目标自适应的可解释立场检测：新任务及大模型实验

蓝伊[1,2],王子豪[1,2],陈波[1,2,3,*],赵小兵[1,2,3,*]
1.中央民族大学,信息工程学院,北京,100081
2.国家语言资源监测与研究民族语言中心
3.中央民族大学国家安全研究院
{lanyi,22302069,chenbomuc}@muc.edu.cn, nmzxb_cn@163.com

摘要

传统立场检测通常假设目标已知，且仅输出立场类别（支持，反对，中立），难以应对目标不确定、立场判断需要有具体依据的情形。为此，本文提出目标自适应的可解释立场检测新任务，定义模型的输出为目标、观点和立场标签。具体地，构建了首个中文高质量立场检测数据集，并设计多维评估标准；评估了多种大语言模型的基线性能。实验发现：DeepSeek-V3在目标识别与立场分类表现最优，GPT-4o在观点生成上领先；大语言模型在目标明确时具备较强目标自适应能力，但处理存在反讽现象的输入时性能下降。

数据集和实验结果公布于https://github.com/Cassieyy1102/TAISD。

关键词： 立场检测 ； 目标自适应 ； 可解释 ； 大语言模型

Target-Adaptive Interpretable Stance Detection: A New Task and Empirical Investigations with LLMs

Yi Lan[1,2],Zihao Wang[1,2],Bo Chen[1,2,3,*],Xiaobing Zhao[1,2,3,*]
1. School of Information Engineering, Minzu University of China, Beijing 100081, China
2. National Language Resource Monitoring & Research Center for Minority Languages
3. Institute of National Security MUC

Abstract

Conventional stance detection typically assumes a predefined target and outputs only a stance label (favor, against, neutral), making it difficult to handle scenarios where the target is unknown and stance prediction requires concrete justification. To address these challenges, this paper introduces a new task of target-adaptive interpretable stance detection, defining the model's output as the target, opinion, and stance label. Specifically, we construct the first high-quality Chinese stance detection dataset and design multidimensional evaluation criteria. We then evaluate the baseline performance of various large language models (LLMs). Experimental findings reveal that: DeepSeek-V3 performs best in target identification and stance classification, while GPT-4o outperforms others in opinion generation. Additionally, large models demonstrate strong target-adaptive capability when the target is explicit, but their performance declines when inputs containing ironic expressions.

Keywords: Stance detection , Target-adaptive , Interpretable , LLMs

* 通讯作者

第二十四届中国计算语言学大会论文集，第298页-第310页，济南，中国，2025年8月11日至14日。

1 引言

立场检测（Stance Detection）是自然语言处理中的关键任务，旨在识别文本作者对特定目标的态度立场(赵小兵et al., 2024)。与情感分析不同，立场检测关注的是作者对某一具体目标的支持、反对或中立态度，而不仅仅是情绪色彩。在社交媒体、新闻评论等场景中，立场检测被广泛应用于公共舆情分析、产品/服务支持度评估等领域(Küçük and Can, 2018; Glandt et al., 2021; Putra et al., 2022)。例如，针对近期火爆话题“北理工教师宫某被免职解聘”，网友评价“导师从来没强迫他什么，甚至支持他所做所有事情，不懂他背刺人家的原因”，就表达了该网友对“校方”的反对。随着社交媒体的普及，用户生成内容日益丰富，立场检测在理解和分析公众意见方面的重要性日益凸显(Gambini et al., 2024)。

当前，立场检测研究主要聚焦于已知目标的场景，利用预训练语言模型（如BERT、RoBERTa）构建目标感知的模型架构，以增强模型对目标的关注度。Stanceformer(Garg and Caragea, 2024)引入了目标感知的注意力机制，在多个数据集上取得了优异的表现。为适应真实场景，研究者提出了开放目标立场检测（Open-Target Stance Detection, OTSD）任务(Akash et al., 2024)，旨在在无预设目标的情况下，从文本中识别出潜在的立场目标及其对应的态度。OpenStance(Xu et al., 2022)定义了开放域零样本立场检测任务，旨在处理无领域限制且无特定话题注释的立场检测问题。此外，ZeroStance(Zhao et al., 2024)通过利用ChatGPT生成合成数据集，以支持开放域立场检测模型的训练。然而，上述工作仍依赖于预定义的目标列表或特定的领域数据，限制了模型在真实场景中的适配能力。并且，当前的研究主要关注于立场分类任务，缺乏对立场判断依据的可解释性建模。

针对当前立场检测任务中目标预设和可解释性不足的问题，本文提出了目标自适应的可解释立场检测新任务，旨在更贴近真实场景中的需求。该任务定义模型在无预设目标的条件下，从文本中自动识别和生成立场目标、观点及其态度，形成（目标，观点，立场）三元组输出，以增强模型的实用性和可解释性。为支持该任务的研究，本文构建了首个中文高质量立场检测数据集，包含5个微博热点话题，共10,513条标注数据；并设计了统一的多维评估标准，涵盖目标识别、观点生成和立场分类三方面的评估。其中目标识别的评估综合采用精确匹配（Exact Match, EM）、BLEU-4、BERTScore 和大型语言模型评分（LLM Score）四种指标，构建等权重的综合得分（Target Comprehensive Score, TCS）；观点生成的评估整合BLEU-4、BERTScore和LLM Score三种指标，形成观点生成的综合得分（Opinion Comprehensive Score, OCS）；立场分类的评估仅在目标识别正确的样本上，采用精确率（Precision）、召回率（Recall）和F1值等传统分类指标。此外，系统地评估了多种大语言模型（如DeepSeek-V3、GPT-4o等）的基线性能，分析了模型在不同场景下的表现。实验结果显示，DeepSeek-V3在目标识别与立场分类方面表现最优，甚至超过了DeepSeek-R1这类推理模型，GPT-4o在观点生成上具有领先优势。新任务直接面向真实场景的需求，能够更自然、更直接地服务于实际应用，同时为后续的相关立场检测研究提供了基准。

本文的主要贡献总结如下：

- 提出了目标自适应的可解释立场检测新任务，定义模型识别生成（立场目标、观点、立场）三元组，以提升模型在真实场景中的适用性和可解释性。
- 构建了高质量中文微博立场检测数据集，并提出了多维度评估标准，对目标识别、观点生成和立场分类分别进行综合评估。
- 全面评估并分析了10个大语言模型的基线性能，从整体表现和细分主题两个维度揭示不同大语言模型在该任务中的优劣。

2 相关工作

2.1 立场检测

立场检测研究历经从规则驱动到数据驱动的演变。早期工作依赖人工规则和浅层特征（如关键词、情感词）检测立场。SemEval-2016 Task 6(Mohammad et al., 2016)首次将立场检测标准化为三分类任务（支持、反对、中立）。基于SVM和n-gram特征的基线模型在监督任务（Task A）中取得了最高68.98%的F1宏平均，但在弱监督任务（Task B）中仅28.43%。此类方法对未明确提及目标的推文或意见指向其他实体的推文泛化能力有限，难以处理复杂

语义场景。随着深度学习技术的快速发展，基于双向循环神经网络（Bidirectional RNN）与注意力机制（Attention Mechanism）的方法逐渐成为立场检测领域的主流方法。Augenstein et al. (2016)提出的双向条件编码模型（BiCond）通过联合建模文本与目标的关系，在SemEval 2016数据集上实现了最优性能，在弱监督设置下的Macro F1值达到0.5803。Du et al. (2017)提出目标特定注意力网络（TAN），通过集成注意力机制与双向LSTM架构，动态分配目标相关上下文的注意力权重，显著提升对隐含语义关联的捕捉能力。后续研究引入图神经网络模型(杨顺成et al., 2020; Zhang et al., 2024)，用于微博或Twitter用户的立场检测。随着预训练语言模型的广泛应用，研究者利用其开展立场检测任务。Yin et al. (2024)等人提出融合显式立场标签与事件背景摘要的BERT-SLEK模型，通过建模“评论-目标-事件-立场”的四元关系，显著提升了社交媒体立场检测的效果。但以上研究中目标依旧为预定义的静态实体。部分研究尝试扩展至多目标场景，Sun et al. (2022)提出多目标立场检测框架，通过整合跨目标的主题与情感信息，有效提升模型对未标注目标的泛化能力，但未验证其对完全未知目标的适应性。可解释性方面，Draws et al. (2023)通过用户实验验证了基于LIME和模型系数的跨主题立场检测可解释性方法在搜索场景中的应用有效性。然而，上述研究仍依赖预定义主题且未解决动态目标推理问题，且无法生成与隐含目标对齐的立场依据。

2.2 基于大模型的立场检测

大语言模型（Large Language Models, LLMs）为突破上述瓶颈提供了新途径。在提示工程方面，Gatto et al. (2023)引入了思维链方法，将大语言模型生成的推理过程嵌入到传统的立场检测流程中，在社交媒体立场检测任务上取得了最佳性能。Weinzierl and Harabagiu (2024)提出了一种名为“反状事实树”的新型零样本立场检测方法，该方法不依赖任何立场标注样例，能够在文本和图像内容上进行推理，并在多个数据集上取得了优于微调系统的性能。Dai et al. (2025)利用思维链从大语言模型中提取一阶逻辑规则，并编码到神经网络结构中用于立场检测。张袁硕et al. (2025)在零样本与少样本场景下，使用结构化、加入额外信息、加入立场标签等8种不同的提示方式探索生成式语言模型在立场检测任务上的能力。然而，此类方法仍需显式提供目标，无法实现目标自适应，且未对最终得到的立场分类进行解释。Akash et al. (2024)提出了开放目标立场检测任务，利用大语言模型实现目标生成与立场检测，并提出了对于目标的三个评估指标：目标生成质量(BTSD)、人工评估(HE)和语义相似度(SemSim)。尽管该工作已针对开放目标，但提出的指标采用间接的方式，合理性不足。针对立场检测任务，尚无工作系统且合理地评估LLMs在开放域场景下识别隐含目标的能力（如从“应减少火力发电”文本中推断目标“碳中和”）；而且所得到的立场分类缺乏可解释性的依据，仅能判断出文本中支持、反对、中立的立场，未能生成反映立场的完整的观点表达。

3 目标自适应的可解释立场检测

为提升立场检测模型在真实场景中的适用性与可解释性，本文提出目标自适应的可解释立场检测新任务。该任务定义模型在无预设目标的前提下，自动识别文本中的立场目标、生成相关观点，并判断其立场态度，输出结构化的（目标，观点，立场）三元组结果。为支撑该任务的研究，本文构建了高质量中文立场检测数据集，并提出了覆盖三元组各部分的多维度评估标准。接下来将依次介绍任务设定、数据集构建与评估标准。

3.1 新任务设定

新任务是目标自适应的可解释立场检测，输入是用户的评论内容，输出为（目标，观点，立场）三元组，目标不预设，由模型识别；观点表达由模型生成，以提升立场判定的可解释性（图 1 展示了该任务的示例）。

考虑到实际表达中存在多目标、目标缺省以及间接立场表达等复杂情况，为降低建模难度并提升任务可控性，本文聚焦于单目标、直接目标的情形，即每条文本仅涉及一个明确的立场目标，并对该目标直接表达明确立场态度。目标通常为名词或名词短语，既包括静态实体（如人物、组织、机构），也包括动态事件或状态（如某一政策、行为或决定）；观点是对目标表达态度的语义内容，不局限于字面评价词语，而是更侧重语用层面的含义表达；立场则分为支持（Favor）、反对（Against）与中立（Neutral）三类。

该任务可采用级联式建模或端到端建模方式实现：级联方式将任务拆分为三个子任务，分

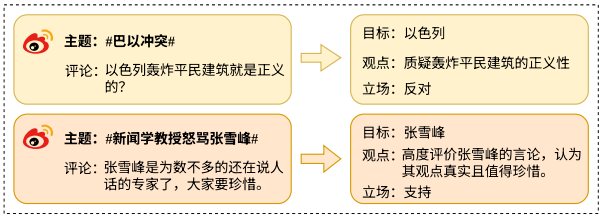


Figure 1: 立场检测新任务示例

别基于用户文本识别立场目标，生成观点表达，并在此基础上判断具体立场；端到端方式则直接基于输入文本生成完整的（目标，观点，立场）三元组。

3.2 中文立场检测数据集构建

3.2.1 数据收集与预处理

为构建面向真实应用场景、具备复杂语义结构的中文立场检测数据集，本文以新浪微博平台为数据来源，围绕社会、教育与国际政治三个具有高度公众关注度的领域选取典型争议性事件。事件选取遵循争议性、多样性与时效性三项标准，旨在捕捉用户在开放域话题中多样化的立场表达。最终所选事件包括“巴以冲突”、“央视记者采访燕郊爆炸被驱离”、“胡锡进谈虐猫事件”、“新闻学教授怒骂张雪峰”、“姜萍数学竞赛违规事件”，多领域的话题能够有效支撑模型在跨领域场景下的泛化能力评估，并为研究提供丰富的立场目标与观点构型。

数据收集过程中，首先选取转发量不低于100的微博主帖，确保文本具备较高的传播性和舆论代表性。随后，基于正则表达式与Unicode编码识别技术，过滤长度小于5字符或仅含表情符号的评论文本，并结合水军账号黑名单及典型行为特征识别方法，剔除非中文内容与疑似机器生成文本，以保证数据的真实性与语言质量。最终，共爬取原始评论35,000条，经预处理后保留22,987条用于后续标注。

3.2.2 数据标注与校验

数据标注过程采用基于大语言模型的自动标注方法，借助提示工程引导DeepSeek-V3完成三元组生成。具体操作中，通过任务描述和高质量示例构造标准化提示模板（如图 2 所示），以统一模型对任务边界和输出结构的理解。

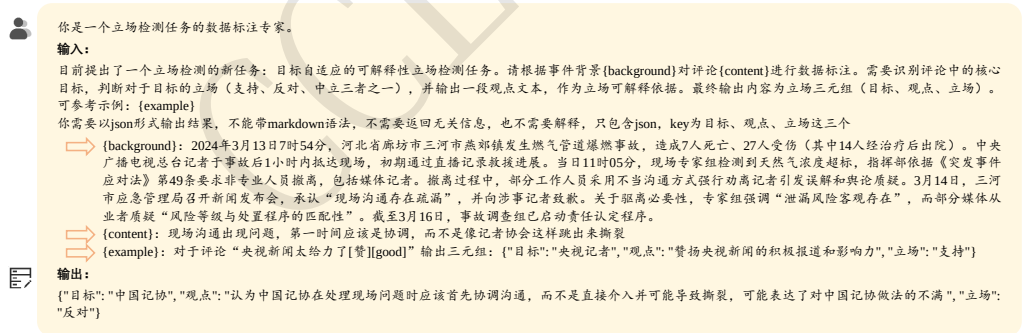


Figure 2: 中文立场检测数据标注的提示模板与示例

为保证标注质量，所有自动生成的标注结果均由三名具备自然语言处理专业背景的研究生进行交叉验证与一致性审校。校验过程中，明确给出“立场目标”与“观点”的判定约束，并结合具体实例进行一致性训练。立场目标限定为评论中直接关联事件主题的核心实体或抽象概念，需满足最小语义单元标准，并剔除过度泛化或模糊指代。例如，在“巴以冲突”相关语料中，需结合上下文对“铁穹”等具象表达进行指称还原，明确其语义指向为“以色列防空系统”；对于“以色列没收土地”一类“实体+行为”复合结构，仅保留“以色列”作为目标，相关行为纳入观点处理范畴。观点的生成聚焦于评论中针对目标的主张性表达，强调语用层面的立场推理能力。标注要求摒弃情感性修饰，保留具有逻辑指向和论证功能的观点核心。例如，在“燕郊爆炸事件”数

据中，需提炼出诸如“质疑信息发布时效性”或“肯定应急处置专业性”等带有明确论据的立场陈述。通过上述策略确保模型输出三元组在语义边界上的清晰性与可解释性。

3.2.3 数据统计与分析

经过人工校验和筛选，最终形成覆盖“巴以冲突”、“央视记者采访燕郊爆炸被驱离”（简称“采访冲突”）、“胡锡进谈虐猫事件”（简称“谈虐猫”）、“新闻学教授怒骂张雪峰”（简称“骂张雪峰”）、“姜萍数学竞赛违规事件”（简称“竞赛违规”）5个主题的、10,513条包含评论和立场三元组的有效数据，数据统计如表 1 所示。

事件主题	包含目标	支持	反对	中立	总计
巴以冲突	以色列（466）、哈马斯（58）、犹太人（54）等	111	907	84	1102
采访冲突	央视记者（258）、相关负责人或部门（213）、现场维护秩序的工作人员（183）等	115	1041	46	1202
谈虐猫	胡锡进（1018）、虐猫的学生（511）、虐猫事件（393）等	173	2959	118	3250
骂张雪峰	张雪峰（670）、新闻学教授（629）、新闻学（488）等	868	2438	272	3578
竞赛违规	姜萍（298）、媒体（227）、阿里巴巴全球数学竞赛组委会（106）等	131	1163	87	1381
总计		**1398**	**8508**	**607**	**10513**

Table 1: 数据各事件主题的目标分布与立场分布（目标后面的数字是目标出现的次数）

数据按照每个主题8:2、立场1:1的比例划分训练集/测试集，最终得到训练集数据为8386条，测试集数据为2127条。

分析发现，数据具有以下两个特点：

1）目标隐含且多样：评论中的立场目标常隐含在上下文中（如使用代词、简称），且对象多样（如可能指向“央视记者”“中国记协”“相关部门”等），表述方式灵活多变。这要求模型具备强大的上下文语义理解和目标实体识别能力。

2）观点复杂与口语化：微博评论语言呈现口语化、网络化特征，常包含讽刺、反语等修辞方法，如称媒体人为“无冕之王”实为讽刺。同时还存在情绪化措辞、非完整句式和对文化背景知识的依赖，如“巴以冲突”事件的评论涉及圣经典故，需将“应许之地”理解为殖民话语。此外，部分评论文本的理解需要进行多层语义推理。因此，准确提炼核心观点并生成流畅表述，对模型的语言理解和生成能力提出了更高要求。

3.3 评估标准

鉴于本任务输出形式为开放生成式的（目标，观点，立场）三元组，其中立场目标与观点内容均不具备唯一标准答案，传统的精确匹配指标难以全面反映模型性能。同时，三元组中各要素具有不同的语义属性与建模挑战，统一指标对整体结构进行评估在可操作性与解释性上均存在局限。因此，本文对三元组的三个组成部分——目标识别、观点生成与立场分类，分别设计评估指标体系。

3.3.1 目标评估指标

针对目标识别任务的开放生成特性，本文综合采用形式匹配与语义对齐结合的多维评估策略，构建目标识别综合得分（Target Comprehensive Score, TCS）。该评估体系由四项指标构成：

1）完全匹配率（Exact Match, EM）：该指标源自机器阅读理解评估范式(Rajpurkar et al., 2016)，本文定义为预测目标与标注目标在文本表面形式上是否完全一致的情况，若完全一致，则该值为1，否则为0。该指标严格衡量模型对立场目标边界的识别精度。

2）N-gram相似度（BLEU-4加权计算）：该指标基于经典机器翻译评估指标(Papineni et al., 2002)进行设计，我们改进权重分配策略以适配目标识别任务的短文本特性。具体而言，采用修正后的权重向量w=(0.5,0.3,0.1,0.1)，分别对应1-gram至4-gram的匹配权重，在确保核心词汇完整性的同时，兼顾基础词组的组合模式识别。

3）深度语义相似度（BERTScore）：该指标基于预训练语言模型的上下文感知能力(Zhang et al., 2019)，我们采用预训练的多语言BERT模型（bert-base-multilingual-cased）对目标文本进行语义一致性度量。通过提取候选文本与参考文本的上下文嵌入向量，计算二者在语义空间中的匹配程度。针对中文语言特性，设置参数lang='zh'以优化语义表征适配性。最终通过最大化词级相似度匹配，生成范围在(0，1)的归一化F1分数，并采用均值聚合获得整体评分结果。

4）大语言模型评分（LLM Score）：我们采用DeepSeek-V3模型作为评估主体，设计结构化提示模板引导大模型从核心实体匹配、语义范围一致、表述清晰度三个维度进行评分，具体提示模板如图 3 所示。

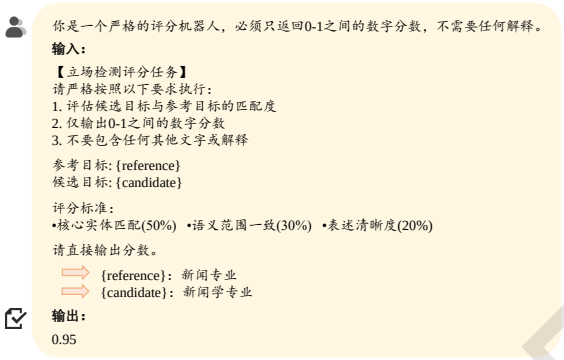


Figure 3: 大模型评估目标质量的提示词模板

综合上述四个维度的评估结果，通过等权平均构建目标综合得分TCS，计算公式如下：

$$\mathrm{TCS} = \frac{1}{4}\left(\mathrm{EM} + \mathrm{BLEU} + \mathrm{BERTScore} + \mathrm{LLM\ Score}\right) \tag{1}$$

3.3.2 观点评估指标

由于观点表述存在高度多样性，且可接受的语义表达通常不唯一，本文采用覆盖形式与语义两个层面的三种评估指标对生成观点进行综合评价：

1）N-gram相似度（BLEU-4均衡加权）：由于观点表达具有长度不定的特点，我们采用等权重分配策略w=(0.25,0.25,0.25,0.25)，分别对应1-gram至4-gram的匹配权重，以平衡词汇覆盖度、结构敏感性与语义完整性需求，避免因文本长度波动引发的评估偏差。

2）深度语义一致性（BERTScore）：沿用目标识别任务的评估方式，通过预训练多语言BERT模型（bert-base-multilingual-cased）提取上下文嵌入，计算候选观点与参考观点间的语义匹配度。

3）大模型评分（LLM Score）：采用DeepSeek-V3模型作为评分主体，通过结构化提示工程引导模型从立场一致性、论证逻辑匹配、表述清晰度、情感倾向一致四个维度进行量化评估。具体提示模板如图 4 所示。

通过等权融合策略构建观点综合得分OCS，计算公式如下：

$$\mathrm{OCS} = \frac{1}{3}\left(\mathrm{BLEU} + \mathrm{BERTScore} + \mathrm{LLM\ Score}\right) \tag{2}$$

3.3.3 立场评估指标

当预测目标错误时，立场的判定无意义，因此仅对目标判断正确的样本的立场进行评估。由于目标评估指标中的BLEU-4、BERTScore、LLM评分和综合得分Target Comprehensive Score（TCS）均为连续值，我们对每个指标分别设定了阈值用于目标正确与否的分类。首先将测试集结果按评估指标（如BLEU-4）降序排列。随后，由三名标注员独立审查排序后的样本，标注其质量等级（“合格”/“不合格”）。通过分析标注结果与指标值的对应关系，选择能够最大限度区分高质量与低质量样本的阈值。最终阈值经三位标注员讨论一致确认，以确保其符合实际应用需求。最终人工确定目标评估指标BLEU-4、BERTScore、LLM评分、TCS的阈值分别为0.1、0.7、0.65、0.3，因此将目标正确的样本定义为同时满足BLEU-4大

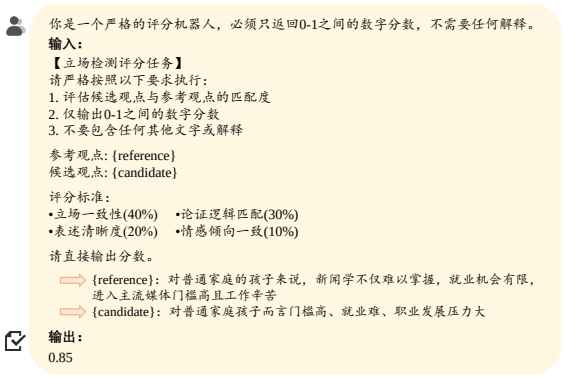


Figure 4: 大模型评估观点质量的提示词模板

于0.1、BERTScore大于0.7、LLM评分大于0.65和目标综合得分TCS值大于0.3这四个条件的样本。

针对目标正确样本，我们采用传统分类指标宏观精确率（Precision）、宏观召回率（Recall）、宏观F1值进行评估。通过独立计算支持、反对、中立三个类别的指标数值后取算术平均，确保类别不平衡场景下的评估公平性。

4 大模型实验

4.1 实验设置

4.1.1 模型选择

我们选取具有代表性的大语言模型对数据集进行评估，其中GPT系列选择GPT-4o(Hurst et al., 2024)和GPT-3.5 Turbo(OpenAI, 2023)，DeepSeek系列选取DeepSeek-R1(DeepSeek-AI, 2025)、DeepSeek-V3(DeepSeek-AI, 2024)，Qwen2.5系列(Team, 2024)选取Qwen2.5-72B-Instruct、Qwen2.5-32B-Instruct和Qwen2.5-7B-Instruct，Gemini系列选择Gemini-1.5-Pro(Team et al., 2024)，以及GLM系列(GLM et al., 2024)选择GLM-4-32B、GLM-4-9B。所有实验均采用API调用的方式完成。

4.1.2 提示方式

为确保大语言模型能够理解新任务，准确输出立场三元组（目标-观点-立场），我们采用结构化的提示方式。首先整合事件背景描述与评论文本，为模型提供上下文。其次，设计思维链的提示框架，明确要求大模型遵循目标识别、观点生成、立场分类的三阶段推理流程，给定目标和观点的约束条件，并限定输出格式为标准化三元组。具体提示模板如图 5 所示。

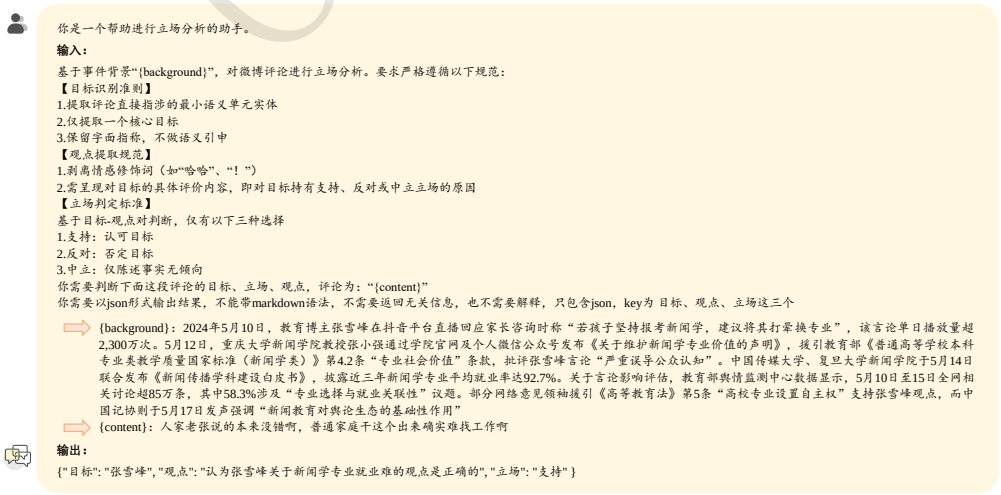


Figure 5: 引导大模型生成立场三元组的提示模板

4.1.3 评估方式

我们评估模型在测试集上的表现，训练数据预留给将来微调模型使用。采用3.3 介绍的评估标准分别评估目标识别、观点生成和立场分类，对于目标识别和观点生成，给出了各维度和综合的指标（BERTScore简写为BERT，LLM Score简写为LLM）。

4.2 实验与分析

4.2.1 整体结果与分析

模型	目标识别					观点生成				立场分类		
	EM	BLEU	BERT	LLM	TCS	BLEU	BERT	LLM	OCS	P	R	F1
GPT-4o	0.341	0.344	0.837	0.697	0.555	0.139	**0.767**	0.803	**0.570**	0.756	0.783	0.767
GPT-3.5 Turbo	0.324	0.337	0.835	0.681	0.544	0.119	0.753	0.774	0.549	0.485	0.601	0.485
DeepSeek-R1	0.348	0.355	0.841	0.699	0.561	0.135	0.753	0.810	0.566	0.789	0.836	0.810
DeepSeek-V3	0.415	0.412	**0.862**	**0.749**	**0.609**	0.139	0.743	0.823	0.568	**0.822**	**0.862**	**0.841**
Gemini-1.5-Pro	0.419	0.413	0.858	0.729	0.605	0.123	0.765	0.817	0.568	0.753	0.853	0.791
Qwen2.5-72B	**0.425**	**0.415**	0.859	0.735	0.608	0.136	0.740	**0.824**	0.567	0.723	0.818	0.742
Qwen2.5-32B	0.337	0.339	0.830	0.688	0.549	**0.140**	0.751	0.794	0.562	0.739	0.787	0.760
Qwen2.5-7B	0.279	0.300	0.823	0.678	0.520	0.108	0.759	0.759	0.542	0.639	0.705	0.666
GLM-4-32B	0.391	0.394	0.853	0.725	0.591	0.135	0.766	0.783	0.561	0.678	0.816	0.692
GLM-4-9B	0.381	0.386	0.846	0.706	0.580	0.124	0.735	0.741	0.534	0.587	0.739	0.615

Table 2: 整体实验结果（加粗表示最优结果，下划线表示次优结果）

10个大语言模型在测试集上的立场检测结果如表2 所示。分析可知：

1）整体上，大模型的参数规模与表现并不成正比，推理模型毫无优势。虽然DeepSeek-V3模型在目标识别、观点生成、立场分类的任务中都表现出最优或次优的性能，但大规模参数的推理模型DeepSeek-R1在目标识别上还不如Qwen2.5-72B（TCS指标：0.561 VS 0.608），在观点生成任务上还不如Gemini-1.5-Pro（OCS指标：0.566 VS 0.568），我们认为这是因为其过度关注推理步骤，简单问题复杂化，导致性能下降。而GPT-4o在目标识别任务中的表现仅达到中等水平，其TCS指标甚至显著低于GLM-4-9B模型。我们发现这是因为GPT-4o存在未有效结合输入的事件背景而仅简单根据评论文本进行目标识别的情况，例如，评论“背着手。好大的官威。”的目标为“工作人员”，而GPT-4o错误识别为“官威”。

2）在目标识别中，BERTScore指标普遍较高（0.823∼0.862），反映大模型对目标语义的强捕捉能力。DeepSeek-V3（TCS=0.609）表现领先，得益于其MoE架构对实体边界的精准捕捉。Qwen2.5-72B模型的识别效果仅次于DeepSeek-V3，其参数量达到720亿，相比之下参数量大于其他大多数模型，说明对于目标识别任务模型规模效应显著。Qwen2.5-7B模型由于参数规模过小，实体关系建模能力弱，目标识别表现最差（TCS=0.52）。

3）观点生成的任务差异较小，模型间得分接近，OCS值在0.534至0.57之间。效果最佳的是GPT-4o模型，BERTScore（0.767）和OCS（0.57）值达到最高，我们认为这是因为GPT-4o在语义理解深度上具有优势，且领域泛化能力强。Qwen2.5-32B在观点生成的BLEU指标中数值达到最佳，反超规模更大的Qwen2.5-72B，表明该模型在观点生成时表面语义匹配程度更高。Gemini-1.5-pro取得了和DeepSeek-V3持平的OCS值，且仅次于GPT-4o。其中Gemini-1.5-Pro在语义一致性（BERTScore）上表现突出，反映其预训练数据对复杂语义关系的强覆盖。GLM-4-9B模型观点生成的BLEU为0.124，但BERTScore和LLMScore值最低，分析发现其观点表述存在机械复制原文的现象，导致虽然表面匹配程度不低但语义提炼情况差。

4）在立场分析中，DeepSeek-V3的MoE架构在复杂推理中展现统治级表现，精确率P、召回率R和宏观F1值达到了最优，其中宏观F1值为0.841，超过GPT-4o0.074。而同系列同架构的DeepSeek-R1可能由于过度推理，立场分析的性能未能超越DeepSeek-V3。GPT-3.5 Turbo在立场分类表现最差，我们发现其预测立场出现三类立场以外的情况，如“质疑”“矛盾”。此外，该模型倾向于将立场判断为“中立”，在立场分类错误的1405个数据当中，有250个非中立立场的数据都被误判为中立，导致其立场分类效果远低于其他模型。

对模型生成的立场三元组进行进一步分析后发现：

1）小模型的指令遵循能力偏弱。虽然我们在提示词中明确要求大模型判断一个核心目标并输出对应立场三元组，但Qwen2.5-7B模型仍多次出现输出多个三元组的情况。比如对于评论“美帝是给一次战争找借口，犹太人是给一次屠杀找借口，更无耻更可恨”，Qwen2.5-7B模型识别出“美帝”和“犹太人”两个目标并对应生成了两个立场三元组。我们将输出多个立场三元组判定为输出错误，因此其结果较差。

2）大模型存在政治偏见。我们发现GPT-3.5 Turbo模型对政治话题过于敏感，无法客观输出内容。在分析“巴以冲突”主题部分数据如“地球上应该没有犹太人才会平安、杀害妇女儿童的就是犹太人”时无法输出有效的立场三元组，导致模型性能下降。

4.2.2 主题差异性分析

主题	目标识别					观点生成				立场分类		
	EM	BLEU	BERT	LLM	TCS	BLEU	BERT	LLM	OCS	P	R	F1
全部主题	0.415	0.412	0.862	0.749	0.609	0.139	0.743	0.823	0.568	0.822	0.862	0.841
巴以冲突	**0.667**	**0.655**	**0.905**	**0.851**	**0.769**	**0.190**	**0.767**	0.819	**0.592**	**0.945**	0.932	**0.938**
采访冲突	0.205	0.240	0.790	0.608	0.461	0.072	0.709	0.789	0.523	0.714	0.697	0.705
谈虐猫	0.347	0.360	0.844	0.666	0.554	0.150	0.751	0.818	0.573	0.649	0.761	0.695
骂张雪峰	0.445	0.399	0.882	0.820	0.637	0.159	0.754	**0.857**	0.590	0.810	0.847	0.826
竞赛违规	0.477	0.521	0.882	0.799	0.670	0.075	0.704	0.781	0.520	0.922	**0.945**	0.933

Table 3: 不同主题的实验结果统计（主题采用 3.2.3 中介绍的简写方式）

我们选择DeepSeek-V3的结果按主题进行统计（如表 3 所示）。分析发现：

1）“巴以冲突”主题在目标识别、观点生成、立场分类的评估结果都是最优的，其中目标综合指标TCS达到了0.769，观点综合指标OCS达到了0.592，而立场宏观F1值高达0.938。我们发现这是因为这一主题的大多数评论都表达出了明确的目标指向和强烈的情感倾向，比如评论“最大的恐怖组织就是以色列”“以色列的军事行动，让人不禁联想到历史上的某些黑暗篇章，真是时代的重演啊”，使得DeepSeek-V3模型能够准确地进行目标识别、观点生成和立场分类。

2）“采访冲突”（“央视记者采访燕郊爆炸被驱离”）主题的立场三元组评估结果都较差，我们分析发现，该主题的评论中存在大量反讽现象，如评论“央视记者被驱离，真是新闻自由的又一里程碑[狗头]”中使用正面词汇“里程碑”讽刺驱离记者进一步限制了新闻自由，“真是为央视记者的敬业精神点赞，燕郊爆炸现场的热度都被你们抢走了”中选择“敬业精神”“点赞”等用词表达因央视记者的行为导致事件重点转移的不满。对于这类样本，大模型未能进行准确分析和识别，导致表现不佳。

3）主题“骂张雪峰”（“新闻学教授怒骂张雪峰”）和主题“竞赛违规”（“姜萍数学竞赛违规事件”）相对于主题“谈虐猫”（“胡锡进谈虐猫事件”）的评估结果要好，这是因为，前两个主题中评论文本中大多直接出现事件相关人物，如“新闻学教授张小强”“张雪峰”“姜萍”等，这些明确的人物指向能够帮助大模型进行动态目标识别，并且进一步提炼观点和判断立场，因此表现较好。但后者的立场逻辑要稍微复杂，涉及到虐猫的学生、该学生的本科学校（东南大学）、该学生的一志愿研究生学校（南京大学），这加大了目标识别的难度，进而影响观点生成和立场分类。如对评论“他有学历，你可以接纳他呀，你不是喜欢广纳人才吗？与其道德绑架社会，为什么不自己接纳他？”，标注的三元组为（胡锡进，批评胡锡进的道德绑架行为，反对），DeepSeek-V3生成的为（南京大学，认为南京大学应接纳徐某而非道德绑架社会，反对），目标识别就存在问题。

4.2.3 案例分析

我们从DeepSeek-V3的结果中随机抽取部分进行案例分析发现：

1）存在目标识别、观点生成正确但立场分类错误的情况，如案例2中对央视记者的支持立场误判为“反对”，说明DeepSeek-V3在分析未明确表明立场的评论时存在困难。而案例7运用了反讽，讽刺媒体人为“无冕之王”，DeepSeek-V3正确识别了目标并生成了准确观点，但预测立场为“支持”，实际立场是“反对”，再次说明大模型在处理反讽语义上存在不足。

2）对简化目标和隐含目标的识别会出现错误。如案例1的评论将“以色列”简称为“以”，导致模型也将目标识别为“以”，未输出目标的指代全称。案例3中隐含目标为“胡锡进”，而模型错

Figure 6: 不同主题评论的案例分析（DeepSeek-V3）

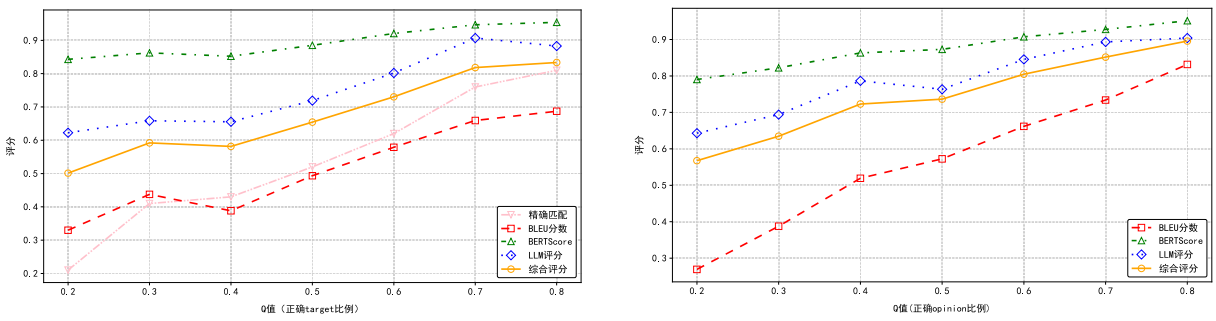
误识别为“其他网友”。

3）评论中存在反讽或反问时，DeepSeek-V3所生成的观点过于表面，未能准确对原评论进行解释。如案例5使用反讽，而DeepSeek-V3生成的观点“学得一手好‘新闻’”仅停留在表面，与标注观点“新闻学培养的媒体人故意制造对立和冷血报道”的表面语义相差甚远。案例6运用反问句，但模型仅简单抽取评论原句“不应该好好反思”作为观点，与实际含义不一致。

4）DeepSeek-V3模型能够生成正确的三元组，进一步发现测试集的2127条数据中少量数据（31条，1.46%）中生成的三元组与标注三元组完全一致，如案例4，其原因可能是数据预标注时选用了DeepSeek-V3模型，且数据质量高，人工审核步骤也无需修改。

4.3 指标有效性验证实验

为验证目标识别的评估指标EM、BLEU-4、BERTScore、LLM评分和Target Comprehensive Score（TCS）的有效性，我们仿照Akash et al. (2024)的做法，通过四种方法系统性地降低目标质量构建实验集来试验不同水平的目标质量：1）修改关键词，2）删除关键语义成分，3）替换为错误目标，4）随机选择无关词汇作为目标。其中，目标质量（Q）定义为测试集中正确目标的比例。如图 7a 所示，实验结果表明，当Q值从0.2提升至0.8时，各指标的值呈现显著线性增长（Pearson系数为0.955~0.986，p值均小于0.01），验证了我们所设定的目标评估指标的有效性。



(a) 目标质量和各评估指标之间的关系　　(b) 观点质量和各评估指标之间的关系

Figure 7: 目标质量/观点质量与各评估指标的关系

对于观点生成评估指标BLEU-4、BERTScore、LLM评分和Opinion Comprehensive Score（OCS），我们通过设计双重扰动验证方案来构建实验集：1）词汇级扰动，即替换核心立场支撑词，如将“显著促进”替换为“轻微改善”，2）生成句子级对抗样本，插入立场反转观点如“反对理由包括...”。实验结果如图 7b 所示，当观点质量Q值逐步提升时，观点的各个评估指标呈现显著线性增长（Pearson系数为0.971∼0.994，p值均小于0.01），证明了观点评估体系的有效性。

5 总结

本文面向现实场景中立场表达的复杂性与多样性，提出了目标自适应的可解释立场检测新任务，通过结构化建模立场目标、观点与态度三元组，提升了立场检测模型的适用性与解释能力。围绕该任务，构建了覆盖社会、教育与国际政治等多领域的高质量中文立场检测数据集，并在数据收集、自动标注与人工校验过程中，制定了严格的规范与准则，确保数据质量与标注一致性。同时，针对目标识别、观点生成与立场分类三个子任务，设计了涵盖形式与语义层面的多维度评估体系，结合自动评估指标与大型语言模型打分方法，为后续模型研究提供了统一、可复现的评估框架。最后，评估了10个大语言模型在新任务上的表现，并从整体和不同主题两个维度分析了评估结果。未来一方面将训练专业的立场分析大模型，另一方面将考虑多立场三元组的复杂场景。

致谢

感谢各位审稿人的细致工作和宝贵意见。本项研究成果受国家社会科学基金重大项目（22&ZD035）、中国语言资源保护工程（课题编号：YB2404A003）资助。

References

Abu Ubaida Akash, Ahmed Fahmy, and Amine Trabelsi. Can large language models address open-target stance detection? *arXiv preprint arXiv:2409.00222*, 2024.

Isabelle Augenstein, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva. Stance detection with bidirectional conditional encoding. *arXiv preprint arXiv:1606.05464*, 2016.

Genan Dai, Jiayu Liao, Sicheng Zhao, Xianghua Fu, Xiaojiang Peng, Hu Huang, and Bowen Zhang. Large language model enhanced logic tensor network for stance detection. *Neural Networks*, 183:106956, 2025.

DeepSeek-AI. Deepseek-v3 technical report, 2024. URL https://arxiv.org/abs/2412.19437.

DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

Tim Draws, Karthikeyan Natesan Ramamurthy, Ioana Baldini, Amit Dhurandhar, Inkit Padhi, Benjamin Timmermans, and Nava Tintarev. Explainable cross-topic stance detection for search results. In *Proceedings of the 2023 conference on human information interaction and retrieval*, pages 221–235, 2023.

Jiachen Du, Ruifeng Xu, Yulan He, and Lin Gui. Stance classification with target-specific neural attention networks. In *26th International Joint Conference on Artificial Intelligence, IJCAI 2017*, pages 3988–3994. International Joint Conferences on Artificial Intelligence, 2017.

Margherita Gambini, Caterina Senette, Tiziano Fagni, and Maurizio Tesconi. Evaluating large language models for user stance detection on x (twitter). *Machine Learning*, 113(10):7243–7266, 2024.

Krishna Garg and Cornelia Caragea. Stanceformer: Target-aware transformer for stance detection. *arXiv preprint arXiv:2410.07083*, 2024.

Joseph Gatto, Omar Sharif, and Sarah Masud Preum. Chain-of-thought embeddings for stance detection on social media. *arXiv preprint arXiv:2310.19750*, 2023.

Kyle Glandt, Sarthak Khanal, Yingjie Li, Doina Caragea, and Cornelia Caragea. Stance detection in covid-19 tweets. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (long papers)*, volume 1, 2021.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. Chatglm: A family of large language models from glm-130b to glm-4 all tools, 2024.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Dilek Küçük and Fazli Can. Stance detection on tweets: An svm-based approach. *arXiv preprint arXiv:1803.08910*, 2018.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, pages 31–41, 2016.

OpenAI. GPT-3.5 Turbo Model Documentation. https://platform.openai.com/docs/models/gpt-3-5-turbo, 2023. URL https://platform.openai.com/docs/models/gpt-3-5-turbo. Accessed: 2025-06-04. Official model documentation including capabilities, limitations, and usage guidance.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

Cornelius Bagus Purnama Putra, Diana Purwitasari, and Agus Budi Raharjo. Stance detection on tweets with multi-task aspect-based sentiment: A case study of covid-19 vaccination. *International Journal of Intelligent Engineering & Systems*, 15(5), 2022.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.

Qingying Sun, Xuefeng Xi, Jiajun Sun, Zhongqing Wang, and Huiyan Xu. Stance detection with a multi-target adversarial attention network. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(2):1–21, 2022.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

Qwen Team. Qwen2.5: A party of foundation models, September 2024. URL https://qwenlm.github.io/blog/qwen2.5/.

Maxwell Weinzierl and Sanda Harabagiu. Tree-of-counterfactual prompting for zero-shot stance detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 861–880, 2024.

Hanzi Xu, Slobodan Vucetic, and Wenpeng Yin. Openstance: Real-world zero-shot stance detection. *arXiv preprint arXiv:2210.14299*, 2022.

Zhaoning Yin, Yuanshuo Zhang, Aohua Li, Bo Chen, and Xiaobing Zhao. Fusion of stance labels and event knowledge for effective stance detection. In *2024 6th International Conference on Natural Language Processing (ICNLP)*, pages 203–207. IEEE, 2024.

Chong Zhang, Zhenkun Zhou, Xingyu Peng, and Ke Xu. Doubleh: Twitter user stance detection via bipartite graph neural networks. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 1766–1778, 2024.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

Chenye Zhao, Yingjie Li, Cornelia Caragea, and Yue Zhang. Zerostance: Leveraging chatgpt for open-domain stance detection via dataset generation. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 13390–13405, 2024.

张袁硕, 李澳华, 陈波, 尹召宁, 王潘怡, and 赵小兵. 基于生成式语言模型的立场检测探究. 中文信息学报, 39(3):139–147, 2025. URL http://jcip.cipsc.org.cn/CN/Y2025/V39/I3/139.

杨顺成, 李彦, and 赵其峰. 基于gcn 和bi-lstm 的微博立场检测方法. 重庆理工大学学报*(*自然科学*)*, 34(6):167–173, 2020.

赵小兵, 尹召宁, 王子豪, 张袁硕, and 陈波. 面向社会媒体的立场检测研究综述. 计算机应用研究, 41(11):3201–3214, 2024.