采用对称 KL 散度进行特征对齐。该方法的理论有效性已在第 4.3.3 节中得到正式证明。

4.3.1 内容编码与模拟传播。为了捕捉每个新闻条目的特征，我们提取两类特征：来自文章文本的内容表示 $\mathbf{X}_C$，以及来自虚拟交互图的传播表示 $\mathbf{X}_G$。

新闻内容特征提取。我们采用层次注意力网络 [71] 对每篇新闻文章的语义特征进行编码。文章被划分为 S 个句子，每个句子最多包含 L 个词。首先，我们利用预训练的 BERT编码器生成上下文化词嵌入，然后通过门控循环单元对其进行处理，以捕捉序列依赖关系。为获得句子表示，我们采用词级注意力机制：

$$\mathbf{s}_j = \sum_{t=1}^{L} \gamma_t^{(j)} \mathbf{h}_t^{(j)}, \quad \gamma_t^{(j)} = \frac{\exp(\mathbf{h}_t^{(j)\top} \mathbf{u}_w)}{\sum_{t'=1}^{L} \exp(\mathbf{h}_{t'}^{(j)\top} \mathbf{u}_w)}, \tag{5}$$

其中，$\mathbf{h}_t^{(j)}$ 表示句子 j在时间步 t 的门控循环单元隐藏状态，$\mathbf{u}_w$ 表示可训练的词级注意力向量。随后，将得到的句子向量 $\mathbf{s}_j$ 输入另一个门控循环单元和句子级注意力机制，以聚合最终的文档级表示 $\mathbf{x}_c$：

$$\mathbf{x}_c = \sum_{j=1}^{S} \delta_j \mathbf{h}_j, \quad \delta_j = \frac{\exp(\mathbf{h}_j^\top \mathbf{u}_s)}{\sum_{j'=1}^{S} \exp(\mathbf{h}_{j'}^\top \mathbf{u}_s)}, \tag{6}$$

其中，$\mathbf{h}_j$ 表示句子级隐藏状态，$\mathbf{u}_s$ 是可训练的句子级注意力向量。这种层次结构确保$\mathbf{x}_c$ 能够捕获局部词语语境和全局文档级语义。

传播特征提取。我们构建一个虚拟传播图，其邻接矩阵为$\mathbf{A}_s$，节点嵌入$\mathbf{v}_i$ 则通过 BERT 编码的代理档案初始化。为捕获结构依赖关系，我们采用图注意力网络 [60],，在第 l层更新节点嵌入，公式如下：

$$\mathbf{v}_i^{(l+1)} = \sigma\Big(\sum\nolimits_{j \in \mathcal{N}(i)} \alpha_{ij}^{(l)} \mathbf{U}^{(l)} \mathbf{v}_j^{(l)}\Big), \tag{7}$$

其中，$\mathbf{U}^{(l)}$是可训练的权重矩阵，$\sigma(\cdot)$ 是激活函数，$\mathbf{N}$ (i) 表示邻居集合。 $\alpha_{(l)ij}$表示遵循标准图注意力网络公式的注意力系数。最后，我们应用图级聚合，得到传播感知表示$\mathbf{x}_g$。

4.3.2 用于检测的去噪引导对齐。为减少模拟虚拟传播中的伪影，我们将新闻内容作为语义锚点，通过对称 KL 散度对齐多模态潜在表示。此对齐充当去噪正则项，在提炼共享的真实性相关语义的同时抑制模态特定噪声。

令$\mathbf{x}_c$ 表示内容表示，$\mathbf{x}_g$ 表示（虚拟）传播表示。由于它们由不同的编码器生成，且可能位于异构的特征空间中，我们首先将它们映射到统一的 d维空间：

$$\tilde{\mathbf{x}}_c = f_c(\mathbf{x}_c), \qquad \tilde{\mathbf{x}}_g = f_g(\mathbf{x}_g), \tag{8}$$

其中，$f_c(\cdot)$ 和 $f_g(\cdot)$ 是线性投影层。此步骤可确保后续的对齐与融合基于具有可比性的表示进行。

随后，我们使用两个变分自编码器（VAE）[29] 将这些对齐后的特征映射到潜在空间中。具体而言，我们将每个模态建模为共享语义信号$\mathbf{s}$与模态特定噪声 $\boldsymbol{\epsilon}$的组合：

$$\tilde{\mathbf{x}}_c = \mathbf{s} + \boldsymbol{\epsilon}_c, \quad \tilde{\mathbf{x}}_g = \mathbf{s} + \boldsymbol{\epsilon}_g, \tag{9}$$

其中 $\boldsymbol{\epsilon}_c$，$\boldsymbol{\epsilon}_g$ 表示平均而言无偏且与$\mathbf{s}$不相关的噪声项。每种模态都被编码为潜变量$\mathbf{z}$上的高斯后验：

$$q_\phi(\mathbf{z} \mid \tilde{\mathbf{x}}_c) = \mathcal{N}\big(\boldsymbol{\mu}_c, \Sigma_c\big), \quad q_\phi(\mathbf{z} \mid \tilde{\mathbf{x}}_g) = \mathcal{N}\big(\boldsymbol{\mu}_g, \Sigma_g\big), \tag{10}$$

其中 $\boldsymbol{\mu}_c$，$\boldsymbol{\mu}_g$ 和 Σ_c，Σ_g 分别表示各模态潜在后验分布的预测均值向量和协方差矩阵。为确保潜变量$\mathbf{z}$ 保留 $\tilde{\mathbf{x}}_c$ 和 $\tilde{\mathbf{x}}_g$中各模态特有的核心语义，我们为每种模态引入重构损失：

$$\mathcal{L}_{\text{rec}}(\tilde{\mathbf{x}}) = \mathbb{E}_{q_\phi(\mathbf{z}|\tilde{\mathbf{x}})}\big[-\log p_\psi(\tilde{\mathbf{x}} \mid \mathbf{z})\big]. \tag{11}$$

该目标促使潜变量$\mathbf{z}$ 保留输入 $\tilde{\mathbf{x}}$中的核心信息，以便进行重构。此外，这一表述还隐式建立了与互信息 $I(\tilde{\mathbf{x}}; \mathbf{z})$ 的联系，该互信息量化了$\mathbf{z}$ 保留了多少关于 $\tilde{\mathbf{x}}$的信息：

$$I(\tilde{\mathbf{x}}; \mathbf{z}) = \sum_{\tilde{\mathbf{x}}, \mathbf{z}} \mathbb{P}(\tilde{\mathbf{x}}, \mathbf{z}) \log \frac{\mathbb{P}(\tilde{\mathbf{x}}, \mathbf{z})}{\mathbb{P}(\tilde{\mathbf{x}})\,\mathbb{P}(\mathbf{z})}. \tag{12}$$

输入与潜变量之间的互信息可以下界为：

$$I(\tilde{\mathbf{x}}; \mathbf{z}) \geq H(\tilde{\mathbf{x}}) - \mathbb{E}_{\tilde{\mathbf{x}}}[\mathcal{L}_{\text{rec}}(\tilde{\mathbf{x}})]. \tag{13}$$

通过最小化 $\mathrm{L_{rec}}$，我们收紧这一下界，从而有效防止潜变量$\mathbf{z}$ 坍缩为无信息表示，并确保在编码过程中保留真实性相关特征。

虽然互信息确保了语义保持，但模态特定噪声 $\boldsymbol{\epsilon}$ 仍可能导致分布偏移。为缓解这种噪声失配问题，我们在模态特定后验之间引入对称 KL（SKL）散度，分别记为 $q_c = q_\phi(\mathbf{z} \mid \tilde{\mathbf{x}}_c)$ 和 $q_g = q_\phi(\mathbf{z} \mid \tilde{\mathbf{x}}_g)$：

$$\mathcal{L}_{\text{skl}} = \frac{1}{2}\Big[\mathrm{KL}(q_c \,\|\, q_g) + \mathrm{KL}(q_g \,\|\, q_c)\Big]. \tag{14}$$

该项作为去噪正则项，用于对齐潜在分布。令$\mathbf{W}$ 表示映射到潜在空间的线性映射矩阵。通过展开高斯后验的 SKL 散度，我们推导出一个二阶下界（详见第 4.3.3 节）：

$$\mathcal{L}_{\text{skl}} \geq \frac{1}{2}\left\|\mathbf{W}\big(\boldsymbol{\epsilon}_c - \boldsymbol{\epsilon}_g\big)\right\|^2_{\Sigma_c^{-1}}. \tag{15}$$

关键在于，共享语义成分$\mathbf{W}\mathbf{s}$ 在均值差异 $(\boldsymbol{\mu}_c - \boldsymbol{\mu}_g)$ 中相互抵消，因此该下界只惩罚由模态特定噪声引起的差异。最小化 $\mathrm{L_{skl}}$ 能够通过将虚拟传播锚定到相对稳定的内容语义上，有效抑制其中固有的噪声。

在完成去噪与对齐阶段后，我们获得了经校准的潜在表示$\boldsymbol{\mu}_c$ 和 $\boldsymbol{\mu}_g$。为了自适应地整合这些模态，我们采用门控融合机制，根据潜在语义动态调整它们的贡献权重：

$$a = \sigma\left(\mathbf{W}_f[\boldsymbol{\mu}_c \oplus \boldsymbol{\mu}_g] + \mathbf{b}_f\right), \tag{16}$$

$$\mathbf{o}_f = a \cdot \boldsymbol{\mu}_c + (1 - a) \cdot \boldsymbol{\mu}_g, \tag{17}$$

其中，$\oplus$ 表示拼接，σ 表示 Sigmoid 激活函数。门控值 $a \in (0,1)$ 充当模态级注意力，在模拟的

传播表现出较高的不确定性或存在伪影时，优先关注可靠的内容锚点。随后将融合表示$\mathbf{o}_f$输入分类网络：

$$\hat{y} = \text{Softmax}\left(\text{MLP}(\mathbf{o}_f)\right). \tag{18}$$

最终目标函数 L_{total} 将交叉熵分类损失 L_{cls} 与两个 VAE驱动的正则项结合起来，以确保语义重建和结构对齐：

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} + \lambda_{\text{rec}} \cdot \mathcal{L}_{\text{rec}} + \lambda_{\text{skl}} \cdot \mathcal{L}_{\text{skl}}, \tag{19}$$

其中， λ_{rec} 和 λ_{skl} 是超参数，用于平衡检测准确率、信息保留以及各模态中的噪声抑制。

4.3.3 去噪引导对齐的理论依据。 本节为第 4.3.2 节中使用的两个正则项提供理论依据。我们证明：(i) 重建目标 L_{rec} 收紧了输入 $\mathbf{x}$与潜变量$\mathbf{z}$之间互信息的下界，从而防止表示坍塌；(ii) 对称 KL 散度（SKL）主要抑制模态特定噪声差异，而非缩减共享语义。

作为互信息下界的重建。 我们引入重构损失 L_{rec}，以促使潜变量$\mathbf{z}$ 保留输入$\mathbf{x}$中的核心语义。最小化该损失具有双重作用：它训练解码器根据$\mathbf{z}$ 重建$\mathbf{x}$ ，同时从理论上收紧二者互信息的可证明下界，从而防止$\mathbf{z}$ 坍塌为无信息表示。具体而言，

$$\begin{aligned} I(\mathbf{x};\mathbf{z}) &= \mathbb{E}_{\mathbf{x},\mathbf{z}}\left[\log \frac{p(\mathbf{x}\mid\mathbf{z})}{p(\mathbf{x})}\right] \\ &= \mathbb{E}_{\mathbf{x},\mathbf{z}}\left[\log \frac{p_\psi(\mathbf{x}\mid\mathbf{z})}{p(\mathbf{x})}\right] + \mathbb{E}_{\mathbf{x},\mathbf{z}}\left[\log \frac{p(\mathbf{x}\mid\mathbf{z})}{p_\psi(\mathbf{x}\mid\mathbf{z})}\right] \\ &= \mathbb{E}_{\mathbf{x},\mathbf{z}}\left[\log \frac{p_\psi(\mathbf{x}\mid\mathbf{z})}{p(\mathbf{x})}\right] + \mathbb{E}_{\mathbf{z}}\Big[\mathrm{KL}\big(p(\mathbf{x}\mid\mathbf{z}) \,\|\, p_\psi(\mathbf{x}\mid\mathbf{z})\big)\Big] \\ &\geq \mathbb{E}_{\mathbf{x},\mathbf{z}}\left[\log \frac{p_\psi(\mathbf{x}\mid\mathbf{z})}{p(\mathbf{x})}\right] = -\mathbb{E}_{\mathbf{x}}\left[\mathcal{L}_{\text{rec}}(\mathbf{x})\right] + H(\mathbf{x}). \end{aligned} \tag{20}$$

因此，最小化 L_{rec} 可收紧式 (20) 中的下界，从而保证$\mathbf{z}$ 保留有关$\mathbf{x}$的关键信息。

为什么L_{rec} 倾向于保留共享语义。为探究噪声抑制机制，我们将带有压缩因子 $\eta \in [0,1]$ 的潜在均值定义为 $\boldsymbol{\mu}_*(\eta) = \eta\,\mathbf{W}\mathbf{s} + \mathbf{W}\epsilon_*$。在高斯解码器假设 $\hat{}\mathbf{x}_* = \mathbf{W}^{-1}\mathbf{z}_*$下，重构损失可简化为均方误差目标。将输入分解为$\mathbf{x} \approx \mathbf{W}(\mathbf{s} + \epsilon_*)$后，关于 η 的损失为：

$$\mathcal{L}_{\text{rec}}(\eta) = \frac{1}{2\sigma_x^2}\,\|\mathbf{x} - \hat{\mathbf{x}}_*\|^2 + \kappa = \frac{\|\mathbf{W}\mathbf{s}\|^2}{2\sigma_x^2}(1-\eta)^2 + \kappa, \tag{21}$$

此处， κ 表示与 η无关的噪声相关项。对式 (21) 求导可得$\partial^2 \mathcal{L}_{\partial \eta^2}^{\text{rec}} > 0$，这表明其具有严格凸性，并在 $\eta = 1$处达到全局最小值。在完整目标函数（式 19）的语境下，由于 L_{skl} 对均值差中的语义成分不敏感，L_{rec} 中的二次惩罚项便充当主要正则项，确保$\mathbf{W}\mathbf{s}$ 得到保留。这一机制有效过滤掉模态特定噪声 ϵ_* ，同时保留共享语义。

通过对称 *KL* 散度进行去噪。鉴于语义信息通过 L_{rec} 得以保留，接下来我们展示 SKL 如何抑制两个潜在后验分布之间残留的噪声差异。

表 2. 数据集统计信息。

数据集	GossipCop	PolitiFact	微博	GossipCop-P	PolitiFact-P
新闻总数	10,350	741	4,704	5,683	489
假新闻	2,387	391	2,772	1,969	199
真实新闻	7,963	350	1,935	3,714	290
图像	10,350	298	4,707	5,683	224

对于高斯分布 后验分布 $q_c = \mathrm{N}(\boldsymbol{\mu}_c, \ \Sigma_c)$ 和 $q_g = \mathrm{N}(\boldsymbol{\mu}_g, \ \Sigma_g)$，我们推导出以下下界：

$$\begin{aligned}
\mathrm{SKL}(q_c, q_g) &= \frac{1}{2}\left[\mathrm{tr}\left(\Sigma_c^{-1}\Sigma_g + \Sigma_g^{-1}\Sigma_c\right) - 2d + \Delta\boldsymbol{\mu}^\top\left(\Sigma_c^{-1} + \Sigma_g^{-1}\right)\Delta\boldsymbol{\mu}\right] \\
&\geq \frac{1}{2}\left[\mathrm{tr}\left(\Sigma_g^{-1}\Sigma_c + \Sigma_c^{-1}\Sigma_g\right) - 2d + \Delta\boldsymbol{\mu}^\top\Sigma_c^{-1}\Delta\boldsymbol{\mu}\right] \\
&= \frac{1}{2}\left[\mathrm{tr}\left(\boldsymbol{\Lambda} + \boldsymbol{\Lambda}^{-1} - 2\mathbf{I}\right) + \|\Delta\boldsymbol{\mu}\|^2_{\Sigma_c^{-1}}\right] \\
&\geq \frac{1}{2}\|\Delta\boldsymbol{\mu}\|^2_{\Sigma_c^{-1}} = \frac{1}{2}\|\boldsymbol{\mu}_c - \boldsymbol{\mu}_g\|^2_{\Sigma_c^{-1}} \\
&= \frac{1}{2}\left\|\mathbf{W}(\epsilon_c - \epsilon_g)\right\|^2_{\Sigma_c^{-1}},
\end{aligned} \tag{22}$$

其中， $\Delta\boldsymbol{\mu} = \boldsymbol{\mu}_c - \boldsymbol{\mu}_g$ 表示高斯均值之差， $\boldsymbol{\Lambda} = \Sigma_c^{-1/2}\Sigma_g\Sigma_c^{-1/2}$刻画相对协方差结构， d 表示潜变量的维度。式（22）表明，共享语义成分**s** 在均值差中完全抵消。因此，最小化对称 KL 散度充当了一种正则项，专门用于缩小模态特定噪声 ϵ_c 与ϵ_g之间的距离，从而有效过滤模拟伪影，同时不损害 $\mathrm{L_{rec}}$所保留的共享语义。

5 实验

5.1 实验设置

5.1.1 数据集。 我们使用了从社交媒体收集的五个真实世界数据集。数据集介绍如下：

基于内容的数据集。 本研究使用了三个不同的数据集：来自 FakeNewsNet 的两个英文数据集 PolitiFact 和 GossipCop [55], , 以及中文微博数据集 [24]。PolitiFact 和 GossipCop 分别包含 741 篇和 10,350 篇新闻文章，这些文章收集自事实核查网站，并标注为真实或虚假的；其中部分文章包含图像。微博数据集收集自微博平台，包含 4,707 条中文新闻，每条新闻均具有独特的配图，这对多模态分析至关重要。

传播数据集。 为了将 AVOID 与基于传播的方法进行比较，我们还使用了两个数据集 PolitiFact-P 和 GossipCop-P [13] （为便于区分，将其命名为 PolitiFact-P 和 GossipCop-P）。每个条目都包含一条新闻条目的完整传播轨迹。数据集统计信息见表 2。

5.1.2 实现细节。 我们使用 PyTorch 库实现所提出的 AVOID框架，并在 NVIDIA GeForce RTX 4090 图形处理器上进行所有实验。对于新闻和评论的编码，我们将特征维度设为 768，以便与预训练 BERT-base [11] 模型的输出保持一致。具体而言，新闻内容最多截取 50 个句子，每个句子包含 25 个词元；用户评论则限制为 15 个句子，每个句子包含 10 个词元

。对于基于智能体的模拟器，我们调用 DeepSeek-V3-Chat [36] API，并将 temperature 和 top-k 值分别固定为 0.7 和 0.9。生成的传播图由一个具有 128 个隐藏维度的双层图注意力网络 [60] 处理。在训练过程中，我们使用 Adam优化器 [28]，初始学习率为 $1e{-}4$，小批量大小为 32。我们通过评估验证集上的性能来确定最优超参数；验证集、训练集和测试集按 7:1:2 的比例划分。为防止过拟合，我们采用提前停止 [72] 策略，耐心值设为 6 个周期，并在所有数据集上将训练周期数上限设为 100。

5.1.3 基线方法。为评估 AVOID 的有效性，我们将其与一系列最先进的假新闻检测模型进行比较。为确保比较公平且严谨，所有基线方法均在相同的实验环境下进行基准测试。具体而言，我们的比较涵盖三类不同的方法：（1）利用新闻文本及相关图像的基于内容的模型；（2）借助大语言模型进行预测的大语言模型增强方法；（3）明确利用新闻结构化传播模式的基于传播的方法。

基于内容的假新闻检测模型：

- BERT [10]：这是一种预训练语言模型，我们对其最后两层进行微调以用于检测。
- HAN [71]：利用新闻文本构建层次注意力网络，以捕捉词级和句子级特征，从而识别假新闻。
- dEFEND [54]：提出了一种可解释的假新闻检测模型，该模型利用正文文本和用户评论，找出 k 条可解释评论，以改进假新闻检测。
- CAFE [8]：构建跨模态对齐模块，将不同模态的特征转换到共享语义空间中，然后评估不同模态之间的歧义。
- HMCAN [51]：设计层次多模态注意力网络，以学习多模态新闻表示。
- BREAK [73]：将新闻文本和图像建模为全连接的句子—图像图，结合序列编码器和图编码器，并通过去噪捕捉广泛范围的语义，用于假新闻检测。
- CSFND [50]：设计一种无监督假新闻检测框架，以捕捉新闻语义特征空间与虚假新闻决策空间之间的关系。
- ALGM [12]：提出一种基于马尔可夫随机场的框架，并根据歧义融合跨模态特征。
- MIMoE-FND [39]：引入一种层次混合专家框架，通过单模态预测一致性和语义对齐显式建模文本—图像模态交互，并将帖子路由至专用融合专家，以实现稳健的多模态检测。

除传统的基于内容的基线方法外，我们还进一步比较以下基于大语言模型的假新闻检测模型：

- ARG [20]：研究大语言模型如何辅助假新闻检测，并引入一种自适应推理引导网络，由小模型利用大语言模型生成的推理依据来提升真实性预测能力。
- L-Defense [62]：利用群体智慧提取证据，并通过提示大语言模型生成解释依据。
- GenFEND [44]：利用大语言模型对多样化的用户档案进行角色扮演，为每条新闻内容生成合成评论，并汇总这些生成的反馈信号，以增强假新闻检测。

• SheepDog [68]：提出一种风格稳健的基于文本的假新闻检测器，利用大语言模型生成的风格多样的新闻改写文本，以及聚焦内容的真实性线索，在不同写作风格下做出稳定预测。

• Adstyle [46]：采用大语言模型生成的对抗式风格转换来扩充训练数据，并学习能够抵御风格迁移攻击的文本假新闻检测器。

基于传播的假新闻检测模型：

• UGGCN [30]：在新闻传播图上应用图卷积网络对节点表示建模，并通过聚合邻域信息进行半监督假新闻分类。

• Bi-GCN [5]：在原始传播图和反向传播图上应用双向图卷积网络，以捕捉社交媒体上的谣言传播和辟谣模式，从而检测谣言。

• GACL [58]：在传播图上采用图对抗式对比学习，通过联合应用对抗扰动和对比目标，学习稳健的新闻表示以检测谣言。

• UPFD [13]：通过用户过去互动过的帖子学习用户偏好，并结合内容与图建模。

• MFAN [80]：在统一框架中融合文本、视觉和社交图特征，用于谣言检测。

• PSGT [82]：设计了一种传播结构感知图变换器，可过滤嘈杂的用户交互，并对多尺度扩散结构进行建模，从而实现鲁棒且可解释的假新闻检测。

5.2 性能比较

我们首先针对仅基于内容的早期假新闻检测，对 AVOID 与基线方法进行性能基准测试，其中包括基于文本的模型、基于多模态内容的方法和大语言模型增强方法。我们在两个英文数据集（PolitiFact 和 GossipCop）以及一个中文数据集（微博）上开展实验，以全面评估检测性能。结果汇总于表3，可得出以下几点主要发现：

• AVOID 在所有数据集上始终取得最佳整体性能。如表3所示，AVOID 在 PolitiFact、GossipCop 和微博上均获得最高分。与各数据集上最强的基线方法相比，AVOID 在 PolitiFact 上的准确率提高了 3.67%，在 GossipCop 上提高了 2.08%，在微博上提高了 1.96%，且精确率、召回率和 F1值均有稳定提升。AVOID 在三个异构平台上的这一优势表明，它能够有效泛化至不同的新闻领域和语言。

• AVOID 的表现优于传统的多模态基于内容的基线方法（例如 CAFE、HMCAN）。与大多数直接融合文本和视觉特征进行分类的基线方法不同，AVOID 利用多模态信号驱动智能体在虚拟传播过程中的决策。结果表明，通过模拟传播图的交互模式隐式编码多模态证据，比静态特征融合更能有效地进行检测。

• 尽管大语言模型增强方法通常通过使用大语言模型生成辅助信号或直接进行推理，整体表现优于传统检测器，但 AVOID 取得了进一步提升。这表明，仅使用大语言模型进行语义理解，不如 AVOID 的策略有效：利用大语言模型赋能的智能体模拟真实世界中的社交互动。其卓越表现验证了从虚拟传播路径中获取的交互证据能够提供关键的额外监督。

表3。AVOID 与基线方法的性能比较。最佳结果以**粗体**显示，次佳结果加下划线。

类别	方法	PolitiFact				GossipCop				微博			
		Acc.	Pre.	Rec.	F1	Acc.	Pre.	Rec.	F1	Acc.	Pre.	Rec.	F1
基于内容	BERT	0.801	0.799	0.801	0.798	0.828	0.820	0.828	0.801	0.863	0.868	0.854	0.859
	HAN	0.857	0.859	0.854	0.856	0.837	0.823	0.837	0.831	0.863	0.863	0.855	0.858
	dEFEND	0.825	0.830	0.821	0.823	0.849	0.846	0.849	0.847	0.857	0.865	0.843	0.850
	CAFE	0.773	0.754	0.780	0.759	0.814	0.824	0.814	0.820	0.836	0.857	0.837	0.847
	HMCAN	0.894	0.898	0.892	0.896	0.836	0.818	0.836	0.820	0.896	0.912	0.881	0.890
	BREAK	0.920	0.921	0.920	0.921	0.864	0.861	0.864	0.863	0.899	0.901	0.897	0.899
	CSFND	0.907	0.907	0.907	0.907	0.837	0.851	0.837	0.848	0.886	0.874	0.882	0.874
	ALGM	0.887	0.895	0.881	0.886	0.831	0.817	0.831	0.805	0.854	0.858	0.861	0.866
	MIMOE-FND	0.878	0.892	0.883	0.886	0.854	0.847	0.854	0.847	0.912	0.917	0.918	0.909
大语言模型增强	ARG	0.926	0.924	0.928	0.924	0.852	0.812	0.852	0.823	0.908	0.913	0.902	0.905
	L-Defense	0.893	0.902	0.899	0.896	0.862	0.858	0.862	0.861	0.875	0.874	0.875	0.875
	GenFEND	0.899	0.897	0.900	0.898	0.850	0.853	0.850	0.848	0.910	0.906	0.910	0.907
	SheepDog	0.913	0.912	0.909	0.911	0.843	0.831	0.843	0.834	0.905	0.908	0.902	0.909
	Adstyle	0.899	0.904	0.899	0.901	0.857	0.851	0.857	0.849	0.918	0.913	0.917	0.916
我们的方法	**AVOID**	**0.960**	**0.956**	**0.960**	**0.958**	**0.882**	**0.876**	**0.882**	**0.879**	**0.936**	**0.930**	**0.932**	**0.934**
	提升（%）	+3.67	+3.46	+3.44	+3.68	+2.08	+1.74	+2.08	+1.85	+1.96	+1.42	+1.53	+1.86

- AVOID 展示了传播层面的社交证据相较于静态评论增强的价值。具体而言，dEFEND 使用真实用户评论，GenFEND 则采用大语言模型生成的评论，而 AVOID 通过模拟异质智能体之间的互动来建模动态社交信号。AVOID 持续优于这些基线方法，说明虚拟传播图中的结构与交互模式比单条评论特征更能为早期假新闻检测提供信息丰富且互补的线索。

5.3 早期检测评估

为了评估 AVOID 在新闻传播早期阶段的表现，尤其是在只能观察到极少部分真实传播甚至完全没有传播数据的情况下，我们在 PolitiFact-P 和 GossipCop-P 上设计了一项早期检测实验。

具体而言，对于每条新闻内容，我们按时间顺序排列所有用户交互，并定义三个早期传播阶段：(1) 仅内容（阶段0），我们仅使用新闻内容本身，并舍弃所有真实传播交互，以模拟一种极端早期检测场景，此时可观察到的传播信息为零；(2) 10%阶段，我们将真实传播截断至时间上最早的10%交互，以模拟传播刚刚开始、系统只能观察到级联中一小段内容的阶段；(3) 30%阶段，同样将真实传播截断至时间上最早的30%交互，以模拟新闻已传播到一定程度、但仍处于相对早期阶段的情形。

基于重建的传播图，我们对AVOID与不同类别中的代表性基线方法进行了比较分析。表4总结了它们的性能，从中可以观察到以下几点：

- 在早期传播设置下，AVOID始终取得更优的检测表现。在PolitiFact-P和Twitter上，无论是仅使用内容（stage0），还是将

表4. 使用时间截断的传播进行早期假新闻检测时，基于传播的方法在 PolitiFact-P 和 Twitter 上的性能。每列中的最佳结果以**粗体** 标出，第二佳结果以下划线标出。

类别	方法	PolitiFact-P						GossipCop-P					
		0		0.1		0.3		0		0.1		0.3	
		Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
传播增强	UDGCN	-	-	0.706	0.705	0.782	0.782	-	-	0.791	0.794	0.814	0.823
	双向图卷积网络	-	-	0.661	0.632	0.777	0.774	-	-	0.824	0.822	0.841	0.836
	GACL	-	-	0.808	0.783	0.810	0.806	-	-	0.810	0.828	0.836	0.842
	UPFD	-	-	0.860	0.835	0.886	0.874	-	-	0.824	0.858	0.867	0.881
	MFAN	-	-	0.842	0.825	0.872	0.860	-	-	0.824	0.865	0.864	0.881
	PSGT	-	-	0.863	0.852	0.872	0.868	-	-	0.869	0.877	0.881	0.885
基于内容	BERT	0.784	0.788	0.784	0.788	0.784	0.788	0.804	0.816	0.804	0.816	0.804	0.816
	L-Defense	0.880	0.871	0.880	0.871	0.880	0.871	0.866	0.841	0.866	0.841	0.866	0.841
	BREAK	0.872	0.868	0.874	0.868	0.874	0.868	0.878	0.874	0.878	0.874	0.878	0.874
提出的方法	**AVOID**	**0.908**	**0.902**	**0.908**	**0.902**	**0.908**	**0.902**	**0.893**	**0.896**	**0.893**	**0.896**	**0.893**	**0.896**
	提升（%）	+3.18	+3.56	+3.18	+3.56	+2.48	+3.20	+1.68	+2.52	+1.71	+2.17	+1.36	+1.24

真实传播截断至最早的10%或30%互动，AVOID在每一列中都取得了最佳准确率和F1值，并且明显优于所有基于传播的基线方法。这表明，在“只能观察早期传播片段”的限制下，相较于依赖完整真实传播路径的方法，AVOID更适用于早期假新闻检测。

• AVOID 通过在真实传播不可用时提供可用的替代方案，有效缓解了冷启动限制。在尚未出现任何可观察级联的极端情况下（阶段0），传播增强基线方法将无法应用。然而，AVOID 利用智能体模拟扩散作为社交反馈的替代指标，始终优于具有竞争力的基于内容的模型。即使无法获取真实传播样本，这一性能提升也证实，我们的框架在新闻传播的最早阶段仍具有很高价值，能够从模拟互动中成功提炼出与真实性相关的语义。

5.4 消融研究

为评估 AVOID 各组成部分的独立贡献，我们将五种变体与完整模型进行比较。首先，*无过滤器* 移除了基于置信度的过滤阶段，并为每条新闻内容模拟传播路径，而不仅限于基检测器识别出的困难样本。其次，*无人物画像* 不将提取的人物画像信息纳入用户画像。第三，*无图像* 仅考虑新闻文章的文本内容，忽略视觉输入。第四，*无验证器* 移除了专门的已认证智能体，仅保留扩散智能体进行互动。最后，为测试特征对齐策略，*无 SKL* 移除了 SKL去噪模块，改为直接拼接文本特征和图特征。

如图3所示，消融实验结果验证了 AVOID 中各个组件的独特贡献。具体分析如下：

• 过滤器的影响（不使用过滤器）：移除基于置信度的过滤器，并为每条新闻内容模拟传播，可略微提升性能——在 PolitiFact 上约提升 1.2%，

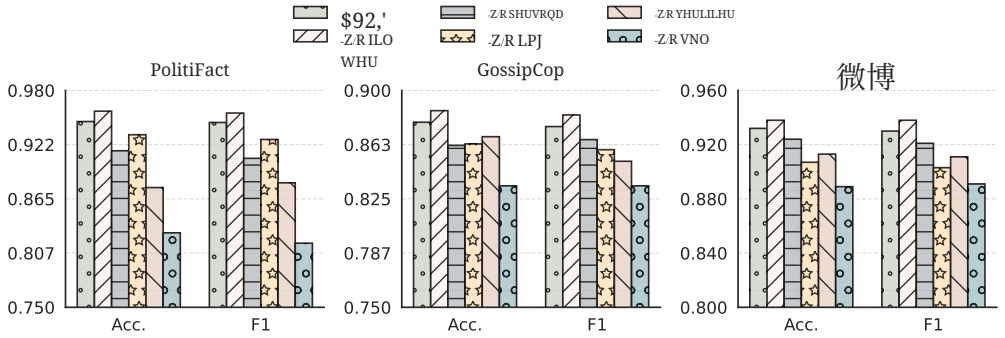

图 3. AVOID 在三个数据集上的消融研究。

在 GossipCop 上提升 0.6%，在微博上提升 0.8%——但所需词元数量会增加三倍以上。这表明，我们默认的 AVOID 配置有意以少量准确率为代价，换取显著的效率提升；当计算资源充足时，可以禁用过滤器，并为所有新闻模拟传播，以获得尽可能好的性能。

- 特征对齐的必要性（无 *SKL*）：移除 SKL 模块后，性能大幅下降，验证了该模块在缓解模态差异方面的关键作用。没有 SKL 时，直接拼接未对齐的特征会引入语义噪声，妨碍稳健决策。
- 已认证智能体的作用（无验证器）：缺少已认证智能体机制会导致准确率明显下降。对高可信度代理进行建模，对于引导传播并有效遏制谣言传播至关重要。
- 用户画像的益处（无人物画像）：移除人格嵌入会导致性能持续下降，这表明使用真实用户数据能够提供有价值的社会语境。融入个性化特征被证明是全面提升模型鲁棒性的有效方法。
- 图像特征的贡献（无图像）：排除图像特征会导致性能略有但明显的下降。视觉信息发挥着补充作用，当模型高度依赖结构和文本线索时，可提供额外信号。

5.5 与真实世界传播级联的比较

尽管初步实验验证了 AVOID 在检测性能方面的有效性，但这仅能间接证明其生成的传播过程是否忠实。为此，我们设计了额外实验，将 AVOID 的模拟传播路径与社交平台上观察到的真实级联直接进行比较，以评估大语言模型驱动的多智能体模拟能够在多大程度上复现真实扩散模式。

具体而言，我们从结构和行为对齐两个角度比较真实级联与模拟级联。在结构方面，我们使用级联深度、平均节点度、边密度、结构病毒性（SV）和聚类系数来刻画传播图，并额外报告度分布之间的詹森–香农散度（JSD），以捕捉平均度之外的分布层面差异。在行为方面，我们通过报告验证者数量以及评论的立场分布（正面/中立/负面），评估模拟是否再现了用户反应的异质性，从而量化群体层面的意见极性。定量结果汇总于表5和表6，据此我们得出以下结论：

表5. 比较　　两个数据集上虚拟传播路径与真实数据的图属性比较。

数据集	Type	级联深度	度	密度	SV	簇JSD	
PolitiFact-P	Real	4.203	1.884	0.016	2.808	0.114	0.078
	虚拟	4.452	2.163	0.022	2.932	0.124	
GossipCop-P	Real	3.086	2.123	0.051	2.302	0.070	0.104
	虚拟	3.221	2.635	0.043	2.186	0.065	

表6. 真实级联与虚拟级联的行为属性比较，包括验证者频率和评论立场分布。真实验证器根据已验证用户元数据统计，而虚拟验证器则为预定义代理；真实评论和模拟评论的立场均使用同一分类器进行标注。

数据集	Type	验证器	正面:中立:负面 (%)
PolitiFact-P	real	4.88	54.6 : 9.3 : 36.1
	虚拟	5.00	57.8 : 10.7 : 31.5
GossipCop-P	real	2.29	75.8 : 18.4 : 5.8
	虚拟	2.00	73.2 : 22.1 : 3.7

• AVOID再现了真实级联的关键结构统计特征，包括级联深度、平均度和边密度，表明模拟图与实证数据在规模和连通性模式上相符。詹森–香农散度（JSD）始终较低，进一步表明整体分布高度对齐。结合结构病毒性（SV）方面的一致性，这些结果说明AVOID捕捉到了通常影响现实世界传播的广度与深度之间的特征性权衡。

• 将智能体设定为从真实评论中提炼出的细粒度角色画像，可以提升模拟级联与观测级联之间的行为对齐。这一点得到了次级验证者数量和立场分布的有力支持：真实评论与虚拟评论在这两方面十分接近，表明模拟反映了异质的反应模式，而非生成统一且不考虑行为差异的互动。
• 在真实数据中，PolitiFact-P 和 GossipCop-P 呈现出不同的传播动态和用户行为特征，而 AVOID 在模拟级联中保留了这些数据集特定差异。这表明，该框架能够感知平台和数据集相关的讨论规范，避免将传播模式简化为一套通用模板，并实现更忠实的跨领域模拟。

5.6 超参数敏感性

在 AVOID 模型中，重构损失 L_{rec} 由 λ_{rec} 加权，用于控制单模态特征重构的保真度；而 SKL 对齐损失 L_{skl} 由 λ_{skl} 加权，决定跨模态语义一致性正则化的强度。为考察这两个组件如何共同影响检测性能，我们对 $\lambda_{rec}, \lambda_{skl} \in \{0.1, 0.2, \ldots, 1.0\}$ 进行网格搜索，步长为 0.1，并固定其他所有超参数。对于每组 $(\lambda_{rec}, \lambda_{skl})$ 组合，我们在相应数据集上训练 AVOID，并记录所得的检测准确率，再将其绘制为图 4 中的热力图。

在三个数据集上，重构损失 L_{rec} 和对称 KL 散度对齐损失 L_{skl} 的最优值并不相同，但呈现出一致的趋势：它们会随着

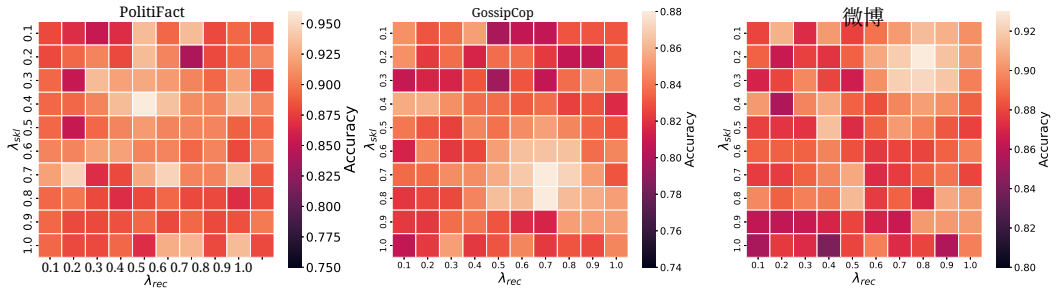

图4. 三个数据集上不同超参数的影响 λ_{rec} 和 λ_{skl} 。

跨模态耦合强度和数据噪声水平而系统性变化，表明这组超参数具有良好的跨数据集泛化能力。具体而言：

- 在 PolitiFact 上，文本主题较为集中，图像主要起辅助作用。模型在中等权重附近（例如 λrec = 0.5, λskl = 0.4）取得最佳性能，这表明适度的跨模态正则化已足够，进一步强化对齐带来的收益有限。相比之下，在 GossipCop 上，最佳性能出现在更大的权重下（约为 λrec = 0.7, λskl = 0.7），这反映出，对于传播模式多样、具有主观性和娱乐性的谣言，需要更强的单模态重构和更强的跨模态一致性，才能抑制文本与图像之间的语义不确定性。

- 在微博上，模型在（λrec = 0.8, λskl = 0.2）处取得最高准确率，并且在周围一个 λrec 较大、 λskl 适中的区域内仍保持较强竞争力（约为λrec $\in [0.5, 0.8]$, λskl $\in [0.2, 0.6]$）。这表明，强单模态重构对该数据集尤其有益，而仅需适度的跨模态对齐即可。该模式与简短且噪声较多的微博帖子相一致，在这类帖子中，图像承担的用途高度多样（例如表情符号、装饰性图片或表情包风格图片）：如果对齐约束过于激进，这些关联较弱的视觉信号可能会被强行映射到共享语义空间中，从而引入额外噪声。

- 由于 λrec 控制单模态鲁棒性，而 λskl 决定模态间语义校准/去噪，因此不应全局固定 AVOID 的 (λrec λskl) 参数。应根据特定数据集固有的文本 - 图像耦合强度和噪声特征进行自适应调整。

5.7 令牌使用量

为更好地了解 AVOID 的资源效率，我们进一步将其词元使用量与其他基于大语言模型的假新闻检测方法进行比较。具体而言，对于每个模型，我们记录评估整个测试集时消耗的词元总数（包括提示词和生成输出），并将其作为推理成本的衡量指标；同时报告相应的检测准确率，以考察性能与效率之间的权衡。

- 在全部三个数据集上，AVOID 始终在基于大语言模型的基线方法中取得最高的检测准确率，同时其令牌消耗量相比其他方法仍然相对较低。在微博数据集上，AVOID 还是令牌使用效率最高的模型，这表明其性能提升并非仅仅依赖于消耗更多词元。

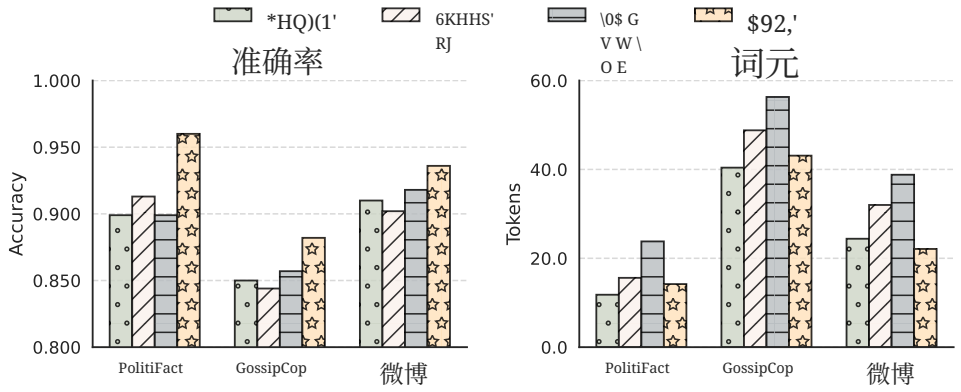

图 5. 三个数据集上各种基于大语言模型的假新闻检测模型在准确率和 token 消耗方面的比较结果。

- 由于 AVOID 仅对通过难度过滤器的项目触发多智能体传播模拟，并使用成本更低的仅基于内容的推理处理简单案例，因此避免了为每个实例模拟完整级联所产生的二次成本。这种选择性策略提升了早期检测性能，同时将总体令牌使用量保持在适中水平，在有效性与资源消耗之间实现了具有实际应用价值的平衡。

5.8 案例研究

我们通过两个互补的案例研究评估 AVOID。首先，对智能体推理进行微观层面分析，以展示其生成机制。其次，通过对比模拟级联与真实级联，验证生成的传播在结构上的保真度。



图6. 模拟错误信息传播的案例研究，展示具有不同角色设定、记忆和社会特征的智能体如何响应并影响传播。

5.8.1 微观层面分析：智能体推理与生成。 图 6 展示了 AVOID 生成的一条模拟传播路径的生命周期，说明单个智能体的认知如何转化为集体扩散的动态过程。模拟首先根据从真实世界数据中提炼出的细粒度人物画像来实例化智能体。如图所示，一个智能体被赋予“强烈支持特朗普者”的个人资料，并继承特定的意识形态先验和行为模式。这种数据驱动的依据确保模拟从具备真实社交属性的智能体开始，而非采用通用的随机行为。

遇到目标虚假信息（例如“凌晨3点投票激增”的说法）时，智能体会触发检索增强推理过程。通过查阅其长期记忆，智能体检索出语境相关信念——例如认为“俄罗斯通俄门是一场政治迫害”的观点。这种检索强化了智能体的确认偏误，使新主张与其既有世界观相一致。因此，这种认知一致性促成了“立即分享”内容的决定，反映出对该刺激的一种合理心理反应。

最后，这些个体微观决策的汇总形成了信息流的宏观层面结构。每次互动都会形成一条有向边，最终构成完整的虚拟传播图。该生成图可作为缺失真实世界数据的结构性代理，为下游检测器提供稳健早期检测所需的丰富、动态的社会语境。

5.8.2 宏观层面分析：扩散保真度与图结构。为了加强关于扩散保真度的宏观层面证据，我们将 AVOID 模拟的级联与其真实世界对应案例进行可视化对比。如图7所示，我们从 PolitiFact 和 GossipCop 数据集中选取具有代表性的假新闻实例，以展示我们的框架能够保留显著的经验结构特征。

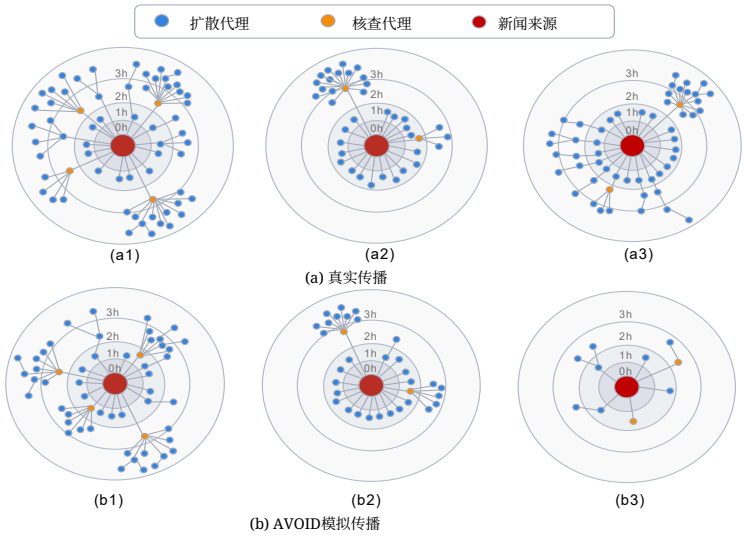

图7。第一行（a1–a3）展示真实传播级联，第二行（b1–b3）展示 AVOID 模拟级联。

在图7的前两列中，AVOID 与真实的传播模式高度吻合。在 PolitiFact 和 GossipCop 数据集中，模拟级联（b1、b2）都与

相应的真实图（a1、a2）在关键结构特征上相似，包括分支深度相近，以及中间层高连接度传播者周围局部簇的形成。这种对齐表明，以现实人物画像和记忆检索为基础来设定智能体行为，可以重现塑造错误信息传播的社会互动模式。

对于第三个示例（PolitiFact 14448），模拟级联远小于真实级联。该主张是：煽动性阴谋论叙事——声称存在由奥巴马领导的政变、“新世界秩序”象征、秘密隧道会面，以及针对特朗普支持者的“人口削减”阴谋。在 AVOID 的人物画像与记忆驱动的决策过程中，既明显不可信又带有公开煽动性的内容会被视为不值得放大，因此传播很快便会消退，只有少数智能体参与其中（b3）。这一行为也与底层大语言模型的安全对齐相一致，后者进一步抑制智能体传播高度耸动且以伤害为框架的叙事。

6 相关工作

6.1 假新闻检测

6.1.1 以内容为中心的早期假新闻检测。早期的假新闻检测方法主要关注新闻文章的文本内容。这些方法结合词嵌入与卷积神经网络或循环神经网络，以学习文章文本的表示 [7, 71]。预训练语言模型的出现标志着一项重大进展。后续研究利用基于 BERT 的 [10] 编码器生成上下文化表示，从而捕捉新闻内容中更丰富的语义信息，并取得了显著的性能提升 [11, 26]。在此基础上，研究人员探索了对文章内部长距离依赖关系进行建模的技术。例如，BREAK [73] 联合建模句子级图结构和序列信息，以获得更具判别性的内容表征，进一步减轻结构级噪声和特征级噪声。与此同时，一些研究通过在文本之外引入视觉模态，扩展了内容建模的范围。具有代表性的多模态方法 [8, 39, 50, 51] 联合编码文本和视觉特征，以学习跨模态交互。通过融合这些异构信号，此类模型旨在捕捉互补证据，并利用跨模态相关性进行更稳健的真实性预测。

6.1.2 大语言模型增强的早期假新闻检测。大语言模型带来了强大的语言理解和推理能力，重塑了错误信息检测。然而，直接使用零样本方法进行细粒度真实性判断往往并不可靠 [49, 81]。为缓解这一问题，早期工作 [48] 探索了监督式适配和大语言模型辅助检测：大语言模型能够提供多角度理由，但可能难以可靠地筛选并整合证据，得出最终判断；后续方法则提炼或整合此类推理依据，以便在成本敏感场景中指导小型检测器 [20, 62]。除了有效性，鲁棒性也日益成为一项关键考量：由大语言模型驱动的风格改写可能掩盖虚假新闻，并削弱依赖风格的检测器，因此促使研究者设计不依赖风格或侧重归因的方法，以突出内容真实性 [46, 68]。

除了依赖大语言模型的参数化知识外，近期研究日益转向基于证据的验证，在此过程中，模型会在推理期间与外部来源进行显式交互。检索增强生成将检索与生成相结合，从搜索引擎或知识库中获取最新证据，减少幻觉并缓解知识截止问题[9, 31]。在这一范式的基础上，基于分解的方法进一步将复杂主张拆分为可验证的子问题，并汇总子结果，以实现更细粒度的评估 [21],；与此同时，工具增强框架支持多步骤验证，并生成可审计的证据链；例如，Self-Checker 提供即插即用模块和一个策略代理，由其规划验证操作并输出可验证的推理依据 [34]。

不同于这些“单模型”增强方法，基于智能体的方法通过角色专门化、辩论和交叉核查，协调多个大语言模型智能体开展协作式验证。Debate-to-Detect [17] 将错误信息检测重新表述为结构化的多阶段对抗性辩论，为智能体分配领域专属档案，并依据多维评判标准生成可解释判定。MARO [32] 通过将分析分解给多个专家智能体，面向跨领域虚假信息检测；该方法进一步引入问题反思机制和自动化决策规则优化，以提升泛化能力。LoCal [42] 提出了一种基于大语言模型的多智能体事实核查框架，强调逻辑与因果一致性，将分解智能体、专门化推理智能体和评估智能体相结合，以迭代验证复杂主张。

6.1.3 社会语境与基于传播的检测。除了内容之外，许多方法还利用社会语境，例如将用户评论和互动作为真实性预测的补充信号 [54]。另一项重要研究方向是将传播轨迹建模为传播图：把用户或帖子视为节点，将转发、回复和评论关系视为边，从而利用基于图的方法或时间聚合模块捕捉谣言相关结构模式，以进行检测 [5, 13, 58, 82]。然而，在早期阶段，这类传播信号往往十分稀疏，甚至完全不可用，导致基于图的检测器在冷启动场景下可靠性不足。为缓解这一局限，近期研究将传播生成视为图合成任务：它们不再只是从观测到的级联中提取结构模式，而是学习生成模型，以生成合理的传播图或扩散路径，并将这些合成结构作为额外的传播证据，从而增强冷启动场景下的下游谣言检测器 [19, 41, 76]。此外，一些方法 [44, 63] 使用基于大语言模型的生成器，并通常通过角色设定多样化来应对早期评论匮乏的问题，生成逼真的讨论式评论，将其作为补充社会证据，以便在真实互动有限时提升早期谣言检测效果。

6.2 基于大语言模型的智能体与社会模拟

将大语言模型融入基于智能体的社会模拟，已成为研究人类行为及其集体动态的一个有前景的方向。早期研究表明，通过精心设计提示词的角色扮演和交互协议，大语言模型能够复现经济学、心理语言学和社会心理学领域经典实验的结果 [2, 18, 69]。在此基础上，近期研究已从孤立的决策任务转向大规模模拟场景：多个由大语言模型驱动的智能体置身于交互式环境中，借助记忆和反思生成长期日常活动，生成式智能体社会及相关模拟器便是其中的例子 [25, 45, 64]。除开放式虚拟社会外，大语言模型智能体也被用作平台级生态系统中基于特定领域的用户替身；例如，RecAgent 对真实用户行为进行建模，以支持推荐系统模拟和干预研究 [66]。此外，重复博弈和社会困境等受控社会博弈测试平台，为探究大语言模型智能体之间的策略性交互、合作与协调提供了严谨的方法 [35, 70]。

特别是，社交媒体为此类验证提供了天然的测试平台，推动近期研究将基于大语言模型的社会模拟扩展至社交网络环境：大语言模型赋能的智能体被置于显式交互图上，用于模拟网络化交互和信息扩散。例如，Gao等人 [16] 提出了 S^3,，该方法基于现实世界网络数据实例化由大语言模型驱动的用户，并对情绪、态度和互动这三个关键行为方面进行显式模拟，从而使群体层面现象（例如信息、态度和情绪的传播）能够由个体层面的感知与行动涌现出来。与此同时，Liu等人 [38] 提出了一种基于大语言模型的假新闻传播模拟框架，其中每个智能体都以人格为条件

并配备短期/长期记忆和反思能力，从而能够逐日进行观点交流、态度更新，并沿着从怀疑到接受的过程评估干预策略。一些研究则将经典扩散过程与基于大语言模型的组件相结合，以提升可控性和可扩展性：它们保留标准扩散机制作为骨干，同时使用大语言模型智能体为部分关键用户生成语言层面的交互和决策信号，并依靠轻量级的基于规则/演绎的智能体处理其余群体。这种设计使传播动态更易于解释和调整，同时将模拟成本控制在可承受范围内 [23, 43]。此外，若干研究 [22, 40] 利用大语言模型智能体框架，在不同网络结构和异质智能体类型下研究谣言或错误信息传播，并通过此类模拟分析图拓扑与依角色设定而异的行为如何共同塑造传播结果。

7 结论

本文介绍了 AVOID，这是一种用于早期假新闻检测的新型框架。为解决早期真实传播数据稀缺的问题，AVOID 构建了人物画像与真实用户数据相匹配的基于大语言模型的智能体，并通过多智能体交互模拟可信的新闻传播路径。为更好地贴合现实世界传播，我们（i）区分验证智能体和扩散智能体，以反映不同的参与行为；（ii）将每个智能体的行动建立在角色设定条件化决策规则之上，从而使所产生的交互模式与数据中观察到的用户层面特征相一致。这些虚拟路径提供了关键的社会语境，以补充内容特征。此外，我们提出了一种采用对称 KL 散度的传播感知融合模块，通过对齐内容与传播表征的潜在分布来减轻模拟噪声。在真实世界数据集上的实验表明，AVOID 显著优于最先进的基线方法，证明了其在早期假新闻检测中的有效性。

展望未来，仍有若干有前景的方向值得探索，以提升传播模拟的深度和实用性。尽管当前框架捕捉了普通用户与意见领袖之间的传播动态，但现实世界的信息生态系统本质上具有对抗性。未来研究的一个关键方向是纳入对抗性智能体，以模拟恶意行为者的策略，例如内容捏造和传播操纵，从而更全面地刻画虚假信息从产生到遏制的整个生命周期。此外，将模拟器发展为用于政策干预的反事实测试平台，也是一个重要机遇。通过对缓解情景开展系统性的反事实分析，该框架可以为主动防御设计提供有原则的指导，推动工作重点从被动观察转向战略性生态系统治理。

致谢

我们感谢匿名审稿人提出的宝贵意见和建设性评论。本研究得到国家自然科学基金委员会（项目编号：62176028）以及澳大利亚研究委员会“发现早期职业研究奖”项目（项目编号：DE250100613）的资助。

参考文献

[1] Josh Achiam、Steven Adler、Sandhini Agarwal、Lama Ahmad、Ilge Akkaya、Florencia Leoni Aleman、Diogo Almeida、Janko Altenschmidt、Sam Altman、Shyamal Anadkat 等。2023年。GPT-4技术报告。*arXiv*预印本 *arXiv:2303.08774*（2023年）。[2] Gati V Aher、Rosa I Arriaga 和 Adam Tauman Kalai。2023年。利用大语言模型模拟多个人类并复现人类受试者研究。载于《国际机器学习会议》。PMLR，337–371。[3] Hunt Allcott 和 Matthew Gentzkow。2017年。社交媒体与2016年大选中的假新闻。《《经济展望杂志》》31卷，第2期（2017年），211–236页。

，第1卷，第1期，文章。出版日期：2018年1月。

[4] Eytan Bakshy、Itamar Rosenn、Cameron Marlow 和 Lada Adamic。2012年。社交网络在信息扩散中的作用。载于《第21届万维网国际会议论文集》。519–528页。[5] Tian Bian、Xi Xiao、Tingyang Xu、Peilin Zhao、Wenbing Huang、Yu Rong 和 Junzhou Huang。2020年。利用双向图卷积网络检测社交媒体上的谣言。载于《AAAI人工智能会议论文集》，第34卷。549–556页。[6] Jin Chen、Zheng Liu、Xu Huang、Chenwang Wu、Qi Liu、Gangwei Jiang、Yuanhao Pu、Yuxuan Lei、Xiaolong Chen、Xingmei Wang 等。2024年。大语言模型与个性化相遇：挑战与机遇展望。《万维网》27卷，第4期（2024年），42。[7] Yahui Chen。2015年。用于句子分类的卷积神经网络。硕士论文。滑铁卢大学。[8] Yixuan Chen、Dongsheng Li、Peng Zhang、Jie Sui、Qin Lv、Lu Tun 和 Li Shang。2022年。用于多模态假新闻检测的跨模态歧义学习。载于《2022年ACM万维网会议论文集》。2897–2905页。[9] Tsun-Hin Cheung 和 Kin-Man Lam。2023年。Factllama：利用外部知识优化指令跟随语言模型，实现自动事实核查。载于《2023年亚太信号与信息处理协会年度峰会暨会议（APSIPA ASC）》。IEEE，846–853页。[10] Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。2019年。BERT：面向语言理解的深度双向变换器预训练。载于《2019年计算语言学协会北美分会会议：人类语言技术论文集》，第1卷（长短论文）。4171–4186页。[11] Jia Ding、Yongjun Hu 和 Huiyou Chang。2020年。BERT心智模型：更出色的假新闻检测器。载于《2020年第6届计算与人工智能国际会议论文集》。396–400页。[12] Yiqi Dong、Dongxiao He、Xiaobao Wang、Yawen Li、Xiaowen Su 和 Di Jin 0001。2023年。用于假新闻检测的广义深度马尔可夫随机场框架。载于 IJCAI。4758–4765页。[13] Yingtong Dou、Kai Shu、Congying Xia、Philip S Yu 和 Lichao Sun。2021年。用户偏好感知的假新闻检测。载于《第44届国际ACM SIGIR信息检索研究与开发会议论文集》。2051–2055页。[14] Matthijs Douze、Alexandr Guzhva、Chengqi Deng、Jeff Johnson、Gergely Szilvasy、Pierre-Emmanuel Mazaré、Maria Lomeli、Lucas Hosseini 和 Hervé Jégou。2024年。FAISS库。arXiv预印本 arXiv:2401.08281（2024年）。[15] Marc Fisher、John Woodrow Cox 和 Peter Hermann。2016年。披萨门：从谣言到话题标签，再到华盛顿特区的枪击事件。《华盛顿邮报》6（2016年），8410–8415页。[16] Chen Gao、Xiaochong Lan、Zhiho ng Lu、Jinzhu Mao、Jinghua Piao、Huandong Wang、Depeng Jin 和 Yong Li。2023年。S3：由大语言模型赋能智能体的社交网络模拟系统。arXiv预印本 arXiv:2307.14984（2023年）。[17] Chen Han、Wenzhen Zheng 和 Xijin Tang。2025年。Debate-to-Detect：将错误信息检测重新定义为由大语言模型参与的现实世界辩论。arXiv预印本 arXiv:2505.18596（2025年）。[18] John J Horton。2023年。作为模拟经济智能体的大语言模型：我们能从 homo silicus 中学到什么？技术报告。美国国家经济研究局。[19] Dongpeng Hou、Chao Gao、Xuelong Li 和 Zhen Wang。2024年。用于社交传播图生成的DAG感知变分自编码器。载于《AAAI人工智能会议论文集》，第38卷。8508–8516页。[20] Beizhe Hu、Qiang Sheng、Juan Cao、Yuhui Shi、Yang Li、Danding Wang 和 Peng Qi。2024年。坏演员，好顾问：探索大语言模型在假新闻检测中的作用。载于《AAAI人工智能会议论文集》，第38卷。22105–22113页。[21] Qisheng Hu、Quanyu Long 和 Wenya Wang。2025年。分解困境：主张分解会提升还是损害事实核查性能？载于《2025年美洲国家计算语言学协会分会会议：人类语言技术论文集（第1卷：长篇论文）》。6313–6336页。[22] Tianrui Hu、Dimitrios Liakopoulos、Xiwen Wei、Radu Marculescu 和 Neeraja J Yadwadkar。2025年。使用大语言模型智能体模拟社交网络中的谣言传播。arXiv预印本 arXiv:2502.01450（2025年）。[23] Yuxuan Hu、Gemju Sherpa、Lan Zhang、Weihua Li、Quan Bai、Yijun Wang 和 Xiaodan Wang。2024年。用于信息传播的大语言模型增强型基于智能体的模拟工具。载于《第三十三届国际人工智能联合会议论文集》，IJCAI-24，济州，大韩民国。3–9页。[24] Zhiwei Jin、Juan Cao、Han Guo、Yongdong Zhang 和 Jiebo Luo。2017年。利用循环神经网络进行多模态融合，以检测微博上的谣言。载于《第25届ACM国际多媒体会议论文集》。795–816页。[25] Chenlu Ju、Jiaxin Liu、Shobhit Sinha、Hao Xue 和 Flora Salim。2025年。TrajLLM：用于逼真人类轨迹模拟的模块化大语言模型增强的基于智能体框架。载于《ACM万维网会议2025伴随论文集》。2847–2850页。[26] Rohit Kumar Kaliyar、Anurag Goswami 和 Pratik Narang。2021年。FakeBERT：采用基于BERT的深度学习方法检测社交媒体上的假新闻。《多媒体工具与应用》80卷，第8期（2021年），11765–11788页。

，第1卷，第1期，文章。出版日期：2018年1月。

[27] Elmar Kiesling、Markus Günther、Christian Stummer 和 Lea M Wakolbinger。2012年。创新扩散的基于智能体仿真：综述。《中欧运筹学杂志》20，2（2012年），183–230。[28] Diederik P Kingma。2014年。Adam：一种随机优化方法。arXiv预印本 arXiv:1412.6980（2014年）。[29] Diederik P Kingma、Max Welling 等。2013年。自编码变分贝叶斯。[30] TN Kipf。2016年。基于图卷积网络的半监督分类。arXiv预印本 arXiv:1609.02907（2016年）。[31] Patrick Lewis、Ethan Perez、Aleksandra Piktus、Fabio Petroni、Vladimir Karpukhin、Naman Goyal、Heinrich Küttler、Mike Lewis、Wen-tau Yih、Tim Rocktäschel 等。2020年。面向知识密集型自然语言处理任务的检索增强生成。神经信息处理系统进展 33（2020年），9459–9474。[32] Hui Li、Ante Wang、Kunquan Li、Zhihao Wang、Liang Zhang、Delai Qiu、Qingsong Liu 和 Jinsong Su。2025年。用于跨领域虚假信息检测的自动化决策规则优化多智能体框架。载于《2025年自然语言处理实证方法会议论文集》。5720–5736。[33] Junkai Li、Yunghwei Lai、Weitao Li、Jingyi Ren、Meng Zhang、Xinhui Kang、Siyu Wang、Peng Li、Ya-Qin Zhang、等。2024年。智能体医院：一个由可进化医疗智能体构成的医院拟真模拟体。arXiv预印本 arXiv:2405.02957（2024年）。[34] Miaoran Li、Baolin Peng、Michel Galley、Jianfeng Gao 和 Zhu Zhang。2024年。Self-Checker：用于大语言模型事实核查的即插即用模块。载于计算语言学协会论文集：NAACL 2024。163–181。[35] Jonathan Light、Min Cai、Sheng Shen 和 Ziniu Hu。2023年。Avalonbench：评估大语言模型玩阿瓦隆游戏的能力。arXiv预印本 arXiv:2310.05036（2023年）。[36] Aixin Liu、Bei Feng、Bing Xue、Bingxuan Wang、Bochao Wu、Chengda Lu、Chenggang Zhao、Chengqi Deng、Chenyu Zhang、Chong Ruan 等。2024年。DeepSeek-V3技术报告。arXiv预印本 arXiv:2412.19437（2024年）。[37] Genglin Liu、Vivian T Le、Salman Rahman、Elisa Kreiss、Marzyeh Ghassemi 和 Saadia Gabriel。2025年。Mosaic：面向多智能体仿真中内容传播与治理的社交人工智能建模。载于《2025年自然语言处理实证方法会议论文集》。6401–6428。[38] Yuhan Liu、Xiuying Chen、Xiaoqing Zhang、Xing Gao、Ji Zhang 和 Rui Yan。2024年。从怀疑到接受：假新闻态度动态的模拟。arXiv预印本 arXiv:2403.09498（2024年）。[39] Yifan Liu、Yaokun Liu、Zelin Li、Ruichen Yao、Yang Zhang 和 Dong Wang。2025年。用于假新闻检测的模态交互混合专家模型。载于《2025年ACM Web会议论文集》。5139–5150。[40] Yuhan Liu、Zirui Song、Xiaoqing Zhang、Xiuying Chen 和 Rui Yan。2024年。从微小失误到巨大飞跃：一种基于大语言模型的假新闻演化仿真。arXiv电子预印本（2024年），arXiv–2410。[41] Yang Liu 和 Yi-Fang Brook Wu。2020年。FNED：用于社交媒体假新闻早期检测的深度网络。《ACM信息系统汇刊》（TOIS）38，3（2020年），1–33。[42] Jiatong Ma、Linmei Hu、Rang Li 和 Wenbo Fu。2025年。LOCAL：基于大语言模型的多智能体逻辑与因果事实核查。载于《2025年ACM Web会议论文集》。1614–1625。[43] Xinyi Mou、Zhongyu Wei 和 Xuanjing Huang。2024年。揭示真相并促进变革：迈向基于智能体的大规模社会运动仿真。arXiv预印本 arXiv:2402.16333（2024年）。[44] Qiong Nan、Qiang Sheng、Juan Cao、Beizhe Hu、Danding Wang 和 Jintao Li。2024年。让沉默发声：利用大语言模型生成的评论增强假新闻检测。载于《第33届ACM信息与知识管理国际会议论文集》。1732–1742。[45] Joon Sung Park、Joseph O’Brien、Carrie Jun Cai、Meredith Ringel Morris、Percy Liang 和 Michael S Bernstein。2023年。生成式智能体：人类行为的交互式拟真模拟体。载于《第36届ACM用户界面软件与技术年度研讨会论文集》。1–22。[46] Sungwon Park、Sungwon Han、Xing Xie、Jae-Gil Lee 和 Meeyoung Cha。2025年。通过大语言模型进行对抗式风格增强，实现鲁棒假新闻检测。载于《2025年ACM Web会议论文集》。4024–4033。[47] Mansheej Paul、Surya Ganguli 和 Gintare Karolina Dziugaite。2021年。数据精简下的深度学习：在训练早期寻找重要样本。神经信息处理系统进展 34（2021年），20596–20607。[48] Bohdan M Pavlyshenko。2023年。使用微调大语言模型分析虚假信息与假新闻检测。arXiv预印本 arXiv:2309.04704（2023年）。[49] Kellin Pelrine、Anne Imouza、Camille Thibault、Meilina Reksoprodjo、Caleb Gupta、Joel Christoph、Jean-François Godbout 和 Reihaneh Rabbany。2023年。迈向可靠的虚假信息缓解：泛化、不确定性与GPT-4。arXiv预印本 arXiv:2305.14928（2023年）。[50] Liwen Peng、Songlei Jian、Zhigang Kan、Linbo Qiao 和 Dongsheng Li。2024年。并非所有假新闻在语义上都相似：用于多模态假新闻检测的上下文语义表示学习。《信息处理与管理》61，1（2024年），103564。[51] Shengsheng Qian、Jinguang Wang、Jun Hu、Quan Fang 和 Changsheng Xu。2021年。用于假新闻检测的层次化多模态上下文注意力网络。载于《第44届国际ACM SIGIR研究会议论文集：信息检索研究与发展》

。153–162。[52] Jon Roozenbeek、Claudia R Schneider、Sarah Dryhurst、John Kerr、Alexandra LJ Freeman、Gabriel Recchia、Anne Marthe Van Der Bles 和 Sander Van Der Linden。2020年。全球范围内对新冠肺炎相关错误信息的易感性。《皇家学会开放科学》7，10（2020年），201199。[53] Yunxiao Shi、Wujiang Xu、Zeqi Zhang、Xing Zi、Qiang Wu 和 Min Xu。2025年。Personax：面向长行为序列的推荐智能体用户建模框架。arXiv预印本 arXiv:2503.02398（2025年）。[54] Kai Shu、Limeng Cui、Suhang Wang、Dongwon Lee 和 Huan Liu。2019年。dEFEND：可解释的假新闻检测。载于《第25届ACM SIGKDD知识发现与数据挖掘国际会议论文集》。395–405。[55] Kai Shu、Deepak Mahudeswaran、Suhang Wang、Dongwon Lee 和 Huan Liu。2020年。FakeNewsNet：一个数据仓库，包含用于研究社交媒体假新闻的新闻内容、社会语境及时空信息。《大数据》8，3，2020，171–188。[56] Kai Shu、Amy Sliva、Suhang Wang、Jiliang Tang 和 Huan Liu。2017年。社交媒体假新闻检测：数据挖掘视角。《ACM SIGKDD探索通讯》19，1（2017年），22–36。[57] Ben Sorscher、Robert Geirhos、Shashank Shekhar、Surya Ganguli 和 Ari Morcos。2022年。超越神经缩放定律：通过数据剪枝突破幂律缩放。神经信息处理系统进展 35（2022年），19523–19536。[58] Tiening Sun、Zhong Qian、Sujun Dong、Peifeng Li 和 Qiaoming Zhu。2022年。利用图对抗对比学习进行社交媒体谣言检测。载于《2022年ACM Web会议论文集》。2789–2797。[59] Hugo Touvron、Thibaut Lavril、Gautier Izacard、Xavier Martinet、Marie-Anne Lachaux、Timothée Lacroix、Baptiste Rozière、Naman Goyal、Eric Hambro、Faisal Azhar 等。2023年。Llama：开放且高效的基础语言模型。arXiv预印本 arXiv:2302.13971（2023年）。[60] Petar Veličković、Guillem Cucurull、Arantxa Casanova、Adriana Romero、Pietro Lio 和 Yoshua Bengio。2017年。图注意力网络。arXiv预印本 arXiv:1710.10903 （ 2017 年）。[61] Soroush Vosoughi 、 Deb Roy 和 Sinan Aral。2018 年。真假新闻在网络上的传播。《科学》359，6380（2018年），1146–1151。[62] Bo Wang、Jing Ma、Hongzhan Lin、Zhiwei Yang、Ruichao Yang、Yuan Tian 和 Yi Chang。2024年。通过竞争性智慧防御，利用大语言模型实现可解释的假新闻检测。载于《2024年ACM Web会议论文集》。2452–2463。[63] Bing Wang、Bingrui Zhao、Ximing Li、Changchun Li、Wanfu Gao 和 Shengsheng Wang。2025年。专家间的协作与争议：通过调优评论生成器进行谣言早期检测。载于《第48届国际ACM SIGIR研究与发展会议论文集》。468–478。[64] Lei Wang、Heyang Gao、Xiaohe Bo、Xu Chen 和 Ji-Rong Wen。2025年。Yulan-onesim：迈向新一代基于大语言模型的社交模拟器。arXiv预印本 arXiv:2505.07581（2025年）。[65] Lei Wang、Chen Ma、Xueyang Feng、Zeyu Zhang、Hao Yang、Jingsen Zhang、Zhiyuan Chen、Jiakai Tang、Xu Chen、Yankai Lin 等。2024年 18，6 2024 186345。《基于大语言模型的自主智能体综述》。《计算机科学前沿》（），。[66] Lei Wang、Jingsen Zhang、Hao Yang、Zhi-Yuan Chen、Jiakai Tang、Zeyu Zhang、Xu Chen、Yankai Lin、Hao Sun、Ruihua Song 等。2025年。基于大语言模型智能体的用户行为模拟。《ACM信息系统汇刊》43，2（2025年），1–37。[67] Jason Wei、Xuezhi Wang、Dale Schuurmans、Maarten Bosma、Fei Xia、Ed Chi、Quoc V Le、Denny Zhou 等。2022年。思维链提示可激发大语言模型的推理能力。神经信息处理系统进展 35 （ 2022 年），24824–24837。[68] Jiaying Wu 、 Jiafeng Guo 和 Bryan Hooi。2024 年。披着羊皮的假新闻：针对大语言模型赋能风格攻击的鲁棒假新闻检测。载于《第30届ACM SIGKDD知识发现与数据挖掘会议论文集》。3367–3378。[69] Chengxing Xie、Canyu Chen、Feiran Jia、Ziyu Ye、Shiyang Lai、Kai Shu、Jindong Gu、Adel Bibi、Ziniu Hu、David Jurgens 等。2024年。大语言模型智能体能否模拟人类信任行为？神经信息处理系统进展 37（2024年），15674–15729。[70] Yuzhuang Xu、Shuo Wang、Peng Li、Fuwen Luo、Xiaolong Wang、Weidong Liu 和 Yang Liu。2023年。探索大语言模型在交流游戏中的应用：一项关于狼人杀的实证研究。arXiv预印本 arXiv:2309.04658（2023年）。[71] Zichao Yang、Diyi Yang、Chris Dyer、Xiaodong He、Alex Smola 和 Eduard Hovy。2016年。用于文档分类的层次注意力网络。载于《2016年计算语言学协会北美分会会议：人类语言技术论文集》，Kevin Knight、Ani Nenkova 和 Owen Rambow（编）。计算语言学协会，美国加利福尼亚州圣迭戈，

1480 1489 10 18653 1 16 1174 –。doi：。/

[72] 2007
v/N– Yuan Yao、Lorenzo Rosasco 和 Andrea Caponnetto。。梯度下降学习中的提前停止。《建设性逼近》26，2（2007年），289–315。[73] Junwei Yin、Min Gao、Kai Shu、Wentao Li、Yinqiu Huang 和 Zongwei Wang。2025年。图与序列：面向假新闻检测的广域语义建模。载于《2025年ACM Web会议论文集》。2838–2849。

，第1卷，第1期，文章。出版日期：2018年1月。

[74] Jiaxuan You、Rex Ying、Xiang Ren、William Hamilton 和 Jure Leskovec。2018。Graphrnn：使用深度自回归模型生成逼真的图。在国际机器学习会议。PMLR，5708–5717。[75] Haifeng Zhang 和 Yevgeniy Vorobeychik。2019。基于经验的创新扩散智能体模型：批判性综述。《人工智能评论》52，1（2019），707–741。[76] Litian Zhang、Xiaoming Zhang、Chaozhuo Li、Ziyi Zhou、Jiacheng Liu、Feiran Huang 和 Xi Zhang。2024年。缓解社会危害：通过扩散引导的传播路径生成实现假新闻早期检测。在第*32*届*ACM*国际多媒体会议论文集。2842–2851。[77] Muhan Zhang、Shali Jiang、Zhicheng Cui、Roman Garnett 和 Yixin Chen。2019。D-vae：一种用于有向无环图的变分自编码器。神经信息处理系统进展32（2019）。[78] Wei Emma Zhang、Quan Z Sheng、Ahoud Alhazmi 和 Chenliang Li。2020年。自然语言处理深度学习模型的对抗攻击：综述。《*ACM*智能系统与技术汇刊》11，3（2020），1–41。[79] Xueyao Zhang、Juan Cao、Xirong Li、Qiang Sheng、Lei Zhong 和 Kai Shu。2021。挖掘双重情绪以检测假新闻。在*2021*年万维网大会论文集。3465–3476。[80] Jiaqi Zheng、Xi Zhang、Sanchuan Guo、Quan Wang、Wenyu Zang 和 Yongdong Zhang。2022。MFAN：用于谣言检测的多模态特征增强注意力网络。在*IJCAI*，第2022卷。2413–2419。[81] Jiawei Zhou、Yixuan Zhang、Qianni Luo、Andrea G Parker 和 Munmun De Choudhury。2023。合成谎言：理解人工智能生成的错误信息，并评估算法和人工解决方案。在*2023*年计算机系统中的人为因素*CHI*会议论文集。1–20。[82] Junyou Zhu、Chao Gao、Ze Yin、Xianghua Li 和 Jürgen Kurths。2024年。面向鲁棒且可解释的假新闻检测的传播结构感知图变换器。在第*30*届*ACM SIGKDD*知识发现与数据挖掘大会论文集。4652–4663。