面向中文细粒度仇恨识别的两阶段多轮问答框架

李文煜，张雪，陈钰枫
（北京交通大学计算机科学与技术学院，北京 100044）

摘要：仇恨言论的精准识别是网络内容安全治理的重要环节，对维护社会舆论环境与公共秩序具有现实意义。然而，现有中文细粒度仇恨识别方法在片段级要素抽取能力不足，难以准确定位文本中的攻击对象、论点与仇恨群体，对隐晦表达与复杂语义结构的处理亦存在明显局限。为解决上述问题，该文提出面向中文细粒度仇恨识别的两阶段多轮问答框架，构建多轮提示结构并引入自检索增强机制，通过多轮累积投票提升输出稳定性与一致性。实验在 STATE-ToxiCN 数据集上表明，该方法所有细粒度元素匹配得分超过当前 CCL2025 评测基线与领先方案，在四元组层面软硬匹配 F1 平均值达 38.85。结果验证了多轮提示与检索增强在提高模型鲁棒性和复杂语义识别方面的有效性，为中文细粒度仇恨识别及多层次信息抽取提供可行方案。

关键词：大语言模型；提示工程；仇恨言论识别
中图分类号：TP391

A Two-Stage Multi-Round Question-Answering Framework for Fine-Grained Chinese Hate Speech Recognition

Li Wenyu, Zhang Xue, Chen Yufeng
(School of Computer Science & Technology, Beijing Jiaotong University, Beijing 100044)

Abstract: Accurate identification of hate speech is a crucial component of online content safety governance and is of practical significance for maintaining a healthy public opinion environment and social order. However, existing Chinese fine-grained hate speech identification methods remain weak in segment-level element extraction: they struggle to precisely locate attack targets, arguments, and hateful groups in text, and show clear limitations in handling implicit expressions and complex semantic structures. To address these issues, this paper proposes a two-stage multi-round question–answering framework for fine-grained Chinese hate speech recognition. The framework designs multi-round prompting and incorporates a self-retrieval augmented mechanism, while using multi-round accumulated voting to improve the stability and consistency of model outputs. Experiments on the STATE-ToxiCN dataset show that the proposed method outperforms the current CCL2025 evaluation baseline and leading approaches across all fine-grained element matching scores, achieving an average soft/hard matching F1 of 38.85 at the quadruple level. These results demonstrate the effectiveness of multi-round prompting and retrieval augmentation in enhancing model robustness and recognizing complex semantics, providing a feasible solution for fine-grained Chinese hate speech identification and multi-level information extraction.

Key words: large language model; prompt engineering; hate speech recognition

作者简介：李文煜（2001-),女，硕士研究生，主要研究方向：自然语言处理和仇恨识别
通信联系人：陈钰枫（1981-），女，研究员、博导，主要研究领域为自然语言处理和机器翻译. E-mail: chenyf@bjtu.edu.cn

0 引言

随着社交媒体的普及，用户生成内容呈爆炸式增长，仇恨言论也随之泛滥。仇恨言论通常指针对特定群体或个人、包含仇恨或煽动伤害的有害表述。这类内容对个体和社会造成严重危害，破坏社会稳定性，因此，对仇恨言论的有效识别与治理已成为亟待解决的重要问题[1]。近年来，研究者们积极投身于仇恨言论检测研究，且该领域的研究重心已逐渐从帖文级[2,3]转向片段级[4-6]，致力于实现更细粒度的仇恨言论检测。尽管该问题在国际学界受到广泛关注，然而相关研究长期以英文语料为主，非英文语种的系统性研究相对滞后。现有的中文仇恨检测工作大多停留在帖文级粗粒度分类[7,8]，即仅对句子整体进行仇恨检测，然而，由于仇恨表达的强度和指向往往取决于所针对的对象和论据[9]，目前这种粗粒度的检测难以深入理解仇恨言论内涵，因而需要开展更细粒度的识别，包括识别出句中的攻击目标、针对该目标的贬损内容，以及仇恨所指向的群体等信息。

细粒度中文仇恨检测在实践中通常面临语言学与工程实现的双重挑战。一方面，中文缺乏显式词界定符且句法灵活，倒装、省略等现象使攻击目标与贬损论据的定位更加困难；另一方面，中文网络语境中广泛存在隐晦表达，如谐音替换、拆字、合字与典故暗喻等“避审”现象，这类表达会显著干扰基于词表或固定模板的方法，并对模型鲁棒性提出更高要求[10]；同时，面向中文的细粒度仇恨要素抽取任务也普遍受到该类隐式表达与上下文依赖的影响[11]。已有研究虽构建了部分词典与数据集[8,12]，并在预训练模型上取得进展，但可解释的精细标注与对隐含表达的系统建模仍显不足，限制了模型对仇恨语义的深入理解与迁移。

近年来，大型预训练生成模型（Large Language Model, LLM）的兴起为上述问题提供了新的解决思路，LLM 以出色的通用能力可以快速适配到不同的任务中。具体来说，以上下文学习为代表的提示学习范式（Prompt Learning），使 LLM 在零样本或少样本场景下展现出强大的潜力[13]；而对模型进行全参数微调（Fine-tuning）或参数高效微调（Parameter-Efficient Fine-Tuning）则能使模型更深入地拟合特定任务的数据分布，从而获得更优的性能。与传统方法相比，这两类范式在快速迁移、跨域应用与性能提升上各具优势。此外需要注意的是，LLM 对不同提示呈现出高度敏感性，同时不同微调策略也可能导致性能出现显著差异，因此本文聚焦于结合提示学习与模型微调解决中文细粒度仇恨识别任务中存在的挑战。

本文提出面向中文细粒度仇恨识别的两阶段多轮问答框架。该框架以“两阶段提示推理”为主线：在阶段一先定位评论对象与论点要素，在阶段二基于候选要素完成目标群体判定；为提升两阶段推理的稳定性与可解释性，框架在提示构造环节引入自检索增强实现动态检索的上下文学习（In-Context Learning, ICL），并在推理环节通过多轮累积投票进行一致性聚合，从而增强模型对复杂句式与隐含仇恨表达的解析能力。本文在 CCL2025 评测提供的 STATE-ToxiCN 数据集上开展实证研究：该数据集包含 8029 条中文帖子与 9533 个“评论对象—论点—仇恨群体—是否仇恨”四元组标注，覆盖性别歧视、种族主义、地域偏见与

反同性恋等多类仇恨类型。实验结果表明，所提出的两阶段多轮问答框架结合适度的全参数微调，能够显著提升细粒度仇恨要素抽取性能，在多个指标上达到或超过 CCL2025 对应评测任务的最优水平，并在低资源场景下保持良好的适应性与可行性。

本文的主要贡献包括：

1. 提出一个面向中文细粒度仇恨识别的两阶段多轮问答框架：将整体识别流程重构为“要素定位”和“群体判定”两阶段，每个阶段中引入自检索增强构造动态上下文提示，通过多轮累积投票进行一致性聚合，从而提升仇恨要素抽取的可解释性与预测稳定性。

2.在细粒度仇恨识别数据集上，本文方法在关键性指标软硬 F1 上平均值为 38.85，其表现优于该评测任务已公开的最佳系统，证明了本方法在中文细粒度仇恨识别任务上的有效性和先进性。

3.探究了多种不同的提示词和推理输出设计策略，相关的经验性结果可以为其它中文自然语言理解任务提供参考。

1 相关工作

1.1 仇恨言论数据集

仇恨言论检测是自然语言处理（Natural Language Processing）领域的重要研究方向之一，近年来受到广泛关注。随着预训练语言模型的兴起，研究者已逐步将其应用于仇恨言论检测任务，以提升模型的泛化与迁移能力。Caselli 等人[14]提出在 BERT 基础上重训练得到 HateBERT，用于英文辱骂性语言检测；Hanu 等人[15]提出 Detoxify 模型，结合 RoBERTa 与多语言预训练，提升了有害评论分类的鲁棒性；Li 等人[13]提出 COVID-HateBERT，在 COVID-19 相关语料上强化仿写能力，实现 F1 值显著提升。

为推动该领域发展，学界已构建多个仇恨言论检测专用数据集[16-19]。Pavlopoulos 等人[4]首次提出片段级仇恨言论检测概念；TBO 数据集[6]则进一步推动该领域发展，开创性地实现了“目标-论点-有害性”三元组的提取。然而，中文仇恨言论检测研究仍显著滞后。

现有中文仇恨言论数据集大都局限于帖子级标注。其中，TOCP 与 TOCA B 数据集[20]主要用于检测亵渎性语言与辱骂性内容；SWSR 数据集[8]聚焦性别歧视识别任务；COLD 数据集[7]将句子划分为个人攻击、反偏见等类别；Zhou 等人[21]提出的 CDial-Bias 是首个针对中文对话中社会偏见的标注数据集；Xiao 等人[10]则提出 ToxiCloakCN 数据集，用于评估大语言模型对伪装干扰的鲁棒性。

CCL2025 的任务 10 中评测的数据集 STATE-ToxiCN 是首个面向中文的片段级仇恨言论数据集，涉及性别、种族、地域等类别，侧重于从中文社交媒体文本中抽取四元组“评论对象—论点—仇恨群体—是否仇恨”。由于中文仇恨言论的微妙性、上下文依赖性、四元组元素的相互依赖以及高质量标注数据的可获得性有限，这一任务尤其具有挑战性。Bai 等[11]突出了这些困难，表明即使是最先进的模型，如 GPT-4o，平均得分也只有 15.63，而微调的开源模型如 Qwen2.5-7B 达到 35.365，但仍需进一步优化。

1.2 仇恨言论识别

随着 ChatGPT、Llama 等大语言模型的迅速发展与广泛应用，其在自动问答、文本生成、情感分析等多个领域的应用潜力受到了极大关注。在应用大语言模型进行仇恨言论检测方面，Chiu 等人[22]利用 GPT-3 实现低资源场景的仇恨识别，在少样本设置下准确率最高可达 85%。Wang 等人[23]通过精调 7B 规模大语言模型、少样本提示和多数投票的策略，取得了 CCL2025 评测任务的第二名。

采用提示词的方式能够在无需微调参数的情况下提升大语言模型的任务适应能力，已成为其重要的应用范式之一[24]。提示词的构造过程通常被称为“提示词工程”，其中思维链推理与上下文学习是两种核心设计策略。

思维链方法（Chain-of-Thought, CoT）[25]旨在通过明确的提示指引模型进行多步推理，以提升其处理复杂任务的能力。当前该方法已衍生出多种变体，如自询问（Self-Ask）[26]、树状思维（Tree-of-Thoughts, ToT）[27]、图结构推理（Graph-of-Thoughts, GoT）[28]、回退式提问（Step-back Prompting）[29]，以及结合程序结构的 Program-of-Thoughts[30]等。

上下文学习允许模型通过输入中提供的少量标注样例直接完成任务，无需额外训练。现有样例选取方法主要包括基于输入相似度的检索方法、基于模型输出分布的选择策略[31]，以及构建分类模型进行监督选择[32]。此外，样例排序则通常考虑样例与测试输入的语义相似性[33]或任务难度[34]。

在标注数据极少的情况下，提示学习能够取得较合理的结果，但在大多数真实数据集上仍略逊于利用充足数据微调得到的专用模型。本研究基于以上认识，在中文细粒度仇恨识别任务中深入探索大语言模型提示学习和模型微调等方法，希望进一步推动这一新兴方向的发展。

2 本文方法

2.1 任务定义

本文研究面向 STATE-ToxiCN 数据集的中文细粒度仇恨言论识别任务，即给定一条用户评论文本，模型需要识别该评论中的评论对象（Target）、论点（Argument）、目标群体（Targeted Group）、是否仇恨（Hateful）四个要素，并将其作为一个四元组输出。

评论对象指评论中被攻击或针对的实体，论点是针对该对象所表达的仇恨或负面观点的文本片段，目标群体则指被攻击对象所属的群体类别，是否仇恨表示评论是否带有仇恨倾向，即“non-hate”与“hate”。这一任务属于细粒度的仇恨言论识别，需要模型在片段级别上定位文本中的仇恨要素，而非仅对整条评论做粗粒度的分类判断。

2.2 方法框架

利用大模型进行细粒度中文仇恨四元组抽取的一个最简单直观的方法是利用简单的提示词，询问大模型从而得到答案。然而，简单的提示词在句子较短且实体关系类型较少的数

据集上是可行的，当句子过长以及四元组过多（比如 3）时，大模型抽取出来的四元组会发生严重的遗漏，此外还存在四元组仇恨判断不准确等问题。经过分析，出现这种现象的原因是四元组类型过多导致任务复杂度提升，大模型对单个四元组类型的关注度不够，导致难以兼顾所有类型的四元组抽取。为此，本文设想“是否可以将细粒度中文仇恨识别拆解成更简单的子问题，再分步骤去解决单个子问题”，以此为启发，本文提出了先抽取评论对象与论点，再判断目标群体的两阶段多轮问答抽取框架。

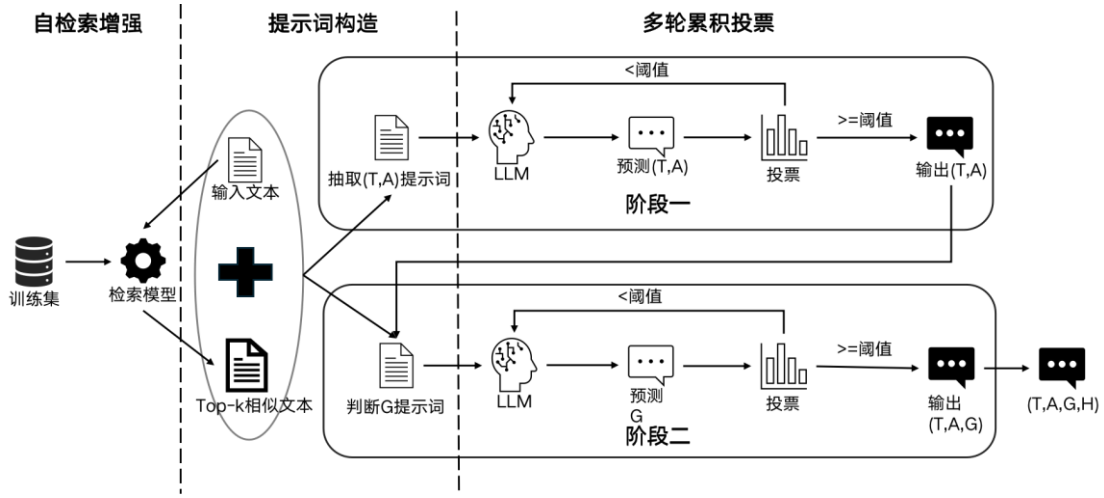

图 1 本文方法框架示意图

此外，本文发现目标群体与是否仇恨标签之间存在很强的相关性，只有当目标群体是非仇恨群体时，是否仇恨被标注为“non-hate”，否则，是否仇恨标注为“hate”。因此，本文将中文细粒度仇恨言论识别任务简化为评论对象、论点、目标群体三元组抽取任务，输出结果时根据目标群体转化为四元组输出。

在总体上，图 1 展示了本文框架在一次输入上的基本处理链路，可概括为三个顺序衔接的部分：先进行自检索增强（见 2.3 节）获取示例，再进行两阶段提示构造（见 2.4 节）与多轮问答推理，最后通过多轮累积投票（见 2.5 节）得到稳定输出。其中，自检索增强模块首先基于输入评论在训练样本库中检索相似标注示例，得到可复用的示例集合；随后框架进入两阶段任务流程，分别在阶段一与阶段二中以不同的格式和模板对该示例集合进行重排与填充，构造对应阶段的提示词并完成多轮推理；最后，多轮累积投票对两阶段推理产生的候选结果进行频次累积与阈值筛选，输出最终的结构化预测。

2.3 自检索增强

为提升少样本示例选取的鲁棒性，缓解语境噪声干扰及高相似度样例依赖导致的过拟合和语义偏置，本文设计并采用了优化后的自检索增强机制（Self-Retrieval Augmented Generation, SRAG）[23,35]，具体流程如下：

1.构建检索库：以训练集中已标注的样本为语料，针对每个样本，使用句向量模型对训练样本编码并存入向量检索库。

2.检索示例：给定输入文本，计算其向量表示，在检索库中计算所有样本与输入文本的

余弦相似度并按照相似度排序，得到候选列表。在依次选取示例的过程中，引入探索概率 ε 判定是否进行探索。判定探索时，将从非 top-k 所构成的探索池中随机选择一个样本作为当前示例，并从探索池中剔除该样本；未判定探索时，则选取下一个的相似度最高文本示例。

3.示例构造：依据任务设定格式化筛选出的 k 个示例，嵌入到多轮提示结构中作为上下文输入。

相较于静态少样本提示，SRAG 方法的核心优势在于可以动态适应每条输入，提升大语言模型对任务分布的适应能力，同时，检索时以一定概率检索出相似性较低的样例作为提示示例，构造个性化的上下文提示，从而提升模型对任务目标的理解能力和抽取精度。

2.4 两阶段提示词构造

本文研究面向 STATE-ToxiCN 数据集的中文细粒度仇恨言论识别任务，即给定一条用户评论文本，模型需要识别该评论中的评论对象（Target）、论点（Argument）、目标群体（Targeted Group）、是否仇恨（Hateful）四个要素，并将其作为一个四元组输出。

本文提出的两阶段框架将细粒度仇恨识别流程分解为“要素定位”和“群体判定”两步，并据此设计了对应的两阶段提示词。

- **阶段一**

阶段一的目标是识别评论中真实出现的评论对象 T 及其对应的论点 A，并输出评论对象和论点二元组集合{(T,A)}。再以结构化格式给出所有(T,A) 对，以便于后续解析与传递。若评论中不包含任何可抽取的对象和论点要素，模型需返回 “NULL” 标记，以保持框架的可执行性与输出对齐。模板如下：

“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(评价对象,论点)两元组，如果实例没有明确的目标，则设置为 NULL，仅需抽取发布者的直接观点，而非转述他人的观点。

示例：

{icl}

句子：

{content}

(评价对象，论点)：”

其中，content 表示输入的文本；icl 表示经过自检索增强挑选出的示例，可选择性加入。

- **阶段二**

阶段二的目标是基于阶段一得到的候选 (T,A) 对，进一步判定其所指向的目标群体 G，输出三元组集合{(T,A,G)}，并在最后映射回四元组结果。与阶段一不同，阶段二通常包含多个问答回合：模型需要对阶段一抽取出的每个 (T,A) 逐一进行群体归因判断。模板如下：

“你是一个内容审查专家，根据给定评论和评论中的(评价对象，论点)进行分析，并判断它们是否表达了针对特定群体的仇恨情绪。如果存在仇恨情绪，请从给定的标签集合["LGBTQ", "Region", "Sexism", "Racism", "others"]中选择一个或多个合适的标签来描

述这种仇恨类型。如果没有检测到任何仇恨情绪，则应为该对文本标记为 ["non-hate"]。

示例：

{icl}

(评价对象，论点)：

{pair}

(评价对象，论点，目标群体)：”

其中，pair 表示经过第一阶段提取出的评价对象和论点；icl 表示经过自检索增强挑选出的示例，可选择性加入。

2.5 多轮累积投票

在生成式大模型的提示推理过程中，单次输出容易受到提示排序、采样随机性与上下文噪声影响而产生波动。为此，本文在两阶段推理后引入多轮累积投票机制（Multi-round Accumulated Voting, MAV），通过“多次独立推断—频次统计—阈值筛选”的方式获得最终结构化结果。MAV 具体步骤包括：

①多样化提示：通过 SRAG，从检索库中为每个输入检索 k 个样本，并将每一个检索到的样本与输入进行拼接，构建 k 个不同的提示。

②多轮推理：对每个提示进行迭代推理，同时在迭代过程中统计所有结果出现的频率，阶段一统计(T,A)二元组频次，阶段二统计(T,A,G)三元组频次。

③投票机制：选择超过频率阈值 τ 的答案输出。

MAV 不需要耗时地重新训练模型，而是根据可用的计算资源动态调整迭代推理次数，通过迭代推理提高了鲁棒性和准确性，使其在资源受限环境下的复杂任务中变得高效。

3 实验

3.1 数据集

为系统评估基于提示学习的细粒度中文仇恨识别方法，本文在 CCL2025 评测任务 10 细粒度中文仇恨言论识别数据集 STATE-ToxiCN[11]上展开实验，数据规模统计见表 1。

表 1 训练集和测试集的统计数据

	帖子	四元组	仇恨四元组	非仇恨性四元组
Train	6424	7631	4842	2789
Test	1605	1902	1221	681
Total	8029	9533	6063	3470

表 2 是来自 STATE-ToxiCN 数据集的带注释帖子示例。对于评论“男同是艾滋高发群体”，模型应当抽取出四元组“男同|艾滋高发群体|LGBTQ, others|hate”，其中“男同”是评论对象，“恶心”是论点，“LGBTQ,”表示该对象属于 LGBTQ 群体，“hate”表示仇恨。若评论不含仇恨倾向，则仇恨立场标记为“non-hate”，所属群体仍需返回“non-hate”以保持输出对齐，但仍需提取相关的评论对象和观点内容，如表 1 中的示例 1。需要注意的是，一条评论

中可能出现多个独立的仇恨目标及观点，此时应当抽取出多个四元组予以表示。

表 2 来自 STATE ToxICN 数据集的带注释帖子示例

示例	帖子	评论对象	论点	目标群体	是否仇恨
1	这花呗干什么的。	花呗	干什么的	non-hate	non-hate
2	雀食男同恶心	男同	恶心	LGBTQ	hate
3	默滚出中国就行了。	默(黑犬)	滚出中国	Racism	hate

3.2 评价指标

由于中文边界存在模糊性，单一的评估指标可能无法准确评估模型在中文片段级仇恨言论检测中的表现。为准确评估模型在片段级细粒度仇恨识别上的效果，本文完全沿用 STATE-ToxiCN 的评测设定：对于评论对象、论点采用了硬匹配和软匹配，而是否仇恨、目标群体则只采用硬匹配，在评论对象（Target）、论点（Argument）、评论对象-论点对（T-A Pair）、评论对象-论点-是否仇恨三元组（T-A-H Tri.）、四元组（Quad.）五个任务上计算 F1 分数作为评价指标，以四元组的软硬匹配 F1 值的平均数作为最终分数。F1 的计算满足式(1)，其中，Precision（P）表示预测为正样本中真实为正样本的比例，Recall（R）表示真实正样本中被正确预测的比例，F1 值为 Precision 与 Recall 的调和平均数。

$$F1=2\times\frac{P\times R}{P+R} \tag{1}$$

软硬匹配规则具体如下：

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：采用 Han 等人[36]提出的算法，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。计算满足式(2)，其中，Similarity 表示预测片段与标准答案片段的字符串相似度；M 表示二者匹配到的字符数；lenpred 与 lengold 分别表示预测片段与标准答案片段的字符长度。

$$Similarity=\frac{M\times 2}{len_{pred}+len_{gold}} \tag{2}$$

3.3 对比模型和实验设置

为验证本文方法的有效性，实验设置了以下对比模型：

基线方法的结果（Baseline Models）包括：(1) STATE-ToxiCN 论文中公开的实验结果；(2) 本文在相同实验配置下复现 STATE-ToxiCN 论文所得的结果，包括微调模型 LLaMA3-8B、Qwen2.5-7B 和 DeepSeek-v3.1 大模型 API；(3) CCL 测评任务第二名在 STATE-ToxiCN 上的实验结果。

本文方法的结果（Our Method）包括应用于大模型 API 的结果和基于微调模型的结果，其中自检索增强使用的句向量模型均为 bge-large-zh-v1.5[37]。

由于 MAV 策略需在每个输入上执行多次独立推理，对 API 调用次数和延迟带来显著

开销。以投票阈值 τ =30 为例，DeepSeek-v3.1 单条样本执行完整 MAV 流程平均需约 50 次推理请求，其推理成本和时间相应提高 50 倍，若使用并行推理时间成本相应缩减对应倍数。考虑到 API 接口的并行限制与预算约束，本研究仅在微调模型设置中使用 MAV。

大模型 API 采用的是 DeepSeek-v3.1 官方的 API，自检索增强参数示例个数为 20、探索概率为 0.3。

微调模型是基于 Qwen2.5-7B-Instruc 模型进行全参数监督微调得到的模型，训练使用 LLaMA-Factory 框架，在 STATE-ToxiCN 的训练集上进行全参数监督微调，和 Wang 等人[23]保持一致，采用 DeepSpeed ZeRO-2 并行优化策略以支持大规模参数高效训练，模型输入的最大长度为 1024，采用 Cosine 学习率调度策略，学习率设为 1e-5，训练 epoch 为 2，每张卡的训练批量大小为 2，梯度累积步数为 4。推理过程中自检索增强参数示例个数为 10、探索概率为 0.3，多轮累积投票参数投票阈值为 200。

3.4 实验结果

表 3 展示了基线方法和本文方法在 STATE-ToxiCN 数据集上的实验结果。可以看到，本文提出的方法应用于微调模型上和调用大模型 API 上均取得了显著优于现有模型或方法的性能。

表 3 STATE-ToxiCN 实验结果表

	Model	Target		Argument		T-A Pair		T-A-H Tri.		Quad.		Score
		Hard	Soft	Hard	Soft	Hard	Soft	Hard	Soft	Hard	Soft	
微调模型	LLaMA3-8B[11]	64.07	73.74	36.72	70.82	31.64	60.88	27.04	51.62	24.27	46.08	35.18
	Qwen2.5-7B[11]	63.96	74.64	35.42	70.36	30.63	60.52	26.51	52.86	23.70	47.03	35.37
	ShieldGemma-9B[11]	63.40	74.31	34.40	71.11	29.99	61.51	25.64	52.70	23.49	47.14	35.32
	LLaMA3-8B	62.10	73.61	37.12	70.90	31.08	59.97	24.05	45.65	24.00	45.50	34.80
	Qwen2.5-7B	64.27	74.81	36.64	70.93	32.03	60.87	24.48	46.03	24.48	45.92	35.20
	Wang 等[23]	-	-	-	-	-	-	-	-	26.66	48.35	37.51
	Ours	66.82	76.21	37.87	72.63	32.81	62.92	28.96	55.33	27.00	50.69	38.85
大模型 API	GPT-4o[11]	46.85	58.19	22.64	62.41	17.21	46.41	13.21	35.68	9.00	23.34	16.17
	DeepSeek-v3[11]	48.16	59.25	22.79	59.38	18.68	46.40	14.95	37.19	11.48	27.38	19.43
	DeepSeek-v3.1	40.15	52.27	20.48	53.61	15.49	40.43	12.41	32.81	9.00	24.45	16.73
	DeepSeek-v3.1+prompt	52.52	62.08	24.68	62.86	20.34	48.77	16.63	39.75	13.56	32.82	23.19
	Ours	61.57	70.96	36.15	71.20	30.62	60.70	26.34	51.76	23.76	46.24	35.00

在全参数微调的 Qwen2.5-7B 上引入本文两阶段多轮问答框架后，模型在所有层级指标上实现一致提升：四元组硬匹配 F1 达到 27.00，软匹配 F1 达到 50.69，平均分为 38.85。这一结果超过 STATE-ToxiCN 中微调基线 Qwen2.5-7B 平均分 35.37 以及 LLaMA3-8B、ShieldGemma-9B 等同规模模型，并高于 Wang 等方法在四元组层面的公开成绩 37.51。说明两步任务重构能够降低端到端四元组抽取的耦合难度，而多轮问答提示在要素定位与群体归因阶段发挥了稳定增益。

直接调用 API 的模型总体弱于微调模型，四元组平均得分多处于[10,20]。在此设定下，

本文方法仍取得显著提升：四元组硬匹配、软匹配分别为 23.76、46.24，平均分达 35.00。相较于 STATE-ToxiCN 的 DeepSeek-v3 结果提升 10.57 个百分点，相较同配置复现的 DeepSeek-v3.1 提升 18.27 个百分点；同时，在 DeepSeek-v3.1 上仅加入本文提出的两阶段问答提示即可带来 6.46 的平均分增益，验证了分步求解在无需额外训练时仍能明显改善抽取质量与输出一致性。

表 4 CCL2025 结果表

方法	Quad.		Score
	Hard	Soft	
第一名	25.41	47.40	36.41
ours	**27.35**	**51.17**	**39.26**

表 4 进一步报告了 CCL2025 官方测试集成绩。本文系统在四元组硬匹配、软匹配 F1 上分别达到 27.35、51.17，最终综合得分为 39.26，较测评第一名系统提升 2.85 个百分点，该结果体现本文框架的优越性。

4 分析与讨论

4.1 消融实验

为探究各子模块对整体性能的影响，本文在微调大模型 Qwen2.5-7B 的基础上，开展消融实验，结果如表 5 所示。整体来看，提示结构的设计、自检索增强策略和多轮累积投票机制三者协同配合，共同保障了模型在复杂场景下的性能与鲁棒性。

表 5 消融研究结果

方法	Quad.		Score
	Hard	Soft	
Base Model	23.70	47.03	35.37
+prompt	25.99	49.36	37.67
+prompt+ SRAG	26.72	49.77	38.25
+prompt+ SRAG+MAV	27.00	50.69	38.85

4.2 上下文示例数量的影响

对于上下文示例数量的实验是通过 DeepSeek-v3.1 的 API 进行的，直接选取相似度前 k 个，不进行探索，结果如表 6 所示。

表 6 Icl 示例个数 k 的影响

k	Quad.		Score
	Hard	Soft	
10	20.93	44.96	32.95
20	22.14	48.85	35.50
30	22.73	47.73	35.23
40	20.85	44.79	32.82

从表 6 可以看出，当样例数量从 10 增加到 20 时模型性能得到提升，这表明多样化示例

有助于模型更好地捕捉输入样本的语义特征，从而更准确地识别评论对象与立场内容。值得注意的是，当样例数量从 30 增加到 40 时，性能提升趋于平缓甚至出现轻微下降，这可能与输入长度限制或上下文干扰有关，提示过多示例可能会稀释模型对当前样本的关注度。

该实验表明，合理控制示例数量对提示学习效果具有重要影响，未来可进一步结合长度约束或内容选择机制优化样例配置。

4.3 探索概率对 SRAG 的影响

为探究 SRAG 中示例选择的超参数探索概率 ε 对模型性能的具体影响，本文在推理阶段开展对比实验。由于在较高 MAV 阈值下，探索概率的变化对结果影响较小，主要表现为轻微波动，故本实验将 MAV 阈值固定为 30，以更清晰地观察 ε 的调节效应。实验结果如表 7 所示。

表 7 探索概率 ε 对 SRAG 影响

ε	Quad.		Score
	Hard	Soft	
0.1	26.92	50.26	38.59
0.2	26.83	50.08	38.46
0.3	27.06	50.12	38.59
0.4	26.74	49.65	38.19
0.5	26.64	49.44	38.04
0.6	26.45	49.28	37.86
0.7	26.84	49.33	38.09
0.8	26.47	48.60	37.54
0.9	26.38	48.73	37.56

整体来看，随着探索概率 ε 的逐步增加，模型性能呈现出先稳中有升、后缓慢下降的趋势。在 ε =0.3 时，模型在指标上取得峰值，说明适当引入一定比例的非最相似示例，有助于提升提示覆盖面与模型的生成鲁棒性；当 ε >0.3 时，模型各项指标开始缓慢下降，表明过多低相关示例的引入可能引起干扰，削弱模型的判断一致性。

值得注意的是，即使在 ε =0.9 的极端配置下，模型整体性能依然维持在合理水平，说明 SRAG 框架具备一定的容错性，但最佳探索概率仍集中于 0.3 附近。

本实验表明，适当调节探索概率可在相关性与多样性之间实现有效平衡。未来可进一步结合句向量质量评估、主题分布稀疏性控制等机制，提升探索样本选择的策略性与可控性。

4.4 MAV 阈值的影响

为分析 MAV 阈值参数对模型性能的影响，本文在 $\tau \in \{1,2,3,5,8,10,15,20,30,40,50,80,100,200\}$上进行实验，并分别报告基于四元组评估的软、硬匹配分数及二者平均分，结果如图 2 所示。整体来看，三项指标均在更高阈值下获得提升，随阈值变化呈现“先升后趋于平稳”的趋势，说明 MAV 在合理阈值范围内能够稳定提升模型在细粒度仇恨言论识别上的鲁棒性。

具体而言，硬匹配分数随阈值 τ 的提升呈现稳健上升趋势：从最小阈值 23.28(τ =1)持续提高至 25.75(τ =200)，整体提升 2.47 分，在 τ ∈[1,40]区间内提升较为平滑，在高阈值区域(τ ≥80)提升更显著。这说明累积更多投票结果（更大的 τ ）能够有效提升严格四元组硬匹配的稳定性与准确性。

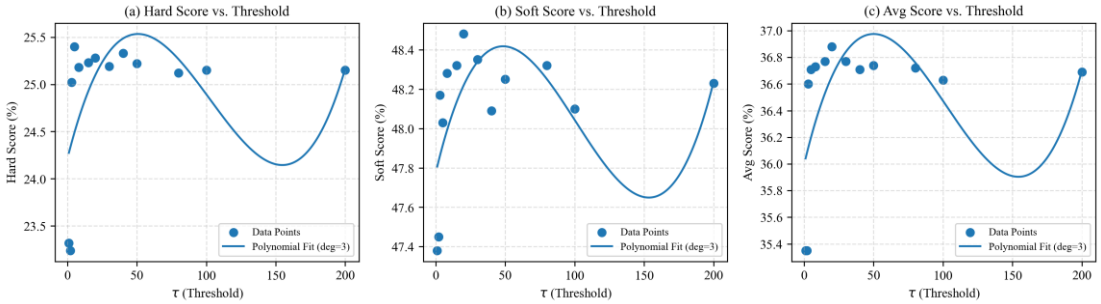

图 2 MAV 阈值参数 τ 的敏感性分析

软匹配分数也随阈值上升表现稳定增长：由 47.38 升至 49.06，总提升 1.68 分。中等阈值(τ =[40,100])附近增幅较缓；高阈值(τ =200)出现明显跃升。说明 Soft 匹配同样受益于更高的投票一致性，且对极高阈值更敏感。

平均分从 35.35(τ =1)增加到 37.45(τ =200)，提升 2.10 分，由于同时考虑软硬匹配，平均分提升也印证了 MAV 在多视角累积投票机制下实现整体性能增强的能力。

因此，本组实验进一步验证了 MAV 机制对预测结果进行稳定化与细粒度纠偏的有效性。这与 Wang 等[23]的研究中的结果一致。

4.5 错误分析

为系统评估基于提示学习的细粒度中文仇恨识别方法，本文在 CCL2025 评测任务 10 细粒度中文仇恨言论识别数据集 STATE-ToxiCN[11]上展开实验，数据规模统计见表 1。

为进一步探讨模型在实际生成过程中的不足与改进空间，本文对测试集中模型输出与标准答案之间的不一致样本进行分类统计与误例分析，归纳出以下几类典型错误类型，如表 8 所示。

表 8 模型典型错误类型及示例

错误类型	标准答案	模型预测
对象识别不完整	黑河南人的 | 离谱 | Region | hate [SEP] 河南 | 一个人代表一亿人 | Region | hate	河南 | 一个人代表一亿人 | Region | hate
仇恨属性误判	作者 | 专业打拳的 | non-hate | non-hate	作者 | 专业打拳 | Sexism | hate
引用立场混淆	你 | 在劳教所里就别祸害别人了 | non-hate | non-hate	你 | 别祸害别人了 | non-hate | non-hate
群体标签分类错误	泼峰 | 会不得已提供一些有色服务 | LGBTQ | hate	泼峰会 | 提供一些有色服务 | Sexism | hate

（1）对象识别不完整或边界偏差

在部分样例中，模型未能准确提取目标对象的完整表达，通常表现为主语被截断或补足语遗漏，影响整体语义准确性。表 8 示例中，模型仅捕捉了后半部分内容，遗漏了情绪较强

的前缀“黑河南人的”，导致表达主体被弱化，从而影响下游判别。

（2）隐性或讽刺性表达的判断

模型对一些隐性或讽刺性表达的判断仍存在困难，表现为对情感极性或攻击意图的误判。表 8 示例中，模型误将“打拳”解读为攻击性行为，进而赋予错误的群体标签与仇恨属性，反映出其对讽刺语境理解能力的局限。

（3）引用/立场混淆，错将“被评价者”作为主语

在涉及引用性表达或转述他人观点的语句中，模型易出现主语识别错位的问题。表 8 示例中，虽然预测结果本质语义相近，但省略了关键限定语“在劳教所里”，在具体应用场景中可能造成歧义，对极端语句的定位不够精确。

（4）群体标签识别错误

部分样本中，模型识别出的目标群体标签与语义不符，群体归因错误或混用多标签是常见问题。表 8 示例中，模型将涉及同性群体的攻击性言论误归入性别歧视范畴，反映出其在多类别判别任务中的边界模糊问题。

总体来看，当前模型在对象边界识别、情绪强度判别、讽刺语义理解与多标签归因等方面仍存在提升空间。未来工作可考虑引入情感词典约束、结构感知解码器或多通道自监督训练等方式，以进一步增强模型对复杂语义结构与隐含攻击意图的识别能力。

5 结论

本文提出面向中文细粒度仇恨识别的两阶段多轮问答框架。针对细粒度的中文仇恨识别，本文构建了多轮提示词结构，并结合自检索增强策略，系统探索如何在无需大规模参数训练的前提下提升模型对细粒度仇恨四元组的识别与理解能力。在构建动态上下文示例与引入多轮累积投票机制后，本文方法在 STATE-ToxiCN 数据集上取得了四元组软硬匹配 F1 均值 38.85 的表现，超过当前评测基线与领先系统，验证了所提出提示词设计策略、自检索增强机制及多轮推理结构的有效性。本文还开展了全面的实验分析与讨论，从提示词结构设计、示例检索策略、探索概率设置、多轮投票阈值等多维度考察不同因素对中文细粒度仇恨识别表现的影响，相关经验性结果也可以为其它中文自然语言理解任务提供参考。

整体来看，本文提出的方法在提高中文仇恨言论的细粒度抽取能力、增强模型在复杂语境中的鲁棒性方面表现突出，为中文内容安全任务提供了有效的新思路，但当前模型在对象边界识别、情绪强度判别、情绪表达判断、群体标签分类及讽刺语义理解等方面仍存在提升空间，后续可从结构建模、常识引入、对比学习等方向开展优化。

[参考文献] (References)

[1] SILVA L, MONDAL M, CORREA D, et al. Analyzing the targets of hate in online social media[C]//Proceedings of the 10th International AAAI Conference on Web and Social Media. Cologne, Germany: AAAI, 2021: 687-690.

[2] ALKHAMISSI B, LADHAK F, IYER S, et al. ToKen: Task decomposition and knowledge infusion for few-shot hate speech detection[C]//Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Abu Dhabi, UAE: ACL, 2022: 2109-2120.

[3] ALKHAMISSI B, LADHAK F, IYER S, et al. ToKen: Task decomposition and knowledge infusion for few-shot hate speech detection[C]//Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Abu Dhabi, UAE: ACL, 2022: 2109-2120.
[4] PAVLOPOULOS J, SORENSEN J, LAUGIER L, et al. SemEval-2021 Task 5: Toxic spans detection[C]//Proceedings of the 15th International Workshop on Semantic Evaluation. Virtual Event: ACL, 2021: 59-69.
[5] MATHEW B, SAHA P, YIMAM S M, et al. HateXplain: A benchmark dataset for explainable hate speech detection[C]//Proceedings of the 35th AAAI Conference on Artificial Intelligence. Virtual Event: AAAI, 2021: 14867-14875.
[6] ZAMPIERI M, MORGAN S, NORTH K, et al. Target-based offensive language identification[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Toronto, Canada: ACL, 2023: 762-770.
[7] DENG J, ZHOU J, SUN H, et al. COLD: A benchmark for Chinese offensive language detection[C]//Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Abu Dhabi, UAE: ACL, 2022: 11580-11599.
[8] JIANG A, YANG X, LIU Y, et al. SWSR: A Chinese dataset and lexicon for online sexism detection[J]. Online Social Networks and Media, 2022, 27: 100182.
[9] COWAN G, HODGE C. Judgments of hate speech: The effects of target group, publicness, and behavioral responses of the target[J]. Journal of Applied Social Psychology, 1996, 26(4): 355-374.
[10] XIAO Y, HU Y, CHOO K T W, et al. ToxiCloakCN: Evaluating robustness of offensive language detection in Chinese with cloaking perturbations[C]//Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. Miami, FL, USA: ACL, 2024: 6012-6025.
[11] BAI Z, YANG L, YIN S, et al. STATE ToxiCN: A benchmark for span-level target-aware toxicity extraction in Chinese hate speech detection[C]//Findings of the Association for Computational Linguistics: ACL 2025. Vienna, Austria: ACL, 2025: 10206-10219.
[12] LU J, XU B, ZHANG X, et al. Facilitating fine-grained detection of Chinese toxic language: Hierarchical taxonomy, resources, and benchmarks[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Toronto, Canada: ACL, 2023: 16235-16250.
[13] LI M, LIAO S, OKPALA E, et al. COVID-HateBERT: A pre-trained language model for COVID-19 related hate speech detection[C]//Proceedings of the 20th IEEE International Conference on Machine Learning and Applications. Pasadena, CA, USA: IEEE, 2021: 233-238.
[14] CASELLI T, BASILE V, MITROVIĆ J, et al. HateBERT: Retraining BERT for abusive language detection in English[C]//Proceedings of the 5th Workshop on Online Abuse and Harms. Virtual Event: ACL, 2021: 17-25.
[15] Unitary. Detoxify[EB/OL]. (2024-02-01). https://github.com/unitaryai/detoxify.
[16] HARTVIGSEN T, GABRIEL S, PALANGI H, et al. ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland: ACL, 2022: 3309-3326.
[17] FOUNTA A, DJOUVAS C, CHATZAKOU D, et al. Large scale crowdsourcing and characterization of Twitter abusive behavior[C]//Proceedings of the 12th International AAAI Conference on Web and Social Media. Palo Alto, CA, USA: AAAI, 2018: 491-500.
[18] DAVIDSON T, WARMSLEY D, MACY M, et al. Automated hate speech detection and the problem of offensive language[C]//Proceedings of the 11th International AAAI Conference on Web and Social Media. Montreal, QC, Canada: AAAI, 2017: 512-515.
[19] WASEEM Z, HOVY D. Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter[C]//Proceedings of the NAACL Student Research Workshop. San Diego, CA, USA: ACL, 2016: 88-93.
[20] CHUNG I, LIN C J. TOCAB: A dataset for Chinese abusive language processing[C]//Proceedings of the 22nd International Conference on Information Reuse and Integration for Data Science. Las Vegas, NV, USA: IEEE, 2021: 445-452.
[21] ZHOU J, DENG J, MI F, et al. Towards identifying social bias in dialog systems: Framework, dataset, and benchmark[C]//Findings of the Association for Computational Linguistics: EMNLP 2022. Abu Dhabi, UAE: ACL, 2022: 3576-3591.
[22] CHIU K L, COLLINS A, ALEXANDER R. Detecting hate speech with GPT-3[J]. arXiv preprint arXiv:2103.12407, 2022.
[23] WANG J, LIU R, ZHANG L, et al. System report for CCL25-eval task 10: SRAG-MAV for fine-grained Chinese hate speech recognition[J]. arXiv preprint arXiv:2507.18580, 2025.
[24] DONG Q, LI L, DAI D, et al. A survey on in-context learning[C]//Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. Miami, FL, USA: ACL, 2024: 1107-1128.
[25] WEI J, WANG X, SCHUURMANS D, et al. Chain-of-thought prompting elicits reasoning in large language models[C]//Proceedings of the 36th International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates, 2022: 24824-24837.
[26] WANG X, WEI J, SCHUURMANS D, et al. Self-consistency improves chain of thought reasoning in language models[J]. arXiv preprint arXiv:2203.11171, 2022.
[27] YAO S, YU D, ZHAO J, et al. Tree of thoughts: Deliberate problem solving with large language models[J]. arXiv preprint arXiv:2305.10601, 2023.
[28] BESTA M, BLACH N, KUBICEK A, et al. Graph of thoughts: Solving elaborate problems with large language models[C]//Proceedings of the 38th AAAI Conference on Artificial Intelligence: Vol. 38. Vancouver, Canada, 2024: 17682-17690.

[29] MADAAN A, TANDON N, GUPTA P, et al. SELF-REFINE: Iterative refinement with self-feedback[C]//Proceedings of the 37th International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates, 2023: 46534-46594.
[30] CHEN W, MA X, WANG X, et al. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks[J]. Transactions on Machine Learning Research, 2023, 10: 1-15.
[31] LI X, QIU X. Finding support examples for in-context learning[J]. arXiv preprint arXiv:2302.13539, 2023.
[32] LI X, LV K, YAN H, et al. Unified demonstration retriever for in-context learning[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Toronto, Canada: ACL, 2023: 4644-4668.
[33] LIU J, SHEN D, ZHANG Y, et al. What makes good in-context examples for GPT-3?[C]//Proceedings of Deep Learning Inside Out : The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures. Dublin, Ireland: ACL, 2022: 100-114.
[34] LIU Y, LIU J, SHI X, et al. Let's learn step by step: Enhancing in-context learning ability with curriculum learning[J]. arXiv preprint arXiv:2402.10738, 2024.
[35] 来雨轩, 王夏菁, 胡文鹏. 基于提示词工程的中文修辞识别与理解方法[J]. 中文信息学报, 2025, 39(6): 22-34.
[36] HAN R, YANG C, PENG T, et al. An empirical study on information extraction using large language models[J]. arXiv preprint arXiv:2409.00369, 2024.
[37] XIAO S, LIU Z, ZHANG P, et al. C-Pack: Packed resources for general Chinese embeddings[C]//Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. Washington, DC, USA: ACM, 2024: 641-649.