



强化学习驱动的有害信息智能检测方法

闫硕^{1,2}, 张兴伟³, 张淳杰^{1,2}, 郑晓龙^{3,4*}

1. 北京交通大学计算机科学与技术学院信息科学研究所, 北京 100044

2. 北京交通大学计算机科学与技术学院视觉智能交叉创新教育部国际合作联合实验室, 北京 100044

3. 中国科学院自动化研究所多模态人工智能系统全国重点实验室, 北京 100190

4. 中国科学院大学, 北京 100190

* 通信作者. E-mail: xiaolong.zheng@ia.ac.cn

收稿日期: 2026-07-03; 修回日期: 2026-08-14; 接受日期: 2026-09-01; 网络出版日期: 2026-09-21

国家自然科学基金 (批准号: 72225011, 72434005, 62476021) 资助项目

摘要 有害信息检测对于维护网络生态安全和防范社会风险具有重要意义, 是数字化社会治理中的关键任务. 近年来, 基于人工智能尤其是大语言模型的有害信息检测方法在自动化识别方面取得了显著进展. 同时, 在隐性表达频繁出现和网络黑话持续演化的真实网络环境中, 引入外部语义信息有助于补充模型对上下文语义和隐含有害意图的理解能力, 进一步提升有害信息识别效果. 然而, 现有外部知识语义增强方法多依赖静态匹配或规则触发式检索, 难以自适应地判断外部知识引入的必要性及具体内容, 容易产生语义噪声, 限制模型的判别效果与泛化能力. 针对上述问题, 本文提出一种基于反事实奖励的主动词典检索方法. 该方法将外部知识检索过程建模为策略学习问题, 通过强化学习机制自适应地决定是否检索以及检索哪些关键语义词项, 并设计反事实检索增益奖励机制, 引导模型学习对判别任务具有实际贡献的高价值检索行为, 从而实现从规则驱动向语义驱动的检索范式转变. 在 SWSR, ToxiCN 与 STATE-ToxiCN 三个具有不同语言复杂度和隐晦有害表达分布特征的数据集上的实验结果表明, 本文方法在不同识别任务、数据分布及模型规模 (3B 与 7B) 下均取得了性能提升, 且在隐喻表达和网络黑话占比较高的场景中优势更为明显, 验证了其在复杂网络语义环境下的有效性和泛化能力.

关键词 数字化社会治理, 有害信息检测, 强化学习, 大语言模型, 反事实奖励

1 引言

随着社交媒体的广泛普及, 网络空间已成为公众表达观点和参与社会互动的重要平台. 与此同时, 网络信息传播的范围和影响力持续提升, 使仇恨言论、群体对立和煽动性内容等有害信息更容易在短时间内快速扩散^[1,2], 导致冲突和重大网络舆情事件频发^[3~5]. 因此, 及时、准确地识别有害信息, 对于遏制其快速传播、维护网络空间的安全稳定具有重要的理论和现实意义, 并已成为数字社会治理、网络空间安全与平台内容治理领域的重要研究问题^[6~8]. 该任务的核心挑战在于, 如何从海量动态变化的社交媒体内容中, 自动、准确地识别具有隐蔽表达和语义变体等复杂语境特征的有害信息. 这不仅需要满足大规模网络内容自动化审核与风险预警

引用格式: 闫硕, 张兴伟, 张淳杰, 等. 强化学习驱动的有害信息智能检测方法. 中国科学: 信息科学, 2026, 56: 2310–2326, doi: 10.1360/SSI-2026-0335

Yan S, Zhang X W, Zhang C J, et al. A reinforcement learning-driven method for intelligent harmful information detection. Sci Sin Inform, 2026, 56: 2310–2326, doi: 10.1360/SSI-2026-0335

表 1 网络平台有害信息示例.

Table 1 Examples of harmful information on online platforms.

Text	Type	Explanation
太多辣鸡媒体宣传默默卖蠢和苦难来博取国人尤其是女性共情	谐音变体	辣鸡: 垃圾的谐音.
半年没来之前反默老哥都不在了, 只看到wwwewecao 刷屏, 说他默吧他又好像钓鱼, 说他钓鱼吧又好像媚默	字体拆解	默: 由“黑”和“犬”组合而成, 常被用作对黑人的蔑称.
楼主说的那种东西真是不言而喻当然指的就是虫族	隐喻表达	虫族: 指代黑人, 用虫子的寄生属性隐喻黑人是寄生虫.
别结束啊, 别当驴啊小仙女独立女性啊!	反讽叙事	驴: 指“婚驴”, 即对所谓适应男权社会女性的蔑称; 该文本通过反讽方式攻击女权或独立女性.

的现实需求, 也要求模型具备对上下文语义、表达意图和群体指向等信息的综合理解能力. 因此, 开展有害信息检测研究, 不仅能够为网络平台治理和风险预警提供技术支撑, 提升网络空间治理效率与风险防控能力, 也有助于推动复杂网络语境下语言理解、语义建模与智能分析等关键技术的发展, 具有重要的理论价值和现实意义.

近年来, 人工智能技术的快速发展为有害信息检测提供了新的技术路径. 相较于传统人工审核和规则匹配方法, 基于机器学习和深度学习的检测方法能够从海量网络文本中自动学习语义表征^[9], 显著提升有害信息识别的效率和准确性. 因此, 人工智能技术逐渐成为有害信息检测的重要支撑手段, 并在互联网平台内容审核、网络舆情治理和公共安全风险防控等场景中得到广泛应用. 此外, 以 BERT, RoBERTa, GPT 系列模型等为代表的预训练语言模型在上下文语义建模、隐含语义捕捉和复杂表达理解方面展现出更强能力^[10], 使模型不再局限于识别显式攻击性词汇或敏感关键词, 而是能够从语境关联、情感倾向和表达意图等层面捕捉隐含立场及潜在有害意图, 推动了有害信息检测从表层特征匹配向深层语义理解的转变.

尽管基于人工智能的有害信息检测方法在上下文建模和自动化识别方面取得了显著进展, 但其在真实网络环境中的适应能力仍然面临诸多挑战. 一方面, 社交媒体中的有害表达具有明显的动态演化特征^[11]. 为了规避平台审核, 用户常常通过谐音变体、字体拆解、隐喻表达和反讽叙事等方式对有害内容进行伪装^[12,13], 使得原本显式的攻击意图转化为更加隐蔽和语境依赖的表达形式, 相关示例如表 1 所示. 另一方面, 新兴黑话和社群内部用语往往缺乏充足有效的训练数据支撑, 导致模型难以准确学习其语义指向和潜在有害倾向. 尤其是在不同平台和社群语境下, 同一词汇可能具有截然不同的语义指向^[14]. 这种表达形式与语义指向之间的非固定映射关系, 进一步削弱了模型对领域特定隐含语义的稳定建模能力, 使其在面对新兴表达和语境转移时容易出现漏检或误检问题. 基于此, 已有研究通过引入外部知识对输入文本进行语义增强^[14,15], 识别文本中的敏感词、群体指称或特定触发词, 检索相关常识关系或背景信息, 并与原始输入融合, 从而为模型提供额外语义线索, 提升其对隐性攻击、语义暗示和语境依赖型有害表达的识别能力.

然而, 现有语义增强方法在场景适配和需求感知方面仍存在明显不足. 首先, 基于规则的语义增强方法通常依赖预设的敏感词或特定触发词进行检索触发, 无法充分适应网络黑话和社群内部用语的快速演化^[16]. 其次, 现有方法缺乏对上下文语义的动态判断能力, 难以区分预设词项在不同语境中的真实含义, 容易引入与当前判别任务不匹配的外部知识, 从而造成语义偏移或错误关联^[17]. 最后, 固定触发式检索策略容易在语义需求不足的场景中引入冗余知识, 使模型引入与当前判别任务弱相关的外部信息, 增加输入噪声并诱发错误语义关联, 从而削弱模型的判别性能^[18]. 总体来看, 现有方法大多将语义增强视为独立于模型判别过程的外部预处理环节, 缺乏面向具体输入语义需求的自适应决策能力. 因此, 模型难以根据文本的语义复杂度和外部知识需求, 自适应地判断是否需要引入外部语义信息以及应当引入哪些具体内容, 从而制约了语义增强方法在复杂网络语境下的判别有效性与跨场景泛化能力.

针对上述问题, 本文提出一种基于反事实奖励的主动词典检索方法 CALR (counterfactual active lexicon retrieval). 与传统基于固定的词典匹配或规则触发的数据增强方式不同, 本文将外部知识检索过程建模为模型

可学习的语义决策行为,使模型能够根据输入文本的上下文语义、潜在歧义程度以及黑话表达等特征,自主判断是否需要调用外部知识,并进一步选择需要检索的关键信息.具体而言,本文首先利用监督微调使大语言模型具备基本的任务理解能力和结构化输出能力;随后引入基于反事实检索增益奖励的强化学习机制,对模型的检索行为与检测结果进行联合优化,使模型在提升有害信息识别性能的同时,减少不必要的检索和错误匹配带来的语义噪声.通过上述方式,模型不再被动依赖固定词典触发,而是能够围绕具体语境进行动态检索,从而增强对隐式有害表达、网络黑话和面向特定群体的隐含语义的识别能力.在3个典型有害信息检测数据集上的实验结果表明,本文方法相较于基线方法取得了明显性能提升.此外,通过进一步的对比实验验证了强化学习检索策略和反事实检索增益奖励机制的有效性,表明该方法能够通过增强正向增益信息获取、抑制负向增益信息和冗余外部知识干扰,提升检索增强方法在复杂网络语境下的判别有效性与泛化能力.

本文的主要贡献如下.

(1) 提出了 CALR 框架,将检索增强建模为基于强化学习的策略优化问题,实现从规则驱动到语义驱动的范式转变.

(2) 设计了反事实检索增益奖励机制,通过强化学习对模型的检索策略进行动态优化,从而提升检索效率,降低冗余外部信息对模型判别过程的干扰.

(3) 在 SWSR^[14], ToxiCN^[15] 与 STATE-ToxiCN^[19] 三个具有不同语言复杂度和隐晦表达占比的数据集上进行系统评估,实验结果表明所提出方法的有效性.

2 相关工作

2.1 基于机器学习和深度学习的方法

早期有害信息检测研究主要依赖统计学习方法与人工特征工程. Warner 等^[20]、Burnap 等^[21] 和 Nobata 等^[22] 利用词法、句法、n-gram 等多种语言特征,并结合支持向量机、逻辑回归等传统机器学习方法,对有害信息进行自动识别.随着深度学习的发展,模型对文本语义的建模能力进一步增强,研究重点逐渐由浅层特征匹配转向深层语义表示学习. Badjatiya 等^[23] 将 CNN, LSTM 和 FastText 用于 Twitter 有害信息分类,证明神经网络表示相比传统特征具有更强的语义建模能力.此外,其他相关工作分别从用户上下文、注意力建模、目标识别、多语言检测和用户级识别等角度扩展了深度学习方法的应用范围^[24~28]. 总体来看,这类研究推动有害信息检测从表层统计特征向神经语义表示学习转变,但仍难以充分建模隐晦表达、反讽语境、跨领域表达以及复杂目标指向关系.

2.2 基于预训练语言模型的方法

随着预训练语言模型的发展, BERT 凭借较强的语义表征与上下文建模能力,逐渐成为有害信息检测的重要基础.早期研究多将 BERT 作为文本编码器,通过微调完成句子级分类任务.例如, Deng 等^[29] 基于中文 BERT 构建了有害信息识别模型,用于中文冒犯性语言识别; Dai 等^[30] 则将 BERT 与多任务学习结合,同时建模冒犯性检测、类型识别和攻击目标识别.随后,相关研究开始从整体文本分类转向有害片段定位, Chhablani 等^[31] 使用 BERT, RoBERTa 和 SpanBERT 进行有害片段检测,以识别文本中触发有害性的具体片段.近年来,大语言模型凭借更大的预训练规模和更强的指令遵循、语义理解与推理能力^[32],被逐渐应用于有害信息检测任务.现有研究主要围绕提示学习、数据增强、监督微调和语义表征增强展开. Guo 等^[33] 比较了 GPT-3.5 在零样本、少样本和 CoT 提示下的检测效果,发现推理式提示有助于理解隐式语义,但在中文等非英语场景下仍不稳定. Deng 等^[34] 和 Jin 等^[35] 分别提出 AugCOLD 方法和 GPT-HateCheck 方法,均借助大模型生成样本进行数据增强,有效提升了模型的泛化性和鲁棒性. Cheremetiev 等^[36] 则通过微调大模型文本嵌入提升隐式有害信息检测的跨数据集泛化能力.总体而言,大语言模型推动了有害信息检测向上下文推理、结构化抽取和语义增强方向发展,但现有方法仍较多依赖人工提示与模型内部参数化知识,对隐喻表达、网络黑话和动态演化语义的适应能力仍然有限.

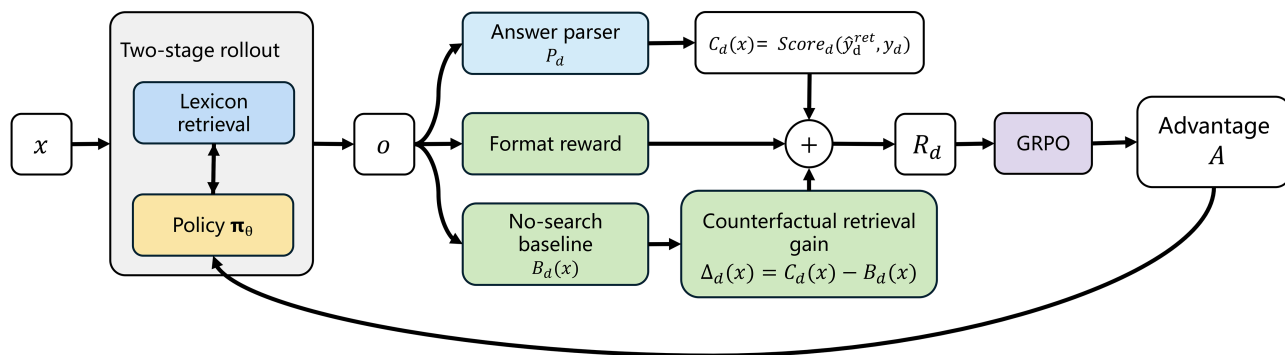


图 1 (网络版彩图) CALR 方法总体流程图.

Figure 1 (Color online) The workflow of CALR.

2.3 基于外部知识检索的方法

上述两类方法的知识主要来源于训练数据或模型内部参数,本质上属于静态知识,因此难以有效应对隐式有害表达的持续演化以及新兴网络黑话的快速出现.为缓解这一问题,已有研究开始引入词典资源、实体知识、知识图谱和检索机制等可动态更新的外部知识,对模型进行语义增强,从而提升其对复杂有害语义的理解与识别能力.词典资源是较容易被应用于有害信息检测的一类外部知识形式,Razavi 等^[37]通过加权攻击性词典与传统分类器结合识别侮辱、歧视和极端主观表达,Bassignana 等^[38]构建 HurtLex 多语言有害词词典,为词典匹配、特征构造和跨语言迁移提供了基础资源, Lee 等^[39]则结合增强型词表、word2vec 和规则模块识别网络文本中变形和新造的攻击性表达.面向中文场景, Jiang 等^[14], Lu 等^[15]和 Bai 等^[19]分别在 SWSR, ToxiCN 和 STATE-ToxiCN 中引入词典信息,以增强模型对中文有害评论的识别能力. Yuan 等^[40]通过匹配词典中关键词,并将检索到词汇的类别信息注入大语言模型提示中,有效提升了有害信息检测的准确率.除词典知识外, Lin^[41]通过实体链接引入维基百科实体摘要, Stranisci 等^[42]利用危险言论本体和知识图谱补充结构化语义信息, Zhao 等^[43]通过毒性知识图谱检索相关路径增强推理, Mnassri 等^[44]利用向量知识库和语义缓存实现 RAG 式检测.总体来看,外部知识增强方法能够为模型补充黑话释义、实体背景、政策语义、知识路径和相似示例信息,有助于缓解模型仅依赖参数化知识和表层词汇判断的问题.

2.4 现有研究存在的问题

现有方法大多依赖词典匹配、实体触发、向量相似度或人工规则进行固定检索,其检索流程通常是预设且静态的,缺乏对检索行为本身的建模与优化.一方面,这类方法使用统一的检索触发规则,无法根据文本语义复杂度、自身知识不确定性或潜在歧义程度,自适应判断是否需要引入外部知识,从而容易导致无效检索或信息冗余.另一方面,现有检索策略通常缺乏对检索内容的精细控制,难以针对隐式表达、跨语境隐喻或新兴黑话动态选择最相关的知识来源,导致检索结果与实际语义需求之间存在偏差.此外,检索与检测任务往往是松耦合的,缺乏端到端的协同优化机制,使得外部知识难以被有效利用.因此,如何将外部知识检索从规则驱动的静态流程转化为语义驱动的动态决策过程,使模型能够根据输入内容自适应地决定检索时机和检索内容以及如何利用检索结果,是当前有害信息检测研究中亟待解决的关键问题.

3 CALR 方法

本文提出 CALR,即一种面向有害信息检测的反事实主动词典检索方法,方法总体流程如图 1 所示.该方法的核心思想是:模型在面对包含黑话、隐喻、缩写、代称或不明确群体称呼的文本时,不必始终直接给出检测结果,而是可以主动发起词汇检索,并在获得检索结果后完成最终判断.通过强化学习训练,模型不仅学习如何完成有害信息检测任务,还学习何时检索、检索哪些词汇,以及检索是否真正有助于提升任务表现.

3.1 任务定义

本文考虑多个中文有害信息检测数据集. 由于不同数据集的标注粒度、输出格式和评价方式可能不同, 本文不将任务限定为某一种固定输出结构, 而是将每个数据集视为一个具体任务实例. 记数据集集合为

$$\mathcal{D} = \{d_1, d_2, \dots, d_N\}. \quad (1)$$

对于任意数据集 $d \in \mathcal{D}$, 给定输入文本 x , 模型首先生成原始输出 o . 随后, 通过数据集特定的解析函数 $P_d(\cdot)$ 将原始输出转换为该数据集所需的预测结果:

$$\hat{y}_d = P_d(o), \quad (2)$$

其中 $\hat{y}_d \in \mathcal{Y}_d$, $\mathcal{Y}_d$ 表示数据集 d 的输出空间, y_d 表示对应的标注标签.

为了统一不同数据集下的奖励计算, 本文定义一个数据集特定的任务得分函数:

$$\text{Score}_d(\hat{y}_d, y_d) \in [0, 1]. \quad (3)$$

该函数用于衡量模型预测 $\hat{y}_d$ 与标注标签 y_d 之间的一致性. 不同数据集的 Score_d 可以根据其官方评价方式或实验设置进行实例化, 具体计算方式将在实验部分介绍. 在方法部分, 本文仅将其作为统一的任务质量度量.

基于上述定义, 本文的目标是训练一个策略模型 π_θ , 使其在必要时通过主动检索获得外部词汇知识, 并最大化最终预测结果的任务得分:

$$\max_{\theta} \mathbb{E}_{o \sim \pi_\theta(\cdot | x, d)} [\text{Score}_d(P_d(o), y_d)]. \quad (4)$$

3.2 主动词汇检索

为了使模型能够主动查询不确定的词汇, 本文构建了词典式检索环境作为外部知识源. 模型在生成过程中可以选择两类动作. 第一类是检索动作:

$$\langle \text{search} \rangle q_1 | q_2 | \dots | q_m \langle / \text{search} \rangle, \quad (5)$$

其中 $q_1, q_2, \dots, q_m$ 表示模型认为需要查询的词汇. 第二类是回答动作:

$$\langle \text{answer} \rangle \hat{y}_d \langle / \text{answer} \rangle, \quad (6)$$

其中 $\hat{y}_d$ 是模型生成的最终检测结果.

给定模型提出的检索词集合 $\{q_i\}_{i=1}^m$, 检索器返回相应的词汇释义:

$$r = \text{Retriever}(q_1, q_2, \dots, q_m), \quad (7)$$

检索结果 r 会被格式化为自然语言文本, 并作为新的上下文反馈给模型. 与直接将全部外部词典内容拼接到输入中不同, 本文将检索设计为模型的主动行为, 使模型能够根据具体输入判断哪些词汇值得查询, 从而减少无关知识干扰, 并增强检测过程的可解释性.

3.3 两阶段词汇增强 Rollout

为了模拟发现未知词汇、查询释义和完成判断的交互过程, 本文采用两阶段词典增强 rollout 机制, 详细展开过程如算法 1 所示.

在第一阶段, 模型接收原始输入文本 x , 并生成初始输出:

$$o_1 \sim \pi_\theta(\cdot | x, d), \quad (8)$$

算法 1 Two-stage lexicon-augmented rollout.

Input: Training dataset $\mathcal{D}$; policy model π_θ ; lexicon retriever $\mathcal{R}$; scoring function Score_d ;

Output: Policy model π_θ ;

```

1: Initialize  $\pi_\theta$  with the SFT model;
2: for each  $(x_i, y_i, d_i) \in \mathcal{D}$  do
3:   Generate the no-retrieval prediction  $\hat{y}_i^{\text{base}}$ ;
4:   Compute the baseline score  $B_i = \text{Score}_{d_i}(\hat{y}_i^{\text{base}}, y_i)$  according to Eq. (16);
5: end for
6: while not converged do
7:   Sample a batch  $\mathcal{B}$ ;
8:   for each  $(x_i, y_i, d_i) \in \mathcal{B}$  do
9:     Generate  $o^{(1)} \sim \pi_\theta(\cdot \mid x_i, d_i)$ ;
10:    if  $o^{(1)}$  contains <search> then
11:      Retrieve lexical evidence  $r = \mathcal{R}(q)$ ;
12:      Generate  $o^{(2)} \sim \pi_\theta(\cdot \mid x_i, o^{(1)}, r, d_i)$ ;
13:       $o \leftarrow (o^{(1)}, r, o^{(2)})$ ;
14:    else
15:       $o \leftarrow o^{(1)}$ ;
16:    end if
17:    Parse the prediction  $\hat{y}_i = P_{d_i}(o)$ ;
18:    Compute the answer quality reward  $R_{\text{answer}}^{d_i} = \text{Score}_{d_i}(\hat{y}_i, y_i)$ ;
19:    Compute the counterfactual retrieval gain reward  $R_{\text{gain}}^{d_i}$  according to Eq. (25);
20:    Compute the total reward  $R_{d_i} = \lambda_f R_{\text{format}}^{d_i} + \lambda_a R_{\text{answer}}^{d_i} + \lambda_g R_{\text{gain}}^{d_i}$ ;
21:  end for
22:  Update  $\theta$  using GRPO with group-relative advantages;
23: end while
24: return  $\pi_\theta$ .
    
```

如果 o_1 是回答动作, 即模型直接输出 <answer>, 则交互过程结束, 系统将 o_1 作为最终输出进行解析和评分. 若 o_1 是检索动作, 即模型输出 <search> 且没有同时输出 <answer>, 系统首先解析检索词:

$$q_1, q_2, \dots, q_m = \text{ParseSearch}(o_1), \quad (9)$$

然后调用词典检索器:

$$r = \text{Retriever}(q_1, q_2, \dots, q_m). \quad (10)$$

随后, 检索结果 r 被追加到对话上下文中, 模型进入第二阶段生成最终答案:

$$o_2 \sim \pi_\theta(\cdot \mid x, o_1, r, d). \quad (11)$$

因此, 最终输出 o 可以表示为

$$o = \begin{cases} o_1, & \text{如果模型直接回答,} \\ (o_1, r, o_2), & \text{如果模型先进行检索.} \end{cases} \quad (12)$$

需要注意的是, 检索结果 r 由外部环境返回, 并不是模型自身生成的 token. 为了避免将环境文本误认为模型动作, 本文在训练中构建环境掩码 m_t :

$$m_t = \begin{cases} 1, & \text{如果该 token 是模型生成的,} \\ 0, & \text{如果是检索返回的 token.} \end{cases} \quad (13)$$

在策略优化时, 模型生成的 <search> 和 <answer> token 会参与优化, 而检索环境返回的文本不会被直接作为模型生成内容进行优化.

3.4 GRPO 训练

本文在监督微调模型的基础上采用 GRPO^[45] 进行强化学习训练. 对于每个输入样本 (x, y_d, d) , 模型根据当前策略采样 G 个候选输出:

$$\{o_i\}_{i=1}^G \sim \pi_\theta(\cdot | x, d). \quad (14)$$

每个候选输出可能采取不同策略, 系统根据奖励函数为每个候选输出计算奖励 R_i . 随后, GRPO 使用同一输入下候选输出之间的相对奖励来估计优势:

$$A_i = \frac{R_i - \text{mean}(\{R_j\}_{j=1}^G)}{\text{std}(\{R_j\}_{j=1}^G) + \eta}, \quad (15)$$

其中 η 是一个很小的平滑项, 用于避免分母为零. 模型通过提高高优势输出的生成概率、降低低优势输出的生成概率, 逐渐学习更优的主动检索策略.

在训练之前, 本文还为每个训练样本构造无检索基线分数. 具体而言, 使用同一基础模型在利用特殊提示词使模型在不检索的条件下生成预测 $\hat{y}_d^{\text{base}}$, 并计算其任务得分:

$$B_d(x) = \text{Score}_d(\hat{y}_d^{\text{base}}, y_d). \quad (16)$$

该基线分数 $B_d(x)$ 将作为反事实参照, 用于判断主动检索是否真正提升了当前样本的检测表现.

3.5 奖励设计

本文的奖励函数由三部分组成: 格式奖励、答案质量奖励和反事实检索增益奖励. 对于数据集 d , 模型输出 o 的总奖励定义为

$$R_d(o, y_d) = \lambda_g R_{\text{gain}}^d(o, y_d) + \lambda_f R_{\text{format}}^d(o) + \lambda_a R_{\text{answer}}^d(o, y_d), \quad (17)$$

其中, R_{gain}^d 表示反事实检索增益奖励, R_{format}^d 表示格式奖励, R_{answer}^d 表示答案质量奖励. λ_g , λ_f 和 λ_a 分别为反事实增益奖励、格式奖励和答案质量奖励的权重系数, 取值范围设置为 $\lambda_g, \lambda_f, \lambda_a \in [0, 1]$.

3.5.1 格式奖励

格式奖励用于鼓励模型生成可解析、符合交互协议的输出. 该奖励包含两个方面. 首先, 模型需要正确使用动作标签: 当模型需要查询词汇时, 应输出 $\langle \text{search} \rangle$; 当模型给出最终检测结果时, 应输出 $\langle \text{answer} \rangle$. 其次, 模型最终答案应能够被数据集对应的解析函数 $P_d(\cdot)$ 转换为合法预测结果.

因此, 格式奖励可以抽象为

$$R_{\text{format}}^d(o) = R_{\text{tag}}(o) + R_{\text{parse}}^d(o), \quad (18)$$

其中, $R_{\text{tag}}(o)$ 衡量模型是否正确使用检索和回答标签, $R_{\text{parse}}^d(o)$ 衡量模型输出是否符合数据集 d 的输出格式要求. 格式奖励的作用是在强化学习训练早期减少无效输出、格式漂移和不可解析回答.

3.5.2 答案质量奖励

答案质量奖励用于衡量模型最终预测与标注标签之间的一致性. 为了适配不同数据集, 本文不在方法部分固定采用某一种具体指标, 而是统一使用数据集特定任务得分函数 Score_d .

给定模型最终输出 o , 首先通过解析函数得到预测结果:

$$\hat{y}_d = P_d(o). \quad (19)$$

答案质量奖励定义为

$$R_{\text{answer}}^d(o, y_d) = \text{Score}_d(\hat{y}_d, y_d). \quad (20)$$

通过这种抽象, 不同数据集可以使用各自的评价方式计算 Score_d , 而强化学习框架本身保持不变.

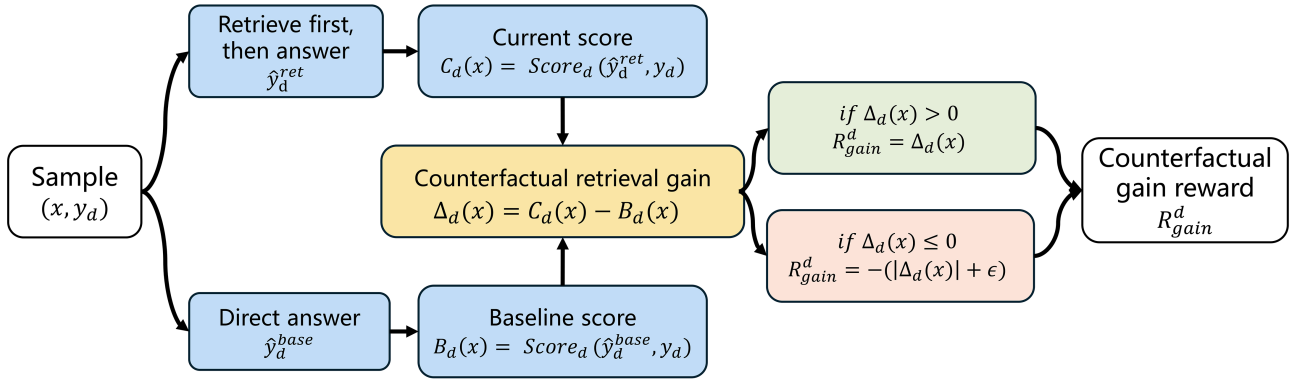


图 2 (网络版彩图) 反事实检索增益计算流程.

Figure 2 (Color online) The computation process of counterfactual retrieval gain.

3.5.3 反事实检索增益

反事实检索增益奖励是本文的核心奖励设计, 其计算流程如图 2 所示. 该奖励用于判断模型发起检索后是否真正提升了当前样本的任务表现, 从而避免模型学习到对所有输入都机械检索的策略.

对于每个样本, 首先使用无检索基线分数:

$$B_d(x) = \text{Score}_d(\hat{y}_d^{\text{base}}, y_d), \quad (21)$$

当模型在当前策略下发起检索并生成最终答案时, 其检索后预测为

$$\hat{y}_d^{\text{ret}} = P_d(o), \quad (22)$$

对应的当前任务分数为

$$C_d(x) = \text{Score}_d(\hat{y}_d^{\text{ret}}, y_d), \quad (23)$$

检索带来的反事实增益定义为

$$\Delta_d(x) = C_d(x) - B_d(x). \quad (24)$$

令 $\mathbb{I}_{\text{search}}$ 表示模型是否发起检索. 反事实检索增益奖励定义为

$$R_{\text{gain}}^d(o, y_d) = \begin{cases} \Delta_d(x), & \mathbb{I}_{\text{search}} = 1 \text{ and } \Delta_d(x) > 0, \\ -(|\Delta_d(x)| + \epsilon), & \mathbb{I}_{\text{search}} = 1 \text{ and } \Delta_d(x) \leq 0, \\ 0, & \mathbb{I}_{\text{search}} = 0, \end{cases} \quad (25)$$

其中, ϵ 是零增益惩罚项. 该项用于惩罚模型的检索未带来性能增益的情况.

该奖励构造了一个反事实比较: 对于同一个输入样本, 比较主动检索后的指标与无检索条件下的基线指标. 如果检索后的分数高于无检索基线, 说明检索为当前样本提供了有效信息, 模型获得正向奖励; 如果检索没有提升任务得分, 模型则受到惩罚. 由此, 模型学习到的不是简单地输出检索标签, 而是一种以任务性能提升为目标的主动信息获取策略.

4 实验结果与分析

4.1 数据集与评价指标

为验证所提出方法的有效性, 本文在 SWSR, ToxiCN 和 STATE-ToxiCN 三个具有代表性的中文有害内容检测数据集上开展实验. 上述数据集均提供了与有害表达相关的黑话词表, 能够为本文构建黑话检索框架并评

估其有效性提供数据支撑. 同时, 三个数据集具有不同平台来源及隐晦表达内容占比, 并涵盖二分类、多分类和多标签分类等多种有害信息检测任务, 有助于从不同数据分布和复杂语义场景下验证本文方法的有效性和泛化能力.

SWSR^[14] 是一个中文性别歧视检测数据集, 包含 8969 条微博评论数据和 3016 个词的有害词词典, 参考数据集所属论文的处理方式, 本文随机选取 90% 的数据作为训练数据, 剩余 10% 的数据为测试数据, 且训练过程中词典全程保持冻结. 该数据集主要用于判断评论是否包含 sexist (性别歧视) 内容, 并进一步识别 category (歧视类型) 和 target (攻击目标), 涉及二分类和多分类任务. 该数据集聚焦中文社交媒体中的性别刻板印象, 微侵犯和性冒犯表达, 适合检验模型对性别相关隐性有害内容的识别能力.

ToxicCN^[15] 是一个中文有害评论分类数据集, 涉及种族、地域、性别等领域, 包含 12011 条来自知乎和贴吧的评论, 以及一个 1032 词的有害词词典, 数据集中包含了训练数据和测试数据文件, 其中训练数据 9600 条, 测试数据 2411 条. 该数据集的标注维度包括 toxic (有害性), type (有害类别), group (目标群体) 和 expression (表达方式), 涉及二分类、多分类和多标签分类任务. 本文分别计算各标签的 Macro-f1 值, 以评估模型在有害性识别、有害类别判断、表达方式区分和目标群体识别方面的性能. 该数据集覆盖显性与隐性有害表达, 并进一步区分一般冒犯性语言和有害信息.

STATE-ToxicCN^[19] 是一个面向细粒度隐性有害信息检测的数据集, 涉及种族、地域、性别等领域, 包含 8029 条来自知乎和贴吧的文本, 以及一个 830 个词的黑话词典, 数据集中原生提供了训练集和测试集文件, 其中训练集 6424 条, 测试集 1605 条. 该数据集的任务目标是抽取 target (目标)、argument (论据)、hateful (有害性) 和 group (目标群体), 涉及二分类、多分类和片段抽取任务. 该数据集强调目标对象、攻击论述和有害类别之间的结构化对应关系, 适合评估模型对隐性有害信息与网络黑话的理解能力.

本文采用 Macro-Precision ($prec$), Macro-Recall (rec) 和 Macro-f1 ($mf1$) 作为主要评价指标. 对于某一识别任务中的类别 c , 其精确率、召回率和 F_1 分别定义为

$$P_c = \frac{TP_c}{TP_c + FP_c}, \quad (26)$$

$$R_c = \frac{TP_c}{TP_c + FN_c}, \quad (27)$$

$$F_{1,c} = \frac{2P_cR_c}{P_c + R_c}, \quad (28)$$

其中, TP_c , FP_c 和 FN_c 分别表示类别 c 的真正例、假正例和假负例. P_c 为精确率, 表示模型预测为类别 c 的样本中实际属于该类别的比例, 其值越高表示误报越少; R_c 为召回率, 表示类别 c 的真实样本中被模型正确识别出的比例, 其值越高表示漏报越少; $F_{1,c}$ 为精确率和召回率的调和平均, 用于综合衡量模型在该类别上的识别性能.

对于包含 C 个类别的识别任务, 首先分别计算各类别的精确率、召回率和 F_1 , 再对各类别指标进行算术平均, 得到 Macro-Precision, Macro-Recall 和 Macro-f1:

$$\text{Macro-Precision} = \frac{1}{C} \sum_{c=1}^C P_c, \quad (29)$$

$$\text{Macro-Recall} = \frac{1}{C} \sum_{c=1}^C R_c, \quad (30)$$

$$\text{Macro-f1} = \frac{1}{C} \sum_{c=1}^C F_{1,c}, \quad (31)$$

其中 Macro-f1 是各类别 F_1 的平均值.

表 2 模型在 SWSR 数据集上的表现. 粗体表示各列的最优结果.

Table 2 Performance of models on the SWSR dataset. The best results in each column are shown in bold.

Method	Sexist			Category			Target		
	<i>prec</i>	<i>rec</i>	<i>mf1</i>	<i>prec</i>	<i>rec</i>	<i>mf1</i>	<i>prec</i>	<i>rec</i>	<i>mf1</i>
BERT _{SFT}	79.1	77.9	78.5	36.3	34.0	35.1	51.2	46.6	48.8
RoBERTa _{SFT}	78.2	79.2	78.7	36.3	35.3	35.8	44.2	51.8	47.7
RoBERTa _{Lexicon} ^[14]	79.1	78.9	79.0	37.8	34.7	36.2	46.7	49.8	48.2
Qwen2.5-3B _{SFT}	79.4	78.0	78.7	40.5	39.1	39.8	52.4	53.3	52.8
Qwen2.5-7B _{SFT}	81.2	79.7	80.4	41.5	39.2	40.3	55.4	54.6	55.0
Qwen2.5-3B _{CALR} (Ours)	81.6	79.4	80.5	41.5	41.7	41.6	53.6	52.6	53.1
Qwen2.5-7B _{CALR} (Ours)	82.7	81.4	82.0	43.0	42.8	42.9	56.3	55.7	56.0

此外, 针对 STATE-ToxicCN 中的片段抽取任务, 本文参考文献 [19] 设置, 进一步采用 Hard 和 Soft 两种匹配方式进行评价. 其中, Hard 要求预测结果与标注结果完全匹配, 对预测结果的完整性和准确性要求较高; Soft 要求 Target 和 Argument 的字符串匹配度达到 0.5, 同时 Hateful 和 Group 与标注结果完全匹配, 更侧重衡量模型是否正确识别了有害表达的核心语义及主要目标. 最终综合分数定义为

$$\text{Score} = \frac{F_{1,\text{Hard}} + F_{1,\text{Soft}}}{2}. \quad (32)$$

4.2 基线方法与参数设置

在三个数据集上, 本文均选取 Qwen2.5-3B 和 Qwen2.5-7B 作为大语言模型基线 ^[19], 并采用监督微调 (supervised fine-tuning, SFT) 进行训练. 为保证不同方法之间的可比性, 所有大语言模型均采用一次性生成全部任务标签的输出方式. 此外, 本文进一步选取 BERT 和 RoBERTa 作为预训练语言模型基线 ^[14]. 对于 SWSR 数据集, 本文增加 RoBERTa Lexicon ^[14] 方法作为引入词典知识的对比基线. 对于 ToxicCN 数据集, 进一步加入 TKE 方法 ^[15] 作为词典增强基线. 对于 STATE-ToxicCN 数据集, 由于该任务涉及多四元组抽取, 传统分类模型难以直接适应该结构化生成任务, 因此本文仅对 Qwen2.5 系列大语言模型进行对比分析.

本文采用 LoRA (low-rank adaptation) ^[46] 对大语言模型进行参数高效微调, 以降低训练成本并提高微调效率. LoRA 的秩 (Rank) 和缩放系数 (Alpha) 分别设置为 32 和 64 ^[47]. 对于其余超参数, 本文从各数据集训练集中随机划分 10% 的样本作为验证集, 其余 90% 用于模型训练 ^[48], 并对不同候选参数配置进行实验比较, 最终综合三个数据集上的验证结果, 选取整体表现最优的参数配置用于后续实验. 经验证集实验确定, SFT 阶段的训练批大小设置为 4, 学习率设置为 2×10^{-4} , 训练轮次为 3; CALR 训练阶段的批大小设置为 4, 学习率设置为 5×10^{-7} , 训练轮次为 5. 对于 CALR 的奖励函数, 反事实增益奖励是本文的主要优化目标, 因此其权重固定为 1. 由于经过 SFT 阶段训练后, 模型已能够较稳定地遵循预设输出格式, 因此将格式奖励权重设置为较小的 0.05. 答案质量奖励的权重则经过验证集调优, 最终设置为 0.1.

4.3 主要实验结果

表 2~4 分别展示了本文方法和基线方法在 SWSR, ToxicCN 和 STATE-ToxicCN 数据集上的整体表现, 每个结果是在当前设置下运行 5 次选取的最优结果. 在 STATE-ToxicCN 数据集的评估中, Hard 和 Soft 分别表示严格匹配与宽松匹配条件下的 F_1 值, 综合指标 Score 为二者的算术平均, 即 $\text{Score} = (F_{1,\text{Hard}} + F_{1,\text{Soft}})/2$. 为便于呈现和比较, 表中 *prec*, *rec* 和 *mf1* 等指标均以百分制形式报告, 后续表格和图中的同类指标亦采用相同表示方式. 从中可以看出, 基于 Qwen2.5 的模型在多个有害信息识别任务中均取得了较为突出的性能表现. 在引入 CALR 方法后, 相较于 SFT, 模型在大多数评价指标上均获得一致性提升, 表明 CALR 能够有效增强模型对复杂有害语义特征的建模能力.

在 SWSR 数据集上, RoBERTa+Lexicon 在有害性 F_1 上达到 79.0, 超过 BERT 和 RoBERTa 等传统基线模

表 3 模型在 ToxiCN 数据集上的表现. 粗体表示各列的最优结果.

Table 3 Performance of models on the ToxiCN dataset. The best results in each column are shown in bold.

Method	Toxic			Type			Group			Expression		
	<i>prec</i>	<i>rec</i>	<i>mf1</i>	<i>prec</i>	<i>rec</i>	<i>mf1</i>	<i>prec</i>	<i>rec</i>	<i>mf1</i>	<i>prec</i>	<i>rec</i>	<i>mf1</i>
BERT _{SFT}	81.9	81.9	81.9	71.9	70.1	71.0	62.4	49.2	55.0	63.2	61.1	62.1
RoBERTa _{SFT}	82.7	82.6	82.6	73.2	74.0	73.6	63.2	50.9	56.4	65.4	65.3	65.3
RoBERTa _{TKE} ^[15]	83.0	82.9	82.9	74.8	73.4	74.1	64.1	51.2	56.9	65.7	65.9	65.8
Qwen2.5-3B _{SFT}	84.4	83.6	84.0	81.3	80.8	81.0	78.2	80.1	79.1	71.0	72.0	71.5
Qwen2.5-7B _{SFT}	84.5	83.5	84.0	82.1	81.0	81.5	77.6	80.2	78.9	72.0	72.6	72.3
Qwen2.5-3B _{CALR} (Ours)	85.2	85.2	85.2	82.7	82.6	82.6	82.5	82.0	82.2	75.2	74.9	75.0
Qwen2.5-7B _{CALR} (Ours)	85.8	85.7	85.7	83.3	83.5	83.4	82.8	83.3	83.0	76.2	76.3	76.2

表 4 模型在 STATE-ToxiCN 数据集上的表现. 粗体表示各列的最优结果.

Table 4 Performance of models on the STATE-ToxiCN dataset. The best results in each column are shown in bold.

Method	Score	Hard			Soft		
		<i>prec</i>	<i>rec</i>	<i>mf1</i>	<i>prec</i>	<i>rec</i>	<i>mf1</i>
Qwen2.5-3B _{SFT}	28.5	18.9	19.1	19.0	37.6	38.1	37.8
Qwen2.5-7B _{SFT}	34.2	25.3	24.4	24.9	44.2	42.3	43.3
Qwen2.5-3B _{CALR} (Ours)	33.4	22.6	24.2	23.4	41.9	44.9	43.4
Qwen2.5-7B _{CALR} (Ours)	36.0	25.4	25.7	25.5	46.2	46.6	46.4

型, 说明引入词典信息在一定程度上有助于提升模型对有害倾向的识别能力. 然而, 与 Qwen2.5 系列模型相比, 传统模型在有害类型与目标识别任务上的表现仍存在明显差距, 难以超越较小规模的 Qwen2.5-3B 模型. 进一步对比 SFT 与 CALR 方法可以发现, CALR 在 SWSR 上带来的性能提升较大, F_1 值的提升幅度约为 0.3~2.6, 表明该方法在不同参数规模模型上均具有一定的鲁棒增益.

在 ToxiCN 数据集中, 存在较多难以理解的网络黑话^[15], 对于模型的语义理解能力要求更高. BERT, RoBERTa 以及引入 TKE (Toxic Knowledge Enhancement, 有害知识增强) 的 RoBERTa 等传统方法在部分指标上取得一定效果, 但整体仍落后于 Qwen2.5 系列模型, 尤其在表达方式与目标群体识别任务上差距更为明显. 这表明传统模型在复杂语义推理与隐含信息建模方面能力有限. 进一步对比 SFT 与 CALR 在 Qwen2.5 系列模型上的结果可以看到, CALR 带来的 F_1 值整体提升幅度为 1.2~4.1, 其中 group 与 expression 识别任务的提升最大 (3.1~4.1). 这一结果表明, CALR 对依赖上下文推理、语义关联与隐含信息理解的细粒度任务具有更强的促进作用.

在 STATE-ToxiCN 数据集中, 同样存在较多难以理解的网络黑话^[19], 同时样本原文中包含目标与论据等复杂结构信息, 对模型的细粒度抽取能力提出了更高要求. 实验结果显示 $F_{1, \text{Soft}}$ 整体高于 $F_{1, \text{Hard}}$, 说明模型在语义近似标签匹配上具有一定能力, 但在严格的精确匹配设置下仍存在一定困难, 这也反映出该数据集对细粒度语义判别能力的较高要求, 尤其是在多维标签组合与隐含语义识别场景下, 模型仍需进一步提升边界判别能力. 在最终性能上, CALR 对 3B 模型的综合 Score 提升幅度为 4.9, 而对 7B 模型的提升为 1.8, 表明 CALR 对较小规模模型的增益效果更好. 这可能是由于小模型内部知识容量有限, 对外部语义增强与检索信息的依赖更强, 从而使 CALR 的作用更加明显.

总体而言, 三个数据集在识别任务、数据分布以及网络平台等方面均存在明显差异, 同时实验涵盖了 3B 与 7B 两种不同规模的基础模型. CALR 在不同数据集和不同模型规模下相较于基线模型均取得了不同程度的性能提升, 表明该方法能够较好地适应不同任务、数据分布与模型规模, 体现出较好的跨领域、跨平台及跨模型泛化能力.

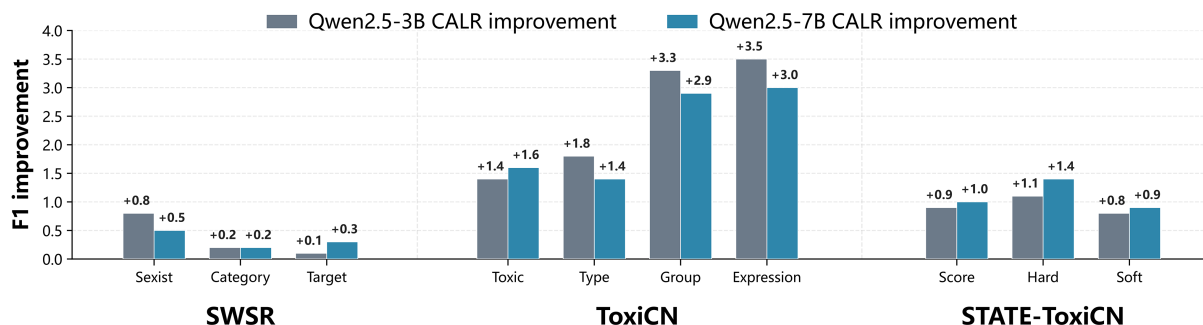


图 3 (网络版彩图) 强化学习策略性能表现.

Figure 3 (Color online) The performance of the reinforcement learning policy.

4.4 强化学习策略有效性验证

为进一步验证强化学习检索策略的有效性,本节构建基于规则匹配检索的基线方法,并在相同外部知识资源下进行对比,旨在排除外部知识来源差异的影响,重点分析强化学习策略相较于规则匹配方法在有害信息识别中的作用.该方法直接利用词库中的词汇在输入文本中进行字符串匹配,并将匹配到的词及其对应释义拼接至原始输入中,来构建基于规则驱动的对比如实验.图3展示了强化学习检索策略相对于匹配规则检索所带来的性能增益.

从整体结果来看,引入强化学习检索策略的模型在大多数指标上优于匹配检索方法,表明基于策略学习的检索机制能够更有效地提升模型性能.然而,在SWSR数据集上,强化学习策略在有害类型和目标识别任务中的提升并不明显,且相较于ToxiCN与STATE-ToxiCN数据集(整体性能提升范围分别为1.4~3.5和0.8~1.4),SWSR上的提升幅度较小.这一差异主要源于不同数据集的语言特性:SWSR中隐喻性表达和复杂黑话较少^[14],文本语义相对直接,LLM本身即可较好地理解其含义,因此匹配检索带来的错误触发和语义噪声影响较弱,使得强化学习策略的优势难以充分体现.相比之下,在包含更多隐晦表达和非规范用语的数据集上,匹配检索更容易引入无关或误匹配信息,从而干扰模型判断.这一结果表明,本文所提出的策略能够根据语义上下文自适应地选择是否以及检索何种词汇,从而有效减少无关词触发带来的冗余检索与噪声干扰,使检索机制从传统的规则驱动范式转变为语义驱动的决策过程,进而提升下游有害信息检测任务的整体性能.

4.5 反事实检索增益奖励有效性验证

在上节验证强化学习检索策略有效性的基础上,本节进一步通过移除反事实检索增益奖励构建对比实验,仅保留格式奖励与答案质量奖励,并基于Qwen2.5-3B和Qwen2.5-7B在三个数据集上进行实验对比,以验证反事实检索增益奖励对强化学习检索策略优化的有效性.图4展示了引入反事实检索增益奖励后,模型在不同数据集与评价指标上的性能提升幅度.从实验结果可以看出,引入反事实检索增益奖励的模型在各数据集及多数评价指标上均优于不使用该奖励的基线模型,表明该奖励机制能够为检索策略学习提供有效的优化信号,从而提升整体检测性能.在ToxiCN与STATE-ToxiCN数据集上,该奖励带来的性能提升较为明显,其中ToxiCN的 F_1 提升幅度为0.3~2.3,STATE-ToxiCN为1.1~2.3.相比之下,在SWSR数据集上,其增益相对较小,尤其在有害类型识别任务中表现不明显.相比其他两个数据集,SWSR中隐喻表达与黑话现象相对较少,文本语义更为显式,大语言模型本身即可较好地完成语义理解,因此性能瓶颈并不主要来自词汇层面的语义缺失.在该情况下,通过反事实检索增益奖励进行额外语义增强所带来的收益有限,使得该奖励机制的优势难以充分体现.总体而言,引入反事实检索增益奖励的模型表现更优,说明该奖励机制能够有效引导模型倾向于选择更具信息价值的检索词汇,从而优化检索决策过程,实现更具针对性的语义增强,并最终提升有害信息检测任务的整体性能.

4.6 检索效率分析

在前面验证所提方法有效性的基础上,本节进一步通过检索次数、正负增益比、净增益率和负增益率四项

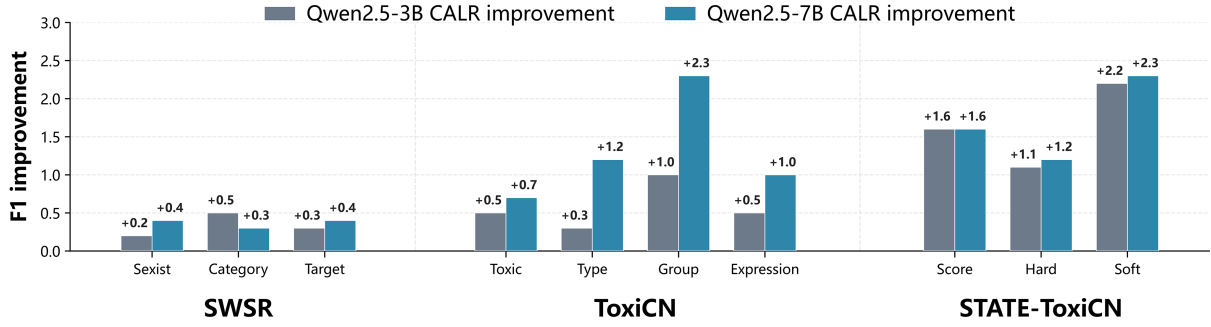


图 4 (网络版彩图) 反事实增益奖励性能表现.

Figure 4 (Color online) The performance of counterfactual retrieval gain.

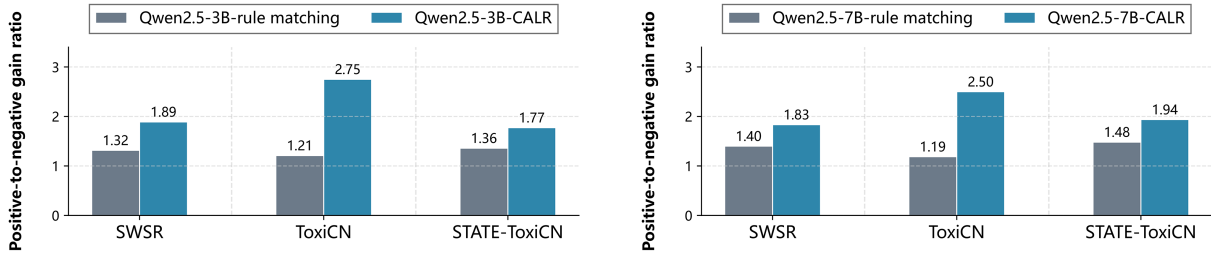


图 5 (网络版彩图) 正负增益比.

Figure 5 (Color online) Positive-to-negative gain ratio.

指标, 从检索行为层面分析其性能提升来源, 揭示所提出的强化学习方法在减少无效检索、提升高价值知识调用比例和抑制负增益检索方面的作用. 其中, 检索次数用于衡量不同方法触发外部知识检索的频率; 正负增益比表示检索后模型性能提升样本与性能下降样本之间的比例, 用于量化检索收益是否能够超过其潜在风险; 净增益率表示正增益样本数与负增益样本数之差占总检索次数的比例, 用于衡量单次检索所带来的实际有效收益; 负增益率则表示负增益样本数占总检索次数的比例, 用于评估检索过程引入噪声并导致模型性能下降的风险. 各个指标的计算方式如下:

$$\text{正负增益比} = \frac{N_{pos}}{N_{neg}}, \quad (33)$$

$$\text{净增益率} = \frac{N_{pos} - N_{neg}}{N_{ret}} \times 100\%, \quad (34)$$

$$\text{负增益率} = \frac{N_{neg}}{N_{ret}} \times 100\%, \quad (35)$$

其中, N_{pos} 表示检索后带来性能提升的正增益次数, N_{neg} 表示检索后导致性能下降的负增益次数, 在极端情况下, 若未出现负增益样本, 正负增益比趋于无穷大, N_{ret} 表示总检索次数. 本文中, 净增益率和负增益率均采用百分制表示, 即首先根据样本计数计算其比例值, 再乘以 100% 转换为百分数.

图 5~8 展示了 CALR 在基准数据集下对比规则匹配方法在上述评价指标下的表现. 从中可以看出, CALR 在不同模型规模下均明显优于基于匹配规则的检索策略. 具体来说, 匹配规则方法检索次数多, 平均正负增益比为 1.32, 平均净增益率为 3.35%, 平均负增益率为 11.54%; 而 CALR 使用了更少的检索次数, 但平均正负增益比提升至 2.11, 平均净增益率提升至 8.13%, 平均负增益率降低至 7.74%. 这说明 CALR 通过更高质量的检索提升了检索带来的正向收益, 还有效减少了错误检索或噪声检索带来的负面影响. 此外, 在 CALR 策略下, 3B 模型的检索次数在三个数据集上均多于 7B 模型, 说明检索对于小模型带来的收益普遍高于规模较大的模型, 小模型对于检索和外部知识的依赖度更高.

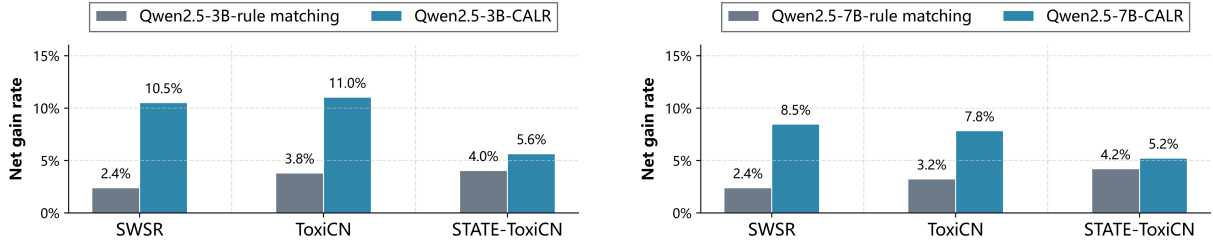


图 6 (网络版彩图) 净增益率.

Figure 6 (Color online) Net gain rate.

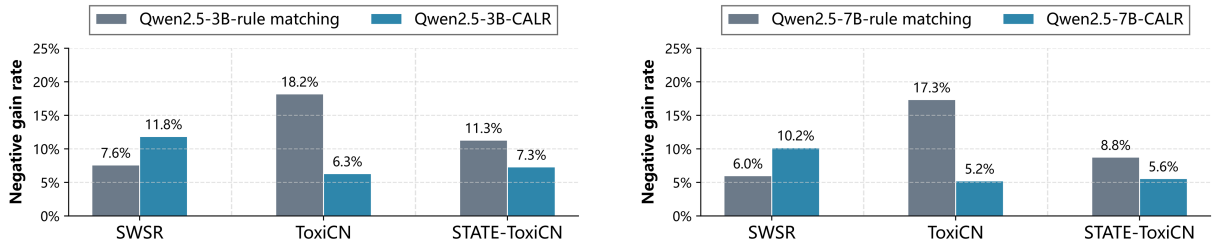


图 7 (网络版彩图) 负增益率.

Figure 7 (Color online) Negative gain rate.

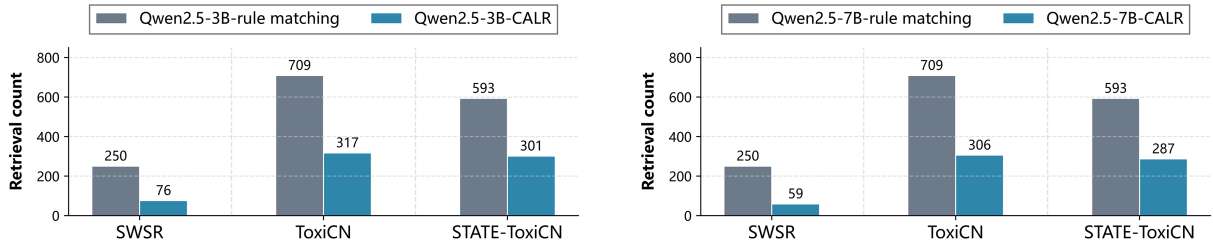


图 8 (网络版彩图) 检索次数.

Figure 8 (Color online) Retrieval count.

5 结论

针对现有语义增强方法在检索触发和知识选择方面的不足, 本文提出一种基于反事实奖励的主动词典检索方法 CALR. 该方法将外部知识检索过程建模为可学习的决策问题, 使大模型能够根据输入文本的上下文语义动态判断是否需要调用外部知识, 并确定需要重点检索的关键信息, 从而实现对外部知识的按需调用与语义增强. 在识别任务 (性别歧视、种族主义等)、数据分布及网络平台 (知乎、百度贴吧等) 均存在差异的 SWSR, ToxiCN 和 STATE-ToxiCN 三个数据集上的实验结果表明, CALR 在不同模型规模下均能取得有效的性能提升, 验证了该方法在不同任务与应用场景下的有效性和泛化能力. 此外, CALR 在包含较多隐喻表达、隐式有害语义和网络黑话的数据集上提升更为明显, 说明该方法能够有效缓解传统匹配式检索可能引入的错误匹配和语义噪声问题.

本文提出的方法在隐晦表达场景下取得了较好的检测效果, 但从实验结果可以看出, 在有害语义较为显式、上下文歧义程度较低的数据集上, CALR 的性能提升幅度仍相对有限, 说明当基础模型已具备较强语义建模能力时, 外部知识检索所带来的增益相对有限. 此外, 当前方法的检索效果仍依赖于词典对待检测表达的覆盖程度. 当面对训练阶段和现有词典中未出现的网络黑话、新兴缩写、谐音变体等表达时, 模型可能由于缺乏有效的检索线索而难以获取相关知识; 同时, 网络隐晦表达具有持续演化的特点, 其语义可能随时间、事件和网络社群发生变化, 静态或更新不及时词典难以充分覆盖此类动态语义. 另一方面, 本文所提的方法主要面向文本形式的有害信息检测, 然而, 在真实网络场景中, 有害意图往往并非仅由文本单独表达, 还会结合表情、图片等表达形

式共同呈现. 因此, 未来工作将进一步研究面向动态网络语言的知识更新与开放词汇检索机制, 在现有反事实检索增益的基础上加强对检索行为、检索知识与性能增益之间细粒度因果关联的建模, 以提升检索决策的准确性与可解释性^[49], 同时开展面向多模态内容的检索增强方法研究, 从而提升模型对持续演化的隐晦表达及跨模态有害意图的识别能力^[50].

参考文献

- Fortuna P, Nunes S. A survey on automatic detection of hate speech in text. *ACM Comput Surv*, 2018, 51: 1–30
- Xue X, Li G, Zhou D, et al. Research roadmap of service ecosystems: a crowd intelligence perspective. *Int J Crowd Sci*, 2022, 6: 195–222
- Zhang D Y, Lu J Y, Min C R, et al. The technologies and practices of detecting online violent speech. *Chin J Lang Policy Plan*, 2024, 9: 64–75 [张冬瑜, 卢俊宇, 闵昶荣, 等. 网络暴力言论检测的技术和实践. *语言战略研究*, 2024, 9: 64–75]
- Zheng X, Zeng D, Wang F Y. Social balance in signed networks. *Inf Syst Front*, 2015, 17: 1077–1095
- Zheng X, Tian H, Wan Z, et al. Game starts at GameStop: characterizing the collective behaviors and social dynamics in the short squeeze episode. *IEEE Trans Comput Soc Syst*, 2021, 9: 45–58
- Mathew B, Dutt R, Goyal P, et al. Spread of hate speech in online social media. In: *Proceedings of the 10th ACM Conference on Web Science*, 2019. 173–182
- Xue X, Shen Y, Yu X, et al. Computational experiments: a new analysis method for cyber-physical-social systems. *IEEE Trans Syst Man Cybern Syst*, 2024, 54: 813–826
- Zheng X L, Bai S R, Zeng D J. Construction and analysis of cognitive graphs for complex decision-making scenarios. *J Manag World*, 2023, 5: 188–199 [郑晓龙, 白松冉, 曾大军. 面向复杂决策场景的认知图谱构建与分析. *管理世界*, 2023, 5: 188–199]
- Wu J X, Gao B B, Wei X S, et al. Resource-constrained deep learning: challenges and practices. *Sci Sin Inform*, 2018, 48: 501–510 [吴建鑫, 高斌斌, 魏秀参, 等. 资源受限的深度学习: 挑战与实践. *中国科学: 信息科学*, 2018, 48: 501–510]
- Zhao W X, Zhou K, Li J, et al. A survey of large language models. *Front Comput Sci*, 2026, 20: 2012627
- Xue X, Wang S, Zhang L, et al. Social learning evolution (SLE): computational experiment-based modeling framework of social manufacturing. *IEEE Trans Ind Inf*, 2019, 15: 3343–3355
- Xiao Y, Hu Y, Choo K T W, et al. ToxiCloakCN: evaluating robustness of offensive language detection in Chinese with cloaking perturbations. In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024. 6012–6025
- Zhang X W, Zheng X L, Mao W J. Adversarial perturbation defense on deep neural networks. *ACM Comput Surv*, 2021, 54: 1–36
- Jiang A, Yang X, Liu Y, et al. SWSR: a Chinese dataset and lexicon for online sexism detection. *Online Soc Netws Media*, 2022, 27: 100182
- Lu J, Xu B, Zhang X, et al. 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*, 2023. 16235–16250
- Lewis P, Perez E, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In: *Proceedings of Advances in Neural Information Processing Systems*, 2020. 9459–9474
- Jin B, Zeng H, Yue Z, et al. Search-R1: training LLMs to reason and leverage search engines with reinforcement learning. In: *Proceedings of the Second Conference on Language Modeling*, 2025
- Asai A, Wu Z, Wang Y, et al. Self-RAG: learning to retrieve, generate, and critique through self-reflection. In: *Proceedings of the International Conference on Learning Representations*, 2024
- Bai Z, Yang L, Yin S, et al. STATE-ToxiCN: a benchmark for span-level target-aware toxicity extraction in Chinese hate speech detection. In: *Proceedings of Findings of the Association for Computational Linguistics: ACL 2025*, 2025. 10206–10219
- Warner W, Hirschberg J. Detecting hate speech on the World Wide Web. In: *Proceedings of the Second Workshop on Language in Social Media*, 2012. 19–26
- Burnap P, Williams M L. Cyber hate speech on Twitter: an application of machine classification and statistical modeling for policy and decision making. *Policy Internet*, 2015, 7: 223–242
- Nobata C, Tetreault J, Thomas A, et al. Abusive language detection in online user content. In: *Proceedings of the 25th International Conference on World Wide Web*, 2016. 145–153
- Badjatiya P, Gupta S, Gupta M, et al. Deep learning for hate speech detection in tweets. In: *Proceedings of the 26th International Conference on World Wide Web Companion*, 2017. 759–760
- Pitsilis G K, Ramampiaro H, Langseth H. Effective hate-speech detection in Twitter data using recurrent neural networks. *Appl Intell*, 2018, 48: 4730–4742
- Zhang Y, Xu B, Zhao T. CN-HIT-MI.T at SemEval-2019 Task 6: offensive language identification based on BiLSTM with double attention. In: *Proceedings of the 13th International Workshop on Semantic Evaluation*, 2019. 564–570
- Ribeiro A, Silva N. INF-HatEval at SemEval-2019 Task 5: convolutional neural networks for hate speech detection against women and immigrants on Twitter. In: *Proceedings of the 13th International Workshop on Semantic Evaluation*, 2019. 420–425
- Sedek N S B M, Ong Y C. Deep learning for multilingual hate speech detection in social media. In: *Proceedings of the 7th International Conference on Interdisciplinary Computer Science and Engineering*, 2025. 82–88

- 28 Khan S, Fazil M, Sejwal V K, et al. BiCHAT: BiLSTM with deep CNN and hierarchical attention for hate speech detection. *J King Saud University-Comput Inf Sci*, 2022, 34: 4335–4344
- 29 Deng J, Zhou J, Sun H, et al. COLD: a benchmark for Chinese offensive language detection. In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022. 11580–11599
- 30 Dai W, Yu T, Liu Z, et al. Kungfupanda at SemEval-2020 Task 12: BERT-based multi-task learning for offensive language detection. In: *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, 2020. 2060–2066
- 31 Chhablani G, Sharma A, Pandey H, et al. NLRG at SemEval-2021 Task 5: toxic spans detection leveraging BERT-based token classification and span prediction techniques. In: *Proceedings of the 15th International Workshop on Semantic Evaluation*, 2021. 233–242
- 32 Che W X, Dou Z C, Feng Y S, et al. Towards a comprehensive understanding of the impact of large language models on natural language processing: challenges, opportunities and future directions. *Sci Sin Inform*, 2023, 53: 1645–1687 [车万翔, 窦志成, 冯岩松, 等. 大模型时代的自然语言处理: 挑战、机遇与发展. *中国科学: 信息科学*, 2023, 53: 1645–1687]
- 33 Guo K, Hu A, Mu J, et al. An investigation of large language models for real-world hate speech detection. In: *Proceedings of the 2023 International Conference on Machine Learning and Applications*, 2023. 1568–1573
- 34 Deng J, Chen Z, Sun H, et al. Enhancing offensive language detection with data augmentation and knowledge distillation. *Research*, 2023, 6: 0189
- 35 Jin Y, Wanner L, Shvets A. GPT-HateCheck: can LLMs write better functional tests for hate speech detection? In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, 2024. 7867–7885
- 36 Cheremetiev V, Ngo Q L H, Kot C Y, et al. Specializing general-purpose LLM embeddings for implicit hate speech detection across datasets. In: *Proceedings of the 2nd International Workshop on Diffusion of Harmful Content on Online Web*, 2025. 23–36
- 37 Razavi A H, Inkpen D, Uritsky S, et al. Offensive language detection using multi-level classification. In: *Advances in Artificial Intelligence*. Berlin: Springer, 2010. 16–27
- 38 Bassignana E, Basile V, Patti V. Hurllex: a multilingual lexicon of words to hurt. In: *Proceedings of the Fifth Italian Conference on Computational Linguistics*, 2018. 52–57
- 39 Lee H S, Lee H R, Park J U, et al. An abusive text detection system based on enhanced abusive and non-abusive word lists. *Decision Support Syst*, 2018, 113: 22–31
- 40 Yuan W K, Zhang Y J, Bao Y Y, et al. A hate speech detection approach based on self-taught reasoning enhancement with large language models. *Data Anal Knowl Discov*, 2026, 10: 90–102 [袁伟康, 张予捷, 包易云, 等. 基于大语言模型自学习推理增强的仇恨言论检测方法. *数据分析与知识发现*, 2026, 10: 90–102]
- 41 Lin J. Leveraging world knowledge in implicit hate speech detection. In: *Proceedings of the Second Workshop on NLP for Positive Impact*, 2022. 31–39
- 42 Stranisci M A, Frenda S, Lai M, et al. O-dang! The ontology of dangerous speech messages. In: *Proceedings of the 2nd Workshop on Sentiment Analysis and Linguistic Linked Data*, 2022. 2–8
- 43 Zhao Y, Zhu J, Xu C, et al. Enhancing LLM-based hatred and toxicity detection with Meta-Toxic knowledge graph. In: *Proceedings of Findings of the Association for Computational Linguistics: ACL 2025*, 2025. 24747–24760
- 44 Mnassri K, Farahbakhsh R, Crespi N. RAG and recall: multilingual hate speech detection with semantic memory. In: *Proceedings of the 9th Workshop on Online Abuse and Harms*, 2025. 219–227
- 45 Tong C, Guo Z, Zhang R, et al. Delving into RL for image generation with CoT: a study on DPO vs. GRPO. In: *Proceedings of Advances in Neural Information Processing Systems*, 2025. 38: 115632–115655
- 46 Hu E J, Shen Y, Wallis P, et al. LoRA: low-rank adaptation of large language models. In: *Proceedings of the International Conference on Learning Representations*, 2022
- 47 Lin F J, Zhang Y W, Huang Y, et al. System report for CCL25-Eval task 10: improving LLMs on fine-grained Chinese hate speech detection. In: *Proceedings of the 24th China National Conference on Computational Linguistics*, 2025. 411–418 [林凡钧, 张晏玮, 黄杨, 等. 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. 见: 第二十四届中国计算语言学会大会, 2025. 411–418]
- 48 Rosenthal S, Atanasova P, Karadzhov G, et al. SOLID: a large-scale semi-supervised dataset for offensive language identification. In: *Proceedings of Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021. 915–928
- 49 Deng Z, Tian H, Zheng X, et al. Deep causal learning: representation, discovery and inference. *ACM Comput Surv*, 2026, 58: 1–36
- 50 Zhang X, Zhou G, Zheng X, et al. Deep supervised adversarial robust hashing for retrieval. *IEEE Trans Pattern Anal Mach Intell*, 2026, doi: 10.1109/TPAMI.2026.3725315

A reinforcement learning-driven method for intelligent harmful information detection

Shuo YAN^{1,2}, Xingwei ZHANG³, Chunjie ZHANG^{1,2} & Xiaolong ZHENG^{3,4*}

1. *Institute of Information Science, School of Computer Science and Technology, Beijing Jiaotong University, Beijing 100044, China*

2. *Intelligence + X International Cooperation Joint Laboratory of MOE, School of Computer Science and Technology, Beijing Jiaotong University, Beijing 100044, China*

3. *State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China*

4. *University of Chinese Academy of Sciences, Beijing 100190, China*

* Corresponding author. E-mail: xiaolong.zheng@ia.ac.cn

Abstract Harmful information detection is of great significance for maintaining the security of the online ecosystem and preventing social risks, and constitutes a key task in digitalized social governance. In recent years, harmful information detection methods based on artificial intelligence, especially large language models, have made substantial progress in automated identification. Meanwhile, in real-world online environments where implicit expressions frequently emerge and internet slang continues to evolve, incorporating external semantic information can help supplement models' understanding of contextual semantics and implicit harmful intent, thereby further improving harmful information detection performance. However, existing external knowledge-based semantic enhancement methods mostly rely on static matching or rule-triggered retrieval, making it difficult for models to adaptively determine whether external knowledge is necessary and what specific knowledge should be introduced. This may introduce semantic noise and limit the discriminative performance and generalization ability of the model. To address these issues, this paper proposes an active lexicon retrieval method based on counterfactual rewards. The proposed method formulates the external knowledge retrieval process as a policy learning problem. Through reinforcement learning, the model adaptively decides whether to retrieve external knowledge and which key semantic terms to retrieve. Moreover, a counterfactual retrieval gain reward mechanism is designed to guide the model toward high-value retrieval behaviors that make actual contributions to the discrimination task, thereby enabling a shift from a rule-driven retrieval paradigm to a semantic-driven retrieval paradigm. Experimental results on SWSR, ToxiCN, and STATE-ToxiCN, which exhibit varying levels of linguistic complexity and different distributions of implicit harmful expressions, demonstrate that the proposed method consistently improves performance across different detection tasks, data distributions, and model scales (3B and 7B). The improvements are particularly pronounced in scenarios with a higher prevalence of metaphorical expressions and internet slang, validating the effectiveness and generalizability of the proposed method in complex online semantic environments.

Keywords digitalized social governance, harmful information detection, reinforcement learning, large language models, counterfactual reward