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

Zewen Bai · Liang Yang · Shengdi Yin · Junyu Lu · Jingjie Zeng · Haohao Zhu · Yuanyuan Sun · Hongfei Lin

Dalian University of Technology, China

dlutbzw@mail.dlut.edu.cn

ACL 2025 VIENNA

Abstract & Introduction

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

The proliferation of hate speech has caused significant harm to society. Detection of Chinese hate speech and hateful slang remains limited :

- **Constructed the first Chinese fine-grained span-level dataset, STATE-TOXICN, supporting target-aware toxicity span extraction ;**
- **We evaluated models' fine-grained detection performance and conducted the first study on Chinese hateful slang, testing large language models' grasp of hate semantics.**

Hateful Slang Understanding

Example 1	*Example 2*
Term: "默" Text: "默我是真的很讨厌。"	**Term: "冉闵" Text: "神州不止一个冉闵。"**
ShieldGemma-9B: I speculate that '默' (mò) is used in this context to express hatred towards a group associated with silence or stillness. Without more context regarding the speaker's background, it's difficult to determine what they mean by '默'.	*ShieldGemma-9B:* '冉闵' (Ran Min) could be coded hostility toward: minority ethnic or racial groups (as historically used to express prejudice); political opponents (for demonization); or religious groups (as a veiled expression of intolerance).
DeepSeek-v3: The term '默' (mò) could be a name, nickname, or designation. More context is needed to determine if it has hateful connotations. If '默' refers to a specific group and expresses animosity, it could have hateful undertones.	*DeepSeek-v3:* Ran Min of the Sixteen Kingdoms is known for the 'Kill the Hu' order and extreme actions against northern minorities, resulting in numerous deaths. He's seen by some as a national hero, but by others as a tyrant or nationalist.
Human Annotator: The word '默' (mò), meaning 'slience', when spelled out as '黑犬' (hēi quǎn), meaning 'black dog,' is a merge-word used to insult Black people.	*Human Annotator:* Ran Min was the founder of the Ran Wei regime. He led a revolt against the 'Hu' people and issued the 'Kill the Hu' order. He is often used as a symbol of nationalism, and has been used to justify racial discrimination.

We selected two terms with distinct Chinese characteristics as case studies:

Case-1: "默" (mò), a typical "merging word" formed by the characters "黑犬" (hēi quǎn, meaning "black dog"), a unique Chinese linguistic phenomenon.

Case-2: "冉闵" (rǎn mǐn), a hateful slang term rooted in Chinese history and culture, referring to an ancient emperor who massacred ethnic minorities, often used for racial discrimination.

Results show that even high-performing models face significant challenges in understanding Chinese hateful slang with cultural specificity and subtle connotations.

Experiments Results

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

Extract T-A pairs:

Fine-tuning enhances span boundary identification, particularly in T-A pair extraction, with soft-match results showing similar semantic understanding for targets and arguments.

Hate detection & categorization:

Fine-tuned models significantly outperform APIs, especially in T-A pair extraction and simultaneous identification of hate speech and groups.

Experiments Results on Hateful-slang-containing Subset

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	$52.69_{4.53}$	$64.25_{4.00}$	$23.79_{1.00}$	$62.10_{2.72}$	$19.22_{0.54}$	$50.40_{4.00}$	$16.13_{1.18}$	$42.07_{4.88}$	$11.69_{0.21}$	$30.11_{2.73}$

The results demonstrate that fine-tuned models struggle with hateful slang due to domain knowledge deficiencies, while LLM APIs exhibit superior performance through contextual comprehension. Incorporating API-derived background knowledge into fine-tuned models may enhance performance, though the performance gap remains substantial.

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

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

Statistics of STATE ToxiCN Dataset

Code

Paper