CCL25-Eval任务10系统报告：基于动态线索增强提示与多阶段渐进优化的中文仇恨言论检测方法

阮禄，翟波，张蕾，鲍烈，王泽宇，危枫，王晨子
中国电信股份有限公司重庆分公司，重庆
{ruanlu.cq, zhaib, zhangl157, baol, wangzy124,
weifeng2.cq, wangchenzi.cq }@chinatelecom.cn

摘要

随着社交媒体的迅速普及，用户生成内容呈指数级增长，同时也助长了仇恨言论的扩散。因此，有效检测仇恨言论已成为自然语言处理研究领域的一项关键挑战。为推动中文仇恨言论检测技术的发展，本文提出了一种新颖的大语言模型微调框架，该框架融合了动态线索增强提示和多阶段渐进优化方法。所提出的方法将复杂的细粒度仇恨言论识别任务分解为两个相辅相成的子任务：仇恨倾向分类和仇恨信息提取。为此采用了两种专门的训练策略：动态线索增强提示微调（DCA-SFT）用于优化模型的分类性能，而动态线索增强强化学习（DCA-RL）则用于提升模型的信息提取能力。具体而言，在DCA-SFT阶段，引入判别式分类并采用多标签独热（Multi-Hot）编码作为输出表示形式，以提高模型的多类别分类准确率。在DCA-RL阶段，通过知识蒸馏的方式，将闭源大语言模型在执行仇恨信息提取任务时的思维链（CoT）知识迁移至小参数模型，同时引入基于规则奖励的强化微调策略来增强小参数模型在信息提取任务中的逻辑推理能力。实验结果证明了该方法的有效性，在CCL25-Eval任务10的初赛排行榜上以0.3864的F1值，排名第二；在决赛排行榜上以0.3591的F1值，位列第三。

关键词： 仇恨言论检测 ； 多阶段渐进优化 ； 动态线索增强提示

System Report for CCL25-Eval Task 10: Chinese Hate Speech Detection Method Based on Dynamic Clue-Augmented Prompting and Multi-Stage Progressive Optimization

Lu Ruan,Bo Zhai,Lei Zhang,Lie Bao,Zeyu Wang,Feng Wei,Chenzi Wang
China Telecom Corporation Limited Chongqing Branch, Chongqing
{ruanlu.cq, zhaib, zhangl157, baol, wangzy124,
weifeng2.cq, wangchenzi.cq }@chinatelecom.cn

Abstract

With the burgeoning popularity of social media, user-generated content has witnessed exponential growth, concurrently fueling the proliferation of hate speech. Effectively detecting hate speech has thus emerged as a critical challenge in natural language processing research. To advance Chinese hate speech detection techniques, this paper introduces a novel large language model (LLM) fine-tuning framework integrating dynamic clue-augmented prompting and multi-stage progressive optimization. The proposed approach decomposes the intricate fine-grained hate speech recognition task into two complementary subtasks: hate tendency classification and hate information extraction. Two specialized training strategies are employed: Dynamic Clue-Augmented Prompting (DCA-SFT) optimizes the model's classification performance, while Dynamic Clue-Augmented Reinforcement Learning (DCA-RL) is adopted to refine its extraction capabilities. Specifically, in the DCA-SFT phase, Multi-Hot encoding is utilized as the output representation to enhance the model's multi-class classification accuracy. During DCA-RL, knowledge distillation is used to transfer the Chain of Thought (CoT) knowledge from closed-source LLMs during information extraction

第二十四届中国计算语言学大会论文集，第419页-第429页，济南，中国，2025年8月11日至14日。

tasks to small-parameter Large Models (LMs), while a rule-based reward strategy is introduced to enhance the logical reasoning ability of small-parameter LMs in extraction tasks. Experimental results demonstrate the effectiveness of this approach, achieving an F1-score of 0.3864 and ranking second on the preliminary leaderboard, and an F1-score of 0.3591, securing third place on the final leaderboard for Task 10 of CCL25-Eval.

Keywords: Hate Speech Detection , Multi-Stage Progressive Optimization , Dynamic Clue-Augmented Prompting

1 引言

随着社交媒体发展与用户生成内容激增，网络有害言论传播问题凸显，移动互联网时代更使种族主义、仇恨言论及煽动歧视暴力言论传播加速。仇恨言论指基于宗教、族裔、国籍、种族、肤色、血统、性别等身份因素，对特定个体或群体进行攻击或使用贬损、歧视性语言的言论等交流，相较其他有害言论，其更具强迫性、欺凌性和煽动性，影响被攻击对象及整个社会。高效精准识别仇恨言论是网络空间治理关键前提，故构建兼具准确性和鲁棒性的仇恨言论检测系统成为自然语言处理领域重要研究方向。

近年来，仇恨言论识别研究逐渐从粗粒度的整体分类转向细粒度的精准定位分析 (György et al., 2021; Lu et al., 2023)。粗粒度仇恨分类通常将整个言论作为分类单位，利用CNN (Badjatiya et al., 2017; Gambäck et al., 2017; Park et al., 2017; Zimmerman et al., 2018)、RNN (Del et al., 2017; Do et al., 2019; Wang et al., 2019)、BERT (Marzieh et al., 2019; Zampieri et al., 2019)、RoBERTa (Bertie et al., 2021)等模型进行多分类任务，例如RoBERTa-CHSD (Rao et al., 2023)方法中的数据集CHSD仅标注文本的整体仇恨倾向。粗粒度仇恨言论分类方法虽能直观判断是否为仇恨言论，却无法提供仇恨的具体攻击对象和论点等信息，难以满足网络空间治理的精细化需求。为此，研究者们开始探索细粒度仇恨言论分析范式 (Lu et al., 2023; Bai et al., 2025)，其基于片段级言论数据集展开，面临两大核心挑战：一是数据覆盖不足。现有片段级仇恨言论数据集多局限于1-2个目标群体，难以全面反映现实网络中复杂的仇恨言论分布特征；二是任务复杂度高。细粒度分析需在完成仇恨倾向分类的同时进行仇恨要素信息抽取，对模型的语义理解和推理能力有更高要求。

针对数据覆盖不足、缺乏跨度级别的细粒度标注等问题，大连理工大学信息检索研究室构建了首个支持细粒度中文仇恨言论检测分析的跨度级别目标感知毒性提取数据集STATE ToxiCN (Bai et al., 2025)，涵盖了性别歧视、种族主义、地区偏见和反LGBTQ情绪等多个目标群体的仇恨言论。并且每个样本被都被标注为四元组<评论对象，论点，目标群体，是否仇恨>，包含了高质量的仇恨倾向二元分类标签，以及细粒度的评论对象、论点和目标群体的片段级标注。

面对细粒度中文仇恨识别任务复杂度高的挑战，本文提出了一种基于动态线索增强（Dynamic Clue-Augmented，DCA）提示与多阶段渐进优化的大模型微调解决方案：首先，将复杂的四元组抽取任务解耦为仇恨倾向分类和仇恨信息抽取两个子任务。其次，设计了一个三阶段渐进式的DCA大模型微调方法，通过动态线索增强提示微调（DCA-SFT）和强化学习（DCA-RL）训练策略分别优化大模型在分类和抽取两个子任务上的效果。在DCA-SFT阶段引入判别式分类并采用多标签独热（Multi-Hot）编码作为输出表示形式来提升模型的多类别分类准确率；在DCA-RL阶段通过知识蒸馏的方式，将闭源大语言模型在执行仇恨信息提取任务时的CoT知识迁移至小参数模型，同时引入基于规则奖励的强化微调策略来增强小参数模型在信息提取任务中的逻辑推理能力。

本文选取了开源的Qwen2.5-7B作为基础模型进行多阶段渐进式的DCA-SFT与DCA-RL微调，并采用闭源大模型Qwen-Plus进行知识蒸馏。实验结果表明，1）与传统监督微调（SFT）相比，本文提出的DCA-SFT在相同实验条件下展现出显著优势；2）将群体标签数字化处理比直接使用字符标签能够为模型带来更优的分类性能；3）通过知识蒸馏整合闭源大模型的先进知识和思维链推理能力，可以有效提升模型的领域知识学习能力。最终，本文提

出的方法在CCL25-Eval任务10的初赛排行榜上以0.3864的F1值，排名第二；在决赛排行榜上以0.3591的F1值，位列第三。

2 方法

2.1 方法总述

本次评测任务需要基于输入的社交媒体文本，通过模型识别并输出仇恨四元组，包括评论对象（Target）、论点（Argument）、目标群体（Targeted Group）、是否仇恨（Hateful）。由于任务复杂程度较高，需要细粒度分析能力与领域知识支撑，单纯依靠提示词工程驱动大模型难以实现完整四元组的准确抽取。经观察发现，仇恨四元组各元素间存在一定规律与联系：1）仇恨评论必然涉及对某一群体的仇恨，否则应界定为非仇恨评论文本；2）论点作用于评论对象可作为构成仇恨的证据；3）评论对象属于目标群体的代表；4）评论对象与论点的内容不固定，而目标群体和是否仇恨的取值固定。基于此，我们对四元组抽取任务进行层次拆解。评论对象与论点内容需模型自动识别抽取，可定义为抽取任务；目标群体和是否仇恨的取值固定，可定义为分类任务。由此，本评测任务被拆解为仇恨分类任务与仇恨信息抽取任务。

针对上述两个细粒度子任务，我们提出了DCA增强提示技术，并设计了多阶段渐进式优化微调方案，以提升子任务模型的精度与泛化能力。具体如图1所示，展示了分类和抽取任务模型的微调和推理过程。首先，基于官方数据集构建分类任务微调数据集，并通过添加DCA 提示对分类任务进行冷启动微调，得到初始分类模型(stage1)。为充分利用无标注的外部数据集，使用微调后的模型预测补充数据样本作为伪标签，通过一致性检验和蒙特卡洛采样(Monte Carlo Dropout,MC dropout) 过滤不可信标签以提升伪标签样本质量，随后将其并入分类任务微调数据集继续优化分类任务模型，经多次增强训练迭代（stage2）逐步提升补充数据伪标签质量与分类任务模型精度；在训练抽取任务模型时(stage3)，先基于闭源超大参数模型Qwen-plus 在官方数据集上生成CoT以构建带分析过程的数据集，再结合DCA 提示与强化学习方法对Qwen2.5-7B 进行微调；推理过程中，同样需添加DCA 提示，先通过分类任务模型预测获得<目标群体,是否仇恨>二元内容，再通过抽取任务模型获得<评论对象,论点>二元内容，最终得到仇恨四元组。

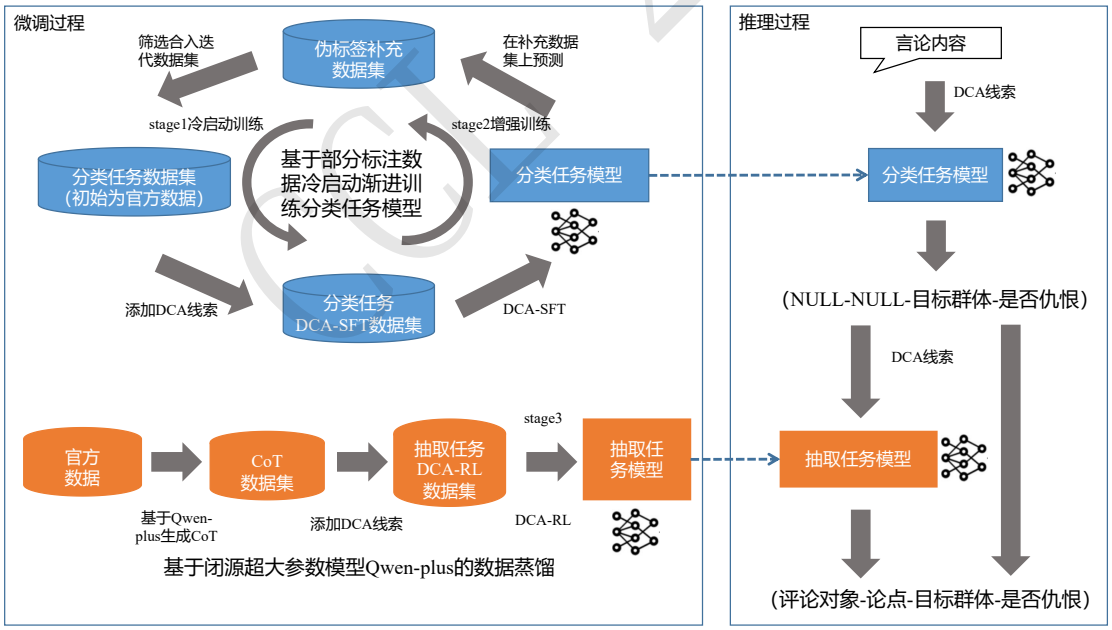

图 1: 基于动态线索增强提示与多阶段渐进优化的方案示意图

2.2 动态线索增强提示

针对仇恨言论检测任务的特点，我们提出了动态线索增强提示提示DCA技术，它是一种面向分类与抽取任务的提示工程与模型参数联合优化技术，旨在通过引入结构化先验知识与动态

推理线索，引导大模型聚焦输入文本的关键语义信息，提升复杂任务下的输出精度与稳定性。其核心思想是将领域知识（如仇恨言论关键词）与任务中间结果（如分类标签）编码为提示词的组成部分，形成“输入文本+动态线索”的联合输入模式，从而显式调控模型注意力分配机制，缓解传统SFT中语义稀疏与推理路径模糊的问题。

（1）分类任务DCA线索设计：我们将关键信息库(来源于ToxiCN)中的关键词以及关键词的类别作为“语义锚点”，通过提示词模板强制模型在编码阶段分配更多注意力至这些位置，帮助模型在进行分类任务的时候重点关注这些信息，分类任务DCA线索如下：

分类任务DCA线索
关键信息：“基佬”可能涉及到：LGBTQ群体。

（2）抽取任务DCA线索设计：我们设计了一个多阶段线索级联的DCA,除了引入（1）中的关键词外，我们还将分类任务中预测的仇恨标签作为二级DCA,形成一个多阶段增强的DCA线索，这里的关键词主要用于指导模型生成<评论对象,论点>二元组，仇恨标签主要用于指导模型生成<群体,是否仇恨>二元组。抽取任务DCA线索如下：

抽取任务DCA线索
关键词：泥鸽、母勾 是否仇恨：hate，主要涉及：Racism, Sexism等群体。

（3）逻辑约束注入：在设计DCA线索时，当仇恨标签为“non-hate” 时，强制群体标签字段为“non-hate”，避免逻辑矛盾，同时当未匹配到关键词时应将DCA线索设置为“无关键信息”，从而让模型学会自主决策并辨认出更隐晦的关键信息。

2.2.1 分类任务DCA-SFT指令微调模板设计

由于大模型的特性，微调过程中损失的实际计算方式是通过比较输出文本和标准答案的一些生成指标得到，如果直接使用群体的英文字符作为标签会导致模型计算损失出现分类边界认知差距，如Region和Region,Sexism这两个字符标签的损失就较为接近，从而导致大模型无法学习类别的边界，因此为了提升模型对类别的边界认知能力，同时鉴于该任务是一个多标签分类任务，我们设计了一个多标签独热编码（Mutil-Hot）[0,0,0,0,0,0],当评论文本涉及到哪一个群体的仇恨则将对应的位置置为1，将分类标签从字符标签映射为数字标签，从而提升模型的分类性能。分类任务DCA-SFT指令微调模板的设计原则如下，详情如附录B所示：

（1）基础指令：任务角色、具体任务、任务要求、基础群体标签的定义；

（2）DCA线索：匹配到的关键词及关键词的群体类别；

（3）输出：仇恨分类对应的Mutil-Hot标签。

2.2.2 抽取任务DCA-RL强化微调模板设计

针对抽取任务，根据四元组定义，首先我们构造了一个one-shot反向蒸馏生成四元组CoT过程的蒸馏指令distill-CoT-instruction，通过该指令我们使用Qwen-Plus通过蒸馏技术萃取出模拟通过输入评论文本生成最终四元组数据的这一结果的中间CoT过程，该模板如附录C所示。然后使用蒸馏得到的数据集、关键信息库、四元组中的仇恨标签构造DCA-RL训练集，具体来说对于蒸馏得到的数据集我们只取CoT中生成论点和评论对象对应过程的CoT信息，然后将四元组中群体标签和是否仇恨、以及匹配关键信息库中匹配到的关键词，作为DCA线索构造四元组抽取训练集。抽取任务DCA-RL强化微调模板的设计原则如下，详情如附录D所示：

（1）基础指令：任务角色、具体任务、仇恨四元组定义、抽取四元组的CoT步骤；

（2）DCA线索：匹配到的关键词、仇恨标签以及群体类别；

（3）输出：抽取四元组的CoT步骤以及最终抽取四元组的结果。

2.3 仇恨分类任务DCA-SFT

在传统生成式微调的基础上我们引入了判别式分类的视角，具体来说我们在大模型的输出

层之后加入线性层，用于将生成任务转化为判别任务，旨在增强模型对仇恨言论中复杂语义的理解和判断能力。

（1）标签计算：模型的输出是未经归一化的分数（logits）。为了将这些分数转换成概率并映射到标签对应的Mutil-Hot位置，我们使用Sigmoid 函数将每个位置对应的logit实数值映射到(0, 1) 范围内，用于表示相应类别的概率，Sigmoid定义如公式（1）所示：

$$\sigma\left(z_i\right)=\frac{1}{1+e^{-z_i}} \tag{1}$$

（2）损失计算：针对本次多标签分类任务我们采用BCEWithLogitsLoss来衡量模型的损失。损失定义如公式（2）所示，其中求和符号$\sum$表示对所有标签的损失进行累加，系数$\frac{1}{n}$表示对所有标签概率取平均值。

$$L=\frac{1}{n}\sum_{i=1}^{n}-\left[y_i\cdot\log\left(\sigma\left(z_i\right)\right)+\left(1-y_i\right)\cdot\log\left(1-\sigma\left(z_i\right)\right)\right] \tag{2}$$

2.4 仇恨抽取任务DCA-RL

我们在抽取任务中引入了基于规则的奖励机制,通过定义明确的奖励函数，引导大模型在生成评论对象和论点时更多地关注与仇恨言论相关的特征。并且，我们注意到预测的结果可能存在包含关系，例如真实的评论对象、论点为<这女人，能活到现在也是个奇迹！>，如果模型预测为<女人，能活到现在>，则论点和评论对象的真实值和预测之间存在重叠关系。因此，我们采用平滑奖励机制，将存在重叠关系的预测结果给予一定的奖励，而不是直接为0，具体来说，我们设计了一个基于规则的综合奖励函数如公式（3）所示，包括格式奖励和<评论对象，论点>二原组重叠比例奖励：

$$R=0.2R_{format}+0.3R_{target}+0.5R_{argument} \tag{3}$$

（1）奖励定义：其中，R_{format}用于确保模型输出符合预定义的CoT格式要求，R_{target}用于衡量模型预测评论对象与真实标签的重叠比例，$R_{argument}$则用于衡量模型预测论点与真实标签的重叠比例。对于格式的奖励我们设定较小的约束值0.2，对最终实际四元组影响较大的R_{target}和$R_{argument}$我们按照实际字符长度给与不同的约束，对较短的目标群体设置了0.3的约束，而较长的论点则设了0.5的约束,3个规则奖励定义如公式（4）、（5）、（6）所示：

$$R_{format}=\begin{cases}1,\text{ 如果输出包含完整的}CoT\text{格式}\\0,\text{ 否则}\end{cases} \tag{4}$$

$$R_{target}=\begin{cases}\rho,\text{若存在重叠，}\rho\text{表示预测结果与真实评论对象中重叠的比例}\\0,\text{ 否则}\end{cases} \tag{5}$$

$$R_{argument}=\begin{cases}\rho,\text{若存在重叠，}\rho\text{表示预测结果与真实论点中重叠的比例}\\0,\text{ 否则}\end{cases} \tag{6}$$

（2）损失计算：在训练过程中，我们使用近端梯度裁剪策略优化（Proximal Policy Optimization Clip, PPO-Clip）并更新模型参数，使得模型在生成仇恨四元组结果时能够最大化预期奖励。其损失定义如公式（7）所示，其中公式（8）$r_t(\theta)$是新旧策略比率，$\widehat{A}_t$是优势估计，ϵ是一个小常数，用于将策略比例限定在$1-\epsilon$和$1+\epsilon$区间，确保更新不会过度偏离策略。

$$L^{CLIP}(\theta)=E_t\left[\min\left(r_t(\theta)\widehat{A}_t,clip\left(r_t(\theta),1-\epsilon,1+\epsilon\right)\widehat{A}_t\right)\right] \tag{7}$$

$$r_t(\theta)=\frac{\pi_\theta\left(a_t\mid s_t\right)}{\pi_{\theta_{old}}\left(a_t\mid s_t\right)} \tag{8}$$

2.5 渐进式优化微调过程

渐进式优化策略贯穿于任务拆解与子任务模型的迭代优化过程，通过数据增强与模型调优的闭环反馈实现性能提升。在分类任务模型中我们构建了一个stage1冷启动+stage2增强训练的多阶段渐进式优化步骤，首先stage1基于测评任务原始训练集通过冷启动训练获得初始分类模型。stage2利用该模型对补充数据集ToxiCN数据集进行伪标签预测，通过一致性检验与MC dropout 采样技术过滤低置信度样本，形成高质量伪标签数据。将筛选后的伪标签数据与原始数据集合并，对分类模型进行增强迭代微调。通过多轮"模型预测-伪标签筛选-数据融合-模型优化"的闭环循环，实现伪标签质量与分类模型精度的协同提升，完成分类任务模型的渐进式优化。针对抽取任务，以训练集原始文本及标注的仇恨四元组为输入，并融合DCA线索与蒸馏得到的CoT 过程数据并采用强化学习算法对模型进行针对性训练，最终实现抽取任务模型的优化，这一步我们称为stage3。

3 实验

3.1 实验设置

数据详情：本次评测分为初赛和复赛2个阶段，提供的训练集包括训练集共4000条数据，初赛和复赛测试集test1、test2各2000条数据，外部数据集ToxiCN共12011条数据，使用Qwen-plus蒸馏的CoT数据集共3992条数据，关键信息库包含了关键词共计537个。测试集test1和test2用于测试模型的泛化性能以及最终的能力。

基线设置：本次评测我们选取Qwen2.5-7B作为基准模型，蒸馏模型选择了闭源超大模型Qwen-plus，此外，为了进行实验对比，本文选取了基于中文STATE ToxiCN的basic prompt SFT作为基线，basic prompt如附录E所示。

参数设置：本文三个阶段均是在4张英伟达H800 GPU上进行训练，训练框架使用Llama-factory，微调方式使用全参数微调、同时微调策略选用ds zero3、每张卡bach_size设置为8，梯度累积设置为2，使用余弦学习率衰减方式来调整学习率，同时我们将wrmup_ratio设置为epoch的1%;对于抽取模型DCA-RL，我们通过完全分片数据并行FSDP加速训练，具体差异化参数如附录A所示：

评估指标：本次测评任务中由于<群体,是否仇恨>二元组即仇恨分类的准确率会直接影响到硬匹配分数（Hard F1）和软匹配分数（Soft F1）的值，当仇恨分类准确率越高，最终的Hard F1和Soft F1值也会越高。因此针对仇恨分类任务，我们使用ToxiCN中的仇恨标签作为标准答案并使用分类准确率（acc）进行评估，针对仇恨抽取任务我们使用Soft F1和Hard F1进行评估。

3.2 实验结果

3.2.1 分类任务stage1-2 实验结果

表1对比了仅stage1微调与stage1冷启动+stage2增强训练两阶段训练在传统SFT和DCA-SFT方法下的acc表现。实验结果表明：

1）两阶段训练策略显著优于单阶段微调，说明多阶段优化能有效提升模型的泛化能力；

2）在相同实验条件下，本文提出的DCA-SFT方法明显优于传统SFT方法，最高准确率达到0.93，较stage1常规SFT提升0.14，证明了DCA线索注入可增强对模型领域知识学习能力；

3）采用Mutil-Hot数字标签映射的方法相比直接使用群体字符标签具有显著优势，平均准确率提升2个百分点，证明数字标签映射能有效增强模型对分类边界的认知能力；

4）对于分类任务，判别式策略能有效模型性能上限，平均acc提高3个百分点。

3.2.2 抽取任务stage3 实验结果

表2展示了在分类任务DCA-SFT最优微调方案stage1-2实验设定下，stage3 DCA-RL采用不同微调数据构造策略下抽取任务的Hard F1和Soft F1性能指标，并与STATE ToxiCN仅采用basic prompt SFT的微调效果进行了对比分析。实验结果表明：

1）蒸馏带来广泛的知识注入：通过蒸馏带有外部大模型先进知识与推理能力的四元组CoT过程，可显著提升模型的领域知识学习效果。相较于无CoT过程的SFT微调，带有蒸馏CoT过程的微调方案在两个测试集上均展现出明显的性能提升，其中,test1数据集上提升最为明显，Soft F1提升幅度（Δ=0.0515）,Hard F1提升幅度（Δ=0.0635）

第二十四届中国计算语言学大会论文集，第419页-第429页，济南，中国，2025年8月11日至14日。

步骤	微调方法	标签构造方式	任务	分类准确率（acc）	
				test1	test2
stage1直接微调	SFT	字符标签	生成式分类	0.79	0.73
	DCA-SFT	字符标签	生成式分类	0.84	0.77
		Mutil-Hot标签	判别式分类	0.87	0.79
stage1冷启动+stage2增强训练	SFT	字符标签	生成式分类	0.88	0.76
	DCA-SFT	字符标签	生成式分类	0.90	0.81
		Mutil-Hot标签	判别式分类	0.93	0.83

表 1: 分类任务stage1-2准确率acc对比

2）强化学习优化效果：强化微调主要优化论点和抽取评论对象抽取性能，最终模型在两个测试集的Soft F1平均值较传统SFT方法均提升了3个百分点，验证了强化微调的有效性。

3）多阶段渐进式优化方案优势：与basic prompt SFT相比，本文所提出的stage1-3渐进式优化方案虽在初始无CoT阶段（数据量4k）时，仅在test2数据集的Hard F1指标上略低于basic prompt SFT方法（数据量优势差异），但通过引入蒸馏CoT过程和DCA-RL优化方案后，最终实现了性能的全面超越，Hard F1和Soft F1分别提升0.0192和0.0211。这一结果充分证明了多阶段渐进式优化策略在仇恨言论四元组抽取任务中的有效性。

步骤	CoT	数据	微调方式	Hard f1&Soft f1值			
				test1		test2	
				Hard f1	Soft f1	Hard f1	Soft f1
stage3	无	4k	SFT	0.2093	0.4117	0.2286	0.4417
			DCA-RL	0.2298	0.44	0.2393	0.4587
	有		SFT	0.2448	0.4752	0.244	0.4677
			DCA-RL	0.2675	0.5053	0.2468	0.4715
basic prompt SFT	无	6k	SFT	/	/	0.2276	0.4504

表 2: 抽取任务stage3不同微调方法Hard F1 和Soft F1 值对比表

3.2.3 四元组抽取榜单对比

表3展现了本研究提出的方法的稳定性和泛化能力，具体而言，在初赛阶段，本方法以0.3864的F1-score排名第二，与第一名仅相差0.0049；在复赛阶段，本方法以0.3591的F1-score位列第三，与第一名差距仅为0.005。值得注意的是，在top6队伍的横向比较中，本方法是唯一在初赛和复赛均保持前三的方案。这一结果充分证明了本方法在该测评任务中具有显著的泛化性能和稳定性优势。

初赛test1		初赛test2	
top6队伍	F1-score	top6队伍	F1-score
Nova-Z	0.3913	过来水个比赛队	0.3641
星辰之力	**0.3864**	_KING_	0.3636
xaxaxa	0.3735	**星辰之力**	**0.3591**
珠科智能2班	0.3708	珠科智能2班	0.3566
我是最棒的	0.3693	BMI	0.3555
zutNLP	0.3693	zutNLP	0.3545

表 3: 比赛成绩对比表

4 结论

本文基于STATE ToxiCN数据集，探索了LLM在细粒度中文仇恨识别任务上的能力，针对仇恨四元组抽取的复杂挑战，我们提出了一种基于动态线索增强提示和多阶段渐进式优化的LLM微调方案，通过将四元组抽取任务解耦为仇恨倾向分类与仇恨信息抽取两部分。对于分类任务，我们在微调指令中添加DCA线索，并优化了模型架构，通过在生成式LLM输出层后添加线性分类层的方式将该任务从生成式转化为判别式DCA-SFT；对于抽取任务，借助

知识蒸馏技术，利用Qwen-Plus补全四元组抽取过程的完整CoT，并引入基于规则的强化微调方案DCA-RL，优化模型抽取仇恨信息的效果。同时，我们设计了一个多阶段渐进式优化微调方案，通过stage1冷启动+stage2增强训练优化分类模型DCA-SFT的性能，再通过stage3提升DCA-RL抽取模型的能力。最后通过对比实验详细的验证了我们提出的DCA技术和多阶段渐进式微调方案的有效性，该方案是唯一一个在CCL25-Eval任务10初赛和复赛阶段均稳定保持在Top3之列的，展现出较强的泛化性能和领域知识学习能力。当然，当前方案仍具优化空间，鉴于对中文仇恨俚语的研究不足等问题，后续可通过在DCA线索中添加中文仇恨俚语的解释进而形成更强大的DCA线索注入来提升LLM对仇恨言论中俚语信息的理解能力。

参考文献

Badjatiya P, Gupta S, Gupta M, Varma V. 2017. *Deep learning for hate speech detection in tweets.* In Proceedings of the 26th International Conference on World Wide Web Companion, WWW '17 Companion.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021. *Learning from the worst: Dynamically generated datasets to improve online hate detection.* In ACL.

Del Vigna F, Cimino A, Dell'Orletta F, Petrocchi M, Tesconi M. 2017. *Hate me, hate me not: Hate speech detection on facebook.* ITASEC.

Do HTT, Huynh HD, Nguyen KV, Nguyen NLT, Nguyen AGT. 2019. *Hate speech detection on vietnamese social media text using the bidirectional-lstm model.* arXiv:1911.03648.

György Kovács, Pedro Alonso, and Rajkumar Saini. 2021. *Challenges of Hate Speech Detection in Social Media: Data Scarcity, and Leveraging External Resources.* SN Comput. Sci. 2, 2 (Apr 2021).

Gambäck B, Sikdar UK. 2017. *Using convolutional neural networks to classify hate-speech.* In Proceedings of the First Workshop on Abusive Language Online. Association for Computational Linguistics, Vancouver, BC, Canada.

Junyu Lu, Bo Xu, Xiaokun Zhang, Changrong Min, Liang Yang, and Hongfei Lin. 2023. *Facilitating fine-grained detection of Chinese toxic language: Hierarchical taxonomy, resources, and benchmarks.* In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, Toronto, Canada. Association for Computational Linguistics.

Marzieh Mozafari, Reza Farahbakhsh, Noel Crespi. 2019. *Hate Speech Detection and Racial Bias Mitigation in Social Media based on BERT model.* PLOS ONE.

Park J, Fung P. 2017. *One-step and two-step classification for abusive language detection on twitter.* In ALW1 1st Workshop on Abusive Language Online.

Wang B, Ding Y, Liu S, Zhou X. 2019. *YNU_Wb at HASOC 2019: Ordered Neurons LSTM with Attention for Identifying Hate Speech and Offensive Language.* Working Notes of FIRE 2019 - Forum for Information Retrieval Evaluation, Kolkata, India.

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, Harbin, China. Chinese Information Processing Society of China.

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.

Zampieri M, Malmasi S, Nakov P, Rosenthal S, Farra N, Kumar R. 2019. *Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval).* In: Proceedings of the 13th International Workshop on Semantic Evaluation.

Zimmerman S, Kruschwitz U, Fox C. 2018. *Improving hate speech detection with deep learning ensembles.* In Proceedings of LREC. European Language Resources Association (ELRA), Miyazaki, Japan.

A stage1-3最佳参数设置

参数	分类模型(DCA-SFT)		抽取模型(DCA-RL)
	stage1	stage2	stage3
sft_learning_rate	1e-05	5e-06	/
sft_epochs	3	2	/
sft_wrmup_ratio	0.03	0.02	/
cutoff_len	4096	4096	2048
ppo_actor_lr	/	/	3e-05
ppo_epchs	/	/	3
ppo_mini_batch_size	/	/	16
ppo_micro_batch_size_per_gpu	/	/	2
ppo_group_num	/	/	4
ppo_adv_estimater	/	/	ppo

B DCA-SFT instruction

DCA-SFT instruction

你是一个互联网社交平台仇恨信息识别领域的专家，现需要对社交平台的用户评论文本进行针对种族、性别、地域、LGBTQ、others等特定群体或个体表达仇恨的不良言论检测，具体任务是根据下面的“要求”、“定义”、“标签定义”和“已知信息”对“输入评论文本”进行分类。
要求1.请仔细理解判断仇恨的定义以及目标群体的定义；
要求2.请仔细理解“输入评论文本”中的每一个字、词及表情符号，很多关键信息被使用中文字符的多音字替换、同音字词替换、中文偏旁部首组合、中文拼音、中英文缩写等隐晦的表达形式进行表达，如默表示黑犬、飞周表示非洲、txl表示同性恋等，请综合这些信息对“输入评论文本”的语义信息进行理解；
定义:
1. 目标群体的定义：该输入评论文本是否构成了对某些群体的仇恨，包括暴力、歧视、嘲讽、攻击、侮辱、煽动伤害、或贬低等，包括“地域（Region）”、“种族（Racism）”、“性别（Sexism）”、“LGBTQ”、“其他（others）”共5个群体；
1.1 Racism：指的是基于个体所属的种族或民族背景而产生的偏见和不公平对待。在仇恨识别中，这可能表现为使用贬低性的语言、刻板印象或者直接攻击某个种族或民族，传播与该群体相关的负面信息或谣言。种族歧视主要针对的是个体的种族背景，包括他们的外貌、文化和历史等，种族歧视的基础往往是生物学上的差异，以及由此衍生出的文化和社会差异；
1.2 Sexism：涉及基于性别的不平等和偏见，通常表现为对某一性别的成员持有固定看法，并给予不公平待遇。例如，在工作场合中的性别歧视可能包括对女性能力的低估或是男性不适合某些“女性化”工作的观念。在仇恨言论中，可能会出现贬低某个性别价值的话语；
1.3 Region：是根据人们来自哪个地区（国家、省份、城市等）而进行区别对待的一种形式。它通常包含了对于特定地区的居民持有的负面刻板印象，如认为他们具有某种不良品质或行为模式。地域歧视可以在不同层面上发生，从轻微的玩笑到严重的社会排斥，地域歧视则更多关注的是地理位置，地域歧视则更多的是基于经济状况、教育水平、方言等与地理区域相关的因素；
1.4 LGBTQ：指对非异性恋者及跨性别者的偏见和歧视，涵盖同性恋、双性恋、跨性别者以及其他性别认同和性取向多样化的人群。LGBTQ歧视可以体现在拒绝承认他们的身份、权利，以及通过言语或行动对其进行侮辱、威胁或暴力行为。这种歧视往往根植于对传统性别角色和社会规范的严格遵循。
1.5 others：Region、Racism、Sexism、LGBTQ以外但属于易导致仇恨歧视的特殊群体，如艾滋梅毒、各种传染病、生理特征、身体缺陷、师生群体等；
已知信息:
DCA线索
输入评论文本:
input

C 反向蒸馏one-shot提示模板distill-CoT-instruction

反向蒸馏one-shot提示模板distill-CoT-instruction

你是一个互联网社交平台仇恨信息识别领域的专家，你擅长对社交平台的用户评论文本进行针对种族、性别、地域、LGBTQ、others等特定群体或个体表达仇恨的不良言论检测，并输出多个以[SEP]分隔最后一个面加[END]的四元组。以下“输出四元组”是你根据“任务定义”和“任务步骤”得到的输出但是缺乏具体的判断依据。现需要根据“输入评论文本”、“输出四元组”以及“四元组定义”补全任务步骤中每一步的具体推理过程，对于步骤1你需要给出你识别评论对象,论点 二元组的依据并解释每个二元组，对于步骤2你需要解释标签为什么是仇恨或者不仇恨，对于步骤3你需要给出该判断评论文本涉及目标群体的依据并结合步骤1-2给出解释。
要求1：每一步的判断依据必须来自于“输出四元组”的结果，请严格忠实于“输出四元组”结果严禁编造。
要求2：只输出中间分析结果不要输出“输出四元组””的任何内容或者格式，该中间结果是假设不知道“输出四元组”结果的一个反向推理过程，只是该中间分析结果可以正确的解释“输出四元组”，最后请严格按照“分析结果输出示例”输出标准json格式”步骤1”:”依据和解释”,”步骤2”:”依据和解释”,”步骤3”:”依据和解释”的输出中间分析过程：
四元组定义：
1.评论对象（Target）：文本中可能的评论对象，如一个人或一个群体。当实例无具体目标时设为NULL
2.论点（Argument）：包含对评论对象的关键论点的信息片段，该片段需尽量简洁无需无关的额外信息。
3.是否仇恨（Hateful）：评论对象-论点二元组是否构成了对某些群体的仇恨言论，只有2个类别，形成仇恨（hate）以及不形成仇恨（non-hate）；
4.目标群体（Targeted Group）：包含仇恨信息的评论对象-论点二元组涉及的目标群体。目标群体包括“地域（Region）”、“种族（Racism）”、“性别（Sexism）”、“LGBTQ”、“其他（others）”共5类，其中只有不包含仇恨的群体用non-hate，others表示Region、Racism、Sexism、LGBTQ以外但属于仇恨的特殊群体；
任务步骤：
步骤1：首先你需要准确的对输入文本进行断句，识别句子的主语、谓语、宾语、定语、状语、补语等内容，根据断句信息找到评论对象和论点的二元组数据。断句的目标是找到评论对象和支持该仇恨言论的论点，并且保证评论对象和论点是强关联的，注意这里的评论对象可能是一个英文缩写，可能是中文字符的多音字替换、中文偏旁部首组合、中文拼音等隐晦的表达形式如“默”表示“黑犬”，你需要将评论对象和论点精炼且准确无误的提取出来，不包含如何无用的信息，如停用词等
步骤2：接着你需要准确的判断输入文本的语义信息，并对输入文本进行仇恨分类，判断每个二元组数据是否构成仇恨，具体任务是对输入文本进行二分类，标签是构成仇恨hate和不构成仇恨non-hate;
步骤3：根据步骤2得到的是否仇恨标签分为两种情况，如果标签是non-hate,那最终评论对象所属目标群体的标签是也是non-hate。如果如果标签是hate，接着需要对二元组中评论对象所属的目标群体进行分类，注意这里的评论对象可能属于一个或多个目标群体，目标群体包括“地域（Region）”、“种族（Racism）”、“性别（Sexism）”、“LGBTQ”、“其他（others）”共5类。但在判断群体前，考虑到步骤2中存在多个二元组数据，这些二元组数据可能有一小部分不构成仇恨，因此需要判断每一个二元组是否构成仇恨，标签也是hate和non-hate，主要分为情况一：一个二元组和情况二：多个二元组等2种情况；
输入评论文本：
input
原始输出四元组：
ori_output
分析结果输出示例：
{”步骤1”:”断句结果与分析：
输入文本为：“没爹的黑孩到处扔”。
通过断句，可以明确：
- **评论对象（Target）**：“没爹的黑孩”
- **论点（Argument）**：“到处扔”
解释每个二元组
- **评论对象**：“没爹的黑孩”：这是一个隐晦的表达形式，其中“黑孩”可能是指具有黑色皮肤的孩子，而“没爹”则暗示某种社会地位或家庭背景。
- **论点**：“到处扔”：这句话表达了对“没爹的黑孩”的负面评价，暗示他们像垃圾一样被随意丢弃。”,”步骤2”:”判断是否构成仇恨言论
语义分析：
- **评论对象**：“没爹的黑孩”涉及种族和家庭背景，尤其是“黑孩”明显指向肤色问题。
- **论点**：“到处扔”是一种非常负面的描述，带有侮辱性和歧视性，暗示这些孩子不值得尊重，甚至被视为垃圾。
基于上述分析，这句话明显包含对特定群体的负面刻板印象和歧视，因此构成仇恨言论。”,”步骤3”:”确定目标群体：
- **种族（Racism）**：评论对象中提到“黑孩”，明显涉及种族歧视。
- **其他因素**：虽然“没爹”涉及家庭背景，但主要的歧视点还是在于肤色，因此种族是主要的仇恨目标。”}
最终输出结果：

D DCA-RL instruction

DCA-RL instruction

你是一个互联网社交平台仇恨信息识别领域的专家，现需要对社交平台的用户评论文本进行针对种族、宗教、性别、地域LGBTQ、others等特定群体或个体表达仇恨的不良言论检测，具体任务是基于“四元组定义”、“任务步骤”和“已知信息”输出“四元组”，该“四元组”中包含评论对象（Target）、论点（Argument）、目标群体（Targeted_Group）、是否仇恨（Hateful），每一个“四元组”输出格式为评论对象—对象观点—是否仇恨—仇恨群体。评论对象可以为’NULL’,对象和观点尽量简洁，多个四元组之间用[SEP]分隔,最后一个四元组后面加[END]，对于非仇恨文本和不包含特定群体的一般攻击性言论，同样需要对评论对象和论点进行提取，并将是否仇恨标签和涉及群体设为non-hate。由于样本中可能有多个评论对象，因此可以包含多个“四元组”。以下是“四元组定义”、“任务步骤”以及具体的“输入社交评论文本”。
要求1.论点和评论对象必须精准和简洁;
要求2.生成的论点和评论对象必须来自于原文，否则会收到严厉的惩罚;
四元组定义:
1.评论对象（Target）：文本中可能的评论对象，如一个人或一个群体。当实例无具体目标时设为NULL;
2.论点（Argument）：包含对评论对象的关键论点的信息片段，该片段需尽量简洁无需无关的额外信息。
3.是否仇恨（Hateful）：评论对象-论点二元组是否构成了对某些群体的仇恨言论，只有2个类别，形成仇恨（hate）以及不形成仇恨（non-hate）;
4.目标群体（Targeted Group）：包含仇恨信息的评论对象-论点二元组涉及的目标群体。目标群体包括“地域（Region）”、“种族（Racism）”、“性别（Sexism）”、“LGBTQ”、“其他（others）”共5类，其中只有不包含仇恨的群体用non-hate，others表示Region、Racism、Sexism、LGBTQ以外但属于仇恨的特殊群体;
已知信息:
DCA线索
任务步骤:
步骤1：首先需要对输入文本进行断句，识别句子的主语、谓语、宾语、定语、状语、补语等内容，根据断句信息找到评论对象和论点的二元组数据。断句的目标是找到评论对象和支持该仇恨言论的论点，并且保证评论对象和论点是强关联的，注意这里的评论对象可能是一个英文缩写，可能是中文字符的多音字替换、同音字词替换、中文偏旁部首组合、中文拼音、中英文缩写等隐晦的表达形式如“默”表示“黑犬”，飞周表示非洲，你需要将评论对象和论点精炼且准确无误的提取出来，不包含如何无用的信息，如停用词等;
步骤2：根据已知信息和步骤1，得出最终的四元组;
输入评论文本:
input

E basic prompt SFT instruction

basic prompt SFT instruction

你是一个内容审查专家，请你分析我的句子并且从中提取出一个或者多个四元组。请从下面的文本抽取一个或多个四元组，每一个四元组输出格式为评论对象—对象观点—是否仇恨—仇恨群体。评论对象可以为’NULL’,对象观点尽量简洁,仇恨群体只包括(LGBTQ、Region、Sexism、Racism、others、non-hate)，同一四元组可能涉及多个仇恨群体，是否仇恨标签为(hate、non-hate),多个四元组之间用[SEP]分隔,最后一个四元组后面加[END]。提取出句子中包含的所有四元组:
我的句子:
input

第二十四届中国计算语言学大会论文集，第419页-第429页，济南，中国，2025年8月11日至14日。