CCL25-Eval任务10系统报告：面向细粒度中文仇恨言论识别的大语言模型增强

林凡钧，张晏玮，黄杨，姚之远
华东师范大学，数据科学与工程学院
51275903052@stu.ecnu.edu.cn

摘要

本文介绍了我们在第二十四届中文计算语言学大会细粒度中文仇恨言论识别任务中的参赛系统。该任务要求构建结构化仇恨四元组（评论对象、论点、目标群体、是否仇恨），提升模型的细粒度检测与可解释性。我们基于大语言模型，首先评估了LoRA参数高效微调效果，优化了超参数配置；其次对标注数据进行结构化处理，增强数据规范性；最后优化提示词设计，引导模型生成准确的结构化输出。实验表明，三阶段优化提升了模型性能。

关键词： 情感分析；大语言模型；提示学习

System Report for CCL25-Eval Task10: Improving LLMs on Fine-grained Chinese Hate Speech Detection

Fanjun Lin, Yanwei Zhang, Yang Huang, Zhiyuan Yao
East China Normal University, School of Data Science and Engineering
51275903052@stu.ecnu.edu.cn

Abstract

This paper presents the system we developed for the fine-grained Chinese hate speech detection task at the 24th China National Conference on Computational Linguistics. The task aims to construct structured hate quadruples, including the target, argument, target group, and hateful, thereby enhancing the model's ability to perform fine-grained detection and improving the interpretability of its decisions. Based on a large language model framework, we first evaluated the effectiveness of parameter-efficient fine-tuning via LoRA, with a focus on optimizing key hyperparameters. Second, we applied structural refinement to the annotation data to improve data consistency and model comprehension. Finally, we further optimized the prompt design during fine-tuning to better guide the model in generating accurate structured outputs. Experimental results demonstrate that this three-stage optimization approach improves model performance.

Keywords: Sentiment analysis , Large language model , Prompt learning

第二十四届中国计算语言学大会论文集，第411页–第418页，济南，中国，2025年8月11日至14日。

1 引言

近年来，随着社交媒体的普及，仇恨言论的传播对社会和谐与个体安全构成了严峻挑战。细粒度中文仇恨言论识别旨在从文本中精准提取仇恨言论的核心要素（如评论对象、论点、目标群体及是否仇恨），从而实现对仇恨言论的结构化建模与可解释性分析。

仇恨言论检测研究早期聚焦多语言基础模型优化(Aluru et al., 2020)，通过分析9种语言的16个数据集，确立低资源语言中轻量级模型（如LASER嵌入+逻辑回归）与高资源语言中BERT类模型的性能差异，并揭示意大利语、葡萄牙语的零样本迁移优势。随着Transformer架构(Vaswani et al., 2017)普及，比较研究(Malik et al., 2022)验证了BERT、ELECTRA等模型在英语数据集上的优越性，同时指出其计算成本显著高于传统方法（如TF-IDF+XGBoost）。近年来，Twitter平台仇恨言论的长尾分布特性被系统揭示(Zhang and Luo, 2018)，其低占比（5.8%-31.6%）和特征模糊性催生CNN+跳过卷积层（sCNN）等新型结构，显著提升检测性能5-8个百分点。最新进展则关注多模态融合与公平性挑战，(Mandal et al., 2024)首创基于Transformer的文本-音频注意力融合框架，宏观F1值达0.927。(Das et al., 2024)则批判性揭示大语言模型（如GPT-3.5/GPT-4）在性别、种族维度的系统性标注偏见。

中文仇恨言论检测面临隐性表达与文化特异性双重挑战。显性检测领域，(Rao et al., 2023)构建首个多主题数据集CHSD（17430条），融合RoBERTa的语义理解、TextCNN的局部特征与BiGRU的全局依赖，实现89.12%的F1值。针对更隐蔽的仇恨形式，(Zhang et al., 2024)等提出领域增强提示学习框架（DePL），结合动态领域向量注入与软编码模板，在多领域数据集MCHID（20000条）上实现83.32%的准确率，显著优化低资源场景性能。最前沿研究(Bai et al., 2025)进一步深入细粒度解析，构建首个中文四元组标注数据集STATE ToxiCN（8029文本和9533标注），揭示谐音拆字、文化隐喻等语言特性对仇恨目标边界识别的挑战。

然而，该研究尚未系统解决多实例情境下的局部情感强化与隐含映射识别问题。本文在此基础上，提出LoRA微调、数据结构化与提示词优化三阶段策略，增强模型对于仇恨言论的识别能力。实验结果表明，1)LoRA微调超参数的合理设置不仅提升了模型的性能，还避免了过拟合问题;2)标注数据结构化增强了模型对各部分语义角色的理解，提升了模型在多实例情境下的判别精度;3)提示词优化显著提升了模型的推理准确性与输出一致性，通过引入少量代表性样例并结合思维链，模型能够更好地理解任务需求，从而显著提升性能。

2 方法

本文围绕提升大语言模型在中文细粒度仇恨言论识别任务中的推理能力，从模型微调、数据结构化与提示工程三个层面展开了系统性优化。首先，在资源受限条件下，采用LoRA微调策略，通过调整秩与缩放因子探索出适用于当前任务的最优参数配置。其次，提出两阶段的标注数据结构化方法，通过引入字段标签与局部情感强化机制，显著提升了模型对四元组语义结构的理解与建模能力。最后，针对提示词进行多轮迭代优化，涵盖输出格式规范、关键字段抽取规则细化、样例引导增强以及推理流程设计，全面提升了模型的推理准确性与输出一致性。

2.1 LoRA微调超参数优化

为了在有限的硬件资源下高效地对大语言模型进行微调，本文采用了低秩自适应(Low-Rank Adaptation, LoRA)方法。LoRA(Hu et al., 2021)是一种参数高效的微调技术，其核心思想是在预训练模型的权重矩阵中引入低秩矩阵来表示微调过程中新增的可训练参数。具体而言，设原始模型权重为$W \in \mathbb{R}^{d\times d}$，LoRA 引入两个低秩矩阵$A \in \mathbb{R}^{d\times r}$ 和$B \in \mathbb{R}^{r\times d}$，其中$r \ll d$，使得更新后的权重为W' = W + AB。通过这种方式，仅需训练A和B中的参数，从而极大地减少了训练所需的计算资源和内存开销。

相较于传统的全参数微调(Devlin et al., 2018)，LoRA在显存占用方面具有显著优势。全参数微调需要保存所有模型参数的梯度与优化器状态，导致显存消耗随模型规模呈线性增长，这在实验室资源配置有限的情况下往往难以实现。而LoRA通过低秩矩阵建模增量信息，仅需训练少量额外参数，因此大幅降低了显存需求，使得在单卡或小规模GPU环境下也能顺利完成微调过程。

第二十四届中国计算语言学大会论文集，第411页-第418页，济南，中国，2025年8月11日至14日。

鉴于本研究实验平台资源受限，本文选择LoRA方法作为主要的微调策略。在此基础上，本文进一步探索了不同超参数设置对模型性能的影响，重点考察LoRA的秩(rank)与缩放因子(alpha)之间的关系，以寻找最适合当前任务的最佳参数组合。

2.2 标注数据结构化

为了提升模型对细粒度中文仇恨言论识别中各要素的理解与建模能力，本文提出了两阶段的数据结构化策略。

第一阶段：显式标注字段名称以增强语义理解。

原始标注格式采用无字段名的简洁形式表示四元组，这种方式虽然节省了序列长度，但在实际训练过程中可能导致模型难以准确区分各部分语义角色，从而影响其抽取和分类效果。为此，本文对该格式进行了结构化改进，在每个字段前明确添加其语义标签，形成如下格式：

评论对象：{评论对象} | 论点：{论点} | 目标群体：{目标群体} | 是否仇恨：{是否仇恨} [END]

第二阶段：局部情感强化以提升细粒度判断能力。

在初步实验中本文发现，当文本中存在多个四元组时，模型倾向于依赖整体文本的情感倾向进行预测，而忽视了每个评论对象-论点对的局部情感判断。这种偏差可能削弱模型在复杂语境下对不同立场或观点的辨识能力。为了解决这一问题，本文在第一阶段的基础上进一步优化了标注格式，将目标群体与是否仇恨的判断绑定到具体的评论对象-论点对上，具体格式为：

评论对象：{评论对象} | 论点：{论点} | 评论对象-论点目标群体：{目标群体} | 评论对象-论点是否仇恨：{是否仇恨} [END]

2.3 提示词优化

在完成标注数据结构化的基础上，本文进一步对模型推理阶段所使用的提示词进行了系统性优化。提示词作为引导大语言模型生成任务相关输出的关键输入，其设计质量直接影响模型的理解能力与推理效果。

初始阶段，本文构建了一条基础提示词，主要包括以下几个部分：

- 角色设定："你是一名仇恨言论识别专家。"
- 任务描述："你的任务是仔细阅读并理解每一条输入的社交媒体文本。你需要从输入文本中准确地识别并抽取以下四个信息：评论对象、论点、目标群体、是否仇恨。"
- 定义说明：对四元组中各字段的具体含义进行了简要解释。
- 输入占位符："输入的文本：{数据集的社交媒体文本}。"

尽管该提示词初步实现了任务引导功能，但实际推理效果并不理想。为此，本文围绕提升模型理解准确性、增强输出一致性以及提高细粒度判断能力等目标，开展了多轮提示词优化实验。所有优化策略均经过验证，能够不同程度地提升模型在仇恨言论识别任务中的表现。

2.3.1 输出格式优化

为解决模型在推理过程中出现的输出结构不完整问题，本文在原有提示词基础上明确加入了输出格式要求，以增强模型对结构化输出的认知。此外，本文观察到模型倾向于将相同评论对象下的多个连续论点拆分为多个独立的四元组输出，导致结果冗余且不符合任务预期。因此，我们在提示词中增加了相同评论对象合并规则。

2.3.2 评论对象与论点部分优化

在文本分析任务中，评论对象与论点之间的紧密语义关联性对模型识别能力提出了更高要求。基于这一观察，我们在定义二者基本范畴后，系统性地制定了针对关键实体的抽取规则体系。实验发现，模型在推理过程中普遍存在对论点内容的语义重构现象，这种改写行为往往导致原始语义的偏移。为解决这一问题，我们在提示机制中特别强调必须严格遵循原文表述的约束条件，要求模型对文本片段进行逐字复制而非语义重述。

在评论对象的识别层面，模型尤其倾向于在无明显名词性评论对象的情况下将行为性描述误判为评论对象。对此，我们通过强化约束条件，明确指出评论对象必须为具体实体而非抽象

行为的判定原则。针对形容词修饰语境下的实体边界模糊问题，我们在提示词中嵌入了专门的决策指引。值得注意的是，在谐音类评论对象的识别环节，系统存在漏检现象，这直接影响了后续群体特征的判定准确性，因此我们补充了识别提示。

对于复杂句式的处理，我们发现模型在面对疑问句和否定句时，往往采取断章取义的抽取策略，这种简化处理容易导致情感判断的系统性偏差。为此，我们在提示机制中建立了完整的句式处理规范，要求模型在解析此类句子时必须完整保留原始表述，确保论点抽取的完整性与准确性。

2.3.3 是否仇恨部分优化

模型在判断文本是否具有仇恨性质时，常常因情感边界模糊而导致误判，尤其是在面对轻微仇恨表达或含有隐含映射的文本时表现不佳。为此，我们在提示词中加入了对“仇恨言论”的粗略定义，明确了判断标准，帮助模型更好地把握情感强度与语义内涵之间的对应关系。

2.3.4 目标群体部分优化

目标群体的识别是判断仇恨性质的重要依据，但由于其语义隐含性强，模型容易忽略其中的深层映射关系。我们在提示词中对对每个目标群体类别进行了更为详尽的定义说明，帮助模型建立清晰的分类边界。

针对模型忽视文本中隐含映射的问题，如中文拼音谐音、同音异字等现象，我们对提示词中进行隐含映射识别强化，加入了提示。

通过对训练集与测试集输出结果的分析，本文发现某些词汇或语句结构频繁出现并与特定目标群体高度相关。因此，在“地域仇恨”、“LGBTQ”、“其他仇恨”等类别下加入了简单的关键词列表，辅助模型做出更精准的判断。

2.3.5 其它优化策略

为进一步提升模型推理的稳定性和可解释性，本文先加入了Few-shot(Brown et al., 2020)的应用。我们引入比赛平台提供的两个样例，但是提升效果不明显。最终，我们在训练集中筛选出七个代表性样本，覆盖所有目标群体类别，并确保样例在抽取方式上具有多样性，显著提升了模型的泛化能力与推理质量。

本文发现模型在四元组整体抽取过程中还存在诸多问题，包括忽略局部细节、处理隐含映射能力不足、评论对象位置敏感等。为此，我们采用Few-shot-CoT(Wei et al., 2022)方法，设计了一个四步骤推理框架，详细描述每一步的推理逻辑，并将其嵌入七个样例之中，形成“思考+输出”的双阶段提示模式。该策略有效增强了模型的推理逻辑性和输出稳定性。

3 实验

3.1 实验设置

数据：本次评测使用的数据由比赛平台提供(Bai et al., 2025)。该数据集抽取数据集收集了贴吧、知乎等国内社交媒体平台的用户评论数据，为每条样本提供了高质量的二元分类标签，并对句子中的评论对象、论点和目标群体进行片段级标注。如表1，该数据集总计8029条中文数据，每条语句均包含一个或多个中文仇恨言论四元组，共计9533个，其中仇恨四元组6063个，非仇恨四元组3470个。

Category	#Posts	Quad.	Hateful	Non-hate
Train	6424	7631	4842	2789
Test	1605	1902	1221	681
Total	8029	9533	6063	3470

Table 1: 训练集和测试集的统计

模型与工具：本次任务限制了模型大小参数量小于10B，不得使用商业LLM API调用。本文通过实验对比和参考其他研究报告，选择了GLM4-9B-chat大模型进行测试和增强。大语言模型的微调工具选择的是SWIFT轻量级微调推理框架。

参数设置：本文所使用的模型在A100 GPU上进行训练，训练批次大小为1，训练轮次为2，每轮训练进行360次参数更新，最大序列长度为2048，使用AdamW(Loshchilov and Hutter, 2019)优化器进行优化，使用LoRA方法进行微调，学习率搜索空间为{1e-5, 3e-5, 6e-5, 1e-4}。

评估指标：本文采用预测和实际结果的硬匹配和软匹配分别的F1分数，以及两种方式的F1分数的平均分作为评价指标。

硬匹配：当且仅当预测四元组的每一个元素都与答案中对应元素完全一致才判断为正确抽取的四元组。

软匹配：当且仅当预测四元组的目标群体、是否仇恨两个元素和标准答案中相对应的两个元素完全一致，并且预测四元组的评论对象、论点两个元素和标准答案中相对应的两个元素的字符串匹配程度超过50%才判断为正确抽取的四元组。

3.2 实验结果

不同LoRA超参数的大模型推理性能：表2对比了模型在不同LoRA超参数微调下在测试集上的评测结果。通过调整LoRA的秩和缩放因子，我们发现rank=32和alpha=64的组合在资源受限的情况下实现了最佳性能。这表明在特定任务中，选择适当的低秩矩阵大小和缩放因子能够显著提升模型的表现。然而，随着rank和alpha的增加，性能并没有持续提升，反而有所下降，这可能是由于过拟合或计算复杂度增加导致的。

Setting	Hard matching			Soft matching			score
	P	R	F1	P	R	F1	
rank=8, alpha=16	26.11	26.18	26.15	49.61	49.74	49.67	37.91
rank=8, alpha=32	25.78	25.40	25.59	49.35	48.26	48.80	37.20
rank=16, alpha=32	25.55	25.60	25.58	46.84	49.95	48.89	37.74
rank=32, alpha=64	26.36	26.18	26.27	49.76	49.42	49.89	**37.93**
rank=64, alpha=128	25.05	25.76	25.40	48.42	49.79	49.09	37.25

Table 2: 模型在不同LoRA超参数微调下的评测结果

标注数据结构化性能：表3对比了模型在对标注数据结构化前后的评测结果。我们先是显式标注字段名称，增强了模型对各部分语义角色的理解，减少了混淆。我们再通过局部情感情感，进一步提升了模型在多实例情境下的判别精度，特别是在处理复杂句式时，模型能够更好地关注每个评论对象-论点对的情感倾向。如表3所示，结构化标注使硬匹配F1从26.15%提高至26.95%，软匹配F1从49.67%提升至50.76%，综合得分达到38.86%。虽然局部情感增强在综合分数上没有较大提升，但是在硬匹配上的F1分数有了较大提升。

Setting	Hard matching			Soft matching			score
	P	R	F1	P	R	F1	
Initial Label	26.36	26.18	26.27	49.76	49.42	49.89	37.93
Structuring Label	26.24	26.76	26.50	50.72	51.74	51.22	38.86
Enhancing hate Detection	26.82	27.08	**26.95**	50.52	51.00	**50.76**	**38.86**

Table 3: 模型在标注数据结构化不同阶段的评测结果

提示词优化性能：表4对比了在提示词中对仇恨四元组的不同部分进行优化后对模型推理的性能影响。我们发现提示词优化显著提升了模型的推理准确性与输出一致性。明确输出格式要求有助于模型生成结构化的输出，而细化关键实体的抽取规则体系则提高了模型对评论对象和论点的识别精度。从表4我们看到针对四元组不同字段的提示词优化策略均带来性能提升，而且“四元组整体优化”策略使硬匹配F1达到27.60%，软匹配F1达到50.80%，综合得分为39.20%。

Setting	Hard matching			Soft matching			score
	P	R	F1	P	R	F1	
Initial Prompt	26.82	27.08	26.95	50.52	51.00	50.76	38.86
Standarding format	26.36	26.97	26.66	49.18	50.32	49.74	38.20
Optimizing Target-Argument	27.05	27.97	27.50	49.82	51.52	50.66	39.08
Optimizing Hateful	27.30	28.08	27.68	49.85	51.26	50.54	39.11
Optimizing Targeted Group	27.03	27.97	27.49	49.75	51.47	50.59	39.04
Optimizing Quadruple	27.20	28.02	**27.60**	50.11	51.52	**50.80**	**39.20**

Table 4: 对仇恨四元组不同部分的提示词优化后的模型评测结果

表5对比了在零样本、少样本和少样本+思维链的不同情况下模型评测结果。在对四元组整体优化后，我们先是引入了两个样本的Few-shot，结果发现模型的综合分数下降到了39.02%，我们推测可能是模型对这两个样本的目标群体部分分类产生了过拟合，导致模型性能下降。为了平衡模型对各种目标群体的推理能力，我们将样本数量增加到了七个，覆盖率所有的目标群体类型。在七个样本的Few-shot的实验结果中，我们发现模型性能有所提升。随后我们加入了思维链，设计了模型对该任务的四步骤推理流程，该策略使得模型的综合得分上升到了39.33%。

Setting	Hard matching			Soft matching			score
	P	R	F1	P	R	F1	
Zero shot	27.20	28.02	27.60	50.11	51.52	50.80	39.20
Few shot (two examples)	26.96	27.94	27.44	50.02	51.20	50.60	39.02
Few shot (seven examples)	27.23	28.17	27.69	50.15	51.55	50.84	39.27
Few-shot-CoT	27.54	27.97	27.75	50.52	51.31	50.91	**39.33**

Table 5: 零样本、少样本的模型评测结果

性能对比：表6列出了我们方法的得分以及与其他论文(Bai et al., 2025)中大模型在该数据集上的得分。这些模型的参数量均在10B以下且经过微调，与我们测试的模型能够较好地进行比较。我们的模型在硬匹配F1上达到了27.75%，明显高于其他模型。这表明我们的方法在精准提取四元组各字段方面具有显著优势。在软匹配F1方面，我们的模型达到了51.31%，远超其他模型。这表明我们的方法在处理模糊匹配和部分匹配任务中表现尤为出色。特别是Qwen2.5-7B和ShieldGemma-9B虽然在某些场景下表现较好，但在整体软匹配F1上仍不及我们的模型。综合得分反映了模型在多种匹配条件下的综合性能。我们的模型以39.33%的高分位居榜首，显示出其在处理复杂句式和多实例情境中的强大能力。

Model	Hard F1	Soft F1	score
Finetuned Models			
mT5-base	16.60	38.61	27.61
Mistral-7B	23.72	45.62	34.67
LLAMA3-8B	24.27	46.08	35.18
Qwen2.5-7B	23.70	47.03	**35.37**
ShieldLM-14B-Qwen	23.59	45.58	34.59
ShieldGemma-9B	23.49	47.14	35.32
Ours	**27.75**	**51.31**	**39.33**

Table 6: 评测结果对比

3.3 分析

本研究提出了一种基于大语言模型的细粒度中文仇恨言论识别方法，通过LoRA微调、数据结构化与提示词优化三阶段策略，显著提升了模型的性能。如3.2实验结果所述，通过LoRA微调策略，我们能够在有限资源条件下实现高效优化。rank=32、alpha=64的配置不仅提升了模型的性能，还避免了过拟合问题。与其他基线模型相比，我们的方法在保持计算成本可控的同时，实现了更高的准确性和鲁棒性。

显式标注字段名称增强了模型对各部分语义角色的理解，减少了混淆。局部情感绑定进一步提升了模型在多实例情境下的判别精度，特别是在处理复杂句式时，模型能够更好地关注每个评论对象-论点对的情感倾向。

提示词优化显著提升了模型的推理准确性与输出一致性，明确输出格式要求有助于模型生成结构化的输出，而细化关键实体的抽取规则体系则提高了模型对评论对象和论点的识别精度。通过引入少量代表性样例并结合Few-shot-CoT推理框架，模型能够更好地理解任务需求，从而显著提升性能。这表明在有限样本条件下，合理的样本选择和推理逻辑设计对于提升模型性能至关重要。

4 结论

在本次细粒度中文仇恨言论识别任务中，本研究提出了一种基于大语言模型的细粒度中文仇恨言论识别方法，通过LoRA微调、数据结构化与提示词优化三阶段策略，显著提升了模型的性能。实验结果表明，适当调整LoRA的秩和缩放因子能够在不显著增加计算成本的前提下提升模型性能。字段标签和局部情感绑定机制有效提升了模型对四元组语义的理解能力，减少了字段混淆问题，增强了模型在复杂句式中的表现。通过逻辑链引导，提升了复杂句式和多实例场景下的推理难题的能力，加强了模型的稳定性和准确性。尽管本方法在硬匹配和软匹配指标上都有所优化，但是仍存在对隐含映射的建模不足的问题，需进一步融合语言学规则或引入多模态信息，以提升对谐音、文化隐喻等现象的识别能力。而且模型的泛化能力待验证，需在跨领域或跨平台数据上测试模型的鲁棒性，确保其在不同应用场景下的有效性。未来对细粒度中文仇恨言论识别问题，我们认为一种解决方案是结合字符级与句法级信息，提升对谐音、文化隐喻的识别能力。一种是扩大优质数据集，通过在更多社交媒体上搜集更大样本量的数据，来扩展模型在面对不同场景下的适用性。

参考文献

Amit Das, Zheng Zhang, Najib Hasan, Souvika Sarkar, Fatemeh Jamshidi, Tathagata Bhattacharya, Mostafa Rahgouy, Nilanjana Raychawdhary, Dongji Feng, Vinija Jain, Aman Chadha, Mary Sandage, Lauramarie Pope, Gerry Dozier, Cheryl Seals. 2024. Investigating Annotator Bias in Large Language Models for Hate Speech Detection. *arXiv preprint arXiv:2406.11109*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin. 2017. Attention is All You Need. In *Advances in Neural Information Processing Systems*.

Atanu Mandal, Gargi Roy, Amit Barman, Indranil Dutta, Sudip Kumar Naskar. 2024. Attentive Fusion: A Transformer-based Approach to Multimodal Hate Speech Detection. *arXiv preprint arXiv:2401.10653*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Lu Wang, Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv preprint arXiv:2106.09685*.

Ilya Loshchilov, Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

Jitendra Singh Malik, Hezhe Qiao, Guansong Pang, Anton van den Hengel. 2022. Deep Learning for Hate Speech Detection: A Comparative Study. *arXiv preprint arXiv:2202.09517*.

Sai Saketh Aluru, Binny Mathew, Punyajoy Saha, Animesh Mukherjee. 2020. Deep Learning Models for Multilingual Hate Speech Detection. *arXiv preprint arXiv:2004.06465*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. 2020. Language Models are Few-Shot Learners. *arXiv preprint arXiv:2005.14165*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems*.

Xiaojun Rao, Yangsen Zhang, Qilong Jia, Xueyang Liu. 2023. Chinese Hate Speech Detection Method Based on RoBERTa-WWM. In *Proceedings of the 22nd Chinese National Conference on Computational Linguistics*.

Zewen Bai, Yuanyuan Sun, Shengdi Yin, Junyu Lu, Jingjie Zeng, Haohao Zhu, Liang Yang, Hongfei Lin. 2025. STATE ToxiCN: A Benchmark for Span-level Target-Aware Toxicity Extraction in Chinese Hate Speech Detection. *arXiv preprint arXiv:2501.15451*.

Yaosheng Zhang, Tiegang Zhong, Tingjun Yi, Haoming Li. 2024. Domain-Enhanced Prompt Learning for Chinese Implicit Hate Speech Detection. In *IEEE Access*.

Ziqi Zhang, Lei Luo. 2018. Hate Speech Detection: A Solved Problem? The Challenging Case of Long Tail on Twitter. *arXiv preprint arXiv:1803.03662*.

第二十四届中国计算语言学大会论文集，第411页–第418页，济南，中国，2025年8月11日至14日。