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 analyze the underlying causes of performance drop in implicit OSD. 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 Chinese OS to evade detection through subtle means (e.g., homophones, irony, and metaphor etc.) (Xiao et al., 2024b). 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 researches on Chinese OSD have made significant progress in identifying explicit speech (Xiao et al., 2024a), but the detection of implicit speech remains exploratory stage. The overall progress of Chinese OSD has been slow, primarily due to the lack of reliable and comprehensive benchmark datasets. There is an urgent need to develop more refined Chinese OSD datasets, particularly 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	-	✓
CDIAL-BIAS (2022)	Zhihu	Real-World	Bias Speech	28,243	-	✓
CHSD (2023)	COLD, etc.	Real-World	Offensive Speech	17,430	-	✓
ToxiCN (2023)	Zhihu, Tieba	Real-World	Toxic Speech	12,011	-	✓
CPCL (2024)	Zhihu, Weibo	Real-World	Patronizing Speech	18,253	unbalanced power, spectator, prejudice appeal, elicit compassion	✓
ToxiCloakCN (2024)	ToxiCN	Generative	Toxic Speech	4,582	homophones, emoji	✓
PANDA (2025)	COLD, etc.	Generative	Hate Speech	26,420	-	✓
SCCD (2025)	Weibo	Real-World	Toxic Speech	677	-	✓
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: Several datasets have been developed to address the tasks of OSD, Hate Speech Detection (HSD), and Toxic Speech Detection (TSD) in Chinese, as detailed in Table 1. COLA (Tang et al., 2020), COLD (Deng et al., 2022) and CHSD (Rao et al., 2023) provide labeled data, detection systems, and interpretability tools for OSD. CDIAL (Zhou et al., 2022) is Chinese dialogue dataset for social bias. SWSR (Jiang et al., 2022) offers a dataset and lexicon for HSD. CPCL (Wang et al., 2024) is dataset for patronizing and condescending language. ToxiCN (Lu et al., 2023) and SCCD (Yang et al., 2025) 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> 像你这样的人能做到这一步, 真的挺意外的. *(People like you to get this far, that's really surprising.)*
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), while research in Chinese is still in the exploratory stage. Inspired by the relevant research (Ocampo et al., 2023) and Chinese linguistics (Arcodia and Basciano, 2021), We propose a set of 10 fine-grained annotation categories specifically designed for Chinese implicit OS, including *circumlocution*, *homophones*, *metonymy*, *context*, *metaphor*, *irony*, *visual signs*, *extra knowledge*, *humiliation* and *black humor*. They account for nearly all prevalent forms of implicit OS on the Chinese internet. Representative examples are provided in Table 2, and detailed definitions of fine-grained labels can be found in the Appendix C.

Table 3 presents the distribution of fine grained implicit labels in SinOffen. Among them, *Circumlocution* (84.1%), *Homophones*

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.

(61.3%), *Metonymy* (36.6%), *Context* (29.5%), and *Metaphor* (28.5%) appear most frequently, while *Humiliation* (4.1%) and *Black humor* (2.8%) are relatively rare. Note that implicit OS may encompass multiple labels.

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. We collected nearly 30,000 tweets between January 2022 and October 2024. Through data filtering (the detailed procedure is in the Appendix A.3) and annotation, we ultimately constructed the SinOffen dataset, consisting of 16,235 Standard Chinese samples (official Mandarin without any dialects). For the Non-OS, to enhance the diversity and difficulty of the samples, we additionally crawled a large number of classic quotations from literary works to expand the dataset. We conducted experiments to demonstrate the impact of these literary samples on our benchmark performance, as detailed in the Appendix A.4. 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 (the detailed specific keywords are in Appendix A.1).

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, bot identification process is in Appendix A.2) 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.

The collected data is distributed as follows: Non-OS Collection (6,001 tweets, 36.96%), Keyword-Based Collection (3,512 tweets, 21.63%), Comment Section Collection (4,007 tweets, 24.68%), and Bot-tweets Collection (2,715 tweets, 16.72%).

4.2 Data Annotation

The annotation task was exclusively conducted by our dedicated research team comprising 10 full-time members (5 men / 5 women) 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 (the annotation details are in Appendix B). 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). To ensure the reliability of annotations, each sample was independently labeled by four annotators. A majority voting scheme was employed, whereby a label was assigned if at least three annotators reached agreement. In case of a tie, a fifth annotator reviewed the sample to determine the final label.

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

[1]https://labelstud.io/

Language	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”).
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”).

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.

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. This annotation process minimized potential annotation errors, ensuring the high quality and reliability of the dataset.

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

5 Experiment

We design three tasks to evaluate SinOffen dataset:

Task1: Three-label classification task (Non-OS/Explicit OS/Implicit OS) using PLMs.

Task2: Binary classification task (Non-OS/Explicit OS, Non-OS/Implicit OS) for LLMs under different prompt templates.

Task3: Fine-grained classification task for Implicit OS using LLMs.

Task 1 and Task 2 use different task formats because LLMs rely on carefully designed prompts (Sahoo et al., 2024). To match the three-class setup of PLMs, the prompt must explicitly specify the classification task. However, overly complex prompts may increase cognitive load and cause classification confusion (the supporting experiments are in Appendix E).

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

Details and configurations of all baseline models are provided in the Appendix F.

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.

Prompted LLMs: In the Task 2 and Task 3, We selected the current advanced models that demonstrate strong performance across various tasks, including Mistral-7B (Jiang et al., 2023), Llama3.1-8B (AI@Meta, 2024a), Qwen2.5-7B (Hui et al., 2024), ERNIE 4.0 (Baidu, 2024), DeepSeek-V3-0324 (DeepSeek-AI, 2024), and ShieldLM-13B (Zhang et al., 2024), tailored for Chinese text safety. In addition, we provide experiments with more LLMs in the Appendix G.

5.3 Results and Discussion

5.3.1 Performance of PLMs

Table 6 presents the experimental results of PLMs on SinOffen. The results show that 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

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**.

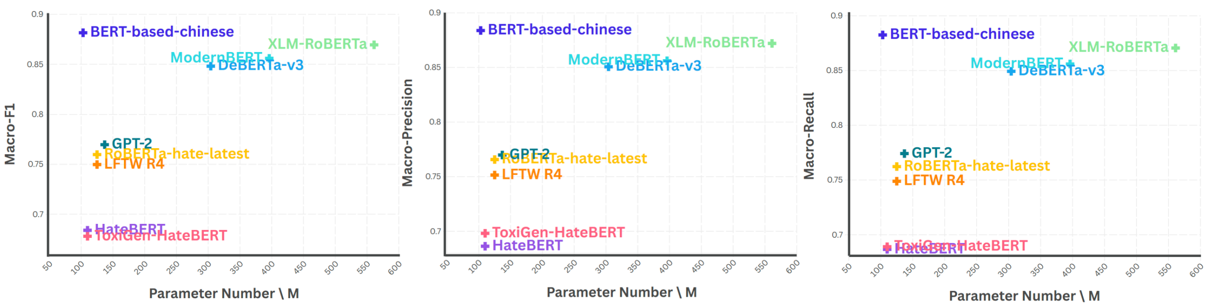

Figure 1: Trend of PLMs' Metrics with Parameter Count. From left to right, the y and x axes represent F1-Parameter, Precision-Parameter, and Recall-Parameter, respectively. See Appendix D for Parameter Count details.

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**
ShieldLM-13B	Chinese	0.8550	0.8363	0.8744	0.8193	0.8426	0.7971	**0.8371**	**0.8395**	**0.8358**
	English	0.7473	0.9079	0.6350	0.7839	0.6809	0.9236	0.7656	0.7944	0.7793
ERNIE 4.0	Chinese	0.8743	0.8122	0.9467	0.8199	0.9207	0.7390	**0.8471**	**0.8665**	**0.8428**
	English	0.8672	0.7920	0.9582	0.8003	0.9336	0.7004	0.8338	0.8628	0.8293
DeepSeek-V3	Chinese	0.9260	0.9005	0.9530	0.9060	0.9398	0.8746	0.9160	**0.9202**	0.9138
	English	0.9249	0.9379	0.9122	0.9132	0.8987	0.9282	**0.9190**	0.9183	**0.9202**

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**
ShieldLM-13B	Chinese	0.7948	0.7264	0.8774	0.7037	0.8144	0.6195	**0.7492**	**0.7704**	**0.7484**
	English	0.6357	0.8005	0.5272	0.7095	0.6095	0.8489	0.6726	0.7050	0.6880
ERNIE 4.0	Chinese	0.8428	0.7594	0.9468	0.7464	0.9072	0.6340	0.7946	0.8333	0.7904
	English	0.8408	0.7482	0.9596	0.7501	0.9310	0.6280	**0.7954**	**0.8396**	**0.7938**
DeepSeek-V3	Chinese	0.9047	0.9015	0.9080	0.8888	0.8925	0.8851	0.8968	0.8970	**0.8965**
	English	0.9102	0.8716	0.9525	0.8848	0.9382	0.8371	**0.8975**	**0.9049**	0.8948

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

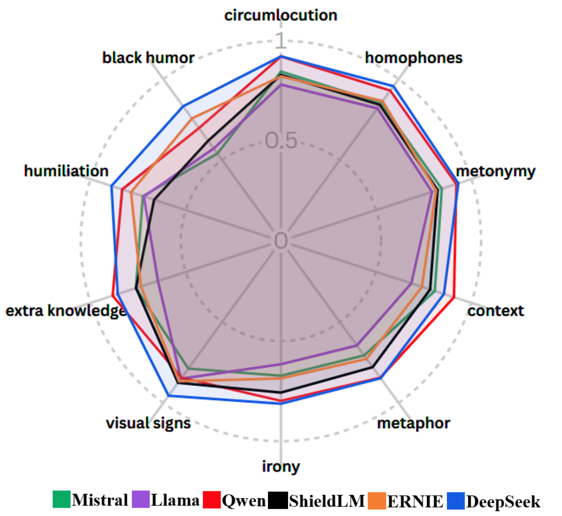

Figure 2: Comparison of Macro-F1 for Different LLMs on Different Fine-Grained Implicit OS Labels. The detailed metric values are provided in Appendix H.2.

suggests that increasing model complexity helps capture 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

The Table 7 and 8 presents the performance of different LLMs in the Chinese OSD task. We observe that DeepSeek-V3 achieves the best performance across both binary classification tasks. However, all models demonstrate a noticeable decline in performance when detecting Implicit OS compared to Explicit OS, highlighting the current limitations of LLMs in capturing subtle and implicit linguistic cues. In addition, a cross-lingual analysis of prompts reveals a noteworthy phenomenon: English prompts outperform their Chinese counterparts across most models.

Discussion: The performance drop may stem from the semantic ambiguity, contextual dependence, and blurred boundaries of Implicit OS, which make it difficult for existing models to detect without tailored mechanisms or annotated data. The difference in results for different prompts may stem from the model's English-centric training and tendency to reason through English internally (such as Mistral, Llama) (Wendler et al., 2024). Additionally, Chinese's high compression and polysemy can obscure task intent, while translating to English often adds semantic and logical clarity, helping models better infer the intended task (for instance, the word *"offensive"* in English has multiple meanings in Chinese, including *offensive*, *aggressive*, *rude*, etc.). This suggests that the optimal instruction language may not be the target language, but rather the one that most effectively activates the model's semantic reasoning.

5.3.3 LLMs in Implicit Offensive Speech

The experimental setup is detailed in the Appendix H.1. Appendix H.2 presents the classification performance of LLMs on fine-grained labels in Implicit OS, with all detailed results included. According to the experimental results, DeepSeek-V3 exhibited the best overall classification performance (as shown in Figure 2). At the same time, for all Implicit OS categories, especially in the *metaphor* (F1-ERNIE 4.0=0.7278, F1-Qwen 2.5=0.8440, F1-DeepSeek-V3=0.8474), *irony* (F1-ERNIE 4.0=0.6865, F1-Qwen 2.5=0.7979, F1-DeepSeek-V3=0.8130), and *black humor* (F1-ERNIE 4.0=0.7569, F1-Qwen 2.5=0.6968, F1-DeepSeek-V3=0.8317), all models showed suboptimal performance.

Discussion: All 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.

Examples of Offensive Tweets
<Non-OS> 幸福三要素: 愚蠢, 自私, 身体健康. *Human:* **Non-OS**, *DeepSeek:* **Offensive**, *ERNIE:* **Offensive**, *Qwen:* **Offensive**, *ShieldLM:* **Offensive**
DeepSeek: This sentence has offensive undertones. Words like “stupid” and “selfish” are clearly negative, directly demeaning a person’s intelligence or morality. While “physical health” is neutral, its placement alongside the negative terms suggests a cynical view that “only numb or indifferent people can be happy.”
Human: This sentence is from the French writer *Flaubert*. The term “stupid” does not refer to true ignorance, but rather to a kind of “wise ignorance” in life. “Selfish” is not extreme egoism, but rather a moderate focus on one’s own needs and a reasonable pursuit of personal happiness.
<metaphor> 他们真的超团结的像金针菇一样紧紧抱在一起. *Human:* **Offensive**, *DeepSeek:* **Non-OS**, *ERNIE:* **Non-OS**, *Qwen:* **Non-OS**, *ShieldLM:* **Non-OS**
DeepSeek: This sentence is neutral and even carries a touch of metaphorical humor. The phrase “hugging each other tightly like enoki mushrooms” is a vivid metaphor that likely emphasizes the characteristics of being close and inseparable.
Human: This sentence is offensive. The reference to enoki mushrooms serves as an derogatory metaphor targeting males, subtly belittling them by comparing them to something that is often perceived as weak or insignificant.
<irony> 都是女的世界才是好文明, 姐妹一思考世界会发笑. *Human:* **Offensive**, *DeepSeek:* **Non-OS**, *ERNIE:* **Non-OS**, *Qwen:* **Non-OS**, *ShieldLM:* **Non-OS**
DeepSeek: This sentence appears to praise a female-dominated society. The latter part draws on the famous saying of the philosopher Hegel, suggesting that when women think, the world smiles easily. It might be intended as self-mockery within the female community.
Human: This sentence is an offensive remark that employs a paraphrase of Hegel’s words to mock and belittle. It implies that women’s thinking is laughable or naive, subtly conveying a clear sense of gender discrimination.
<black humor> 建议设立广岛长崎烧烤节. *Human:* **Offensive**, *DeepSeek:* **Offensive**, *ERNIE:* **Offensive**, *Qwen:* **Non-OS**, *ShieldLM:* **Non-OS**
DeepSeek: This sentence may be offensive. It literally suggests holding a barbecue festival in Hiroshima and Nagasaki, but it could actually link “barbecue” to the nuclear explosions, subtly alluding to the cruelty of the disaster.
Human: This statement is extremely aggressive and offensive, trivializing a massive civilian disaster by turning it into a festival. Referring to the suffering of atomic bomb victims, who were mostly innocent civilians, as a “barbecue festival” is extremely callous and disrespectful.

Table 9: Examples of Offensive Tweets. The label is in red, the tweets are in blue, and the DeepSeek and manual annotations are in black. The classification results of human annotators and LLMs are highlighted in **bold**.

6 Case Study

In this section, we conduct a case study to examine why the model struggles in categories like metaphor, irony, and black humor. We compare the SOTA DeepSeek, with human annotations to highlight differences in interpreting implicit OS. Specifically, we randomly sample sentences from the Non-OS, metaphor, irony, and black humor (as shown in Table 9), and have two team members provide detailed human interpretations.

Non-OS: In our dataset, the Non-OS includes many literary sentences that use critical language to reflect on social issues. For example, Flaubert’s quote was misclassified by LLMs due to words like “stupid” and “selfish”, highlighting the model’s tendency to associate negative terms with offense. Additionally, as LLMs are typically trained to avoid harmful content (Chua et al., 2024), they tend to be overly cautious when handling borderline cases, leading to false positives. This reveals the limitations of current LLMs in emotion analysis and contextual understanding.

Metaphor: In this case, DeepSeek misinterpreted the metaphor “enoki mushroom” as a positive expression of unity due to a lack of cultural background knowledge. In contrast, human annotators who draw on local cultural and internet experience easily recognized its offensive implication. This difference underscores the limitations of LLMs in handling dynamic cultural content and evolving online language.

Irony: In this case, DeepSeek failed to identify the gender discrimination implied beneath an otherwise neutral sentence, while human annotators accurately perceived the underlying meaning. This exposes the model’s continued shortcomings in emotion analysis and contextual comprehension.

Black Humor: In this case, most LLMs failed, but DeepSeek was able to detect OS, and its analysis largely aligned with human interpretation.

In summary, while LLMs have made notable progress in semantic understanding, there is still room for improvement in handling dynamic cultural content, internet language, and deep complex context. More complete implicit OS cases are provided in the Appendix H.3

7 Conclusion

In this paper, we propose an OS taxonomy system with labels for *Non-OS*, *Explicit OS*, and *Implicit OS*, with the Implicit OS further divided into 10 distinct categories. Based on this, we construct the most comprehensive Chinese OSD dataset to date, with a focus on implicit OS. Our goal is to bridge the gap in detecting Chinese implicit OS. Evaluation with strong baselines shows persistent challenges. Future work may explore sarcasm detection (Liu et al., 2024; Zhu et al., 2024; Lin et al., 2024), improved prompt engineering (Lee et al., 2024), and dataset expansion.

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) *Annotation Cost:* Due to the high semantic implicitness and strong context dependence of the content being processed, along with the finer and more subjective category distinctions, annotators need to invest more time and cognitive effort in understanding the context, interpreting intent, and identifying subtle offensive cues, which makes the annotation process considerably time-consuming. Our work highlights the limitations of current advanced models in Chinese OSD, particularly in handling implicit cases. However, it does not yet fully address all the challenges involved in implicit OSD. Future research should explore more in-depth approaches to tackle these issues.

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.

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.

Baidu. 2024. ERNIE 4.0 Turbo model card.

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.

DeepSeek-AI. 2024. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

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 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*.

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.

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.

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*.

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.

Hongbo Wang, LiMingDa LiMingDa, Junyu Lu, Hebin Xia, Liang Yang, Bo Xu, Ruizhu Liu, and Hongfei Lin. 2024. PclGPT: A large language model for patronizing and condescending language detection. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6913–6928, Miami, Florida, USA. 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.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in English? on the latent language of multilingual transformers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15366–15394, Bangkok, Thailand. Association for Computational Linguistics.

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.

Qingpo Yang, Yakai Chen, Zihui Xu, Yu-ming Shang, Sanchuan Guo, and Xi Zhang. 2025. SCCD: A session-based dataset for Chinese cyberbullying detection. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9533–9545, Abu Dhabi, UAE. Association for Computational Linguistics.

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

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, Abu Dhabi, United Arab Emirates. 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 10, 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, 钕铜, 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 10: The keywords used for each theme.

A.2 Bot-Tweets Collection

To identify the Bot-accounts, we developed a dual detection system:

- Semantic Feature Recognition: We built a regular expression library that includes inducive keywords such as “tg”, “投稿”,“稿主” and“高柱”.
- Account Attribute Recognition: We established a feature dictionary based on username suffix characteristics, which often contain the “bot” field.

We then conducted targeted data collection on the platform to gather a set of candidate accounts, which were manually screened by a team of three members to identify bots that incite online polarization.

A.3 Data Filtering and Cleaning

To ensure high-quality data and accurate annotations, we implemented a rigorous data filtering and cleaning process, as described below:

First, we removed invalid content, including posts consisting solely of emojis, which do not provide meaningful textual information for downstream tasks. We also eliminated duplicate samples to avoid redundancy and potential bias in model training and evaluation. In addition, nonsensical or incoherent sentences—such as random character strings and machine-generated spam—were manually identified and excluded. To further enhance data quality, we applied basic normalization steps, such as unifying character encodings and standardizing punctuation marks. After this multi-stage cleaning process, we obtained a high-quality dataset comprising 16,235 unique and coherent samples, laying a solid foundation for subsequent experiments.

A.4 Literary samples for benchmark performance

In our dataset, the number of literary references is 3,798, accounting for approximately 63.28% of the total Non-OS. To show the impact of literary samples on benchmark performance, we conducted a systematic validation through controlled experiments. Specifically, we replaced the literary references in our dataset with Non-OS from the COLD dataset (Deng et al., 2022) while maintaining the overall data distribution. Comparative experiments were then performed using three representative models: BERT-base-Chinese (BERT-C), XLM-RoBERTa (XLM-R), Qwen2.5, ERNIE and DeepSeek-V3.

Models	F1	Precision	Recall
XLM-R (Original)	0.9681	0.9614	0.9750
XLM-R (COLD)	0.9819	0.9815	0.9823
BERT-C (Original)	0.9701	0.9804	0.9600
BERT-C (COLD)	0.9869	0.9861	0.9877
Qwen (Original)	0.8461	0.9279	0.7778
Qwen (COLD)	0.8641	0.7825	0.9647
ERNIE (Original)	0.6277	0.5586	0.7163
ERNIE (COLD)	0.7543	0.6229	0.9562
DeepSeek (Original)	0.7285	0.7376	0.7197
DeepSeek (COLD)	0.8617	0.7837	0.9568

Table 11: Model classification metrics for Non-OS class before and after replacement.

The experimental results indicate that the detection rate of the replaced Non-OS is even higher, as shown in Table 11. These findings align closely with our initial hypothesis: the dense presence of metaphorical expressions, multiple negation structures, and emotionally ambiguous statements in literary texts essentially creates an “adversarial training” environment that challenges models to develop higher-order language understanding. The incorporation of this specialized literary genre enables the models to deeply parse latent semantic layers, thereby significantly enhancing their robustness in handling complex linguistic phenomena.

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 Appendix C.

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*

B.2 Annotation Process

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.

Particularly, for ***Comment Section Collection*** Method which offensive Speech is closely related to the complete conversation context, annotators had access to the complete context to make informed decisions.

To illustrate our annotation process more clearly, we define the root post (R) as the original post and its derived comments as (C1, C2, ..., Cn). In Label Studio, each comment to be annotated (C1, C2, ..., Cn) was presented together with its corresponding root post (R), forming a complete "root post-comment" conversation chain. This ensures that the original content and derived comments (R → C1 → C2 → ... → Cn) are fully displayed during the annotation process.

For example, when annotators encountered the comment "*Support the German military, the German military is the most civilized, outstanding, just, and moral army in the world.*" (C1), they had already obtained the relevant root post (R) information in advance —a news report on the Middle East situation, which included data on military conflict casualties.

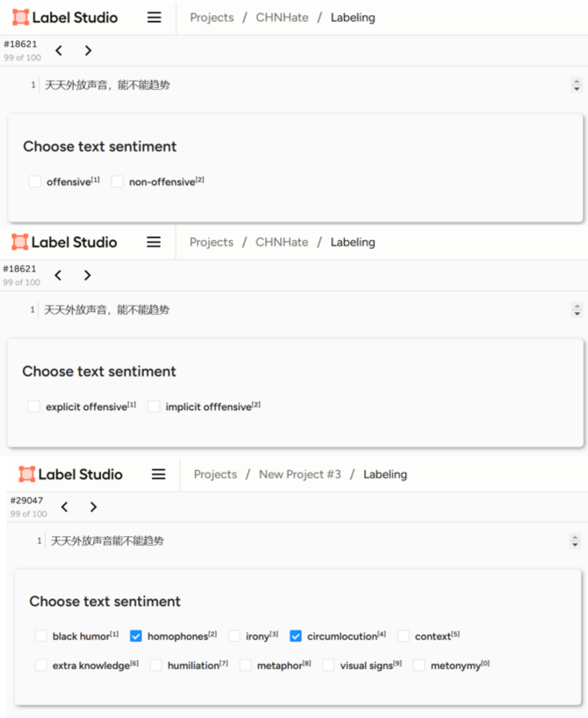

Figure 3: Data annotation on Label Studios.

B.3 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 Implicit-OS Properties

Inspired by related work (Ocampo et al., 2023) and Chinese linguistics, we propose 10 fine-grained categories tailored for Chinese implicit OS. Below, we provide detailed definitions for each category.

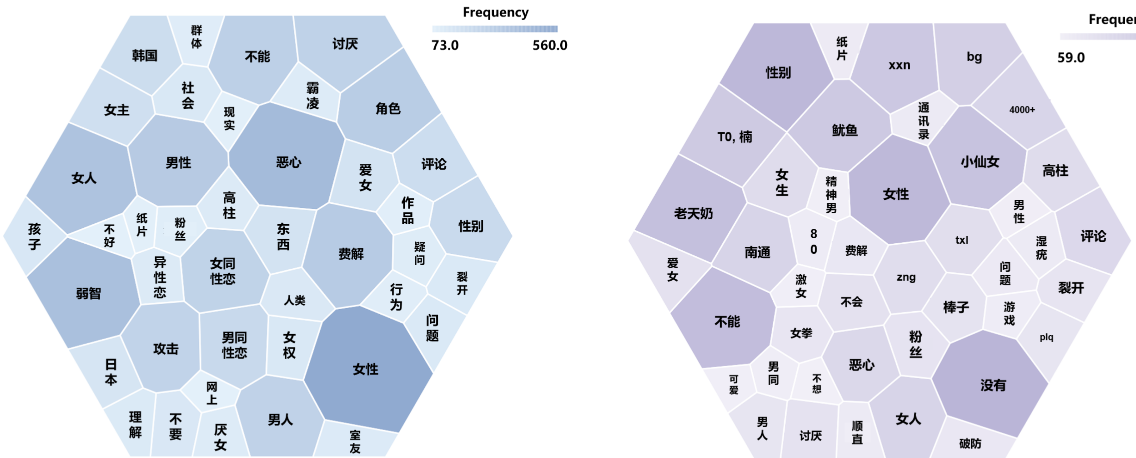

Figure 4: Word Cloud Distribution of Implicit OS (Right) and Explicit OS (Left).

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 convey 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: The feeling of shame elicited by making the target appear foolish or magnifying their errors in a public context (Ocampo et al., 2023).

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.

D PLMs Parameter Display

Table 12 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 12: Detailed PLMs Parameter Numbers.

E Reasons for Using Binary Classification with 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. 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.

Model	All-Macro			FPR		
	F1	Precision	Recall	Non-OS	Explicit-OS	Implicit-OS
Qwen2.5-7B	0.4579	0.5904	0.5014	0.0238	0.0502	0.6516
ERNIE 4.0	0.5398	0.5378	0.5634	0.2014	0.2599	0.1674
DeepSeek-V3	0.5379	0.5573	0.5885	0.0534	0.4312	0.1016

Table 13: Overall Metrics for LLMs Three-Classification Task.

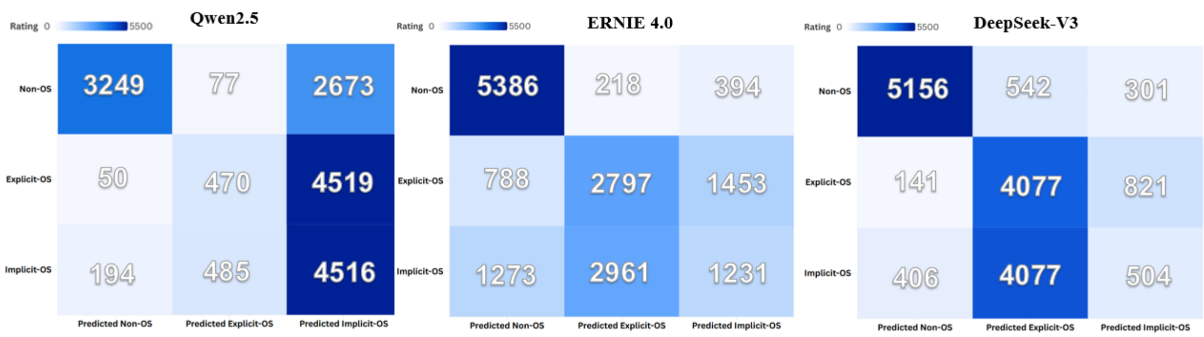

Figure 5: Confusion Matrix for LLMs Three-Classification Task.

To validate our argument, we conducted a three-class classification task experiment using Qwen, ERNIE and DeepSeek as examples, the prompt template used in this experiment is shown below. We conducted a series of experiments to validate aforementioned conclusions.

> ***Prompt Template:*** Please determine whether the following sentence contains offensive content. “0” indicates that the sentence is not offensive speech, “1” indicates that it is explicit offensive speech, and “2” indicates that it is implicit offensive speech. For implicit offensive speech, look for subtle or indirect insults, stereotypes, or harmful implications that aren’t directly offensive but still promote discrimination or prejudice. Please provide your response (only “0”, “1”, or “2”).

The rationale behind our dual-binary classification design stems from the observed semantic asymmetry in offensive language. Explicit OS typically contains strong negative vocabulary or emotions, whereas Implicit OS often depends on external knowledge or nuanced contextual understanding. If a three-classification strategy is employed, both types of offensive speech are forced to share the same decision boundary, which can lead to detection errors for Implicit OS and subsequently contaminate the classification of Explicit OS. As shown in Table 13 and Figure 5, an analysis of the confusion matrix further reveals that when the two types are trained jointly, the False Positive Rate (FPR) of Implicit OS and Explicit OS increases significantly. These findings validate the necessity of our design. Accordingly, we propose decomposing the task into two separate binary classification tasks (*Non-OS vs. Explicit OS* and *Non-OS vs. Implicit OS*), simplifying the learning objectives and allowing the model to more effectively distinguish between offensive and non-offensive content.

F Baseline Models Configurations

HateBERT: A domain-adapted BERT model pre-trained on Reddit hate speech data to better detect abusive and offensive language. We selected two HateBERT models, namely HateBERT and ToxiGen-HateBERT.
GPT-2: A large-scale, generative transformer model developed by OpenAI, trained to predict the next token in diverse internet text.
LFTW R4: A domain-adapted language model pre-trained on Reddit hate speech data.
RoBERTa: An optimized version of BERT with enhanced performance for NLP tasks. We selected two RoBERTa models, namely RoBERTa-hate-latest and XLM-RoBERTa.
DeBERTa: An improved DeBERTa model incorporating disentangled attention and enhanced masked language modeling for better language representation.

ModernBERT: A re-engineered BERT model incorporating modern pretraining advances to achieve stronger performance with fewer parameters.

BERT-based-Chinese: A Chinese-specific BERT model trained on large-scale Chinese corpora, designed to capture semantic nuances in Chinese text.

Mistral-7B: A dense transformer model optimized for efficiency and scalability, offering strong performance on a wide range of language tasks.

Llama3.1-8B: A next-generation open-weight LLM developed by Meta, fine-tuned for improved reasoning, instruction-following, and multilingual capabilities.

Qwen 2.5-7B: A powerful Chinese-English bilingual large language model series developed by Alibaba, emphasizing advanced reasoning and generation.

ShieldLM-13B: A model alignment framework that integrates safety-enhancing techniques during supervised fine-tuning to improve robustness against harmful or unsafe outputs, especially for Chinese.

ERNIE 4.0: Baidu's knowledge-enhanced closed-source LLMs integrating structured knowledge and large-scale pretraining for stronger understanding and generation.

DeepSeek-V3-0324: A large-scale multimodal LLM developed by DeepSeek, featuring strong language understanding and generation capabilities, and achieving outstanding performance on various Chinese benchmark evaluations.

To better adapt English-oriented OSD models such as HateBERT, ToxiGen-HateBERT, and RoBERTa-hate-latest to the task of Chinese offensive language detection, we replaced their original vocabularies with BERT-based-Chinese which is more suitable for Chinese text. For LLMs, since Llama natively does not support Chinese, we specifically chose the Llama3.1 model[2] fine-tuned for Chinese. Among the models evaluated, all except ShieldLM[3] and ERNIE[4] are open-sourced and accessible via the Hugging Face[5].

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

[3]https://github.com/thu-coai/ShieldLM

[4]https://wenxin.baidu.com/

[5]https://huggingface.co

G LLMs Performance Details in Binary Classification Task

Tables 14 and 15 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 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.

H LLMs Performance Details in fine-grained Implicit OS

H.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.

H.2 Results of fine-grained Implicit OS

Table 16 presents detailed classification results of LLMs on different fine-grained Implicit OS categories, with metrics including F1, Precision, and Recall.

H.3 Complete Examples of Offensive Tweets

Table 17 is the examples of a complete case study. We randomly selected sentences from each category and then had them explained and annotated in detail by DeepSeek and two members of the team.

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

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

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

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 (hfl)	Chinese	0.7370	0.8264	0.6026	0.7378	0.6498	0.8535	0.7374	0.7381	0.7281
	English	0.7939	0.8466	0.7474	0.7897	0.7425	0.8432	**0.7918**	**0.7945**	**0.7953**
Llama3.1-8B (Meta)	Chinese	0.7468	0.6125	0.9565	0.4420	0.8551	0.2980	0.5944	0.7338	0.6273
	English	0.7904	0.6701	0.9633	0.6031	0.9136	0.4501	**0.6968**	**0.7919**	**0.7067**
Llama3.1-8B (shenzhi-wang)	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**
Llama3.2-8B (Meta)	Chinese	0.6998	0.5422	0.9865	0.0708	0.7044	0.0373	0.3853	0.6233	0.5119
	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**
ShieldLM-13B	Chinese	0.7948	0.7264	0.8774	0.7037	0.8144	0.6195	**0.7492**	**0.7704**	**0.7484**
	English	0.6357	0.8005	0.5272	0.7095	0.6095	0.8489	0.6726	0.7050	0.6880
ERNIE 4.0	Chinese	0.8428	0.7594	0.9468	0.7464	0.9072	0.6340	0.7946	0.8333	0.7904
	English	0.8408	0.7482	0.9596	0.7501	0.9310	0.6280	**0.7954**	**0.8396**	**0.7938**
DeepSeek-V3	Chinese	0.9047	0.9015	0.9080	0.8888	0.8925	0.8851	0.8968	0.8970	**0.8965**
	English	0.9102	0.8716	0.9525	0.8848	0.9382	0.8371	**0.8975**	**0.9049**	0.8948

Table 14: 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 (hfl)	Chinese	0.7124	0.8718	0.6022	0.7556	0.6539	0.8946	0.7340	0.7629	0.7484
	English	0.8072	0.8873	0.7403	0.8083	0.7418	0.8881	**0.8078**	**0.8145**	**0.8142**
Llama3.1-8B (Meta)	Chinese	0.7736	0.6496	0.9563	0.5350	0.8805	0.3842	0.6543	0.7650	0.6703
	English	0.8059	0.6937	0.9615	0.6417	0.9150	0.4941	**0.7238**	**0.8043**	**0.7278**
Llama3.1-8B (shenzhi-wang)	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**
Llama3.2-8B (Meta)	Chinese	0.7099	0.5543	0.9868	0.1063	0.7842	0.0570	0.4081	**0.6692**	0.5219
	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**
ShieldLM-13B	Chinese	0.8550	0.8363	0.8744	0.8193	0.8426	0.7971	**0.8371**	**0.8395**	**0.8358**
	English	0.7473	0.9079	0.6350	0.7839	0.6809	0.9236	0.7656	0.7944	0.7793
ERNIE 4.0	Chinese	0.8743	0.8122	0.9467	0.8199	0.9207	0.7390	**0.8471**	**0.8665**	**0.8428**
	English	0.8672	0.7920	0.9582	0.8003	0.9336	0.7004	0.8338	0.8628	0.8293
DeepSeek-V3	Chinese	0.9260	0.9005	0.9530	0.9060	0.9398	0.8746	0.9160	**0.9202**	0.9138
	English	0.9249	0.9379	0.9122	0.9132	0.8987	0.9282	**0.9190**	0.9183	**0.9202**

Table 15: 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
ShieldLM-13B	F1	0.8295	0.8407	0.8230	0.7837	0.7791	0.7572	0.8754	0.7614	0.6664	0.6168
	Precision	0.8296	0.8393	0.8234	0.7873	0.7878	0.7680	0.8737	0.7648	0.6679	0.6233
	Recall	0.8294	0.8436	0.8229	0.7822	0.7777	0.7556	0.8818	0.7590	0.6742	0.6410
ERNIE 4.0	F1	0.8225	0.8638	0.8153	0.7405	0.7278	0.6865	0.8664	0.7357	0.7879	0.7569
	Precision	0.8780	0.9040	0.8676	0.8429	0.8323	0.8163	0.9080	0.8473	0.8585	0.8464
	Recall	0.8172	0.8529	0.8159	0.7433	0.7351	0.7001	0.8540	0.7280	0.7752	0.7323
DeepSeek V3	F1	0.9224	0.9560	0.9320	0.8562	0.8474	0.8130	0.9553	0.8576	0.8903	0.8317
	Precision	0.9365	0.9625	0.9410	0.8950	0.8883	0.8735	0.9636	0.9045	0.9088	0.8798
	Recall	0.9173	0.9516	0.9300	0.8498	0.8433	0.8095	0.9498	0.8428	0.8808	0.8084

Table 16: Results of LLMs on different fine-grained Implicit OS categories.

Examples of Offensive Tweets

<Non-OS> 幸福三要素: 愚蠢, 自私, 身体健康. *Human:* **Non-OS**, *DeepSeek:* **Offensive**, *ERNIE:* **Offensive**, *Qwen:* **Offensive**, *ShieldLM:* **Offensive**

DeepSeek: This sentence has offensive undertones. Words like "stupid" and "selfish" are clearly negative, directly demeaning a person's intelligence or morality. While "physical health" is neutral, its placement alongside the negative terms suggests a cynical view that "only numb or indifferent people can be happy."

Human: This sentence is from the French writer *Flaubert*. The term "stupid" does not refer to true ignorance, but rather to a kind of "wise ignorance" in life. "Selfish" is not extreme egoism, but rather a moderate focus on one's own needs and a reasonable pursuit of personal happiness.

<metaphor> 他们真的超团结的像金针菇一样紧紧抱在一起. *Human:* **Offensive**, *DeepSeek:* **Non-OS**, *ERNIE:* **Non-OS**, *Qwen:* **Non-OS**, *ShieldLM:* **Non-OS**

DeepSeek: This sentence is neutral and even carries a touch of metaphorical humor. The phrase "hugging each other tightly like enoki mushrooms" is a vivid metaphor that likely emphasizes the characteristics of being close and inseparable.

Human: This sentence is offensive. The reference to enoki mushrooms serves as an derogatory metaphor targeting males, subtly belittling them by comparing them to something that is often perceived as weak or insignificant.

<irony> 都是女的, 都是女孩子的世界才是好文明, 姐妹一思考世界会发笑. *Human:* **Offensive**, *DeepSeek:* **Non-OS**, *ERNIE:* **Non-OS**, *Qwen:* **Non-OS**, *ShieldLM:* **Non-OS**

DeepSeek: This sentence appears to praise a female-dominated society. The latter part draws on the famous saying of the philosopher Hegel, suggesting that when women think, the world smiles easily. It might be intended as self-mockery within the female community.

Human: This sentence is an offensive remark that employs a paraphrase of Hegel's words to mock and belittle. It implies that women's thinking is laughable or naive, subtly conveying a clear sense of gender discrimination.

<black humor> 建议设立广岛长崎烧烤节. *Human:* **Offensive**, *DeepSeek:* **Offensive**, *ERNIE:* **Offensive**, *Qwen:* **Non-OS**, *ShieldLM:* **Non-OS**

DeepSeek: This sentence may be offensive. It literally suggests holding a barbecue festival in Hiroshima and Nagasaki, but it could actually link "barbecue" to the nuclear explosions, subtly alluding to the cruelty of the disaster.

Human: This statement is extremely aggressive and offensive, trivializing a massive civilian disaster by turning it into a festival. Referring to the suffering of atomic bomb victims, who were mostly innocent civilians, as a "barbecue festival" is extremely callous and disrespectful.

<circumlocution> 她的尿是哑光的吧. *Human:* **Offensive**, *DeepSeek:* **Offensive**, *ERNIE:* **Offensive**, *Qwen:* **Offensive**, *ShieldLM:* **Non-OS**

DeepSeek: This sentence constructs an absurd aesthetic evaluation system by forcibly associating a female physiological phenomenon (urine) with the art term "matte." Essentially, this is an objectification and humiliation of the female body.

Human: This sentence adapts the internet slang phrase "go take a piss and look at yourself in the mirror," which indirectly implies that someone is ugly in a more circumlocutory way.

<homophones> 蝻的破防比泰迪发青来的还快. *Human:* **Offensive**, *DeepSeek:* **Offensive**, *ERNIE:* **Offensive**, *Qwen:* **Offensive**, *ShieldLM:* **Offensive**

DeepSeek: The derogatory term "蝻的" directly stigmatizes male groups. By likening "破防" to "泰迪发青" it animalizes male emotional reactions. The exaggerated and contrasting sentence structure further reinforces stereotypes, implying that men's emotional control is worse than animal instincts.

Human: This sentence is offensive. The reference to enoki mushrooms serves as an derogatory metaphor targeting males, subtly belittling them by comparing them to something that is often perceived as weak or insignificant.

<extra knowledge> 东郭先生与狼郝建和老太太, 人类与小仙女. *Human:* **Offensive**, *DeepSeek:* **Offensive**, *ERNIE:* **Non-OS**, *Qwen:* **Non-OS**, *ShieldLM:* **Non-OS**

DeepSeek: This sentence constructs a metaphorical chain of "women = ungrateful wolves/extortionists" through the forced juxtaposition of fable characters and online labels.

Human: <Mr. Dongguo and the Wolf> is a classic traditional Chinese anecdote that conveys profound lessons about ingratitude. Here, these anecdotes are referenced to subtly express offensive remarks.

<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. *Human:* **Offensive**, *DeepSeek:* **Offensive**, *ERNIE:* **Non-OS**, *Qwen:* **Non-OS**, *ShieldLM:* **Non-OS**

DeepSeek: This sentence may carry offensive and historical revisionist undertones. Such rhetoric could violate laws in multiple countries regarding the denial of Nazi crimes and also contradicts the universally accepted moral standards of human society.

Human: Based on the context, if the surrounding text includes references to Jewish people or similar topics, it could be considered offensive.

<humiliation> 像你这样的人能做到这一步, 真的挺意外的. *Human:* **Offensive**, *DeepSeek:* **Offensive**, *ERNIE:* **Offensive**, *Qwen:* **Non-OS**, *ShieldLM:* **Non-OS**

DeepSeek: This sentence labels a specific group, implying their inferiority and suggesting that certain groups are inherently undeserving of success, even stripping them of their right to strive for progress.

Human: This sentence is offensive. Indirectly expressing hostility or discrimination towards a target group by belittling, insulting, or degrading someone's dignity.

<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. *Human:* **Offensive**, *DeepSeek:* **Offensive**, *ERNIE:* **Offensive**, *Qwen:* **Offensive**, *ShieldLM:* **Non-OS**

DeepSeek: In this sentence, "xxn" is a specific term for women used in online gender confrontation, which reinforces the stereotype of "women falsely complaining."

Human: <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.

<visual signs> 下辈子要当公 ÷ 的都是 4000+ 的 东西. *Human:* **Offensive**, *DeepSeek:* **Offensive**, *ERNIE:* **Offensive**, *Qwen:* **Non-OS**, *ShieldLM:* **Non-OS**

DeepSeek: In this sentence, "4000+" is a numerical pun from vulgar internet culture, where "4" is a homophone for "death,"000" intensifies the meaning, and "+" implies an extension, forming an extreme curse.

Human: < 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 17: Examples of Offensive Tweets. The label is in red, the tweets are in blue, and the DeepSeek and manual annotations are in black.