STATE ToxiCN: A Benchmark for Span-level Target-Aware Toxicity Extraction in Chinese Hate Speech Detection

Zewen Bai[a], Liang Yang[*a,b], Shengdi Yin[a], Junyu Lu[a], Jingjie Zeng[a]
Haohao Zhu[a], Yuanyuan Sun[a], Hongfei Lin[a]
[a]School of Computer Science and Technology,
[b]Key Laboratory of Social Computing and Cognitive Intelligence
Dalian University of Technology, China
dlutbzw@mail.dlut.edu.cn, (liang,hflin)@dlut.edu.cn

Abstract

The proliferation of hate speech has caused significant harm to society. The intensity and directionality of hate are closely tied to the target and argument it is associated with. However, research on hate speech detection in Chinese has lagged behind, and existing datasets lack span-level fine-grained annotations. Furthermore, the lack of research on Chinese hateful slang poses a significant challenge. In this paper, we provide two valuable fine-grained Chinese hate speech detection research resources. First, we construct a **S**pan-level **T**arget-**A**ware **T**oxicity **E**xtraction dataset (STATE TOXICN), which is the first span-level Chinese hate speech dataset. Secondly, we evaluate the span-level hate speech detection performance of existing models using STATE TOXICN. Finally, we conduct the first study on Chinese hateful slang and evaluate the ability of LLMs to understand hate semantics. Our work contributes valuable resources and insights to advance span-level hate speech detection in Chinese. [1]

Disclaimer: *The samples presented by this paper may be considered offensive or vulgar.*

1 Introduction

With the popularity of social media, user-generated content has experienced explosive growth, and hate speech has also flourished. Hate speech refers to harmful statements that express hatred or incite harm against specific groups or individuals based on race, religion, gender, region, sexual orientation, physical characteristics, and other factors (Bilewicz and Soral, 2020). Due to its damaging impact on individuals and society, hate language is now widely considered as a problem of increasing importance (Silva et al., 2016). Recently, researchers have been actively engaged in the study of hate speech detection, and this research has gradually shifted from post-level (Ahn et al., 2024; AlKhamissi et al., 2022b) to span-level (Pavlopoulos et al., 2021; Mathew et al., 2021; Zampieri et al., 2023).

Approximately 941 million people, or 12% of the global population, speak Mandarin Chinese as their first language (Eberhard et al., 2024). However, research on Chinese hate speech detection lags significantly behind. There are two key issues that remain unresolved. Firstly, existing research (Deng et al., 2022; Jiang et al., 2022; Zhou et al., 2022) is limited to the post-level, leaving span-level Chinese hate speech detection unexplored. The intensity and directionality of hate speech are closely tied to the target and argument it is associated with (Cowan and Hodge, 1996). In the Chinese linguistic context, challenges in span-level hate speech detection arise due to flexible word order and the absence of word delimiters like spaces. These issues complicate the identification of targets and arguments. For example, as shown in Exp.1 in Table 1, the use of inversion in Chinese disrupts the conventional subject-verb-object structure. To fill this gap, we construct a span-level Chinese hate speech dataset by annotating the Target-Argument-Hateful-Group quadruples in posts.

Secondly, although Chinese hate lexicons have been developed (Jiang et al., 2022; Lu et al., 2023), the absence of interpretable annotations hinders a deeper understanding of hate speech. As an ideographic language, Chinese has a wealth of synonyms and near-synonyms (Mair, 1991), making the forms and types of hateful slang diverse and difficult to capture. Chinese hateful slang often evades model detection through techniques such as homophonic substitution, character splitting and merging, and historical allusions (Xiao et al., 2024). For example, in Exp. 3 from Table 1, the target uses the merging technique to evade detection. To address this issue, we collect commonly used hateful slang from real-world online forums, providing

* Corresponding Author

[1]Code and datasets are publicly available at https://github.com/shenmeyemeifashengguo/STATE-ToxiCN.

Findings of the Association for Computational Linguistics: ACL 2025, pages 10206–10219

Exp.	Post	Target	Argument	Hateful	Group
1	你这头蠢驴，没人会喜欢。 *No one will ever like you, you idiot.*	你 *you*	蠢驴 *idiot*	non-hate	non-hate
2	男同是艾滋高发群体。 *Gay people are a high-risk group for HIV.*	男同 *Gay people*	艾滋高发群体 *a high-risk group for HIV*	hate	LGBTQ, others
3	默我是真的很讨厌。 *Silence, I really hate it.*	默(黑犬) *Slience(black dog)*	讨厌 *hate*	hate	Racism

Table 1: Examples of annotated posts from STATE TOXICN dataset with corresponding annotations of Target-Argument-Hateful-Group quadruples.

detailed annotations to construct the first annotated lexicon for Chinese hateful slang. This lexicon serves as a foundational resource for research into the understanding of Chinese hate semantics.

To address the lack of resources for span-level Chinese hate speech research, we introduce a **S**pan-level **T**arget-**A**ware **T**oxicity **E**xtraction dataset (STATE TOXICN), a novel dataset containing 8029 posts and 9533 Target-Argument-Hateful-Group quadruples addressing sexism, racism, regional bias, and anti-LGBTQ sentiments. Using this dataset, we evaluate the performance of LLMs in span-level Chinese hate speech detection. Specially, we annotate: 1) extraction of the targets and arguments from the post, 2) determination of whether each Target-Argument pair constitutes hate speech, and 3) classification of the groups for hateful Target-Argument pairs.

Moreover, we summarize and annotate commonly used hateful slang from real-world online forums to address the challenges of identifying hateful slang. We compile a comprehensive annotated lexicon, labeling each hateful slang with its frequent groups and providing explanations of its usage and contextual meaning. This is the first annotated lexicon of Chinese hateful slang with interpretative annotations. This resource not only helps to understand how hateful slang is used to disguise hate speech, but also provides valuable annotated data to evaluate and improve the ability of LLMs to detect hateful slang. The main contributions of this work are summarized as follows:

- We provide a span-level target-aware toxicity extraction dataset, containing 8k posts and 9.5k quadruples, filling the gap in span-level resources for Chinese hate speech detection.
- We construct an annotated lexicon of commonly used hateful slang in real-world online forums to evaluate LLMs' capability in understanding Chinese hate speech semantics.
- We evaluate models on STATE TOXICN, assessing their span-level performance and ability to detect hateful slang, highlighting key challenges and insights for improvement.

2 Related Work

Hate Speech Detection. Hate speech detection is a critical task in Natural Language Processing (NLP) that has attracted considerable attention recently. Researchers have increasingly turned to pre-trained models to address this issue (Caselli et al., 2020; Hanu and Unitary team, 2020; Zhou et al., 2021; AlKhamissi et al., 2022a; Ali et al., 2022). To facilitate progress in this field, several datasets tailored to hate speech detection have been developed (Waseem and Hovy, 2016; Davidson et al., 2017; Founta et al., 2018; Hartvigsen et al., 2022). Pavlopoulos et al. (2021) introduced span-level hate speech detection, while the TBO dataset advanced the field by pioneering the extraction of Target-Argument-Harmful triples (Zampieri et al., 2023). However, Chinese hate speech detection remains significantly underdeveloped.

Chinese Hate Speech Dataset. While some Chinese hate speech datasets exist, these efforts remain limited to the post-level. TOCP and TOCAB, derived from Taiwan's PTT platform, focus on detecting profanity and abusive language (Chung and Lin, 2021). The Sina Weibo Sexism Review (SWSR) centers on identifying sexism (Jiang et al., 2022). The Chinese Offensive Language Dataset (COLD) categorizes sentences into types such as individual attacks and anti-bias (Deng et al., 2022). Zhou et al. (2022) introduce CDial-Bias, the first annotated dataset specifically addressing social bias in Chinese dialogues. Lu et al. (2023) present TOXICN, a dataset encompassing both explicit and implicit toxic language samples. Xiao et al. (2024) introduce a dataset to evaluate LLMs' robustness against cloaking perturbations.

Work	Platforms	Language	#Posts	Span	Tar.	Arg.	Group	Lex.
Davidson et al. (2017)	Twitter	English	24,802					✓
Founta et al. (2018)	Twitter	English	80,000					
Toxic Spans (Pavlopoulos et al., 2021)	Civil Comments	English	10,629	✓				
HateXplain (Mathew et al., 2021)	Twitter, Gab	English	20,148	✓			✓	
TBO (Zampieri et al., 2023)	Twitter	English	4,673	✓	✓	✓		
COLD (Deng et al., 2022)	Zhihu, Weibo	Chinese	37,480				✓	
SWSR (Jiang et al., 2022)	Weibo	Chinese	8,969				✓	✓
Cdial-Bias-Utt (Zhou et al., 2022)	Zhihu	Chinese	13,394				✓	
Cdial-Bias-Ctx (Zhou et al., 2022)	Zhihu	Chinese	15,013				✓	
TOXICN (Lu et al., 2023)	Zhihu, Tieba	Chinese	12,011				✓	✓
TOXICLOAKCN (Xiao et al., 2024)	Zhihu, Tieba	Chinese	4,582				✓	
STATE TOXICN (Ours)	Zhihu, Tieba	Chinese	8,029	✓	✓	✓	✓	✓

Table 2: Comparison of hate speech datasets based on *Platforms*, *Language*, *#Posts*, span-level annotations (*Span*), inclusion of Target (*Tar.*), Argument (*Arg.*), *Group*, and Lexicon information (*Lex.*).

While previous works offer quality corpora, span-level research remains unexplored. STATE TOXICN is the first such dataset, annotating Target-Argument-Hateful- Group quadruples. Unlike existing lexicons, we offer interpretative annotations and targeted group labels, creating a comprehensive Chinese hateful slang lexicon.

3 Dataset Construction

3.1 Overview

In this section, we introduce the construction process of the STATE TOXICN dataset and the annotated lexicon of Chinese hateful slang. First, we describe the data sources and filtering procedures. Next, we detail the annotation process and the measures implemented to ensure high annotation quality. We then examine the Inter-Annotator Agreement (IAA) at different levels of granularity. Finally, we present relevant statistics for the STATE TOXICN dataset.

3.2 Data Source and Filtering

Our dataset construction is based on the post-level Chinese hate speech dataset TOXICN (Lu et al., 2023). We develop a high-quality span-level Chinese hate speech dataset through data filtering and annotation processes. During the data filtering phase, we extract potential samples from the original dataset and systematically clean low-quality, incomplete, or irrelevant content. For instance, we remove meaningless text such as advertisements, random character combinations, and overly short (less than 5 characters) or overly long (more than 500 characters) text fragments.

Building on this, we conduct a comprehensive review of the hate speech annotations, particularly regarding the clarity of the target and argument components of hate speech. For example, in cases describing strongly discriminatory or biased content, we further examine whether the target was specific, removing ambiguous or undefined samples. In particular, for texts lacking a clear Target-Argument structure, we opt to exclude them to ensure the dataset could accommodate the requirements of span-level analysis.

3.3 Data Annotation

3.3.1 Annotation Guidelines

During the annotation process, we establish guidelines and implement multi-stage quality control to ensure the consistency of the annotations, aiming to label Target-Argument-Hateful-Group quadruples in the posts. First, we develop detailed annotation guidelines, including standards for extracting targets and arguments (Target-Argument Pair), criteria for determining hatefulness (Hateful), and methods for classifying groups (Group):

Target-Argument Pair A span consists of both the target and its corresponding argument extracted from the post. A single post may contain more than one Target-Argument Pair.

Hateful If the Target-Argument Pair explicitly or implicitly conveys harm towards the target or other groups, it is labeled as "**hateful**"; otherwise, it is labeled as "**non-hate**."

Group Building on the Target-Argument Pair, this category identifies specific groups targeted by hateful expressions, with a single pair probably involving multiple groups.

Target Span	Argument Span	If Hateful	Targeted Group
0.65	0.61	0.68	0.75

Table 3: Fleiss' Kappa for different granularities.

For example, in Exp. 2 from Table 1, Target-Argument Pair is "男同（*Gay people*）" and "艾滋高发群体(*a high-risk group for HIV*)". This Target-Argument Pair constitutes hate against the LGBTQ community and people living with AIDS, so the Hateful label is "*hate*" and the Group label is "*LGBTQ, others*". The quadruple is [男同| 艾滋高发群体| *hate* | *LGBTQ, others*].

Regarding the annotation of the Chinese hateful slang lexicon, we also establish annotation guidelines for identifying, categorizing, and labeling the terms, with a focus on their frequent groups (Group) and contextual explanations (Explanation):

Group Each hateful slang term is categorized by the group it targets, such as sexism, racism, regional bias, anti-LGBTQ, or others. Some terms may target multiple groups.

Explanation An explanation of hateful slang is provided, including its literal meaning, extended meanings, the reasons for hatred towards targeted groups, and common usage patterns.

3.3.2 Mitigating Bias.

To mitigate bias, we assemble annotators with diverse backgrounds, including differences in gender, age, ethnicity, region, and educational level. The detailed information of annotators can be found in Appendix G. All annotators possess linguistic expertise and have undergone systematic training. During the annotation process, we primarily employ regular cross-validation and expert arbitration to ensure the objectivity and consistency of annotation results. Additionally, we maintain an online document to record and update Chinese hateful slang identified in the posts. This annotated lexicon serves not only as a research resource but also helps to align annotators' standards.

3.3.3 Annotation Procedure

Cross-validation and Expert Arbitration. Each text is independently annotated by at least two annotators, who followed a unified annotation guideline to extract Target-Argument pairs, determine the hatefulness, and classify the groups. After the initial annotation phase, 20% of the samples are regularly selected for cross-validation, during which other annotators re-annotate these samples to ensure a consistent understanding of the rules and standards across annotators. This approach allows us to identify and resolve potential biases or discrepancies in the annotations in a timely manner.

Category	Subcategory	Count	Percentage (%)
Groups	Gender	1663	17.44
	Race	1232	12.92
	Region	1323	13.88
	LGBTQ	628	6.59
	Others	351	3.68
	Multi-group	866	9.08
Hateful	Hate	6063	63.60
	Non-Hate	3470	36.40
Total	-	9533	100.00

Table 4: Statistics of annotated posts from the STATE TOXICN dataset, including Target, Argument, Group, and Hatefulness classifications.

For disputed samples with significant annotation differences, an arbitration team of domain experts reviews the samples, considering the textual context and annotation guidelines to determine the most appropriate annotation. We explore inter-annotator agreement (IAA) on STATE TOXICN, with Fleiss' kappa scores (Fleiss, 1971) for each hierarchy detailed in Table 3. The detailed analysis can be found in Appendix D.

Lexicon Annotation. To ensure annotation quality, we maintain a shared online document for recording and updating detailed information on Chinese hateful slang. This document includes explanations of each slang term and the specific groups they typically reference. Team members could add newly discovered slang terms, which are then evaluated and annotated by the expert team. This dynamic maintenance ensures a consistent understanding of emerging language and slang among all team members.

Additionally, the shared online document serves as a knowledge-sharing platform, providing the annotation team with consistent references and concrete examples for handling complex or ambiguous posts. This mechanism improves annotation efficiency and enhances the consistency of the annotations. This is the first Chinese hateful slang lexicon with interpretable annotations. This lexicon provides valuable research resources for span-level Chinese hate speech semantic understanding.

3.4 Data Description

STATE TOXICN dataset contains a total of 8,029 annotated posts, among which 4,942 posts include hateful content, accounting for 61.55%. A total of 9,533 quadruples are annotated, with 6,034 of them containing hateful information, making up 63.60%. We present the statistical details of STATE TOXICN in Table 4. Gender, Region, and Race are the three most common group types in the dataset. Additionally, "multi-group" refers to Target-Argument pairs that involve hatred directed at multiple target groups, with a total of 854 such instances, accounting for 8.96%. In addition, the annotated lexicon includes 830 Chinese hateful slang terms collected from real online forums. More annotated examples can be found in Appendix A.

4 Experiment

4.1 Baselines

To evaluate the performance of LLMs with varying parameter sizes and capabilities in detecting span-level Chinese hate speech, we choose twelve well-known models across three categories:

Open-source Models: mT5-base (Xue, 2020), Mistral-7B (Jiang et al., 2023), LLaMA3-8B (AI@Meta, 2024), Qwen2.5-7B (Team, 2024); **Safety-domain Models:** ShieldLM-14B-Qwen (Zhang et al., 2024), ShieldGemma-9B (Zeng et al., 2024), and **Closed-source LLMs:** LLaMA3-70B (AI@Meta, 2024), Qwen2.5-72B (Team, 2024), Gemini-1.5-Pro (Team et al., 2023), Claude-3.5-Sonnet (Anthropic, 2024), GPT-4o (OpenAI, 2024), DeepSeek-v3 (Liu et al., 2024). Detailed information of fine-tuning is provided in Appendix F.

4.2 Evaluation metrics.

Due to the ambiguity of Chinese span boundaries, a single evaluation metric may not accurately assess the performance of models in span-level Chinese hate speech detection. To obtain more accurate evaluation results, we therefore utilize both hard and soft matching metrics. We employ the Macro-F1 scores as the main evaluation metrics.

Hard-matching: A predicted quadruplet, particularly its target and argument components, is deemed correct only if it perfectly matches its corresponding gold label.

Soft-matching: We adopt the algorithm proposed by Han et al. (2023), where a prediction is considered correct if the Target and Argument scores achieve a threshold of 0.5.

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

Table 5: Statistics of train and test datasets, including *#Posts*, total quads (*Quad.*), hateful quads (*Hateful*), and non-hateful quads (*Non-hate*).

4.3 Experiment Settings

We conduct fine-tuning on open-source and safety-domain models, with training and testing set sizes detailed in Table 5. The fine-tuning process was performed using only a basic prompt, which included task definitions, output formats, and specific prediction requirements for all elements. We evaluate the performance of closed-source LLMs by calling their APIs. In addition to the aforementioned basic prompt, we also provided a hate speech example and a non-hate speech example. Further details regarding these requirements are provided in Appendix B.

5 Results and Analysis

5.1 *RQ1: Can LLMs identify the spans of targets and arguments in Chinese text?*

Finetuned Models In the tasks of identifying targets, arguments, and target-argument pairs (*T-A Pair*), fine-tuned models significantly outperform direct usage of LLM APIs. LLaMA3-8B, Qwen2.5-7B, and the Shield series models all achieve hard match scores of over 63%, strongly demonstrating the superior ability of fine-tuned models in identifying text span boundaries. Soft match metrics further confirm the advantage of fine-tuned models in target and argument identification, with F1 score approaching 70% or higher. These results indicate that fine-tuned models on task-specific data show a substantial advantage in span identification. LLaMA3-8B and Qwen2.5-7B perform the best. ShieldGemma-9B achieves the highest scores in soft-match metrics for two tasks.

LLM APIs Compared to fine-tuned models, LLMs access through APIs performed significantly worse in all tasks, regardless of whether hard or soft match metrics were used. Even with few-shot prompting strategies, their performance still lag far behind that of fine-tuned models. This suggests that LLMs, without being fine-tuned on Chinese hate speech data, are not adept at identifying text

Model	Target		Argument		T-A Pair		T-A-H Tri.		Quad.	
	Hard	Soft	Hard	Soft	Hard	Soft	Hard	Soft	Hard	Soft
Finetuned Models (with Basic Prompt)										
mT5-base	59.15	70.55	28.63	67.03	23.33	55.90	17.76	43.34	16.60	38.61
Mistral-7B	62.97	73.69	35.58	70.90	30.55	60.49	26.15	51.01	23.72	45.62
LLaMA3-8B	**64.07**	73.74	**36.72**	70.82	**31.64**	60.88	**27.04**	51.62	**24.27**	46.08
Qwen2.5-7B	63.96	**74.64**	35.42	70.36	30.63	60.52	26.51	**52.86**	23.70	47.03
ShieldLM-14B-Qwen	63.83	73.45	34.80	70.23	30.20	59.81	26.18	51.24	23.59	45.58
ShieldGemma-9B	63.40	74.31	34.40	**71.11**	29.99	**61.51**	25.64	52.70	23.49	**47.14**
LLM APIs (with Basic Prompt and 2 Examples)										
LLaMA3-70B	30.54	41.03	14.39	47.96	8.16	27.34	6.03	20.70	3.69	11.93
Qwen2.5-72B	40.94	50.44	21.10	56.36	15.66	39.49	12.48	30.92	8.74	20.29
Gemini-1.5-Pro	29.80	37.29	18.43	54.96	9.37	26.22	7.71	21.88	5.45	14.81
Claude-3.5-Sonnet	37.61	50.72	15.45	57.24	9.72	36.16	7.94	29.82	6.29	22.45
GPT-4o	46.85	58.19	22.64	62.41	17.21	46.41	13.21	35.68	9.00	23.34
DeepSeek-v3	48.16	59.25	22.79	59.38	18.68	46.40	14.95	37.19	11.48	27.38

Table 6: Performance comparison of various models across different levels of annotated tasks, including *Target*, *Argument*, Target-Argument Pair (*T-A Pair*), Target-Argument-Hateful Triple (*T-A-H Tri.*), and Target-Argument-Hateful-Group Quadruple (*Quad.*) under both Hard and Soft evaluation metrics.

span boundaries, which is consistent with the fact that their training objectives are not directly related to this task. Specifically, the hard match scores for LLMs on "Target" and "Argument" were only between 40-50%, while soft match scores were around 60%. GPT-4o and DeepSeek-v3 exhibit the best performance.

Comparison and Summary The fine-tuning technology can improve models' abilities in identifying text span boundaries, demonstrating a clear advantage in target and argument extraction tasks. Joint extraction tasks are more challenging than single element extraction tasks, and all models perform worse in argument extraction than in target extraction. We believe this is primarily because in Chinese text, argument spans tend to be longer and structurally more complex, leading to poor hard match performance. However, the soft match scores for argument extraction were similar to those of target extraction, indicating that the models' semantic understanding capabilities are comparable in both tasks. This suggests that despite difficulties in identifying precise boundaries, the models remain effective in semantic understanding.

5.2 *RQ2: Can LLMs classify Chinese target-argument pairs as hateful and identify the groups?*

Finetuned Models Fine-tuned models demonstrate a clear advantage in both the triplet and quadruplet tasks. In the Target-Argument-Hateful triplet task (T-A-H Tri.), the fine-tuned models achieve hard-match metrics of around 26% and soft-match metrics of around 52%. In the more complex quadruplet task, the hard-match metrics of the fine-tuned models are mostly between 25% and 27% when simultaneously identifying hateful content and target groups, while the soft-match metrics are approximately 45% - 47%. This indicates that the fine-tuned models' performance is mainly limited by the target-argument pair identification. Overall, although the fine-tuned models perform poorly in hard-match metrics, the results in soft-match metrics reach approximately 50%. Among these, LLaMA-3-8B performs best in hard-match, while Qwen2.5-7B and ShieldGemma-9B perform better in soft-match metrics.

LLM APIs LLM APIs perform poorly on both the triplet and quadruplet tasks. In the T-A-H triplet task, LLMs achieve hard-match metrics of only 6% - 15% and soft-match metrics of 20% - 38%. In the quadruplet task, the hard-match metrics of LLMs further drop to 3% - 12%, and the soft-match metrics, though slightly better, are still only 11% - 27%. We believe that the performance limitations also primarily stem from the difficulty in identifying accurate span boundaries. These data clearly show that LLMs, without being fine-tuned on Chinese hate speech data, are significantly inadequate at performing Chinese hate speech triplet and quadru-

Model	Target		Argument		T-A Pair		T-A-H Tri.		Quad.	
	Hard	Soft	Hard	Soft	Hard	Soft	Hard	Soft	Hard	Soft
Finetuned Models (with Basic Prompt)										
mT5-base	56.83$_{2.32}$	68.33$_{2.22}$	27.17$_{1.46}$	64.17$_{2.86}$	21.33$_{2.00}$	51.17$_{4.73}$	18.17$_{1.46}$	44.67$_{1.33}$	16.33$_{0.27}$	36.17$_{2.44}$
Mistral-7B	61.03$_{1.94}$	72.62$_{1.07}$	36.71$_{1.13}$	70.05$_{0.85}$	30.27$_{0.28}$	58.62$_{1.87}$	26.25$_{0.10}$	51.05$_{0.04}$	22.22$_{1.50}$	43.00$_{2.62}$
LLaMA3-8B	61.26$_{2.81}$	72.12$_{1.62}$	35.17$_{1.55}$	70.83$_{0.01}$	28.53$_{3.11}$	59.00$_{1.88}$	24.47$_{2.57}$	51.38$_{0.24}$	19.77$_{4.50}$	42.46$_{3.62}$
Qwen2.5-7B	61.79$_{2.17}$	73.33$_{1.31}$	36.26$_{0.82}$	69.59$_{0.77}$	29.92$_{0.71}$	57.72$_{2.80}$	27.64$_{1.13}$	53.66$_{0.80}$	22.76$_{0.94}$	45.04$_{1.99}$
ShieldLM-14B-Qwen	64.08$_{0.25}$	74.48$_{1.03}$	34.19$_{0.61}$	69.86$_{0.37}$	28.90$_{1.30}$	58.30$_{1.51}$	25.60$_{0.58}$	51.69$_{0.45}$	21.30$_{2.29}$	43.60$_{1.98}$
ShieldGemma-9B	62.50$_{0.90}$	74.35$_{0.04}$	35.71$_{1.31}$	71.10$_{0.01}$	29.87$_{0.12}$	60.71$_{0.80}$	26.95$_{1.31}$	55.03$_{2.33}$	23.21$_{0.28}$	46.10$_{1.04}$
LLM APIs (with Basic Prompt and 2 Examples)										
LLaMA3-70B	30.87$_{0.33}$	41.45$_{0.42}$	14.80$_{0.41}$	46.68$_{1.28}$	8.29$_{0.13}$	25.38$_{1.96}$	7.40$_{1.37}$	22.58$_{1.88}$	4.72$_{1.03}$	13.14$_{1.21}$
Qwen2.5-72B	45.58$_{4.64}$	54.67$_{4.23}$	22.15$_{1.05}$	57.36$_{1.00}$	16.77$_{1.11}$	41.61$_{2.12}$	12.42$_{0.06}$	30.35$_{0.57}$	8.83$_{0.09}$	19.59$_{0.70}$
Gemini-1.5-Pro	31.52$_{1.72}$	39.07$_{1.78}$	18.94$_{0.51}$	54.57$_{0.39}$	9.01$_{0.36}$	27.95$_{1.73}$	8.61$_{0.90}$	25.83$_{3.95}$	6.23$_{0.78}$	17.35$_{2.54}$
Claude-3.5-Sonnet	41.45$_{3.84}$	54.06$_{3.34}$	15.80$_{0.35}$	55.80$_{1.44}$	10.43$_{0.71}$	36.96$_{0.80}$	9.28$_{1.34}$	33.04$_{3.22}$	7.10$_{0.81}$	25.22$_{2.77}$
GPT-4o	49.28$_{2.43}$	61.30$_{3.11}$	21.71$_{0.93}$	59.66$_{2.75}$	16.52$_{0.69}$	46.55$_{0.14}$	12.01$_{1.20}$	33.72$_{1.96}$	8.46$_{0.54}$	22.80$_{0.54}$
DeepSeek-v3	52.69$_{4.53}$	64.25$_{4.00}$	23.79$_{1.00}$	62.10$_{2.72}$	19.22$_{0.54}$	50.40$_{4.00}$	16.13$_{1.18}$	42.07$_{4.88}$	11.69$_{0.21}$	30.11$_{2.73}$

Table 7: Performance comparison of various models on a hateful-slang-containing subset of the test set. Green indicates a decrease in F1 score on the subset compared to the full dataset, while red indicates an increase.

plet predictions, even with few-shot prompting failing to bridge the performance gap. DeepSeek-v3 performs significantly better than other APIs, particularly in the quadruplet prediction task.

Comparison and Summary The fine-tuned model significantly outperforms the LLM APIs, with the performance gap primarily stemming from T-A pair extraction. The fine-tuned model outperforms the LLM APIs in both hate speech the triplet and quadruplet prediction. Specifically, the fine-tuned models show better performance in tasks requiring the simultaneous identification of hate speech and its target group. Although there is still room for improvement in terms of precise matching, the application of fine-tuning technology undoubtedly provides a feasible path for handling such complex tasks. At the same time, the performance of the LLM APIs in these tasks is disappointing, indicating that they need targeted fine-tuning to be applied to these types of scenarios.

5.3 *RQ3: Can LLMs understand Chinese hateful slang?*

5.3.1 Impact of Hateful Slang on Chinese Hate Speech Detection

To evaluate the impact of hate slang on span-level Chinese hate speech detection models, we remove the posts without hateful slang from the test set and obtain a subset of 502 posts containing hateful slang for focused testing. By comparing the model performance on this subset of data with its performance on the full test set, we can gain a clearer understanding of how hate slurs affect model performance. The experimental results are presented in Table 7. The superscript numbers in the table indicate the difference in F1 score between the model's performance on the selected subset and its performance on the full test set. Green indicates a decrease in F1 score on the subset compared to the full dataset, while red indicates an increase.

Finetuned Models Across most tasks, the hard and soft matching metrics for the models show a decreasing trend. However, the metric for the hate speech triplet prediction task (T-A-H Tri.) is an exception, exhibiting an increase. We attribute this primarily to two factors: Firstly, posts containing hateful slang generally demonstrate a more explicit hateful intent, while the posts we remove include many types that are difficult to identify, such as implicit hate expressions. This makes it easier for the T-A-H Tri. task to capture the relationships within hate speech triplets, thereby leading to the observed increase in the metric. Secondly, the experimental results also confirm that the main challenges posed by hateful slang lie in extracting the target and argument, as well as accurately classifying the targeted group. When the training data is not sufficiently comprehensive, fine-tuned models struggle to effectively handle the complexities introduced by hateful slang.

LLM APIs LLM APIs demonstrate superior performance on the subset of the test set containing hateful slang. With the exception of a slight decrease in the soft matching metric for argument identification, nearly all other metrics show an improvement. We believe that this is primarily due to two factors: First, hateful slang often appears as the target, and the explicit target terms effectively compensate for the limitations of LLM APIs in span boundary recognition. Second, the identifica-

Figure 1: Examples of Chinese hateful slang understanding analysis with LLM. The hateful slang terms and texts appear in black, ShieldGemma-v3 explanations are in green, DeepSeek-v3 explanations are in red, and human annotator explanations are in blue.

tion of hateful slang often requires rich background knowledge; for example, Example 2 in the figure requires knowledge of Chinese history and related common usages, which is precisely where LLM APIs excel.

Comparison and Summary Fine-tuned models and LLM APIs exhibit distinct characteristics when processing hateful slang. Fine-tuned models, when lacking sufficient training and understanding of specific domain knowledge, struggle to effectively address the challenges posed by hateful slang, particularly in areas such as fine-grained entity recognition and complex language pattern understanding. In contrast, LLM APIs leverage their robust background knowledge and contextual understanding to better comprehend and handle these complex situations, thus achieving superior performance. However, the performance gap between LLM APIs and fine-tuned models on this task remains significant. Infusing the background knowledge of large models into fine-tuned models may offer a promising avenue for improvement.

5.3.2 Understanding Capabilities of LLMs on Chinese Hateful Slang

To investigate the understanding capabilities of large language models regarding Chinese hateful slang, we conducted experiments using two relatively high-performing models, ShieldGemma-9B and DeepSeek-v3. In these experiments, we provided the models with hateful terms and their surrounding sentence context, asking them to generate explanations of the hateful slang and identify the potentially affected groups. The experimental results are detailed in Figure 1, with specific experimental details provided in Appendix C.

For a more nuanced analysis of the models' understanding abilities, we selected two terms with distinct Chinese characteristics as case studies. In Case 1, "默" (mò) is a typical "merging word," whose literal meaning is 'silence,' but it is composed of the characters "黑犬" (hēiquǎn), meaning "black dog." This type of character combination is a unique linguistic phenomenon in Chinese. The experimental results showed that neither model could accurately understand the hateful information contained within it. In Case 2, "冉闵" (Rǎn Mǐn) is a hateful slang term rooted in Chinese history and culture, referring to an ancient emperor who massacred ethnic minorities and is often used for racial discrimination. In comparison, while ShieldGemma-9B could understand the background information, it failed to generate an accurate explanation of the hateful information and clearly identify the potentially affected groups. DeepSeek-v3's generated information, however, closely matched the results given by human annotators. This demonstrates that even relatively high-performing models face significant challenges in understanding Chinese hateful slang that possesses cultural specificity and subtle connotations.

6 Conclusions

With the advancement of hate speech detection, recent research has shifted from post-level detection to more fine-grained span-level detection. In this work, we focus on building resources for fine-grained Chinese hate speech detection. Firstly, we present the first span-level target-aware toxicity extraction dataset (STATE ToxiCN). Based on

this dataset, we evaluate the performance of LLMs in span-level Chinese hate speech detection. Secondly, we provide the first Chinese hateful slang lexicon with interpretable annotations. All the hateful slang terms are sourced from real-world online platforms. Using this lexicon, we evaluate the impact of hateful slang on model detection capabilities and assess LLMs' understanding of hateful slang. Experimental results show that current LLMs still face challenges in effectively addressing both span-level hate speech detection and the understanding of hate semantics. We hope that our resources and benchmarks will be valuable for researchers in this field.

Limitation

Despite implementing rigorous quality control measures in constructing the span-level Chinese hate speech detection dataset and the annotated lexicon, we acknowledge several limitations. First, although we have taken steps to minimize labeling bias, subjective differences in annotators' understanding of toxic language may still result in mislabeled data. Additionally, while we strived to ensure accuracy and consistency in annotating target-argument pairs, the inherent characteristics of the Chinese language, such as flexible grammar and ambiguous boundaries, pose challenges for completely precise annotations.

Our Annotated Lexicon covers a wide range of commonly used Chinese hateful slang; however, due to the dynamic nature and rapid evolution of the Chinese internet language, certain emerging or less widely recognized hateful slang (e.g., more complex word transformations or domain-specific slang) might not be fully captured. Finally, this study focuses primarily on textual features and does not consider non-textual elements, such as images, videos, or metadata about the authors, which may limit the model's ability to comprehensively detect multimodal hate speech.

In the future, we aim to expand the scope of our dataset to include more diverse contexts and hateful slang, while exploring multimodal and cross-domain hate speech detection methods to improve overall performance and applicability.

Ethics Statement

We adhere strictly to the data usage agreements of all public online social platforms and conduct thorough reviews to ensure that no user privacy information is included in our dataset. The opinions and findings reflected in the samples of our dataset do not represent the views of the authors, either explicitly or implicitly. We aim to ensure that the benefits of our proposed resources outweigh any potential risks. All resources are provided exclusively for scientific research purposes.

To minimize the psychological impact of evaluating harmful content, we have implemented a multi-faceted approach. This includes obtaining informed consent after thoroughly explaining the nature of the content they may encounter, carefully managing exposure by limiting weekly evaluation volumes, and empowering annotators to immediately cease work should they experience any discomfort. Furthermore, we proactively monitor their mental health through regular check-ins, ensuring a supportive and responsible working environment.

Acknowledgment

This research is supported by the Natural Science Foundation of China (No. 62376051, 61702080, 62366040), the Key R&D Projects in Liaoning Province award numbers (2023JH26/10200015), the Fundamental Research Funds for the Central Universities (DUT24LAB123).

References

Hyeseon Ahn, Youngwook Kim, Jungin Kim, and Yo-Sub Han. 2024. Sharedcon: Implicit hate speech detection using shared semantics. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 10444–10455.

AI@Meta. 2024. Llama 3 model card.

Raza Ali, Umar Farooq, Umair Arshad, Waseem Shahzad, and Mirza Omer Beg. 2022. Hate speech detection on twitter using transfer learning. *Computer Speech & Language*, 74:101365.

Badr AlKhamissi, Faisal Ladhak, Srini Iyer, Ves Stoyanov, Zornitsa Kozareva, Xian Li, Pascale Fung, Lambert Mathias, Asli Celikyilmaz, and Mona Diab. 2022a. Token: Task decomposition and knowledge infusion for few-shot hate speech detection. *arXiv preprint arXiv:2205.12495*.

Badr AlKhamissi, Faisal Ladhak, Srinivasan Iyer, Veselin Stoyanov, Zornitsa Kozareva, Xian Li, Pascale Fung, Lambert Mathias, Asli Celikyilmaz, and Mona Diab. 2022b. ToKen: Task decomposition and knowledge infusion for few-shot hate speech detection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2109–2120, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Anthropic. 2024. Claude 3.5 sonnet. https://www.anthropic.com/news/claude-3-5-sonnet.

Michał Bilewicz and Wiktor Soral. 2020. Hate speech epidemic. the dynamic effects of derogatory language on intergroup relations and political radicalization. *Political Psychology*, 41:3–33.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2020. Hatebert: Retraining bert for abusive language detection in english. *arXiv preprint arXiv:2010.12472*.

I Chung and Chuan-Jie Lin. 2021. Tocab: A dataset for chinese abusive language processing. In *2021 IEEE 22nd International Conference on Information Reuse and Integration for Data Science (IRI)*, pages 445–452. IEEE.

Gloria Cowan and Cyndi Hodge. 1996. Judgments of hate speech: The effects of target group, publicness, and behavioral responses of the target. *Journal of Applied Social Psychology*, 26(4):355–374.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515.

Jiawen Deng, Jingyan Zhou, Hao Sun, Chujie Zheng, Fei Mi, Helen Meng, and Minlie Huang. 2022. Cold: A benchmark for chinese offensive language detection. *arXiv preprint arXiv:2201.06025*.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2024. *Ethnologue: Languages of the World*, twenty-seventh edition. SIL International, Dallas, Texas. Online version available at http://www.ethnologue.com.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Proceedings of the international AAAI conference on web and social media*, volume 12.

Ridong Han, Tao Peng, Chaohao Yang, Benyou Wang, Lu Liu, and Xiang Wan. 2023. Is information extraction solved by chatgpt? an analysis of performance, evaluation criteria, robustness and errors. *arXiv preprint arXiv:2305.14450*.

Laura Hanu and Unitary team. 2020. Detoxify. Github. https://github.com/unitaryai/detoxify.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. *arXiv preprint arXiv:2203.09509*.

Sanne Hoeken, Sophie Spliethoff, Silke Schwandt, Sina Zarrieß, and Özge Alacam. 2023. Towards detecting lexical change of hate speech in historical data. In *Proceedings of the 4th Workshop on Computational Approaches to Historical Language Change*, pages 100–111, Singapore. Association for Computational Linguistics.

Robin Jeshion. 2013. Slurs and stereotypes. *Analytic Philosophy*, 54(3).

Aiqi Jiang, Xiaohan Yang, Yang Liu, and Arkaitz Zubiaga. 2022. Swsr: A chinese dataset and lexicon for online sexism detection. *Online Social Networks and Media*, 27:100182.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

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 (Volume 1: Long Papers)*, pages 16235–16250, Toronto, Canada. Association for Computational Linguistics.

Victor H Mair. 1991. What is a chinese" dialect/topolect"?: Reflections on some key sino-english linguistic terms.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 14867–14875.

OpenAI. 2024. Hello gpt-4o. https://openai.com/index/hello-gpt-4o/.

John Pavlopoulos, Jeffrey Sorensen, Léo Laugier, and Ion Androutsopoulos. 2021. Semeval-2021 task 5: Toxic spans detection. In *Proceedings of the 15th international workshop on semantic evaluation (SemEval-2021)*, pages 59–69.

Leandro Silva, Mainack Mondal, Denzil Correa, Fabrício Benevenuto, and Ingmar Weber. 2016. Analyzing the targets of hate in online social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 10, pages 687–690.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie

Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.

Yunze Xiao, Yujia Hu, Kenny Tsu Wei Choo, and Roy Ka-Wei Lee. 2024. 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*, pages 6012–6025, Miami, Florida, USA. Association for Computational Linguistics.

L Xue. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Marcos Zampieri, Skye Morgan, Kai North, Tharindu Ranasinghe, Austin Simmons, Paridhi Khandelwal, Sara Rosenthal, and Preslav Nakov. 2023. Target-based offensive language identification.

Wenjun Zeng, Yuchi Liu, Ryan Mullins, Ludovic Peran, Joe Fernandez, Hamza Harkous, Karthik Narasimhan, Drew Proud, Piyush Kumar, Bhaktipriya Radharapu, Olivia Sturman, and Oscar Wahltinez. 2024. Shieldgemma: Generative ai content moderation based on gemma.

Zhexin Zhang, Yida Lu, Jingyuan Ma, Di Zhang, Rui Li, Pei Ke, Hao Sun, Lei Sha, Zhifang Sui, Hongning Wang, et al. 2024. Shieldlm: Empowering llms as aligned, customizable and explainable safety detectors. *arXiv preprint arXiv:2402.16444*.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand. Association for Computational Linguistics.

Jingyan Zhou, Jiawen Deng, Fei Mi, Yitong Li, Yasheng Wang, Minlie Huang, Xin Jiang, Qun Liu, and Helen Meng. 2022. Towards identifying social bias in dialog systems: Framework, dataset, and benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3576–3591.

Xianbing Zhou, Yang Yong, Xiaochao Fan, Ge Ren, Yunfeng Song, Yufeng Diao, Liang Yang, and Hongfei Lin. 2021. Hate speech detection based on sentiment knowledge sharing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7158–7166.

A Details of the Annotated Lexicon

We use JSON files to store the annotated lexicon, which is a mainstream coding specification to facilitate machine-readability. The data structure is defined as Sample = (Term, [Groups], Definition), where Groups indicates the groups that are commonly targeted by the hateful slang. If a term does not have a specific targeted group, then Groups is set to "others".

- **Term**: 默(Mò)
  Meaning: Silence
  Groups: Racism
  Definition: “默”是“黑犬”的拼字，指黑色的狗，对黑人群体的侮辱性称呼。
  Definition in English: The term “默” (Mò, meaning silence) is a spelling variant of “黑犬” (black dog), an insulting term for the Black community.

- **Term**: 金针菇
  Meaning: Enoki mushroom (metaphor for male genitalia)
  Groups: Sexism
  Definition: 用于对男性生殖器官的比喻，常含贬义。
  Definition in English: “Enoki mushroom” is a derogatory metaphor for male genitalia.

- **Term**: 小仙女
  Meaning: Little fairy (refers to selfish, hypocritical, or annoying women)
  Groups: Sexism
  Definition: 在某些情况下被用来指代那些自私自利、双标、让人厌恶的女性。
  Definition in English: “Little fairy” is used in some contexts to refer to women perceived as selfish, hypocritical, or annoying.

- **Term**: 4000+
  Meaning: Homophone for “死全家” (may your whole family die)
  Groups: Others
  Definition: “4000+”是“死全家”中文谐音，常用于恶意辱骂。
  Definition in English: “4000+” is a homophone for “死全家” (may your whole family die), often used in malicious insults.

B Details of the Prompt

The basic prompt and the version with 2 examples are shown in Figure 2.

Basic Prompt

Chinese:
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：
English:
You are a content review expert, please analyze my sentences and extract one or more quadruples from them. Please extract one or more quadruples from the following text, with the output format of each quadruple being 'Target|Argument|Whether Hateful|Group'. The target can be 'NULL', the argument should be as concise as possible, targeted groups only include (LGBTQ, Region, Sexism, Racism, others, non-hate), the same quadruple may involve multiple groups, the 'whether hate' tag is (hate, non-hate), multiple quadruples are separated by [SEP], and the last quadruple is followed by [END]. Extract all quadruples contained in the sentences:

Basic Prompt with 2 Examples

Chinese:
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]。
仇恨样本输入：谁喜欢小黑啊，黑不拉几的看一眼都嫌脏，也就某些连汽车排气口都不放过的男人会喜欢黑妞
仇恨样本输出：小黑|黑不拉几的看一眼都嫌脏|Racism|hate [SEP] 某些连汽车排气口都不放过的男人|喜欢黑妞|Sexism, Racism|hate [END]
非仇恨样本输入：由同性伴侣抚养长大的孩子，在学校表现比异性伴侣抚养的孩子更好，并且比异性伴侣的孩子高中毕业率高出4.8%。
非仇恨样本输出：由同性伴侣抚养长大的孩子|在学校表现比异性伴侣抚养的孩子更好|non-hate|non-hate [END]
提取出句子中包含的所有四元组：
English:
You are a content review expert, please analyze my sentences and extract one or more quadruples from them. Please extract one or more quadruples from the following text, with the output format of each quadruple being 'Target|Argument|Whether Hateful|Group'. The target can be 'NULL', the argument should be as concise as possible, targeted groups only include (LGBTQ, Region, Sexism, Racism, others, non-hate), the same quadruple may involve multiple groups, the 'whether hate' tag is (hate, non-hate), multiple quadruples are separated by [SEP], and the last quadruple is followed by [END].
Hate sample input: Who likes black people? They are so dirty that even a glance makes you feel unclean. Only some men who don't even mind the exhaust pipes of cars would like black girls.
Hate sample output: black people | so dirty that even a glance makes you feel unclean | Sexism, Racism | hate [SEP] some men who don't even mind the exhaust pipes of cars | like black girls | Sexism, Racism | hate [END]
Non-hate sample input: Children raised by same-sex couples perform better at school than those raised by opposite-sex couples, and have a 4.8% higher high school graduation rate than children raised by opposite-sex couples.
Non-hate sample output: Children raised by same-sex couples | perform better at school than those raised by opposite-sex couples | non-hate | non-hate [END]
Extract all quadruples contained in the sentences:

Figure 2: Illustration of the basic prompt and the version with 2 examples.

C Experimental Details of Hateful Slang Understanding Generation

We prompt the non-fine-tuned ShieldGemma-9B and DeepSeek-v3 to generate hateful slang understanding. The prompt used is as follows.

Prompt: [Text]中的[Term]是否包含特定群体的仇恨信息，如果是，请说明对哪些群体构成仇恨并说明原因。

Prompt in English: In [Text], does the [Term] constitute hate speech directed at specific groups? If yes, please identify the targeted groups and the reasons behind the hatred.

D Analysis of Inter-annotator Agreement

To explore inter-annotator agreement (IAA), we calculate Fleiss' kappa scores (Fleiss, 1971) for the Target, Argument, Hateful, and Group labels separately. Due to the direct relationship between "Hateful" and "Group", if the "Hateful" annotation is inconsistent, the "Group" label will be discarded. The experimental results are shown in Table 3. The score for argument span is the lowest, mainly due to the ambiguity of span boundaries in Chinese and the added complexity of argument spans. Despite having established relevant guidelines, it is not possible to fully standardize annotators' labeling practices. The Target span boundaries are clearer, resulting in a higher score. The Kappa score for the Group label is 0.75, with "Region" and "Racism" being the most easily confused.

E Derivative Rules of Chinese Hateful Slang

Lu et al. (2023) propose seven types of derivative rules: **Deformation** (separating and combining individual characters), **Homophonic** (similar pronunciation), **Irony** (positive words used ironically to insult), **Abbreviation** (shortening and contracting sensitive words), **Metaphor** (degrading targets into something sarcastic), **Code Mixing** (use of non-Chinese codes to emphasize tone), and **Borrowed Words** (foreign phonetic words with specific toxic cultural connotations). We add two more derivative rules, **Historical Allusions** and **Stereotypes**.

Historical Allusions: China has a rich history of allusions, and events or figures from these allusions are sometimes used as insults in certain contexts (Hoeken et al., 2023). The name "冉闵" (an ancient Chinese emperor) is often used as a symbol of racial discrimination, as he once ordered the massacre of other races.

Stereotypes: Internet users often resort to stereotypes to insult their targets, degrading them into negative or prejudiced representations (Jeshion, 2013). Due to an incident in the 20th century where some people from Henan province were involved in stealing manhole covers to sell for money, the term "井盖" (manhole cover) has become a stereotype for people from Henan, often used in regional discrimination to label them as thieves or poor.

F Detailed Information of the Fine-tuning

We use LLaMA-Factory[2] (Zheng et al., 2024) for fine-tuning, which is a widely used framework for fine-tuning LLMs. To avoid overfitting, we monitor the trends of training and test losses and confirm through preliminary experiments that the performance stabilizes after 10 epochs. Therefore, all models are trained for 10 epochs. We adopt the LoRA method and evaluate the results.

To reduce hyperparameter sensitivity, we train models for each learning rate and select the result with the highest F1 score on the test set, then calculate the final performance by weighted averaging. The hyperparameters are presented in Table 8.

Hyperparameters	Value
Epochs	10
Batch size	2
Learning rate	1e-5, 2e-5, 3e-5,, 4e-5, 5e-5
Cutoff length	1024
Compute type	fp16
Gradient accumulation	8
Maximum gradient norm	1.0

Table 8: Annotators Demographics.

G Detailed Information of the Annotators

To mitigate bias, we assemble a diverse group of annotators, encompassing a variety of genders, ages, ethnicities, regions, and educational backgrounds. The statistical information is presented in Table 9.

Characteristic	Demographics
Gender	8 male, 8 female
Age	7 age $<$ 25, 9 age $\geq$ 25
Race	10 Asian, 6 others
Region	From 9 different provinces
Education	5 BD, 6 MD, 5 Ph.D.

Table 9: Annotators Demographics.

[2]https://github.com/hiyouga/LLaMA-Factory

H License

We confirm that the dataset is licensed under the Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) license.