宣传检测的混合标注：整合 LLM 预注释与人类智能

Ariana Sahitaj[1,2]* **Premtim Sahitaj**[1,2]* **Veronika Solopova**[1,2]
Jiaao Li[1,2] **Sebastian Möller**[1,2] **Vera Schmitt**[1,2]
[1]Quality and Usability Lab, Technische Universität Berlin, Germany
[2]German Research Center for Artificial Intelligence (DFKI), Berlin, Germany
ariana.sahitaj@campus.tu-berlin.de

Abstract

由于任务的复杂性和高质量标注数据的有限性，在社交媒体上进行宣传检测仍然具有挑战性。本文介绍了一种新颖的框架，该框架结合了人类专业知识和大型语言模型（LLM）的协助，以提高注释的一致性和可扩展性。我们提出了一个分层的分类法，将 14 种细化的宣传技巧 (?) 组织成三个更广泛的类别，在 HQP 数据集 (?) 上进行了一项人工注释研究，揭示出对细致标签的标注者间一致性低的问题，并实施了一个 LLM 辅助的预注释流程，该流程提取宣传性区段，生成简明解释，并分配局部标签和全局标签。次级人工验证研究表明，在一致性和时间效率上都有显著改善。在此基础上，我们微调了较小的语言模型（SLMs）以执行结构化注释。我们不是在人工注释上进行微调，而是基于高质量 LLM 生成的数据进行训练，让大型模型产生这些注释，并通过知识蒸馏让较小模型学会生成这些注释。我们的工作有助于开发可扩展且稳健的宣传检测系统，支持透明和负责任媒体生态系统的理念，这与可持续发展目标 16 一致。代码在我们的 GitHub 库 [1] 上公开可用。

内容警告：本文包含俄罗斯宣传的示例，其中一些包含误导性或冒犯性的声明。这些提供用于学术分析，并不反映作者的观点。

1 介绍

假新闻和虚假信息已成为一个重大挑战，特别是在像俄罗斯-乌克兰战争这样的地缘政治冲突中 (?) 。虚假信息活动战略性地操控公众舆论和塑造叙事 (??) ，与亲俄偏见相关联的是降低识别宣传的能力 (?) 。宣传被定义为“为了实现宣传者的期望意图而有目的和系统地尝试塑造感知、操纵认知以及引导行为的行为” (??) ，这些活动的核心就是宣传。检测这种操控性内容对于维护公众信任和保护民主进程至关重要 (?) 。尽管长篇文本中的宣传已被广泛研究 (?) ，短篇宣传由于注释数据有限、上下文稀疏、以及非正式语言、缩写和标签的使用而更具挑战性 (?) 。尽管虚假信息和宣传检测的自动化方法已经取得了进展 (?) ，但这一任务仍然困难。微妙的语言线索、依赖上下文的解释以及低注释者间一致性突显了人工注释的复杂性 (??) ，特别是在细粒度分类中，因为宣传常常利用认知偏见并削弱批判性思维，使个人更容易接受阴谋叙述 (??) 。宣传检测符合联合国可持续发展目标（SDG）16 [2] ，该目标促进和平、包容性的社会和有效的机构。错误信息和宣传通过加剧社会分裂、侵蚀对机构的信任、阻碍透明沟通来破坏这些愿望，特别是在被自动化机器人放大时。在这项工作中，我们提出了一种推进宣传检测的方法，具有以下五个关键贡献：首先，我们开发了一个细粒度的宣传分类法，将 14 种不同技术根据其意图分为三大类：那些引发情绪反应的，那些简化或扭曲复杂问题的，以及那些通过权威和群体动态破坏信任的。其次，我们对 HQP 数据集中统计意义上的宣传推文子集进行了一项初步的人类标注研究。该研究强调了手动细粒度标注的挑战，揭示出过程是高度主观的、耗时的，并且容易出现低标注者间一致性。第三，为了克服这些限制，我们提出了一种新的 LLM 辅助标注方法。在我们的流程中，LLM 首先从文本中提取相关的宣传段落，解释为何这些段落被认为是宣传性的，然后在段落级别分配细粒度标签，最后确定整篇文章的全局标签。第四，我们对 LLM 标注的帖子的分层样本进行了第二次人类验证研究。在这个阶段，给人类标注者呈现提取出来的段落及其局部标签，并负责标注全局宣传标签。我们观察到，通过引入 LLM 作为预标注工具，标注一致性增加，时间投入减少。最后，我们微调小型语言模型在 LLM 生成的标注基础上进行结构化的基于段落的标签和解释，从而通过知识蒸馏实现无需

* Equal contribution

[1] https://github.com/XplaiNLP/NLP4PI_2025_submission

[2] https://sdgs.un.org/goals/goal16

依赖人工标注数据的可扩展训练。

2 相关工作

对自动宣传检测的早期研究在文档层面上处理问题，旨在对整个新闻文章进行分类 (?) 。例如，一些系统将文本标记为四大类（可信、讽刺、骗局或宣传）(?) ，而其他系统将其框定为二元任务（宣传、非宣传）(?) ，这限制了细粒度性和可解释性 (?) 。一个进步来自于 ? 的工作，他们使用 PTC 语料库引入了片段级别分析，该语料库包含在句子级别和片段级别用 18 种不同的宣传技术注释的新闻文章。这一方案被 SemEvak-2020 共享任务所采用 (?) ，该任务将 18 种技术整合为一组 14 种广泛使用的标签 (???) ，我们在我们的工作中也遵循这一方案。早期模型使用了基于 BERT 的结构来执行片段识别和技术分类 (?) 。在此基础上，最近的研究探索了大规模语言模型（LLMs）如何进一步增强宣传检测，减少注释时间和成本，同时提高分类任务中的标签一致性和质量 (???) 。然而，使用 LLMs 可能比人工注释者表现出更强的系统偏见，尤其是在政治敏感的背景中 (?) ，并可能遭受如幻觉生成等相关问题的影响 (?) 。在宣传检测中，? 评估了 GPT-3.5、GPT-4 和 Claude 在识别新闻文章中的六种宣传技术方面的表现。? 采用 GPT-4 作为大规模语言模型标注工具的方法，使用多标签和序列标记任务来标注阿拉伯语文本片段的 23 种宣传技术，并在生成的标注上训练基于 BERT 的模型。同样地，? 研究了多种 GPT-3 和 GPT-4 变体在使用 SemEval-2020 任务 11 数据集 (?) 进行文章级别的 14 种宣传技术的多标签分类上的表现，采用了一系列提示工程和微调策略。他们的结果表明，GPT-4 可以达到接近最先进的表现。我们的工作以这些努力为基础，将 14 种细粒度技术 (?) 分组为三大类的新粗粒度分类法，以支持人工标注员的清晰性并实现层次化建模。通过使用完全开源的大规模语言模型（LLaMA3-70B），我们从推文中提取宣传片段，并基于 ? 的 14 种技术分配细粒度局部标签。此外，它分配一个全局宣传标签以捕捉推文的整体框架。尽管大规模语言模型也生成解释以说明每个片段为何被分类为宣传性的，但这些解释不向人工标注员展示，而是作为中间推理步骤来指导模型进行预测。此外，我们在较大模型生成的输出上蒸馏了四个小型学生模型，以通过开源建模管道在资源受限的环境中启用宣传片段标注。

在本节中，我们概述了结合人类专业知识与计算技术的新方法，如图 1 所示，以及他们的结果。我们首先在第 2.1 节定义粗粒度和细粒度类别的标记框架。接下来，我们描述在 HQP 数据集 (?) 上的人类注释研究（研究 1，见第 2.2 节）。然后，我们在第 2.3 节详细介绍我们的 LLM 小样本推理和注释方法，以自动提取宣传范围、生成解释并分配细粒度标签，接着进行第二次人类验证研究（研究 2，见第 ?? 节）。最后，我们在第 2.4 节通过知识蒸馏微调 SMLs。

2.1 宣传标签分类法

针对宣传技术进行文本标注是一项非常复杂的任务，因为它受到主观性、认知偏见、个人经历以及由于不同文化和语言背景引起的微妙意义变化的影响。此前的研究强调，区分多个细致的技术可以特别具有挑战性，导致标注者间一致性低，使得跨标注保持一致性变得困难。为了研究这个问题，我们对文献进行了调查，并汇总了先前工作中的定义，尤其是 ? 引入的 14 种宣传技术，这些技术是对 ? 早期提出的 18 种技术的精细化，随后被 ? 和 ? 用于分析和标注文本中的宣传技术。在我们的框架中，根据其操控意图和修辞功能，这些细致的宣传技术被组织成更广泛的粗略类别。技术的详细定义可以在附录 A.1 中找到。这个分层框架旨在通过先将宣传按概念分组，然后再应用细致分类，来减轻标注者的认知负担并提高标注一致性。它还使我们能够在随后的分析中评估细致预测在粗略标注系统中的背景。三个粗略类别如下：

2.2 研究 1：人工标注

在这项初步研究中，我们旨在复制来自 ? 的先前研究结果，这些结果强调了标注细粒度宣传技术的挑战，特别是此类任务中观察到的低标注者间一致性（IAA）。为了研究细粒度标签的标注，我们使用了 HQP 数据集 (?) ，该数据集包括了 29,596 条推文，这些推文在俄罗斯宣传的背景下被标注为二元宣传检测。在这些推文中，4,534 条此前被识别为具有宣传性质。
假设二元分类（宣传与非宣传）是可靠的，我们将分析范围限制在被标注为宣传的推文子集中。此重点使我们能够在没有二元错误分类的干扰下，专注于分配详细的细粒度标签。基于在 95% 置信水平下 5% 的误差范围，并遵循既定的样本量估计方法 (?) ，从 4,534 条被标注为宣传的推文中选取了一个规模为 $n = 355$ 的样本。虽然此样本在统计上足以估计比例，但我们将其视为一项试验性研究，旨在探索标注的可行性和定性模式，而不是声称对整个语料库具有完全的代表性。

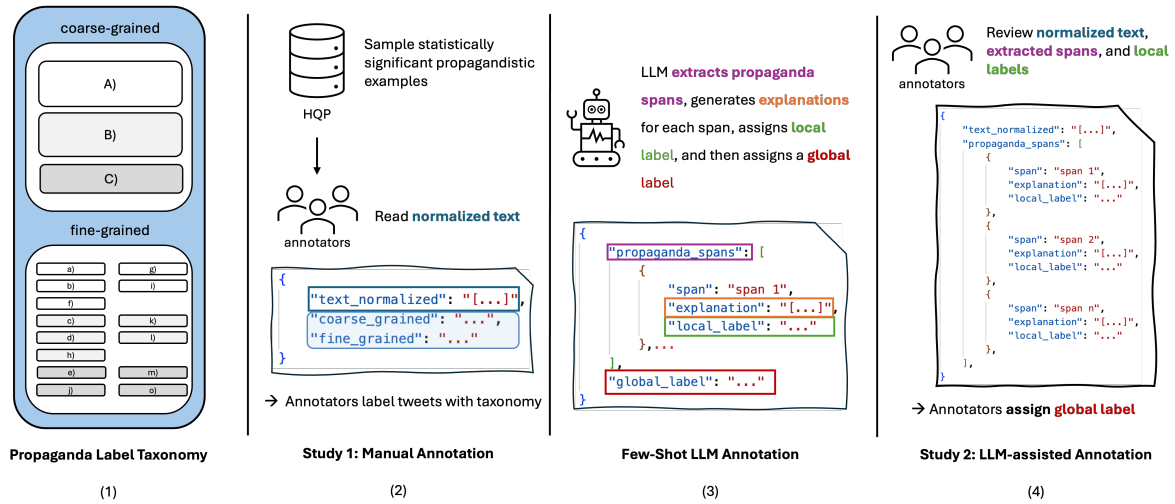

Figure 1: 方法论概述

2.2.1 设置

起初，注释员们被提供了 HQP 注释指南 (**?**)，该指南将宣传定义为旨在影响意见的故意表达，特别关注在俄乌冲突背景下的俄罗斯宣传。这确保了对推文进行宣传性二元分类的共同理解。随后，他们收到了一份补充注释指南，其中包括之前介绍的定义以及粗粒度和细粒度宣传类别的具体例子。注释员被指示先选择最合适的粗粒度类别，然后为每条推文分配单个最重要的细粒度标签。

2.2.2 结果

第一次人工标注研究需要三名标注员使用预定义的粗粒度类别和更详细的细粒度标签对每条推文进行标注。如表 1 所示，粗粒度标签达到了中等水平的一致性。

Metric	Coarse	Fine
Raw Agreement 2/3	0.8845	0.4761
Raw Agreement 3/3	0.2789	0.0761
Krippendorff's Alpha	0.2065	0.1233

Table 1: 第一轮中用于粗粒度和细粒度宣传标注的标注者间一致性度量。

具体而言，粗粒度注释的原始一致性达到 88.45 %，并且有 2/3 的多数，但当要求完全的 3/3 共识时，下降到 27.89 %。细粒度标注表现出更大的挑战，原始一致性（2/3）为 47.61 %，而完全一致性仅达到 7.61 %。粗粒度和细粒度标签对应的 Krippendorff's Alpha 值进一步强调了获得一致细粒度注释的局限性。表 2 中的更详细分析显示，当标注者已经在粗粒度类别上达成一致时，细粒度一致性会大幅度提高。

在 HQP 数据集 (**?**) 的标注指南中，要求标注人员将整个推文标记为宣传性内容，即使只有

Subset	2/3 Fine	3/3 Fine
2/3 Coarse	0.4372	0.0000
3/3 Coarse	0.7475	0.2727

Table 2: 在粗略标签上先前多数 2/3 或完全 3/3 一致情况下的细粒度一致率。

部分文本段包含宣传性内容。尽管我们在自己的细粒度标签标注中遵循了这一概念，但我们的分析显示，许多推文由多个段落组成，每个段落可能与不同的宣传标签相关联。这种复杂性使得对整个文档应用单一明确标签变得具有挑战性，因为标注人员不仅需要在 14 个可能的标签中进行区分，还需要根据其影响对标签进行排名，以便选出最突出的一个。这一额外的主观性和特异性层面，也导致了每个实例平均标注时间为 151.70 秒，强调了探索替代标注策略的必要性，例如以下章节讨论的 LLM 辅助预标注。

2.3 少样本大模型标注

根据研究 1 的发现，我们通过实施一个大型语言模型（LLM）来扩展注释方法，该模型用于提取潜在宣传内容的片段，并在两个层面上分配标签。在这种方法中，LLM 需要完成三个子任务：（i）从所呈现的推文中提取可能包含宣传语言的跨度，（ii）生成简明的解释，说明每个跨度被归类为宣传的原因，以及（iii）为每个提取的跨度分配细粒度的局部标签，并为整个推文分配一个全局标签。XMATHX 我们使用 llama3.3-70B-Instruct 模型进行少样本推理 (**?**)。具体而言，我们为每个细粒度的宣传标签创建一个合成的少样本示例，并将相应的标签定义融入系统提示中。每个示例都是手动构建的，以反映各自技术的典型使用。三位作者

审阅每个示例以确保其清晰度和适用性。我们使用结构化生成以确保输出可以被轻松解析和评估 (**?**) 。没有提供关于该情境内容的额外背景知识，因此 LLM 仅依靠少样本示例和标签定义来执行任务。提示在附录 A.4 的图 4 和 5 中呈现。

2.3.1 结果

LLM 被应用于 HQP 数据集中的所有被标记为宣传性的推文 (**?**) 。在 94 个案例中，模型未检测到任何宣传性跨度。经过手动分析，我们发现其中 30 个案例确实表现出了相当明显的宣传技巧或框架。然而，没有具体的背景知识，这些案例经常可能被误认为是评论文章或新闻。剩余的大多数是新闻报道、讨论或评论文章，不包含明确的宣传。对于接下来的分析，我们过滤掉这些案例。

预测的全局标签分布在附录 A.2 中的表 7 中总结。最常见的标签是 `loaded_language` 、`doubt` 、`reductio_ad_hitlerum` 和 `name_calling` 。先前工作指出 `reductio_ad_hitlerum` 是俄罗斯宣传中常用的技巧 (**?**) 。在我们的设置中，这个标签与类似分类如 `loaded_language` 和 `name_calling` 共同出现，暗示这些技巧在使用中的经验重叠。接下来，我们检查了每条推文中检测到的宣传跨度数量（分布在表 3 中）。我们的实证结果表明，多数宣传性推文包含多个宣传性片段。仅依靠赋予全局标签，如之前的工作所关注的，可能导致重要细节的丢失，这表明未来的工作应保持对片段及其局部标签的提取作为主要目标。

spans	1	2	3	4	5+
count	289	1,119	1,663	1,002	367

Table 3: 检测到的宣传跨度的分布。

我们关注至少有三个提取的宣传片段的推文，即 3,032 个案例，观察到其中 76.65 % 的实例中，分配给第一个提取片段的局部标签与整个推文的全局标签匹配。这表明最具影响力的宣传内容往往出现在推文的开头。此外，大约 30 % 的至少有三个提取的案例呈现出多数局部标签。在这些案例中，83.55 % 的情况中，这些多数局部标签也与全局标签一致。因此，当有大多数提取的局部标签时，我们观察到可以推断出主要的宣传技术。

2.3.2 消融

为了评估我们方法的稳健性，我们进行了几项消融研究。在第一次分析中，我们比较了通过标准化文本（即去除用户名、链接和类似元素后的文本）生成的推文注释与未标准化推文生成的注释。为了在统计上评估这些配对分类观察之间的差异，我们采用了 Stuart-Maxwell（边际同质性）检验。在零假设 H_0 下，标准化变体中每个预测的全局标签的比例等于原始推文文本中的比例。Stuart-Maxwell 检验产生了一个检验统计量为 15.32，自由度为 16，p 值为 0.5014。因此，我们得出结论，归一化文本与非归一化文本获得的标注全局标签之间没有显著差异。接下来，我们通过重复实验 $k = 5$ 次来评估 LLM 输出的稳定性。最初，在具有静态少样本示例、一致任务描述和引导解码的标准条件下，我们的方法在 5/5 个案例中为提取的跨度、分配的局部标签和全局标签提供了稳定的结果。为了进一步挑战模型的鲁棒性，我们通过对每个数据点的少样本示例和提示中的标签定义顺序进行打乱来引入最大的随机性。我们注意到，在五次运行中，每次运行中为每个数据点进行随机化时的一致性（表 4 ）。这些结果表明，即使在最大提示随机性下，我们的方法仍然相当稳健。然而，少样本示例和标签定义的顺序变化对局部标签预测有边际影响，而提取的跨度和全局标签预测则保持较为稳定。这个观察结果强化了我们最初的发现，即某些提取的跨度可能对应多个适当的标签，但仍与一致的全局标签相关联。

Aggreement	≥3/5	≥4/5	5/5
Local Label	100.00 %	95.46 %	81.48 %
Extract. Spans	100.00 %	97.74 %	89.86 %
Global Label	100.00 %	98.58 %	94.17 %

Table 4: 在 5 次随机化运行中的一致性。

在这第二个人工注释研究中，我们旨在评估将大型语言模型（LLM）生成的注释与人工验证相结合是否能提高注释的一致性和效率。与第一项研究不同，第一项研究中标注者在没有帮助的情况下分配粗粒度和细粒度的标签，本研究为他们提供了 LLM 生成的预注释作为可选建议。标注者会看到原始规范化的推文、提取的跨度和相应的标签，但他们不会修改或验证单个跨度。相反，他们从预定义的一组选项中为整个推文选择最合适的粗粒度类别和细粒度技术。在注释过程中，LLM 预测的全局标签保持隐藏，以确保人类决策不受模型最终分类的过度影响和偏见。为尽量减少因任务熟悉度带来的潜在偏见，我们排除了经验最丰富的标注者，并将其替换为没有参与第一次研究的标注者。这一做法旨在引入一种正则化效果，并确保评价更加均衡。

本研究中的注释过程遵循了与第 2.2.1 节设置中描述的相同的结构化方法。然而，与随机选择推文不同，我们采用了一种基于由 LLM

预测的全局标签的分层抽样方法。由于真实世界数据中宣传技术的分布往往不平衡，随机抽样可能导致某些类别的过度代表和其他类别的代表不足。为了确保每个全局标签得到充分覆盖，我们根据 LLM 预测的全局宣传标签对样本进行了分层。LLM 预测的大多数全局标签在数据集中频繁出现，使得能够在各类别中均衡分配。然而，类似随大流和重复这样的技术在 4,534 条宣传推文的完整数据集中相当少见，分别只出现了 8 次和 6 次。基于此，这些全局标签的所有出现都被包含在样本中，以确保它们在分析中得到充分代表。

2.3.3 结果

在第二个人工标注研究中，标注员被提供了 LLM 生成的预标注，其中包括提取的宣传性质的片段以及相应的局部细粒度标签。然而，标注员没有看到 LLM 预测的全局标签，他们仍然完全负责独立选择每条推文的全局粗粒度和细粒度标签。与研究 1 相比，这种方法在标注者间一致性（IAA）和标注效率方面都显著提高。

如表 **??** 所示，在研究 1 中，粗粒度标签的原始一致性从 88.45 %（2/3 多数）和 27.89 %（完全一致性）增加到研究 2 中的 97.46 %（2/3 多数）和 62.25 %（完全一致性）。对于细粒度标签，原始一致性从研究 1 中的 47.61 % 2/3 和 7.61 % 3/3 分别提高到研究 2 中的 90.14 % 2/3 和 47.89 %。同时，Krippendorff's Alpha 从研究 1 中的 0.2065（粗粒度）和 0.1233（细粒度）增加到研究 2 中的 0.6059（粗粒度）和 0.5941（细粒度）。在表 5 中根据粗粒度标签一致性水平对细粒度一致性率的详细检查进一步确认了这些改进。在研究 2 中，对于粗粒度一致性为（2/3）的推文，这些比例提高到 80 %，而对于粗粒度完全一致的推文，（2/3）细粒度一致性增加到 99.55 %，完全（3/3）细粒度一致性为 76.02 %。图 **??** 展示了 LLM 辅助标注的有效性。在这种情况下，LLM 成功识别了关键的宣传性部分，分配了适当的细粒度标签，并提供了与人类解释非常一致的连贯解释。在这一案例中，标签为 "# IStandWithPutin" 被标记为口号，加固了意识形态团结，而 "Russia is our true friend" 被归类为扬旗，展示了俄罗斯作为一个可信赖的盟友。这些解释清楚地证明了每个部分的宣传性质，并且全局标签（"口号"）特别恰当，因为口号，特别是当用作标签时，是简洁且易于分享的，增加了它们在社交媒体上的传播，并比描述性陈述更有效地加强了群体身份。此标注实现了完整的 3/3 IAA，证实了其可靠性。

Subset	2/3 Fine	3/3 Fine
2/3 Coarse	0.8000	0.0160
3/3 Coarse	0.9955	0.7602

Table 5: 在第二轮中，基于之前对粗略标签的多数 2/3 或完全 3/3 一致意见的条件下，细粒度一致率。

此外，计算了 Cohen's Kappa 来衡量人工多数投票标签与 LLM 生成的全局标签之间的一致性。如果未达到 2/3 的多数，则使用一个随机的 LLM 预测作为人工标签。结果得出的 Cohen's Kappa 分数为 0.8438，表明人工注释与 LLM 生成的全局标签之间具有很高的一致性。此外，每条推文的平均注释时间从研究 1 中的 151.70 秒减少到研究 2 中的 41.14 秒。总之，研究 2 中将 LLM 生成的预注释与人工验证相结合，导致 IAA 提高，并减少了相对研究 1 中完全手动方法的注释时间，表明在可靠性、效率和可扩展性方面总体上有所改进。

2.4 知识蒸馏

基于我们的研究结果，我们接下来旨在通过在生成的监督下微调一系列结构化语言模型（SLMs），以扩展结构化宣传注释并在资源受限的环境中实现高效推理。在这一受知识蒸馏启发的设置中，2.3 节描述的 70B 模型作为教师模型，为每个数据点提供结构化的宣传注释。我们训练四个学生模型，两个基于 LLaMA3 的变体（3B 和 8B 参数），标记为 L，以及两个 Qwen2.5 的变体（3B 和 7B 参数），标记为 Q。为了最小化内存使用并加速训练，我们采用参数高效微调（PEFT），结合 4 比特量化。我们使用标准的序列到序列的交叉熵损失，不添加额外的正则化项或显式的教师-学生 logit 匹配，以生成结构化的响应。我们利用分层的 80/20 数据分割，并在训练数据集上学习三个周期。

2.4.1 结果

我们在看不见的测试集上报告了六个评估指标，如表 6 所示。其中，G 表示测试集上的宏观和微观平均全球 F1 分数。Span $_e$ 描述了精确跨度检测的 F1，而 Span $_f$ 在严格的 0.8 相似度阈值下规定了模糊跨度 F1，以考虑由 (**?**) 引入的部分匹配概念所导致的轻微变化。类似地，Local $_e$ 需要精确跨度文本和正确的局部标签分类，而 Local $_f$ 结合了模糊跨度匹配和正确的局部标签分配。

四个学生模型在每个指标上都达到了合理的性能。较大的模型显示出适度的增益，同尺寸的 L 和 Q 变体表现相似。跨 14 个宣传类别的全局标签预测 (**?**) 得到的 F1 分数可以接受，这表明选择全局标签相对简单。跨度检测在精确匹配和模糊匹配标准下表现良好。相比之下，

Model	G $_{macro}$	G $_{micro}$	Span $_{e}$	Span $_{f}$	Local $_{e}$	Local $_{f}$
L $_{3b}$	0.49	0.36	0.40	0.60	0.22	0.32
L $_{8b}$	0.58	0.47	0.47	0.67	0.29	0.40
Q $_{3b}$	0.48	0.34	0.40	0.61	0.21	0.31
Q $_{7b}$	0.51	0.34	0.45	0.66	0.25	0.36

Table 6: 学生模型评估结果。

分配局部标签仍然困难。模型能够可靠地找到宣传跨度，但对标注哪种具体技术的确定性较低。我们假设这源于两个关键因素：(1) 用于精细局部标签预测的训练数据量有限，(2) 由于某些宣传技术定义重叠所导致的固有模糊性，而一般概念上的宣传跨度似乎更加明确。

在本文中，我们介绍了一种由大型语言模型 (LLM) 辅助的标注框架，该框架结合了宣传成分的自动提取与人工验证。我们的实验表明，将 LLM 辅助的预标注与人工验证结合起来，可以显著提高宣传检测的一致性和效率。在研究 1 中，手动进行细粒度标注遭受了低标注者之间的一致性和长时间的标注时间。研究 2 中，结合基于提取的宣传成分的 LLM 生成的预标注，产生了更高的一致性指标并减少了标注时间，尽管部分效率提升可能来自标注者对任务的熟悉程度。值得注意的是，我们的结果表明，单一的全局标签有时不足以捕捉宣传内容的复杂性，因为我们的分析显示，大多数推文包含一个以上的提取宣传成分。这种细化的视角可能提供比传统序列级分类更好的洞见，并且在不同文本长度上更具可扩展性。这些发现与新兴趋势一致，例如在 SemEval-2023 任务 3 中强调的趋势，表明未来的工作应考虑重新定义问题以强调替代的宣传检测策略。探索多标签和层次化的标注策略可能更好地适应宣传技术的重叠性质。最后，整合更丰富的上下文信息和实时事实核查模块可以进一步优化检测性能。我们还提倡发展迭代的人体循环系统，不断更新少样本实例和标签定义，以尽量减少偏见并增强模型的鲁棒性。

虽然很有前景，但我们的方法存在几个局限性。首先，我们的研究仅限于涉及俄罗斯宣传的英文推文，这可能限制了其在其他语言或领域中的适用性。其次，尽管进行了基于局部跨度的分析，但依赖单一的全局标签可能会简化多种宣传技术共存的情况。第三，注释效率的某些提高可能归因于注释者的学习效应，而不仅仅是由大型语言模型辅助的预注释造成的。第四，LLM 生成的预注释的质量取决于提供的少数示例和定义，这可能引入偏见或不一致性。后续工作应涉及更大和更为多样化的注释者群体，以进一步验证和完善框架。此外，从包含多种语言和平台的各种宣传环境中自行收集数据将提供更广泛的评估，并有助于缓解当前数据集中潜在的偏见。另一个限制涉及我们的蒸馏设置。由于预训练，70B 教师模型中存在的偏见可能会传递到学生模型中。因为学生模型仅基于模型生成的监督进行训练，教师模型中的任何意识形态或地缘政治偏见都可能在没有校正的情况下持续存在。虽然使用开源模型改善了透明度和可审查性，但它并不能从根本上防止偏见传播。未来的工作应系统地研究开源宣传检测管道中继承的偏见。

3

伦理和社会影响

在宣传检测中整合 LLM 辅助注释会引发关于偏见、自动化依赖、滥用和公众信任的伦理问题。虽然提高了注释效率，但 LLM 生成的标签可能引入系统性偏见，反映出其训练数据中的主流叙事。这可能影响人工注释员的决策，导致偏见得以强化而不是中立的分类。另一个风险是自动化偏见，即注释员过于依赖 LLM 的建议，降低了他们的批判性思维能力。此外，这些模型可能被用于反宣传，政府或其他行为者可能会利用它们来压制异议声音，并塑造对其有利的公共话语。错误或过于简单的宣传检测可能无意中削弱对媒体和公共机构的信任，从而破坏 SDG 16 所倡导的民主理想。因此，必须确保这些系统的开发和部署保持透明，进行严格的偏见审查，并维持强有力的人类监督，以确保它们支持而不是限制民主话语。

这项研究由联邦研究、技术和空间部（BMFTR，参考编号：03RU2U151C）在新闻测谎的研究项目范围内资助。

A 附录

A.1 细粒度标签

这里介绍的宣传技巧定义基于由 **?** 介绍的 14 个类别，这些类别对 **?** 提出的 18 种技巧进行了改进。这 14 个类别也被用于后来的作品中，例如 **?** 和 **?** ，以分析和标记文本中的宣传技巧。

a) 情感语言涉及使用带有强烈正面或负面情感含义的词语或短语，以影响听众的感知并影响他们的观点。

b) 贴标签、命名涉及为目标对象分配特定标签，意在激发听众的正面或负面情感，如恐惧、仇恨、钦佩或赞美。

c) 重复是信息或观点的持续重复，以随着时间的推移增加听众的接受度。

d) 夸大或缩小涉及以夸张的方式表现某事以放大其重要性，或淡化其重要性以使其看起来不如实际影响大。

e) 怀疑涉及质疑或质疑个人、团体或实体的可信度以破坏信任。

f) 恐惧/偏见诉求旨在通过引发焦虑、恐惧或恐慌来为某个观点建立支持，通常针对一种替代方案或基于现有偏见。

g) 挥舞旗帜涉及利用强烈的国家或群体认同感，如与种族、性别或政治派别相关的认同感，为某一行动、观点或个人辩解或推动其成为整个群体的代表。

h) 因果简化涉及将问题归因于单一原因，而忽视其复杂性或存在多个促成因素。这也可能包括在未充分探讨问题复杂性的情况下将责任归于某个人或群体。

i) 口号是简洁而有力的短语，通常包含标签化或刻板印象，作为情感或认知诉求来影响信仰或感知。

j) 诉诸权威涉及仅基于权威或专家的支持宣称某个主张是真实的，而没有提供额外的证据。这还包括引用的个人缺乏真正专业知识但仍被呈现为权威的情况。

k) 非黑即白的谬误涉及将两种对立的选项呈现为唯一可能的选择，忽视了其他替代方案的存在。在其极端形式中，被称为独裁，观众被明确指向一个特定的行动，有效地消除了所有其他选项。

l) 止思陈词是指用简短、泛泛的语句抑制批判性思考和有意义的讨论，通常是通过提供对复杂问题的过于简化的答案或将注意力从对某话题的深层探讨中转移出来。

m) 扣帽子、稻草人、红鲱鱼结合了三种不同的技巧，由于它们相对较少单独使用，常被归为一组。扣帽子通过指责对方伪善而不直接回应他们的主张来削弱对方的论点。稻草人通过用一个较弱或夸张版本代替对方的立场，以便于否定，从而歪曲或曲解对方的立场。红鲱鱼通过引入不相关的信息或话题偏移对主要论点的注意力。

o) 从众效应、希特勒归谬结合了两个通常被一起讨论的技巧，因为它们具有相似的说服性质。从众效应尝试通过强调“其他人也都在做”来说服观众接受某个想法或行动。希特勒归谬试图通过将某个想法或行动与观众不喜欢或鄙视的群体或个人联系起来来贬低它。

A.2 全局标签分布

表 7 提供了模型在数据集中预测的全球宣传标签分布的概述。如图所示，最常出现的技巧包括 `loaded_language` 、`doubt` 、`reductio_ad_hitlerum` 和 `name_calling` 。

Table 7: 预测的全局标签分布

Global Label	Count
loaded_language	1384
doubt	647
reductio_ad_hitlerum	641
name_calling	519
whataboutism	333
appeal_to_fear_prejudice	250
causal_oversimplification	160
exaggeration	150
flag-waving	122
appeal_to_authority	106
straw_man	54
red_herring	54
thought-terminating_cliches	35
slogans	29
black-and-white_fallacy	25
repetition	17
bandwagon	8

A.3 示例

在 HQP 数据集 (**?**) 中，弱标注被用于分类某些宣传技巧。HQP 数据集 (**?**) 最初将图 2 中的这条推文弱标注为口号。然而，对文本跨度的更详细分析揭示了多种宣传技巧的存在，包括充满感情色彩的语言、夸张和希特勒归谬。
在未来，通过将事实核查整合到宣传检测中，我们可以补充现有的标注方法，并评估所提出的论断是否具有事实依据。这很重要，因为宣传通常通过错误信息传播，而弱标注本身无法验证真实性。事实核查通过区分具有说服力的修辞和彻头彻尾的错误信息来加强检测过程，使其成为对宣传内容进行更精确且可靠分析的必要组成部分。(**?**)
另一个弱标注在宣传检测中的挑战实例在图 3 中展示。这个例子最初在 HQP 数据集 (**?**) 中被人工标注者以二元方式注释为宣传。在后来的改进中，它被弱标注为口号，加强了作为宣传内容的分类。然而，在我们的质性分析过程中，我们发现这是一个原始标注可能不合理的案例。

```
{
    "input": "Ukrainian soldiers, the Nazi Azov battalion are shooting
    their own Ukrainian citizens in the streets with snipers from
    apartment windows HTTPURL",
    "propaganda_spans": [
        {
            "span": "Ukrainian soldiers, the Nazi Azov battalion are
            shooting their own Ukrainian citizens",
            "explanation": "Makes a specific, unverified claim of
            violence to demonize Ukrainian forces.",
            "local_label": "loaded_language"
        },
        {
            "span": "the Nazi Azov battalion",
            "explanation": "Associates the Azov battalion with Nazism
            to evoke strong negative emotions.",
            "local_label": "reductio_ad_hitlerum"
        },
        {
            "span": "shooting their own Ukrainian citizens in the
            streets with snipers from apartment windows",
            "explanation": "Uses vivid, disturbing imagery to create
            outrage without providing evidence.",
            "local_label": "exaggeration"
        }
    ],
    "global_label": "loaded_language"
}
```


Figure 2: 这个例子突出了事实核查的重要性，因为检测到的部分包括无法验证的暴力宣称（“煽动性语言”）、历史歪曲（“希特勒化”）和夸张的形象（“夸张”）。如果不加以验证，这些陈述可能会导致错误信息和对公共认知的操控。


```
{
    "input": "logic of the events now: 1. Russian media has been
    telling its citizens for years that Ukraine is aggressive state 2.
    Intensified this narrative in past months 3. Says Ukraine makes
    military buildup, not Russia 4. Moscow taken by surprise by West's
    reaction, evacuations etc.",
    "propaganda_spans": [
        {
            "span": "Russian media has been telling its citizens for
            years that Ukraine is aggressive state",
            "explanation": "Creates a long-standing narrative to shape
            public opinion about Ukraine.",
            "local_label": "repetition"
        },
        {
            "span": "Intensified this narrative in past months",
            "explanation": "Increases the frequency or intensity of the
            message to reinforce the narrative.",
            "local_label": "repetition"
        },
        {
            "span": "Says Ukraine makes military buildup, not Russia",
            "explanation": "Reverses the reality of military buildup to
            shift blame.",
            "local_label": "causal_oversimplification"
        },
        {
            "span": "Moscow taken by surprise by West's reaction,
            evacuations etc.",
            "explanation": "Presents Russia as the victim, implying the
            West's reaction is unwarranted or disproportionate.",
            "local_label": "loaded_language"
        }
    ],
    "global_label": "repetition"
}
```


Figure 3: 一个例子最初在 HQP 数据集中被标记为宣传，并被弱标记为“标语”（**?**）。在定性分析中，我们发现该例子并不一定表现出明显的宣传意图。

在我们的分析中，我们特别检查了那些最初被标记为宣传但后来收到弱标签的例子。这个例子就是其中之一，但在仔细检查后，我们没有发现明显的宣传意图。相反，文本似乎是对现有叙述的分析性反思或讽刺性评论。这个案例突出了区分真正的宣传和批判性介入叙述的论述的困难。自动化或弱标记的方法可能会错误分类与宣传共享语言模式但具有不同沟通功能的内容。这些发现强调了需要更为细致的标注方法，包括对语境的理解，确保内容不因表层文本特征而被误分类。另一个说明宣传检测复杂性的例子展示在图 **??** 中。该陈述最初在 HQP 数据集（**?**）中被标注为宣传，并随后采用基于 LLM 的方法重新标记。LLM 没有执行二元分类，而是分配了细粒度的修辞标签，识别出思维终止陈词滥调、障眼法和激烈语言。然而，在进行定性审查时，我们注意到这个例子提出了关于什么应该和不应该被视为宣传的重要问题。这个案例特别有趣，因为虽然陈述中包含了反西方的批评，可以用于宣传，但它也可以出现在正常的讨论和政治话语中。即使它运用了通常与宣传相关的修辞技巧，也不一定显示出明显的操控意图。这里的关键挑战在于，仅凭修辞策略不足以自动化地使一个陈述成为宣传。背景和意图很重要。在我们对推文的定性调查中，我们发现这是一个边界案例，人们可以同时主张将其标记为宣传或反对将其标记为宣传。一方面，通过加强关于西方衰落的叙述，其框架可以作为宣传的工具。另一方面，这种批评本身独立于宣传的努力存在。这个例子很有价值，因为它展示了大型语言模型正确地分配了修辞策略，而没有将陈述过于泛化为宣传，突显了在操控性内容和批判性讨论之间划清界限的困难。

A.4 提示

该提示建立了一个结构化框架，用于在宣传检测中协助 LLM 进行标注，定义了识别、解释和分类宣传内容的系统方法。如图 4 和 5 所示，助手被设计为提取指示宣传的特定区段，根据预定义的分类标准提供理由，并分配细致的局部标签和全局标签。该框架（图 4）首先引导助手检测关键宣传区段，根据预定义的宣传技术进行分类，并解释为何每个区段应被视作宣传。

第二部分（图 5）通过强制执行一个结构化的 JSON 输出格式来扩展此过程，确保注释的一致性并促进与人工验证工作流程的集成。通过以这种方式结构化注释过程，我们的方法旨在提高标注效率，减少标注者之间的差异性，并增强大规模数据集中文宣检测的可扩展性。修辞技术的明确分类提供了关于文宣在文本中如何表现的更详细的理解，而标准化的输出格式则确保注释可以被解释和重现。

Prompt

```
SYSTEM:
You are an intelligent annotation assistant specializing in
detecting propaganda. Your task is to analyze, explain, and
pre-annotate the presented text based on a set of potential
propaganda classifications. You MUST return the output in
valid JSON following the defined schema.

**Setting**: Detection of propaganda that is against the
main opposition (i.e., Ukraine), against other oppositions
(e.g., Western countries), or in favour of the Russian
government.

1. **Identify specific words or text spans
     that indicate propaganda.**:

2. **Explain for each extracted span why it
     should be considered propaganda.**:

3. **For each span, determine the dominant
    propaganda technique from the following list**:
    - Loaded language: ...
    - Name calling: ...
    - Appeal to fear/prejudice: ...
    - Flag-waving: ...
    - Slogans: ...
    - Repetition: ...
    - Exaggeration/minimization: ...
    - Causal oversimplification: ...
    - Black-and-white fallacy: ...
    - Thought-Terminating Cliches: ...
    - Doubt: ...
    - Appeal to authority: ...
    - Whataboutism: ...
    - Straw man: ...
    - Red herring: ...
    - Bandwagon: ...
    - Reductio ad hitlerum: ...
4. **Finally, assign the global label of the span that
     is most representative for the full sequence.**
```

Figure 4: 提示（第一部分）：关于宣传检测任务的初始指令，包括跨度提取、解释以及局部和全局标签的分类。

```
**Output Format:**
Respond in **valid JSON** with the structure:
{
    "       $defs": {
        "FineLabelVerdict": {
        "description": "Fine-grained categorization of
            propaganda techniques.",
            "enum": [
                $        {LABELS}
            ]
        },
        "PropagandaSpan": {
         "description": "An identified propaganda span
        within the original text with an explanation.",
            "properties": {
                "span": {
             "description": "The exact propaganda span
                  extracted from the original text.",
                   "title": "Span",
                   "type": "string"
                },
                "explanation": {
              "description": "The explanation why this
                    span is considered propaganda.",
                    "title": "Explanation",
                    "type": "string"
                },
                "local_label": {
              "     $ref": "#/$     defs/FineLabelVerdict",
                   "description": "The appropriate label
                   assigned towards the detected label."
                }
            },
            "required": [
                "span",
                "explanation",
                "local_label"
            ]
        },
        "global_label": {
        "     $ref": "#/$     defs/FineLabelVerdict",
            "description": "The label for the dominant
            propaganda technique in the statement."
        }
    },
   "description": "Schema for structured LLM output after
    propaganda detection and normalization."
}
USER:
${TWEET}

ASSISTANT:
```

Figure 5: 提示（第 2 部分）：我们的宣传检测任务的 JSON 输出格式定义。