Border of Speech: A Benchmark Dataset for Understanding Chinese Offensive Speech

Warning: This paper contains content that could be offensive to some readers.

Anonymous ACL submission

Abstract

Offensive Speech Detection (OSD) has been a prominent research topic in NLP. However, the development of Chinese OSD is constrained by the lack of sufficient benchmark datasets. Moreover, Chinese OSD faces challenges such as ambiguity, context dependence, and particularly the identification of *Implicit Offensive Speech*. To address these challenges, we introduce a fine-grained labeling system for 10 categories of implicit offensive speech, grounded in linguistic principles, and present **SinOffen**, a comprehensive real-world Chinese offensive speech dataset constructed based on this system. We evaluate the performance of mainstream pre-trained language models (PLMs) and generative large language models (LLMs) on this task, and investigate the impact of different prompt templates on model performance. Our work highlights the urgent need to develop more refined detection methods that can accommodate Chinese implicit speech, in order to counter the evolving evasion strategies.

1 Introduction

OSD has become a focal point of attention in both academia and industry, particularly in the context of maintaining a healthy ecosystem on social media platforms (Fetahi et al., 2023). The development of automated detection technologies for **O**ffensive **S**peech (OS) holds significant importance in this regard. In recent years, the rapid advancements in NLP have opened up numerous new possibilities for OSD (Lai et al., 2023). Alongside this progress, reliable and generalizable benchmark datasets serve as a foundation for in-depth research. Several OSD datasets (Ranasinghe et al., 2024; Delbari et al., 2024) have been introduced in recent years, providing valuable resources for advancing research in this field.

However, OSD in Chinese still faces multiple challenges. (1) *Dataset Scarcity:* Compared to OS datasets in other languages, Chinese datasets are significantly lacking in both quantity and scale (Jiang and Zubiaga, 2024). (2) *Linguistic Features:* Unlike English, Chinese, as a logographic language, lacks explicit word boundaries. Its vocabulary is highly polysemous and context-dependent, with flexible word order and loose grammar (Arcodia and Basciano, 2021). These characteristics make it easier for OS to evade detection through subtle means (e.g., homophones, irony, and metaphor etc.). Traditional detection methods that rely on explicit keywords have limited effectiveness in this context. (3) *Annotation Difficulty:* The scarcity of Chinese corpora and the difficulty of annotation exacerbate this issue. Annotators must possess a deep understanding of language, culture, and context to accurately differentiate between offensive and non-offensive. Therefore, Chinese OSD demands higher levels of semantic comprehension and contextual modeling capabilities.

Existing research on Chinese OSD primarily focuses on explicit speech (Deng et al., 2022; Lu et al., 2023; Xiao et al., 2024a), while the detection of implicit speech remains at an exploratory stage. The progress of Chinese OSD has been slow, largely due to the lack of reliable and comprehensive benchmark datasets. There is an urgent need to develop more refined Chinese OSD datasets, especially those capable of capturing implicit OS.

To address above issues, we introduce SinOffen, a comprehensive dataset for Chinese OSD, aimed at understanding the diversity and complexity of Chinese OS, particularly implicit OS. We collected real-world tweets from Weibo and Douyin between January 2022 and October 2024. Annotators with advanced Chinese language proficiency and cultural expertise were employed to conduct manual annotation. A series of annotation strategies were applied to reduce errors, resulting in a dataset comprising 16,235 samples. Each tweet was labeled as

Work	Source	Type	Domain	Size	Implicit Labels	Public
COLA (2020)	YouTube, Weibo	Real-World	Offensive Speech	18,707	-	✗
SWSR (2022)	Weibo	Real-World	Hate Speech	8,969	-	✓
COLD (2022)	Zhihu, Weibo	Real-World	Offensive Speech	37,480	-	✓
CHSD (2023)	COLD, etc.	Real-World	Offensive Speech	17,430	-	✓
ToxiCN (2023)	Zhihu, Tieba	Real-World	Toxic Speech	12,011	-	✓
ToxiCloakCN (2024)	ToxiCN	Generative	Toxic Speech	4,582	homophones, emoji	✓
PANDA (2025)	COLD, etc.	Generative	Hate Speech	26,420	-	✓
SinOffen (ours)	Weibo, Douyin	Real-World	Offensive Speech	16,235	homophones, circumlocution, metonymy extra knowledge, humiliation, black humor metaphor, irony, visual signs, context	✓

Table 1: Summary of Chinese Offensive Speech Detection Datasets.

Non-OS, *Explicit OS*, or *Implicit OS* based on its content. Additionally, we performed fine-grained categorization of all Implicit OS tweets according to linguistic research and defined a label system with 10 categories. Based on the SinOffen dataset, we systematically evaluated the performance of the most popular PLMs and generative LLMs in Chinese OSD. We also explored the impact of different prompt templates on generative LLMs and analyzed their performance differences in fine-grained classification of implicit OS. The experimental results highlight the challenges in Chinese OSD and suggest future research directions.

The contributions of our paper are as follows: (1) We proposed an open-source Chinese OSD dataset containing 16,235 samples, with *Non-OS* accounting for 36.9%, *Explicit OS* for 31.1%, and *Implicit OS* for 32.0%. (2) For Implicit-OS, we introduced a labeling system with 10 categories (*circumlocution, homophones, metonymy, extra knowledge, humiliation, metaphor, irony, context, visual signs, and black humor*) and conducted fine-grained annotation for all Implicit-OS samples. To the best of our knowledge, this dataset is the most comprehensive real-world Chinese dataset of implicit OS with fine-grained labels. (3) Based on the SinOffen dataset, we evaluated the performance of existing mainstream PLMs and LLMs in Chinese OSD, providing an in-depth analysis of their effectiveness and limitations in the task of detecting OS.

2 Related Work

Real-world Datasets: There exist numerous datasets focused on OSD, hate speech detection (HSD), and toxic speech detection (TSD) across various languages, including English (Ocampo et al., 2023), French (Salaam et al., 2022), Spanish (Monnar et al., 2022), Hindi (Paul et al., 2023), Portuguese (Fortuna et al., 2019), and Korean (Jeong et al., 2022). In recent years, several Chinese datasets for OSD have also been proposed, as shown in the Table 1 . COLA (Tang et al., 2020) provides labeled data, detection systems, and interpretability tools for OSD. SWSR (Jiang et al., 2022) offers a dataset and lexicon for HSD, focusing on sexist content. COLD (Deng et al., 2022) supplies annotated data for OSD to aid model development and evaluation. CHSD (Rao et al., 2023) is created by expanding multiple Chinese OS datasets using both semi-automatic and manual annotation. ToxiCN (Lu et al., 2023) provides a hierarchical taxonomy and resources for TSD.

Generative Datasets: Some researchers have used generative methods to create OS samples for datasets (Hartvigsen et al., 2022), addressing the high cost of manual annotation. This approach enables automatic generation of representative OS samples, expanding dataset size efficiently. ToxiCloakCN (Xiao et al., 2024b) is generated by applying semantic perturbations to the OS samples in ToxiCN, resulting in a dataset with two implicit attributes. PANDA (Bennie et al., 2025) is a dataset constructed using an LLM, zero-shot generation, simulated annealing, and a round-robin algorithm, followed by manual verification.

Our SinOffen dataset is built on real-world data, which more accurately captures the complexities of linguistic and social contexts. Moreover, given the scarcity of Chinese implicit OS datasets, generating high-quality samples for this category using LLMs is challenging. Furthermore, real-world data helps mitigate biases, ensuring greater label consistency and accuracy.

Existing Chinese OS datasets are still limited, particularly in terms of the diversity of implicit categories. Our dataset fills this gap by offering a fine-grained classification of implicit OS, covering a wide range of categories and providing valuable resources for Chinese OSD.

Example Implicit Offensive Tweets
<circumlocution> 原来是要给自己过中元节了, 怪不得你这么兴奋憧憬. *(So it turns out you're celebrating the Zhongyuan Festival for yourself, no wonder you are so excited and looking forward to it.)*
Annotation: The Zhongyuan Festival, also known as the Ghost Festival, is a traditional Chinese holiday dedicated to honoring the dead and expressing mourning. Here, the term <Zhongyuan Festival> is used as a subtle and indirect way to convey offensive speech.
<homophones> 石油就是一个该四的件货, 是一个只顾自己利益的唇珠. *(Oil is a commodity that deserves four, a lip bead that only cares about its own interests.)*
Annotation: In Chinese, < 石油 (oil)> is a homophone for < 室友 (roommate)>, < 件货 (commodity)> is a homophone for < 贱货 (bitch)>, < 四 (four)> is a homophone for < 死 (death)>, and < 唇珠 (lip pearl)> is a homophone for < 蠢猪 (foolish pig)>.
<metonymy> 我看 T0 不知道自己是版本之王, xxn 可以说自己是哺乳期有产后抑郁症家人关心不够哈. *(I see T0 doesn't realize they're the king of the version, and xxn can claim they're in the postpartum period with postnatal depression and not getting enough family care.)*
Annotation: <T0> is an internet slang term used to refer to women. And <xxn> is the abbreviation of < 小仙女 > (little fairy) in pinyin. Above words are often used sarcastically or to mock women.
<context> 支持德军, 德军是世界上最文明最优秀最正义最有道德的军队. *(Support the German military, the German military is the most civilized, outstanding, just, and moral army in the world.)*
Annotation: Based on the context, if the surrounding text includes references to Jewish people or similar topics, it could be considered offensive.
<metaphor> 我看你两耳之间夹的是回族的禁忌. *(I see that what's stuck between your two ears is the Hui people's taboo.)*
Annotation: The Hui people are an ethnic group in China, and due to their religious beliefs, they do not eat pork. Here, the <Hui people's taboo> is used as an indirect metaphor for pigs, conveying offensive remarks.
<irony> 小仙女去一趟隆江就老实了, 她急需找到白菜和粉条. *(The little fairy became docile after a trip to Longjiang. She urgently needs to find cabbage and vermicelli.)*
Annotation: This sentence uses <little fairy> to belittle and mock women, while <Longjiang> is a city in China known for its pig's feet. Meanwhile, <pork, cabbage, and vermicelli stew> is a traditional Chinese dish. By linking <Longjiang> with <find cabbage and vermicelli>, the phrase sarcastically suggests that she is as lowly or vulgar as a pig. This use of language, through the connection of food and regional culture, conveys disrespect and insult toward women.
<extra knowledge> 古人就讲过东郭先生和狼农夫与蛇的故事你还能比老祖宗聪明. *(As the ancients have told us the stories of Mr. Dongguo and the wolf, and the farmer and the snake. Could you be smarter than our ancestors?)*
Annotation: <Mr. Dongguo and the Wolf> and <The Farmer and the Snake> are two classic traditional Chinese anecdotes that convey profound lessons about ingratitude. Here, these anecdotes are referenced to subtly express offensive remarks.
<humiliation> 你们这些人就该待在角落里，别指望能和我们一样成功. *(People like you should stay in the corner and never expect to succeed like we do.)*
Annotation: Indirectly expressing hostility or discrimination towards a target group by belittling, insulting, or degrading someone's dignity.
<black humor> 奥斯维辛水上乐园呢, 快用犹皂洗洗说不定能闻到祖先的味道. *(What about Auschwitz Water Park? Wash it with Jewish soap and maybe you'll smell the scent of your ancestors.)*
Annotation: Here, the <Auschwitz> concentration camp, a historical tragedy, is linked with the lighthearted and entertaining activity of a <water park>. Additionally, the use of the term <Jewish soap>, which is associated with Nazi persecution of Jews, along with the phrase <the scent of ancestors>, further intensifies the offensive and malicious tone. Overall, this sentence mocks and employs black humor of a traumatic historical event, expressing severe disrespect for the Jewish people and their history.
<visual signs> 下辈子要当公 ÷ 的都是 4000+ 的 🗡 🖊 东西. *(Those who want to be male-division in their next life are all 4000+ sword-pen.)*
Annotation: < 公 ÷> refers to <male animal (公畜)>, <4000+> refers to <death of your entire family (死全家)>, and <sword-pen> refers to <bitch (贱逼)>.

Table 2: Examples of Implicit OS. The implicit OS label is in red, the tweets are in blue, and the manual annotations are in black.

3 Taxonomy of Offensive Speech

3.1 Explicit Offensive Speech

Explicit OS involves the direct use of aggressive language to clearly express hostility, discrimination, or insult towards specific groups or individuals. Such speech typically employs offensive vocabulary, derogatory labels, or language imbued with overtly negative emotions (Fortuna and Nunes, 2018). As shown in the example below:

> 穷人就是不配生孩子 (*Poor people are not deserving of having children.*)

3.2 Implicit Offensive Speech

Implicit OS subtly attacks specific groups or individuals without using direct offensive language, yet still aims to belittle, exclude, or incite hostility. Implicit OSD in English has developed rapidly, with fine-grained classifications already in place (ElSherief et al., 2021; Ocampo et al., 2023), while research in Chinese is still in the exploratory stage. Drawing on relevant studies and Chinese linguistics (Arcodia and Basciano, 2021), we identified 10 fine-grained labels for implicit OS (as shown in Table 2), along with examples and detailed definitions.

Circumlocution: Using indirect or roundabout expressions to replace direct insults or attacks, subtly conveying offensive emotions.

Homophones: Leveraging the dual meaning of homophones or near-homophones to make the speech appear harmless while conveying negative or hostile implications.

Metonymy: Substituting symbolic words or things associated with the target group to indirectly con-

Implicit labels	#	%
Circumlocution	4,367	84.1
Homophones	3,186	61.3
Metonymy	1,900	36.6
Context	1,534	29.5
Metaphor	1,481	28.5
Irony	1,005	19.3
Visual signs	762	14.6
Extra knowledge	700	13.5
Humiliation	213	4.1
Black humor	148	2.8
Total	5,195	-

Table 3: Statistics on Implicit OS labels distribution. Implicit OS may encompass multiple labels.

vey discriminatory or derogatory intentions.

Context: Setting a specific context or situational background to make the negative meaning of certain words or phrases more concealed and difficult to detect.

Metaphor: Using metaphors to compare a group to a negative thing or phenomenon, indirectly expressing hostility or exclusion.

Irony: Expressing emotions opposite to the literal meaning through sarcasm, indirectly conveying hostility or belittlement toward the target group.

Extra knowledge: Relying on the audience's understanding of specific background knowledge to convey discriminatory or insulting information that only informed individuals can recognize.

Humiliation: Using belittling, insulting, or dignity-stripping tactics to indirectly express hostility or discrimination toward the target group.

Black humor: Employing black humor or mockery to mask offensive emotions through absurdity, teasing, or sarcasm, implying negative views.

Visual signs: Conveying implicit discrimination or insult through visual elements like images, symbols, or emojis, extending beyond verbal expression.

4 Dataset Construction

4.1 Data Collection

We chose Weibo and Douyin as our data sources due to their status as major social platforms in China, with a wide user base and diverse content. These platforms can provide a variety of Chinese OS data, and their coverage of multiple areas such as society and entertainment makes them ideal for collecting implicit OS. We collected nearly 30,000 tweets between January 2022 and October 2024. Through manual screening and annotation, we ultimately constructed the SinOffen dataset, consisting of 16,235 samples. For Non-OS, to enhance the diversity and challenge of the samples, we collected a large number of classic quotes from literary works to expand the dataset. For OS, the data collection followed three strategies:

Keyword-based Collection: Initially, we identified several core themes through preliminary research, such as fan conflicts, gender disputes, political issues, and anti-LGBTQ. Based on these themes, we compiled a list of relevant keywords and conducted data scraping under each theme (see Appendix A.1 for specific keywords).

Comment Section Collection: For certain controversial Weibo posts and Douyin videos that generated malicious discussions, we scraped the content of their comment sections to further enrich the dataset. This approach focuses on uncovering implicit OS in social interactions, offering a diverse range of emotional expressions.

Bot-Tweets Collection: For automated accounts (Weibo bots) that incite online conflicts on the Weibo, we collected all posts made by these automated accounts through authorized API to capture the group-based divisive speech they provoke. This method helps identify organized malicious behavior, providing more consistent and systematic data support.

4.2 Data Annotation

The annotation task was exclusively conducted by our dedicated research team comprising 10 full-time members (5 male/5 female) recruited from different Chinese provinces. All members have strong Chinese language skills and NLP expertise, ensuring scientific accuracy. Annotators were provided with detailed guidelines and compensated at 150 CNY per hour. The annotation work was conducted using the Label Studios[1] platform for efficient and standardized processing. (See Appendix B.1 for annatation details.) Annotation process was divided into two stages:

Coarse-grained annotation The first stage involves classifying tweets into Non-OS and OS (with Fleiss' s $\kappa = 0.798$ for inter-rater agreement (Fleiss, 1971)), and then further dividing OS into Explicit OS and Implicit OS (with Fleiss' s $\kappa = 0.744$ for inter-rater agreement).

[1]https://labelstud.io/

Prompt Template	Content
Chinese	请根据下列句子的内容判断其是否包含冒犯性内容. '0' 表示该句子不包含冒犯性内容, '1' 表示该句子包含冒犯性内容. 请输出结果 (仅回答'0' 或'1'). 句子: {text}
English	Please determine whether the following sentence contains offensive content. '0' indicates that the sentence is not offensive, and '1' indicates that it is offensive. Please provide your response (only '0' or '1'). Sentence: {text}

Table 4: Different Prompt Templates for Detecting Chinese OS with LLMs

Label	Train		Dev		Test		Total	
	#	%	#	%	#	%	#	%
Non-OS	4,201	36.9	900	36.9	900	36.9	6,001	36.9
Explicit OS	3,527	31.1	756	31.1	756	31.1	5,039	31.1
Implicit OS	3,637	32.0	779	32.0	779	32.0	5,195	32.0

Table 5: Statistics on SinOffen dataset distribution.

Fine-grained annotation: The second stage focused on more detailed annotation of tweets labeled as Implicit OS, covering 10 distinct implicit OS attributes. To ensure the consistency and accuracy of the annotations, each tweet was independently annotated by three different annotators. For consistency evaluation, we randomly selected 200 tweets, and the calculated Fleiss' κ was 0.62, indicating substantial agreement. The final label was determined by the intersection of the annotations from the three annotators (as shown in Table 3). This annotation process minimized potential annotation errors, ensuring the high quality and reliability of the dataset.

Finally, we annotated 6,001 *Non-OS* tweets, 5,039 *Explicit OS* tweets, and 5,195 *Implicit OS* tweets.

5 Experiment

We design three primary tasks to evaluate SinOffen dataset:

Task1: Can PLMs effectively distinguish Explicit OS, Implicit OS, and Non-OS in Chinese OSD?

Task2: How is the performance of LLMs? To what extent does prompt design influence the performance of LLMs in Chinese OSD?

Task3: Can LLMs accurately identify and classify various forms of implicit OS in fine-grained classification tasks?

5.1 Experiment Setup

All experiments in this paper were conducted on the NVIDIA H20, with evaluation metrics including macro-F1, macro-Precision, and macro-Recall. The training, validation, and test set splits used for the experiments are shown in the Table 5. For the PLMs, we fine-tuned for $e \in (3, 4)$ epochs, with learning rates of $lr \in (2e-5, 3e-5)$, and a batch size of 8. For LLMs, we conducted zero-shot experiments and designed two prompt templates in different languages, as shown in the Table 4.

5.2 Baselines

PLMs: In the Task 1, we selected models specifically designed for OSD, including HateBERT (Caselli et al., 2021), ToxiGen-HateBERT (Hartvigsen et al., 2022), RoBERTa-hate-latest (Loureiro et al., 2023), and LFTW R4 (Vidgen et al., 2021). We also chose models suitable for Chinese classification tasks, such as XLM-RoBERTa (Conneau et al., 2019) and BERT-based-chinese (Devlin et al., 2019). Additionally, we selected GPT-2 (Radford et al., 2019), DeBERTa-v3 (He et al., 2021), and ModernBERT (Warner et al., 2024), which are currently among the most comprehensive models with strong overall capabilities. Above models are open-source on Hugging Face[2].

Prompted LLMs: In the Task 2 and Task 3, We selected the current advanced open-source models that perform well in various tasks, including Mistral-7B[3] (Jiang et al., 2023), Llama3.1-8B (AI@Meta, 2024a), and Qwen2.5-7B[4] (Hui et al., 2024). Since Llama natively does not support Chinese, we specifically chose the Llama3.1[5] model fine-tuned for Chinese (for more experimental comparisons of other LLMs, see the Appendix C).

5.3 Results and Discussion

5.3.1 Performance of PLMs

Table 6 presents the experimental results of PLMs on SinoOffen. The results show that for Non-

[2]https://huggingface.co

[3]https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3

[4]https://huggingface.co/Qwen/Qwen2.5-7B-Instruct

[5]https://huggingface.co/shenzhi-wang/Llama3.1-8B-Chinese-Chat

Model	Non-OS			Explicit OS			Implicit OS			All Macro		
	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall
HateBERT	0.8329	0.8096	0.8575	0.6327	0.6034	0.6650	0.5869	0.6462	0.5376	0.6841	0.6863	0.6867
ToxiGen-HateBERT	0.8851	0.8945	0.8758	0.6573	0.5624	0.7909	0.4920	0.6375	0.4006	0.6781	0.6981	0.6891
GPT-2	0.9165	0.9282	0.9050	0.7056	0.6532	0.7671	0.6871	0.7279	0.6506	0.7697	0.7698	0.7742
LFTW R4	0.9226	0.9133	0.9042	0.6922	0.6279	0.7711	0.6345	0.7133	0.5714	0.7498	0.7515	0.7489
RoBERTa-hate-latest	0.9373	0.9511	0.9335	0.6920	0.6429	0.7493	0.6501	0.7034	0.6042	0.7598	0.7657	0.7623
XLM-RoBERTa	0.9681	0.9614	**0.9750**	0.8366	0.7971	0.8801	0.8041	0.8578	0.7568	0.8695	0.8720	0.8706
DeBERTa-v3	0.9639	0.9611	0.9667	0.8071	0.7665	0.8523	0.7736	0.8242	0.7288	0.8482	0.8506	0.8493
ModernBERT	0.9571	0.9653	0.9492	0.8092	0.7952	0.8236	0.8016	0.8078	**0.7954**	0.8560	0.8561	0.8560
BERT-based-Chinese	**0.9701**	**0.9804**	0.9600	**0.8518**	**0.8062**	**0.9029**	**0.8229**	**0.8649**	0.7847	**0.8816**	**0.8838**	**0.8825**

Table 6: Results of Three-Class Chinese OSD with PLMs. The best results are highlighted in **bold**.

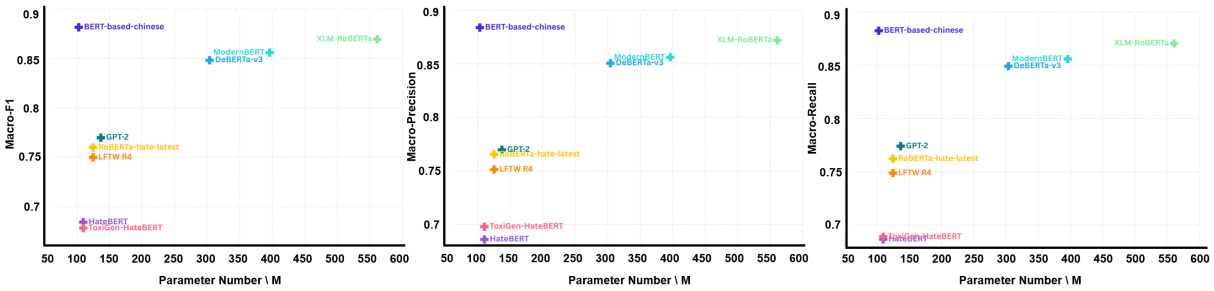

Figure 1: Trend of PLMs' Metrics with Parameter Count. From left to right, the y and x axes represent F1-Parameter, Precision-Parameter, and Recall-Parameter, respectively. See Appendix E for Parameter Count details.

Model	Template	Non-OS			Explicit OS			All Macro		
		F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall
Mistral-7B	Chinese	0.7407	0.9114	0.6239	0.7811	0.6745	0.9278	0.7609	0.7929	0.7759
	English	0.7606	0.8962	0.6606	0.7859	0.6923	0.9089	**0.7733**	**0.7942**	**0.7848**
Llama3.1-8B	Chinese	0.8432	0.7434	0.9740	0.7352	0.9508	0.5993	0.7892	0.8471	0.7867
	English	0.8640	0.7886	0.9555	0.7952	0.9292	0.6950	**0.8296**	**0.8589**	**0.8252**
Qwen2.5-7B	Chinese	0.8694	0.8303	0.9123	0.8254	0.8808	0.7765	0.8474	0.8555	0.8444
	English	0.8543	0.9476	0.7778	0.8573	0.7820	0.9488	**0.8558**	**0.8648**	**0.8633**

Table 7: Results of Binary Non-OS & Explicit OS with LLMs. The best results are highlighted in **bold**.

Model	Template	Non-OS			Implicit OS			All Macro		
		F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall
Mistral-7B	Chinese	0.7199	0.8513	0.6236	0.7565	0.6671	0.8737	0.7382	**0.7592**	0.7487
	English	0.7344	0.8298	0.6586	0.7568	0.6801	0.8431	**0.7456**	0.7550	**0.7509**
Llama3.1-8B	Chinese	0.8101	0.6934	0.9740	0.6540	0.9432	0.5005	0.7320	0.8183	0.7372
	English	0.8428	0.7538	0.9555	0.7557	0.9253	0.6386	**0.7992**	**0.8396**	**0.7971**
Qwen2.5-7B	Chinese	0.8413	0.7347	0.9123	0.7774	0.8727	0.7009	0.8093	0.8037	0.8066
	English	0.8380	0.9083	0.7778	0.8392	0.7794	0.9091	**0.8386**	**0.8438**	**0.8434**

Table 8: Results of Binary Non-OS & Implicit OS with LLMs. The best results are highlighted in **bold**.

OS, all models performed well, with BERT-based-chinese significantly outperforming other models in terms of F1 and Precision scores, indicating its ability to effectively capture subtle semantic differences in Chinese text. For Explicit OS, BERT-based-chinese outperformed all other models across the three metrics. For Implicit OS, BERT-based-chinese excelled in F1 and Precision, while ModernBERT led all models in Recall. Overall, BERT-based-chinese significantly outperforms all baseline models in the Chinese offensive language classification task.

Additionally, we explored the relationship between the number of parameters in PLMs and classification performance. As shown in the Figure 1, except for BERT-based-chinese, the number of parameters in the other models is positively correlated with all metrics—larger parameter sizes lead to higher classification accuracy. This trend suggests that increasing model complexity helps cap-

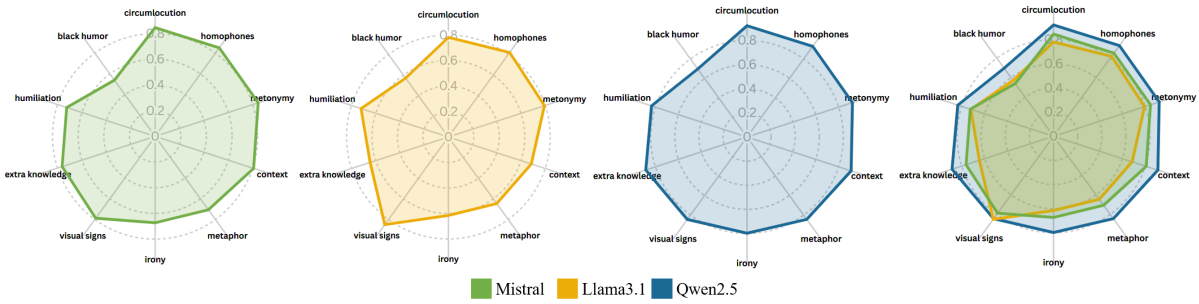

Figure 2: Comparison of Macro-F1 for Different LLMs on Different Fine-Grained Implicit OS Labels. A higher value indicates better classification performance.

ture more linguistic features and semantic information. Despite having fewer parameters, BERT-based-chinese still performs excellently in multiple tasks, demonstrating its specific advantage in Chinese classification tasks.

Discussion: Our experimental results show that PLMs with extensive Chinese corpus pre-training (e.g., BERT-based-chinese, DeBERTa-v3, ModernBERT, XLM-RoBERTa) achieve superior performance in this task. This advantage stems from their optimized handling of Chinese's high-context isolating nature, where other models struggle with tokenization and semantic parsing due to cross-linguistic structural discrepancies. While cross-lingual models exhibit inadequate recognition of implicit OS through insufficient incorporation of Chinese cultural corpora, Chinese-pre-trained models optimized for local linguistic features show greater domain-specific performance.

5.3.2 Performance of LLMs

PLMs perform excellently in the three-classification task. Through multi-task learning, they can deeply explore the semantic differences between Explicit OS, Implicit OS, and Non-OS content, thereby enhancing discriminative ability. In contrast, generative LLMs excel at task-solving under carefully designed prompts (Sahoo et al., 2024). To conduct the same experimental task as with PLMs, the prompt must specify the requirements of the three-class task. However, if the prompt is too complex (e.g., requiring examples of implicit OS for each category), it may increase the classification burden and lead to confusion in the results. Therefore, we propose decomposing the task into two binary classification tasks (*Non-OS & Explicit OS, Non-OS & Implicit OS*), simplifying the learning objectives so that the model can focus more on distinguishing between OS and Non-OS content.

Results of Binary: The Table 7 and 8 presents the performance of different LLMs in the Chinese OSD task. Notably, the Mistral model demonstrates weaker classification performance in the Non-OS category compared to its performance in the OS category, with the macro-F1 score being 2.5% lower for Explicit OS and 2.2% lower for Implicit OS (English Template). A similar trend is observed with Qwen2.5 in certain cases. Additionally, the overall classification metrics for implicit OS are consistently lower than those for explicit OS, underscoring the difficulty in classifying implicit OS. Among all models, Qwen2.5 consistently outperforms the others in both tasks, demonstrating its superior classification ability.

Discussion: In our dataset, the Non-OS contains a large number of sentences from literary works, where critical language is often used to reflect on social phenomena. For example,

> 你嘴上不诉苦就没有人可怜你 (*No one will pity you if you do not complain.*)

from Jane Austen' s *Pride and Prejudice* was classified as offensive speech by all models. This may be because the models associate words like *'complain'* and *'pity'* with negative emotions, incorrectly interpreting them as insulting or offensive. Additionally, generative models are typically trained to avoid producing offensive or harmful content (Chua et al., 2024), which may lead them to be overly cautious when processing text with ambiguous boundaries, resulting in the misclassification of texts that do not fully meet the definition of OS. This phenomenon highlights the limitations of current LLMs in sentiment analysis, context understanding, and handling cultural differences. The models fail to accurately capture the subtle emotional and critical connotations of sentences in the

Non-OS category, reflecting the challenges faced by sentiment analysis and NLP models when dealing with complex contexts.

Results of Different Prompt Template: The Table 7 and 8 presents the classification results under different language prompt templates. The results show that for Implicit OS, the prompt classification results using English templates outperformed those using Chinese templates for all three models. A similar pattern was observed for Explicit OS.

Discussion: This could be because most open-source LLMs are typically trained on English corpora, which are often more prevalent than corpora in other languages. This means that the models may have a stronger semantic understanding of English than of Chinese. Additionally, Chinese, with its inherent high compression and polysemy, may make it more difficult for models to accurately understand intent when expressing complex tasks (for instance, the word *"offensive"* in English has multiple meanings in Chinese, including *offensive*, *aggressive*, *rude*, *unpleasant*, etc.). When translating to an English template, the Chinese expression might naturally be supplemented with more semantic details or logical information, making it easier for the model to infer the task's intent.

5.3.3 LLMs in Implicit Offensive Speech

The experimental setup is detailed in the Appendix D.1. Appendix D.2 presents the classification performance of LLMs on fine-grained labels in implicit OS, with all detailed results included. According to the experimental results, Qwen exhibited the best overall classification performance, while Llama performed particularly well in the Visual signs category (as shown in Figure 2). At the same time, for all implicit OS categories, especially in the *metaphor* (F1-Llama=0.6456, F1-Mistral=0.7065, F1-Qwen=0.8440), *irony* (F1-Llama=0.6156, F1-Mistral=0.6729, F1-Qwen=0.7979), and *black humor* (F1-Llama=0.5714, F1-Mistral=0.5401, F1-Qwen=0.6968), all three models showed suboptimal performance.

Discussion: The results show that Qwen2.5 excels in implicit OS classification, likely because Qwen implements Chinese-specific optimizations through vocabulary expansion with an extensive set of Chinese-centric tokens, enhanced subword regularization trained on large web-crawled Chinese corpora containing modern slang, and adaptive segmentation rules for Chinese morphology. In contrast, although Llama3.1 and Mistral employ Chinese fine-tuning, their linguistic limitations persist. Llama3.1 excels in the Visual Signs category, benefiting from its multimodal training and optimized feature extraction and decision boundaries. All three models show poor performance in the metaphor, irony, and black humor categories, which require a deep understanding of the ironic contradiction between literal meaning and actual intent. The shortcomings of LLMs in these tasks mainly lie in their ability to understand complex cultural contexts and puns. Implicit OS is closely tied to specific cultural and linguistic habits, with certain expressions (such as black humor) being common in some cultures but difficult to understand in others. Although LLMs are trained in multilingual and multicultural contexts, they still face limitations in capturing culturally specific implicit expressions.

6 Conclusion

In this paper, we introduce a OS taxonomy system. All samples are labeled with *Non-OS*, *Explicit OS*, and *Implicit OS*, with Implicit OS further categorized into 10 different labels (*circumlocution, homophones, metonymy, context, metaphor, irony, visual signs, extra knowledge, humiliation* and *black humor*). Based on this taxonomy system, we have constructed the most comprehensive Chinese OSD dataset to date, particularly focusing on implicit OS labels. Our objective is to advance the development of Chinese OSD, particularly by addressing the existing gap in the detection of Chinese implicit OS.

Furthermore, we present several of the most popular and advanced baseline models for offensive speech classification and discussion. Experimental results show that pre-training and fine-tuning in Chinese corpora significantly improve the classification accuracy of PLMs for this task. In our exploration of LLMs, we investigate the impact of prompt engineering on the performance of zero-shot tasks in generative LLMs. At the same time, we found that although LLMs show some potential in handling implicit OS, their ability to process more subtle types of OS remains limited, leading to unsatisfactory results. Future work can focus on several directions, including the sarcasm detection (Liu et al., 2024; Zhu et al., 2024; Lin et al., 2024), the refinement of prompt engineering (Lee et al., 2024), and the expansion of our dataset.

Limitations

The limitations of this paper primarily lie in the following aspects. (1) *Annotation Errors:* Since our annotations are subjective, although various strategies were employed to minimize annotation errors, there remains a possibility of inaccuracies in the labeling. (2) *Baseline Models:* The baseline models we employed are all open-source models. However, we acknowledge that incorporating other advanced closed-source models (such as GPT4, etc.) would strengthen the comparison.

Ethical Considerations

Data Collection & Privacy Compliance

This study complies with China's Personal Information Protection Law (PIPL). The dataset was constructed from publicly accessible content on Weibo and Douyin. Data acquisition strictly followed the platforms' Developer API terms of service and privacy policies (e.g., Weibo Open API). Only text content explicitly marked as public by users was collected, excluding private messages, geolocation tags, or biometric data. All personally identifiable information (PII), including usernames, user IDs, and profile links, was permanently removed using regular expression matching. No sensitive attributes (e.g., ethnicity, political affiliation) were inferred or stored.

Annotation Process

The dataset contains content that may include disturbing or offensive materials, but no sensitive personal identifiers were involved in the annotation process. All annotation work was exclusively conducted by trained research team members who voluntarily participated after thorough protocol orientation. Prior to engagement, each annotator signed informed consent forms specifically detailing: 1) the non-personal nature of the data content, 2) potential exposure to objectionable material patterns, and 3) their unconditional right to pause or terminate participation. To ensure ethical practice, we implemented three safeguard measures: mandatory cool-down intervals between annotation sessions, real-time access to counseling support, and anonymous well-being check-ins conducted weekly by project supervisors.

Intended Use

The dataset was created solely for academic research purposes. Our work is not aimed at any specific group or individual, but rather focuses on providing reliable research outcomes to promote social harmony and public safety.

We are committed to open-sourcing our dataset in order to foster the advancement of Chinese OSD research. We believe that by sharing this resource, we can provide more opportunities for academic and applied research, thus promoting innovation and development in the field. While we are aware that open-sourcing the dataset may present certain risks, we firmly believe that the potential benefits far outweigh these risks.

Accountability

Users may submit the following requests through the designated contact: Correction of labeling errors; Reporting of abusive behavior (processed within 72 hours upon official verification of account ownership by the platform).

References

AI@Meta. 2024a. Llama 3.1 model card.

AI@Meta. 2024b. Llama 3.2 model card.

Giorgio Francesco Arcodia and Bianca Basciano. 2021. *Chinese linguistics: An introduction*. Oxford University Press.

Michael Bennie, Demi Zhang, Bushi Xiao, Jing Cao, Chryseis Xinyi Liu, Jian Meng, and Alayo Tripp. 2025. Panda–paired anti-hate narratives dataset from asia: Using an llm-as-a-judge to create the first chinese counterspeech dataset. *arXiv preprint arXiv:2501.00697*.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2021. HateBERT: Retraining BERT for abusive language detection in English. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 17–25. Association for Computational Linguistics.

Jaymari Chua, Yun Li, Shiyi Yang, Chen Wang, and Lina Yao. 2024. Ai safety in generative ai large language models: A survey. *arXiv preprint arXiv:2407.18369*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Zahra Delbari, Nafise Sadat Moosavi, and Mohammad Taher Pilehvar. 2024. Spanning the spectrum of hatred detection: a persian multi-label hate speech dataset with annotator rationales. In *Proceedings of*

the AAAI Conference on Artificial Intelligence, volume 38, pages 17889–17897.

Jiawen Deng, Jingyan Zhou, Hao Sun, Chujie Zheng, Fei Mi, Helen Meng, and Minlie Huang. 2022. COLD: A benchmark for Chinese offensive language detection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11580–11599, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Endrit Fetahi, Mentor Hamiti, Arsim Susuri, Visar Shehu, and Adrian Besimi. 2023. Automatic hate speech detection using natural language processing: A state-of-the-art literature review. In *2023 12th Mediterranean Conference on Embedded Computing (MECO)*, pages 1–6. IEEE.

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

Paula Fortuna, Joao Rocha da Silva, Leo Wanner, Sérgio Nunes, et al. 2019. A hierarchically-labeled portuguese hate speech dataset. In *Proceedings of the third workshop on abusive language online*, pages 94–104.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):1–30.

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. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3326, Dublin, Ireland. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *Preprint*, arXiv:2111.09543.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Kai Dang, et al. 2024. Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186*.

Younghoon Jeong, Juhyun Oh, Jongwon Lee, Jaimeen Ahn, Jihyung Moon, Sungjoon Park, and Alice Oh. 2022. KOLD: Korean offensive language dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10818–10833, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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.

Aiqi Jiang and Arkaitz Zubiaga. 2024. Cross-lingual offensive language detection: A systematic review of datasets, transfer approaches and challenges. *arXiv preprint arXiv:2401.09244*.

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*.

Mirko Lai, Fabio Celli, Alan Ramponi, Sara Tonelli, Cristina Bosco, and Viviana Patti. 2023. Haspeede3 at evalita 2023: Overview of the political and religious hate speech detection task.

Unggi Lee, Haewon Jung, Younghoon Jeon, Younghoon Sohn, Wonhee Hwang, Jewoong Moon, and Hyeoncheol Kim. 2024. Few-shot is enough: exploring chatgpt prompt engineering method for automatic question generation in english education. *Education and Information Technologies*, 29(9):11483–11515.

Yucheng Lin, Yuhan Xia, and Yunfei Long. 2024. Augmenting emotion features in irony detection with large language modeling. *arXiv preprint arXiv:2404.12291*.

Hao Liu, Runguo Wei, Geng Tu, Jiali Lin, Cheng Liu, and Dazhi Jiang. 2024. Sarcasm driven by sentiment: A sentiment-aware hierarchical fusion network for multimodal sarcasm detection. *Information Fusion*, 108:102353.

Daniel Loureiro, Kiamehr Rezaee, Talayeh Riahi, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and Jose Camacho-Collados. 2023. Tweet insights: a visualization platform to extract temporal insights from twitter. *arXiv preprint arXiv:2308.02142*.

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.

Ayme Arango Monnar, Jorge Pérez, Barbara Poblete, Magdalena Saldaña, and Valentina Proust. 2022. Resources for multilingual hate speech detection. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 122–130.

Nicolás Benjamín Ocampo, Ekaterina Sviridova, Elena Cabrio, and Serena Villata. 2023. An in-depth analysis of implicit and subtle hate speech messages. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1997–2013, Dubrovnik, Croatia. Association for Computational Linguistics.

Sayanta Paul, Sriparna Saha, and Jyoti Prakash Singh. 2023. Covid-19 and cyberbullying: deep ensemble model to identify cyberbullying from code-switched languages during the pandemic. *Multimedia tools and applications*, 82(6):8773–8789.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Tharindu Ranasinghe, Isuri Anuradha, Damith Premasiri, Kanishka Silva, Hansi Hettiarachchi, Lasitha Uyangodage, and Marcos Zampieri. 2024. Sold: Sinhala offensive language dataset. *Language Resources and Evaluation*, pages 1–41.

Xiaojun Rao, Yangsen Zhang, Qilong Jia, and Xueyang Liu. 2023. Chinese hate speech detection method based on roberta-wwm. In *Proceedings of the 22nd Chinese National Conference on Computational Linguistics*, pages 501–511, Harbin, China. Chinese Information Processing Society of China.

Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. 2024. A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*.

Cesa Salaam, Franck Dernoncourt, Trung Bui, Danda Rawat, and Seunghyun Yoon. 2022. Offensive content detection via synthetic code-switched text. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6617–6624, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Sara Sekkate, Safa Chebbi, Abdellah Adib, and Sofia Ben Jebara. 2024. A deep learning framework for offensive speech detection. In *2024 IEEE 12th International Symposium on Signal, Image, Video and Communications (ISIVC)*, pages 1–6.

Xiangru Tang, Xianjun Shen, Yujie Wang, and Yujuan Yang. 2020. Categorizing offensive language in social networks: A chinese corpus, systems and an explanation tool. In *Chinese Computational Linguistics: 19th China National Conference, CCL 2020, Hainan, China, October 30–November 1, 2020, Proceedings 19*, pages 300–315.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021. Learning from the worst: Dynamically generated datasets to improve online hate detection. 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 1667–1682, Online. Association for Computational Linguistics.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *Preprint*, arXiv:2412.13663.

Yunze Xiao, Houda Bouamor, and Wajdi Zaghouani. 2024a. Chinese offensive language detection: Current status and future directions. *arXiv preprint arXiv:2403.18314*.

Yunze Xiao, Yujia Hu, Kenny Tsu Wei Choo, and Roy Ka-Wei Lee. 2024b. 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.

Zhihong Zhu, Xianwei Zhuang, Yunyan Zhang, Derong Xu, Guimin Hu, Xian Wu, and Yefeng Zheng. 2024. Tfcd: Towards multi-modal sarcasm detection via training-free counterfactual debiasing. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 6687–6695. International Joint Conferences on Artificial Intelligence Organization. Main Track.

A Data Collection

A.1 Keyword-based Collection

We identified four main themes for keyword search: *Gender*, *LGBTQ+*, *Fans Conflict*, and *Politics*. Based on the keywords listed in the Table 9, we conducted searches on Weibo and Douyin and collected the relevant data.

Topic	Keywords
Gender	小仙女, 女性, 女人, 男性, 男人, 国男, 女权, 女拳, 楠, 男权, 男拳, 老天奶, 老天爷, 爱女, 爱男, 厌女, 厌男
LGBTQ+	同性恋, 男同, 女同, 南通, 钦铜, 通讯录, txl, 给子, 拉子, gay, les, 跨性别
Fans Conflict	饭圈, 体育圈, 电竞圈, 哈圈, 欧美圈, 内娱, Kpop, 韩圈, 说唱圈, 粉丝, 爱豆, 歌手, 歌迷, 难听, 难看
Politics	棒子, 鬼子, 鱿鱼, 犹太人, 以色列, 美国, 哈马斯, 伊斯兰, 日本, 韩国, 俄罗斯, 乌克兰

Table 9: The keywords used for each theme.

B Data Annotation

B.1 Annotation Guidelines

We provided annotators with annotation guidelines. In the first stage, all tweets were annotated as either Non-OS or OS, with the definition of Offensive as follows:

> ***Offensive:*** OS generally denotes verbal expressions that are likely to cause discomfort, anger, humiliation, or other adverse emotional responses from others. Such expressions may encompass content that involves belittlement, insult, and discrimination directed at individuals or groups, spanning various dimensions including race, gender, religion, sexual orientation, and physical characteristics (Sekkate et al., 2024).

Subsequently, all instances of OS were further annotated as either Explicit OS or Implicit OS, with the definitions of Explicit OS and Implicit OS as indicated in Section 3.1. The second stage involved fine-grained label annotation of Implicit OS, with the definitions provided in Section 3.2 of the main text.

In our dataset, the data format is as follows:

- 原来是要给自己过中元节了, *implicit*, *[circumlocution, extra knowledge]*
- 记住我这张死后会来找你索命的脸, *explciit*, *none*
- 爱就是任何理智的高墙也抵挡不了那个人的一声叫唤, *non-offen*, *none*

Figure 3 is the Label Studio interface used during the two-stage annotation process. In the first stage, we first annotate *Non-OS* and *OS* content, and then classify OS into *Explicit OS* and *Implicit OS*. In the second stage, we perform fine-grained annotation of Implicit OS into 10 categories.

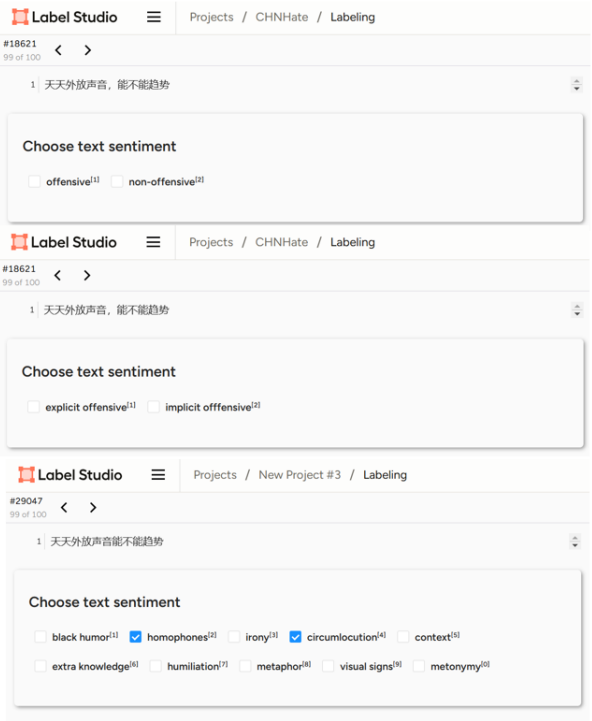

Figure 3: Data annotation on Label Studios.

B.2 Word Cloud Distribution

To investigate the differences between annotated Implicit OS and Explicit OS, we plotted word clouds for both categories based on word frequency, as shown in the Figure 4. It can be observed that Implicit OS often includes abbreviations, euphemisms, and metaphors, while Explicit OS tends to involve specific groups and insulting language.

C LLMs Performance Details in Binary Classification Task

Tables 11 and 12 present additional model classification results for Task 2, including models not mentioned in the main text, such as hfl-Llama3-8B[6], Meta-Llama3.1-8B[7] (AI@Meta, 2024a), and

[6]https://huggingface.co/hfl/llama-3-chinese-8b-instruct-v3

[7]https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

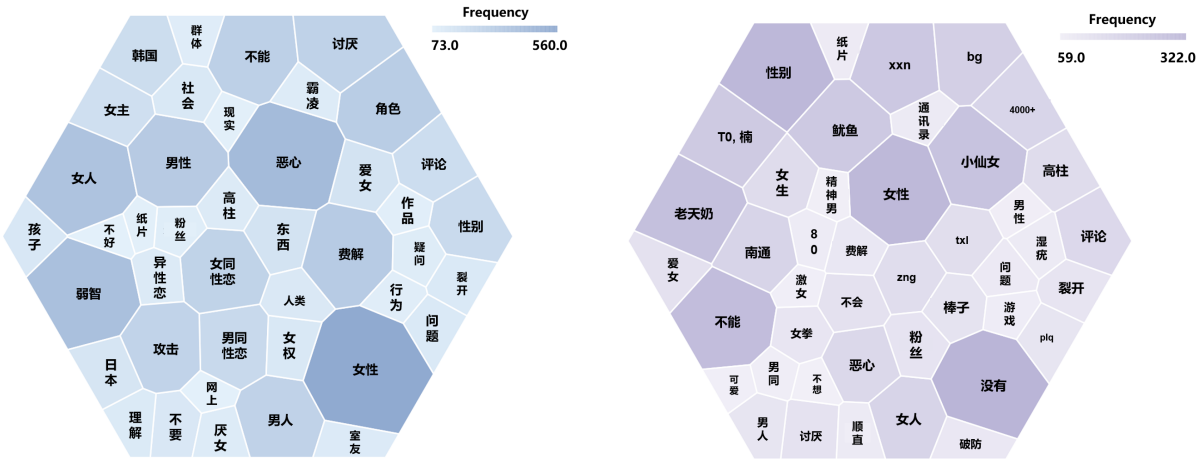

Figure 4: Word Cloud Distribution of Implicit OS (Right) and Explicit OS (Left).

Meta-Llama3.2-3B[8] (AI@Meta, 2024b). Among them, we selected the Llama3.1-8B (fine-tuned by shenzhi-wang) model, which showed the best classification performance, for inclusion in the main text experiments.

D LLMs Performance Details in fine-grained Implicit OS

D.1 Experiment Setup

For this experiment, we first divided Implicit OS into 10 subcategories based on different fine-grained labels, with each subcategory representing a specific type of implicit OS. Next, we combined the OS data from these subcategories with Non-OS data to form 10 sub-datasets. Given that different sub-datasets may have issues with sample imbalance, particularly with relatively fewer OS samples, we applied undersampling to the Non-OS data within these sub-datasets to balance the number of samples between the OS and Non-OS categories. Undersampling was implemented by randomly removing some of the Non-OS samples, ensuring that the class distribution in each sub-dataset remained as balanced as possible.

D.2 Results of fine-grained Implicit OS

Table 13 presents detailed classification results of LLMs on different fine-grained Implicit OS categories, with metrics including F1, Precision, and Recall. To provide a clearer and more intuitive presentation of the results, we have plotted bar charts (as shown in the Figure 5-Figure 7).

[8]https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct

E PLMs Parameter Display

Table 10 illustrates the specific parameter quantities of the PLM utilized in this paper.

Models	Parameter Number
BERT-based-chinese	103M
HateBERT	110M
ToxiGen-HateBERT	110M
LFTW R4	125M
RoBERTa-hate-latest	125M
GPT-2	137M
DeBERTa-v3	304M
ModernBERT	396M
XLM-RoBERTa	561M

Table 10: Detailed PLMs Parameter Numbers.

Model	Template	Non-OS			Implicit OS			All Macro		
		F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall
Mistral-7B	Chinese	0.7199	0.8513	0.6236	0.7565	0.6671	0.8737	0.7382	**0.7592**	0.7487
	English	0.7344	0.8298	0.6586	0.7568	0.6801	0.8431	**0.7456**	0.7550	**0.7509**
Llama3-8B	Chinese	0.7370	0.8264	0.6026	0.7378	0.6498	0.8535	0.7374	0.7381	0.7281
(hfl)	English	0.7939	0.8466	0.7474	0.7897	0.7425	0.8432	**0.7918**	**0.7945**	**0.7953**
Llama3.1-8B	Chinese	0.7468	0.6125	0.9565	0.4420	0.8551	0.2980	0.5944	0.7338	0.6273
(Meta)	English	0.7904	0.6701	0.9633	0.6031	0.9136	0.4501	**0.6968**	**0.7919**	**0.7067**
Llama3.1-8B	Chinese	0.8101	0.6934	0.9740	0.6540	0.9432	0.5005	0.7320	0.8183	0.7372
(shenzhi-wang)	English	0.8428	0.7538	0.9555	0.7557	0.9253	0.6386	**0.7992**	**0.8396**	**0.7971**
Llama3.2-3B	Chinese	0.6998	0.5422	0.9865	0.0708	0.7044	0.0373	0.3853	0.6233	0.5119
(Meta)	English	0.6264	0.6884	0.5747	0.6382	0.5870	0.6991	**0.6323**	**0.6377**	**0.6369**
Qwen2.5-7B	Chinese	0.8413	0.7347	0.9123	0.7774	0.8727	0.7009	0.8093	0.8037	0.8066
	English	0.8380	0.9083	0.7778	0.8392	0.7794	0.9091	**0.8386**	**0.8438**	**0.8434**

Table 11: Results of Binary Non-OS & Implicit OS with LLMs. The best results are highlighted in **bold**.

Model	Template	Non-OS			Explicit OS			All Macro		
		F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall
Mistral-7B	Chinese	0.7407	0.9114	0.6239	0.7811	0.6745	0.9278	0.7609	0.7929	0.7759
	English	0.7606	0.8962	0.6606	0.7859	0.6923	0.9089	**0.7733**	**0.7942**	**0.7848**
Llama3-8B	Chinese	0.7124	0.8718	0.6022	0.7556	0.6539	0.8946	0.7340	0.7629	0.7484
(hfl)	English	0.8072	0.8873	0.7403	0.8083	0.7418	0.8881	**0.8078**	**0.8145**	**0.8142**
Llama3.1-8B	Chinese	0.7736	0.6496	0.9563	0.5350	0.8805	0.3842	0.6543	0.7650	0.6703
(Meta)	English	0.8059	0.6937	0.9615	0.6417	0.9150	0.4941	**0.7238**	**0.8043**	**0.7278**
Llama3.1-8B	Chinese	0.8432	0.7434	0.9740	0.7352	0.9508	0.5993	0.7892	0.8471	0.7867
(shenzhi-wang)	English	0.8640	0.7886	0.9555	0.7952	0.9292	0.6950	**0.8296**	**0.8589**	**0.8252**
Llama3.2-3B	Chinese	0.7099	0.5543	0.9868	0.1063	0.7842	0.0570	0.4081	**0.6692**	0.5219
(Meta)	English	0.6348	0.7099	0.5741	0.6477	0.5878	0.7213	**0.6413**	0.6488	**0.6477**
Qwen2.5-7B	Chinese	0.8694	0.8303	0.9123	0.8254	0.8808	0.7765	0.8474	0.8555	0.8444
	English	0.8543	0.9476	0.7778	0.8573	0.7820	0.9488	**0.8558**	**0.8648**	**0.8633**

Table 12: Results of Binary Non-OS & Explicit OS with LLMs. The best results are highlighted in **bold**.

Model	Metric	Circumlocation	Homophones	Metonymy	Context	Metaphor	Irony	Visual signs	Extra Knowledge	Humiliation	Black humor
Mistral-7B	F1	0.8464	0.8515	0.8453	0.8079	0.7065	0.6729	0.7881	0.7621	0.7254	0.5401
	Precision	0.8474	0.8187	0.8355	0.8275	0.6811	0.6615	0.6721	0.7599	0.6436	0.4815
	Recall	0.8454	0.8870	0.8553	0.7892	0.7340	0.6846	0.9527	0.7643	0.8310	0.6149
Llama3.1-8B	F1	0.7812	0.8191	0.7937	0.6848	0.6456	0.6156	0.8502	0.6465	0.7209	0.5714
	Precision	0.9843	0.9786	0.9814	0.9782	0.9715	0.9660	0.9722	0.9553	0.9466	0.8986
	Recall	0.6476	0.7043	0.6663	0.5268	0.4835	0.4517	0.7554	0.4886	0.5822	0.4189
Qwen2.5-7B	F1	0.9217	0.9289	0.9210	0.9093	0.8440	0.7979	0.8447	0.8845	0.8341	0.6968
	Precision	0.9366	0.9247	0.9307	0.9401	0.8406	0.7761	0.7503	0.8996	0.7796	0.6667
	Recall	0.9073	0.9331	0.9116	0.8805	0.8474	0.8209	0.9662	0.8700	0.8967	0.7297

Table 13: Results of LLMs on different fine-grained Implicit OS categories.

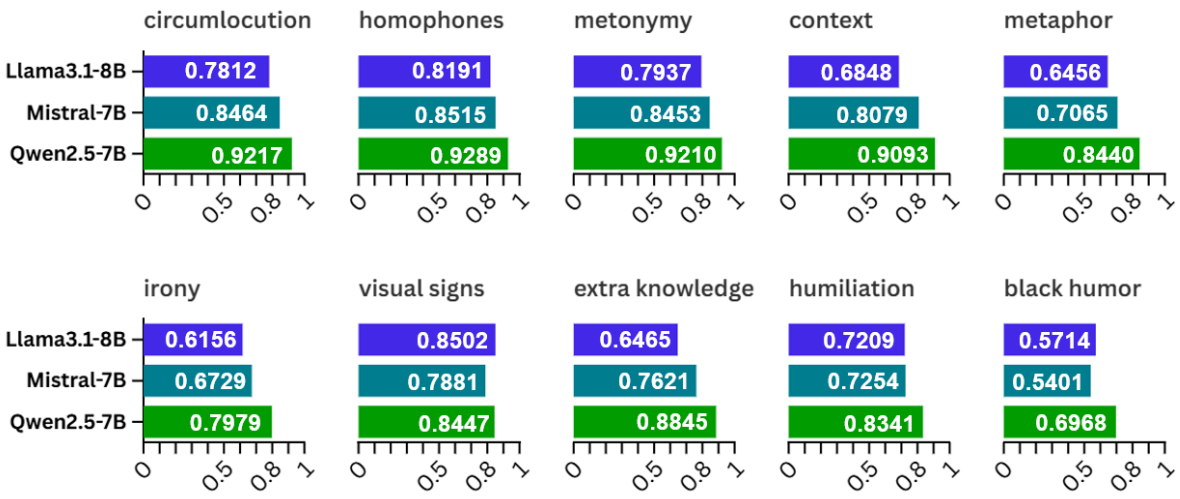

Figure 5: Results on Different Fine-grained Implicit OS labels with Macro-F1 as the Evaluation Metric.

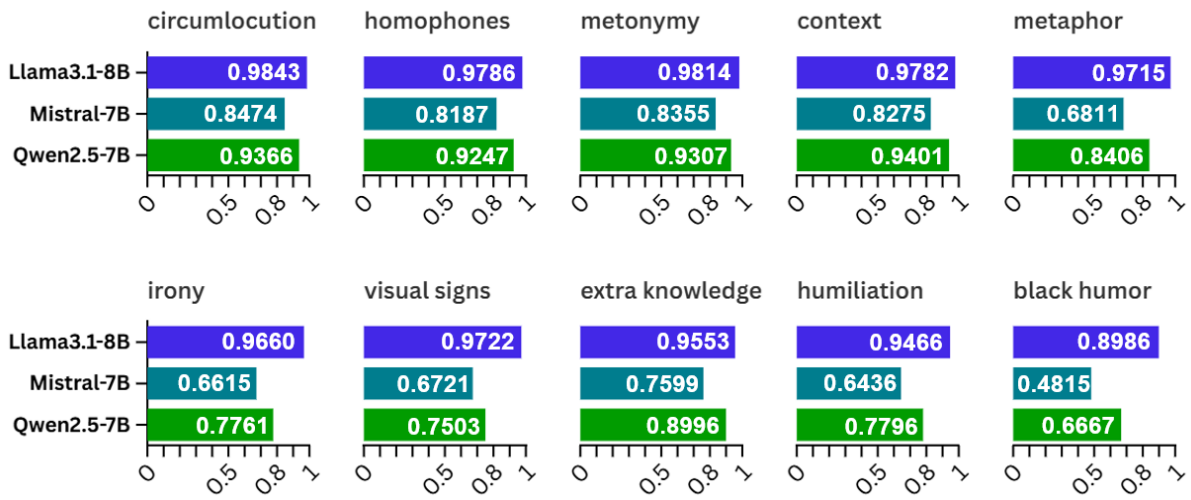

Figure 6: Results on Different Fine-grained Implicit OS labels with Macro-Precision as the Evaluation Metric.

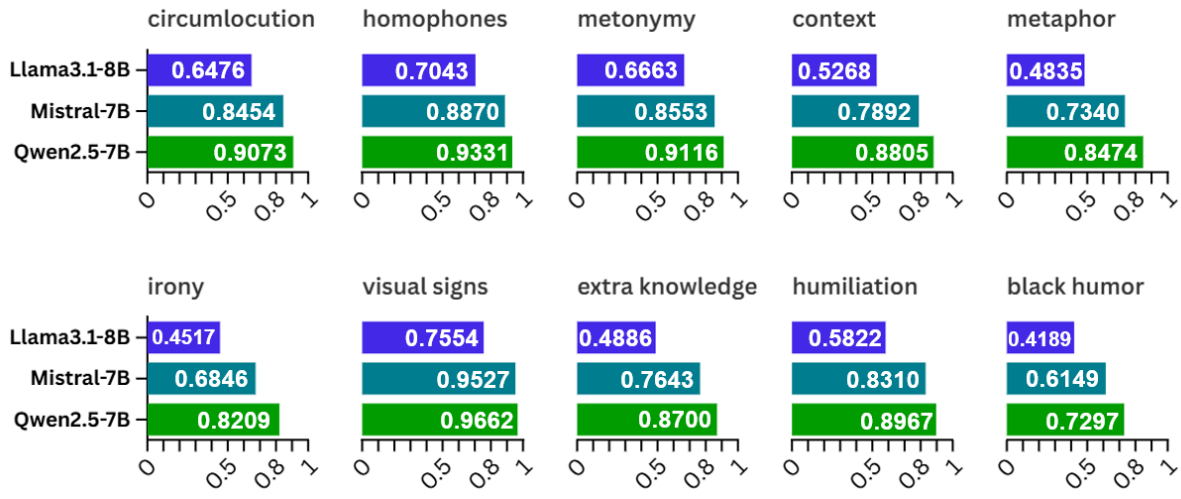

Figure 7: Results on Different Fine-grained Implicit OS labels with Macro-Recall as the Evaluation Metric.