PclGPT: A Large Language Model for Patronizing and Condescending Language Detection

Anonymous ACL submission

Abstract

***Disclaimer**: Samples in this paper may be harmful and cause discomfort!*

Patronizing and condescending language (PCL) is a form of speech directed at vulnerable groups. As an essential branch of toxic language, this type of language exacerbates conflicts and confrontations among Internet communities and detrimentally impacts disadvantaged groups. Traditional pre-trained language models (PLMs) perform poorly in detecting PCL due to its implicit toxicity traits like hypocrisy and false sympathy. With the rise of large language models (LLMs), we can harness their rich emotional semantics to optimize PCL detection. In this paper, we introduce PclGPT, a comprehensive LLM benchmark designed specifically for PCL. We collect, annotate, and integrate the Pcl-PT/SFT dataset, and then develop a bilingual PclGPT-EN/CN model group through a comprehensive pre-training and supervised fine-tuning staircase process to facilitate cross-language detection. Group detection results and fine-grained detection from PclGPT and other models reveal significant variations in the degree of bias in PCL towards different vulnerable groups, necessitating increased societal attention to protect them.

1 Introduction

Patronizing and condescending language (PCL) specifically targets vulnerable groups. As an important but underexplored branch of toxic language, timely detection of PCL is crucial for protecting disadvantaged communities from further exclusion and inequality. Unlike traditional toxic languages such as hate speech (Cao and Lee, 2020; Caselli et al., 2020) and offensive language (Fortuna et al., 2020; Zampieri et al., 2019; Deng et al., 2022), PCL expressions are more subtle and implicit (e.g., *"These poor children! It's truly admirable how they keep striving despite their humble beginnings."*). This example is interesting because the original intention of PCL might have been to positively describe efforts to improve the lives of disadvantaged groups. However, it ultimately conveys subtle arrogance and discrimination, harming the individuals being sympathized with.

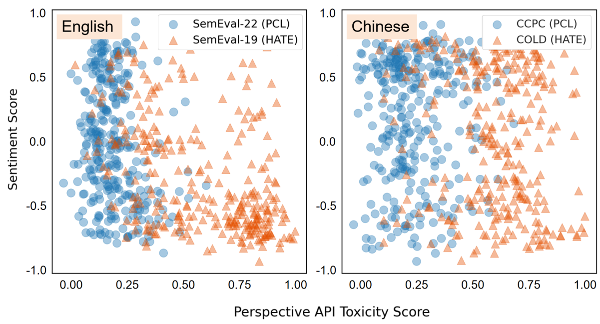

Figure 1: Scatter plots for the scores using the Perspective API on the hate and PCL datasets. The left plot shows the English datasets SemEval-19 (HATE) and SemEval-22 (PCL), while the right plot shows the Chinese datasets COLD (HATE) and CCPC (PCL). The toxicity score ranges from 0 to 1, with increasing toxicity as discrete values.

The subtle toxicity of PCL is further illustrated through toxicity scores. We compared the PCL and HATE datasets in both English and Chinese domains. As shown in Figure 1, the toxicity scores from the Perspective API indicate that, in both Chinese and English corpora, the toxicity scores of PCL are much lower than those of hate speech. This is due to the ambiguous toxic semantic features of PCL, which often lack explicit attacking vocabulary, leading to PLMs struggling to achieve optimal detection performance. The absence of high-quality data further constrains this field (Wang et al., 2023). Large language models (LLMs) offer new opportunities with their extensive pre-trained knowledge and enhanced capability in revealing toxicity (Wen et al., 2023). However, they still lack essential domain-specific knowledge for condescending language and effective guidance, lead-

English Task	PCL Category	PLMs	GPT4.0	PclGPT-EN
Since the elderly have been placed in a nursing home, they are undoubtedly left unattended most of the time.	Unbalanced-Power-Relations	✗	✗	✓
Chinese Task	**PCL Category**	**PLMs**	**GPT4.0**	**PclGPT-CN**
战斗在火焰中激烈进行：茫然、饥饿的非洲难民在燃烧的大门中迷失方向。 *The fighting raged among the flames: Dazed, starving African refugees wandered lost through the burning portals.*	Compassion	✗	✗	✓

Table 1: PclGPT and other models' detection examples for ambiguous PCL. ✗ indicates incorrect prediction results, and ✔ indicates correct prediction results.

ing to incomplete development for implicit toxic detection.

To address these challenges, we focus on three main questions: (1) How can we efficiently construct high-quality pre-training (PT) and supervised fine-tuning (SFT) datasets? (2) How can we design a new LLM benchmark that incorporates PT and SFT to enhance recognition of implicit toxicity? (3) Can we build a multilingual model group for cross-lingual tasks like Chinese PCL detection to support vulnerable non-English-speaking communities?

To solve these issues, we introduce PclGPT, a comprehensive LLM benchmark for PCL detection, exploring the LLM's understanding of implicit toxicity. First, we collect community data from mainstream internet platforms (Reddit for English and Sina Weibo for Chinese) and process it to construct the Pcl-PT dataset for domain-adaptive pre-training. Next, we annotate, restructure, and filter high-quality data to construct the Pcl-SFT dataset, employing the instruction data paradigm to impose additional constraints on both input and output. Subsequently, we undertake the complete process of pre-training and supervised fine-tuning to construct our bilingual model, PclGPT-EN/CN. This model represents the first known LLM designed explicitly for PCL detection. Our results, shown in Table 1, illustrate the testing results on difficult-to-distinguish ambiguous examples. The model demonstrates superior performance compared to other PLMs and LLMs in both English and Chinese tasks. Further group detection and fine-grained toxicity analysis reveal significant differences in the degree of bias in PCL towards various vulnerable groups. The ambiguity of bias also varies among different PCL subcategories. These findings necessitate increased societal attention to effectively protect vulnerable groups.

The main contributions of this paper are summarized as follows:

- We construct the Pcl-PT/SFT datasets to enhance domain-specific knowledge for PCL. Pcl-PT is used for pre-training, covering over 1.4 million data entries from vulnerable communities. Pcl-SFT is used for fine-tuning, with over 100k high-quality bilingual instruction samples.

- We propose a pre-training and fine-tuning framework to build our bilingual model, PclGPT-EN/CN. PclGPT is the first LLM designed to detect PCL and other implicit toxic languages, surpassing all advanced PLMs and LLMs and achieving state-of-the-art (SOTA) results on three public datasets.

- Through group detection and fine-grained toxicity analysis, we demonstrate the differentiated nature of group biases in PCL, with PclGPT laying a foundation for managing biases and protecting vulnerable groups.

2 Related Work

Which is Toxic: Hate Speech or PCL? Toxic language is perceived as an impolite, disrespectful, or irrational statement that may prompt someone to withdraw from a discussion (Dixon et al., 2018). Existing research (Deng et al., 2022; Cao and Lee, 2020; Tekiroglu et al., 2020; Caselli et al., 2020; Mathew et al., 2021) equates toxic language with hate speech, focusing only on direct and explicit offenses and insults, while overlooking implicit forms of toxicity such as stereotypes and irony (ElSherief et al., 2021). Recent studies on hate speech still ignore many direct victims of toxicity (Ocampo et al., 2023; Bourgeade et al., 2023; El-Sayed and Nasr, 2024). Hate speech often targets religion, race, ethnicity, and gender, but neglects other disadvantaged groups like single-parent families, child laborers, and disabled individuals. This gap led to the emergence of patronizing and condescending language (PCL). Pérez-Almendros et al. (2020) integrated categories of implicit toxicity and introduced PCL. Unlike traditional hate speech, PCL focuses on implicit toxicity aimed at marginalized and vulnerable groups. Such ambiguous implicit toxicity is less aggressive and has lower toxicity scores compared to hate speech, making it more difficult to detect (Figure 1). Wong et al. (2014) noted that PCL is often unconscious, driven by good intentions, and uses embellished language. Xu (2022) identified that such unjust treatment of vulnerable groups can exacerbate societal exclusion and inequality, causing users to leave communities or reduce online participation. Wang and Potts (2019); Pérez-Almendros et al. (2020); Wang et al. (2023) collected high-quality PCL corpora from mainstream social media platforms and annotated them with grading. In detection, Pérez-Almendros et al. (2022); Lu et al. (2022) utilized modified BERT networks and adversarial training for PCL detection. While these methodologies are pioneering, their efficacy is significantly compromised by inadequate pre-training and the implicit nature of toxicity within PCL.

LLM for Toxicity Detection. In recent years, decoder-only LLMs, such as ChatGPT (OpenAI, 2022), GPT-4 (OpenAI, 2023), and LLaMA (Touvron et al., 2023), have revolutionized text generation. LLMs have increasingly been applied in toxic language detection and prevention. Shaikh et al. (2022) demonstrated that zero-shot CoT significantly increases LLMs' toxic output. Wen et al. (2023) proved that supervised fine-tuning and reinforcement learning further induce toxic outputs. Zhu et al. (2023); Huang et al. (2023) used ChatGPT to map answers to binary labels through prompt engineering for hate detection. Roy et al. (2023) enhanced hate speech classification accuracy by including additional victim information. However, no systematic LLM engineering is currently used to detect PCL or prevent harmful expressions in such texts. Additionally, LLMs' fine-grained discrimination of implicit toxicity remains vague. To address these gaps, we introduce PclGPT, a new LLM benchmark for PCL detection, using pre-training and supervised fine-tuning to achieve SOTA results on three public datasets.

3 PclGPT

The overall approach is illustrated in Figure 2. Our PclGPT model group consists of two sub-models: PclGPT-EN and PclGPT-CN, using LLaMA-2-7B and ChatGLM-3-6B (Du et al., 2022) as their base architectures, respectively. LLaMA, one of the foremost English open-source LLMs today, has been pre-trained on over 20 trillion tokens. ChatGLM, among the most advanced Chinese LLMs, is built upon the Generalized Linear Model (GLM) architecture and has been extensively optimized for Chinese question-answering and dialogue tasks, exhibiting outstanding performance in the Chinese domain. Both models have a context length of up to 4096 tokens, ensuring a thorough understanding of the context. Detailed descriptions of the pre-training and fine-tuning stages will be provided in the subsequent sections.

3.1 Pre-training

To facilitate the pre-training process, we introduced the Pcl-PT dataset, comprising the RAL-P and WEB-C datasets. Specifically, we employed separate corpora in English and Chinese to pre-train our PclGPT-EN/CN model group. Our pre-training followed a standard paradigm, where the model predicted the next token based on existing input history. For both PclGPT-EN and PclGPT-CN, we utilized the same vocabulary as the base models and employed AdamW as the optimizer. The initial learning rate was set to 2×10^{-4} with a weight decay of 0.1. We also employed efficient training strategies, including mixed precision training with bf16 (Micikevicius et al., 2017). The specific parameters are detailed in Appendix A. Below, we provide detailed insights into the datasets. More details are shown in Table 2. The design of our structure is modeled after the hierarchical format of (Tian et al., 2023).

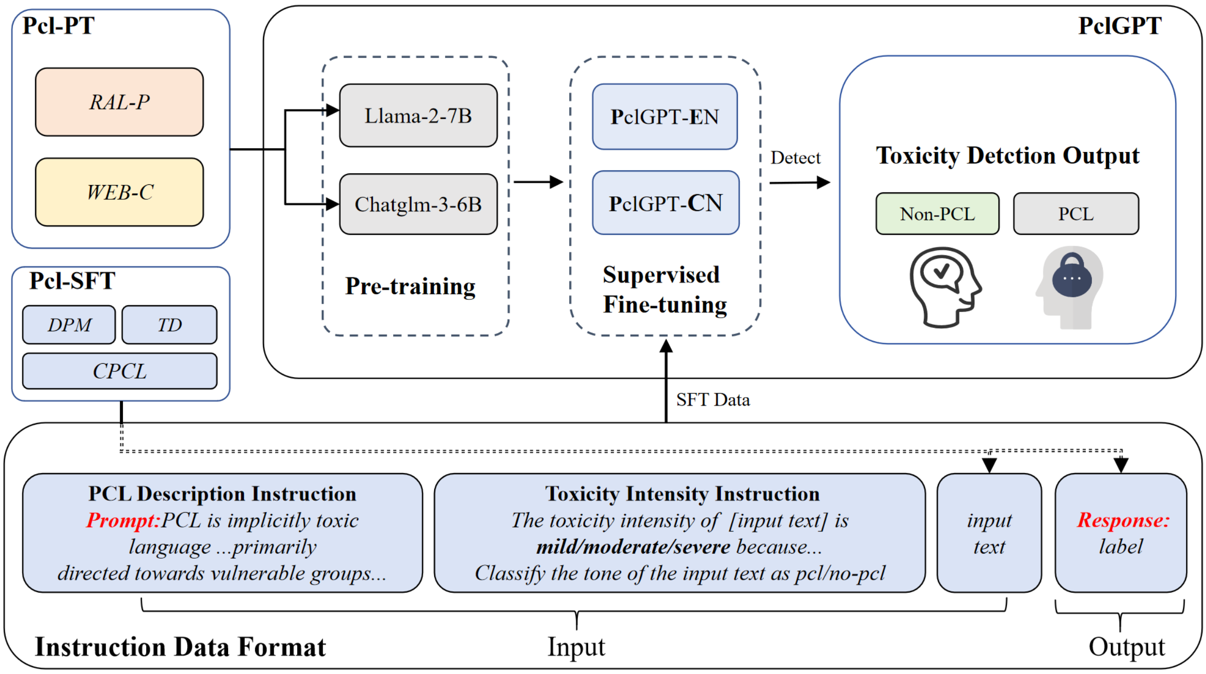

Figure 2: An illustration of the overall PclGPT. We establish Pcl-PT/SFT datasets and build a bilingual model group through pre-training and supervised fine-tuning. Instruction Data Format demonstrates the data construction format for SFT.

- **RAL-P** is derived from the RAL-E dataset. The RAL-E dataset (Caselli et al., 2020) includes offensive, abusive, and hateful content from the Reddit community, comprising 43M tokens collected from December 2005 to March 2017. However, RAL-E predominantly features explicit hate speech, which hinders the accurate identification of PCL, as the toxicity of PCL is often not directly correlated with explicit intensity, positive examples may also convey biased intentions. Therefore, based on the criteria established by Pérez-Almendros et al. (2020), we used LLM to generate a dictionary of over 300 condescending English terms, which was manually calibrated. We used this dictionary to exclude explicitly offensive and hateful sentences from RAL-E, while retaining 30% of non-PCL entries to ensure balanced pre-training data. RAL-P ultimately retained 1091945 data entries. Detailed processes are presented in Appendix B.

- **WEB-C**. The scarcity of data in the Chinese domain constrains the task of PCL detection. To address this, we designed a framework to systematically gather bullying, violent, and discriminatory content from marginalized communities on Sina Weibo, a mainstream Chinese media platform. We initially limited the search scope to seven major disadvantaged group categories based on PCL criteria (Wang et al., 2023), and expanded the keyword list accordingly. We then crawled Weibo posts from July 2022 to January 2024 using these keywords and performed data filtering and user-sensitive information replacement. Ultimately, we collected 315074 instances. The detailed keyword list and data collection process are presented in Appendix B.

3.2 Instruction Data Format

Recent studies have underscored the critical role of supervised fine-tuning (SFT) in shaping the cognitive capabilities of LLMs, with properly formatted instruction data aiding in fully leveraging the knowledge potential of LLMs (Taori et al., 2023; Chiang et al., 2023; Ouyang et al., 2022). It has been pointed out that incorporating fine-grained toxicity intensity can further enhance the robustness of PCL recognition (Wang et al., 2023). The instruction templates we constructed include both ***PCL Description Instruction*** and ***Toxicity Intensity Instruction***, designed to more accurately capture the implicit semantic characteristics of PCL, as shown in Figure 3.

PCL Description Instruction. Since PCL is a

Stage	Dataset	Language	Method	#Instances
Pcl-PT	RAL-P	EN	**Self-built**	1091945
	WEB-C	CN	**Self-built**	315074
Pcl-SFT	Don't Patronize Me (DPM)	EN	Public	10469
	TalkDown (TD)	EN	Public	74865
	CPCL	CN	**Self-built**	18253
Test	DPM/TD/CCPC	EN,CN	Public	5500

Table 2: Statistics of the datasets used in training PclGPT under different stages. Pcl-PT is used in the pre-training stage, and Pcl-SFT is used in the supervised fine-tuning stage. "Method" means we construct our own dataset / modify a public corpus. "Instances" represents the number of sentences or texts.

subjective toxic category, first, we need a complete description of PCL to guide the model to respond in a standardized format. The description includes the definition and subcategories. This part of the content is fixed and descriptive.

(PCL Description Instruction) Suppose you are a linguist and you are asked to judge whether a given text is patronizing and condescending. *<definition of PCL>*
Main Subcategories and Criteria:
<definition of Subcategories1><definition of Subcategories2>...

(Toxicity Intensity Instruction) The toxicity intensity of this sentence is mild/moderate/severe because *input reason*

Your return: Based on the following conversation, make a decision and return your choice.
Here is the text: *input text*

Output: *label*

Figure 3: A template for supervised fine-tuning instructions, including definitions of PCL and its subcategories, as well as toxicity intensity.

Toxicity Intensity Instruction. Next, we focus on the potential influence of the intensity of toxicity on implicit emotions. We incorporated the toxicity intensity labels from the original data (Commonly annotated by numerical levels), using LLM to assist in generating explanatory text and constructing instructions that describe the intensity of text toxicity.

3.3 Supervised Fine-tuning

Following the instruction format outlined in Section 3.2, we constructed the Pcl-SFT dataset for the SFT process, comprising English datasets: Don't Patronize Me (DPM) and TalkDown (TD), as well as the Chinese dataset CPCL. We adhered to the same bilingual training rules described in 3.1 to ensure the multilingual detection capability of PclGPT. In the following sections, we present detailed information regarding the Pcl-SFT dataset. More details are shown in Table 2.

- **Don't Patronize Me (DPM)** (Pérez-Almendros et al., 2020) contains 10,469 English paragraphs about potentially vulnerable groups, extracted from the News on the Web (NoW). The dataset was annotated hierarchically with numerical labels ranging from 0 to 4, indicating the toxic intensity of PCL. In SFT, we only utilized information from community texts and their corresponding labels.

- **TalkDown (TD)** (Wang and Potts, 2019) is a Reddit community dataset containing 74K English comment/reply pairs. The collected information comes from disadvantaged groups from 2006 to 2018. Each pair is marked as one of three categories: PCL, non-PCL, and unsure. In SFT, we concatenated the comment/reply pairs and manually filtered a subset to serve as training data.

- **CPCL** is a Chinese dataset we manually collected and annotated from Chinese social media platforms. We conducted hierarchical structured annotations on the data according to the toxicity definition of PCL (Pérez-Almendros et al., 2020; Wang et al., 2023). The annotations include toxicity existence, fine-grained PCL categories, and considerations for vulnerable groups. The corpus now has more than 18K two-level structured annotations. For toxicity categories, we used Wang's standard (Wang et al., 2023) to categorize Chinese PCL statements into the following subcategories: *"Unbalanced Power Relations"*, *"Spectator"*, *"Prejudice"*, *"Appeal"*, and *"Elicit Compassion"*. The annotation process involved specialized training, with two annotators for the initial annotation and one annotator for proofreading, to minimize subjective errors in marginal cases. Additionally, we performed a subjective consistency review on the annotation results to ensure the reliability of our annotated data. The detailed annotation process is described in Appendix C.

We transformed the union of the original datasets into the SFT data format, combining PCL descriptions with toxicity intensity as described in Section 3.2. We connected pairs of Enhancement-Response to form long input texts, maximizing the sequence length of LLMs. During training, we used sequence-to-sequence loss exclusively and map the final generated output to binary label pairs. We performed SFT on 8 RTX 4090 GPUs, conducting 5 epochs of full-parameter tuning with the AdamW optimizer at a learning rate of 2e-5. The specific parameters are detailed in Appendix A.

3.4 Bias Detection for PCL

Inspired by Zhang et al. (2023), we further investigated the effectiveness and fairness of our PclGPT model through group detection and fine-grained classification tasks.

Group Detection. Group detection helps us address bias issues in the model against different demographics. We conducted experiments using the DPM dataset, which balances coverage across various minority groups. We compared fine-tuned BERT series models with PclGPT-EN in these experiments.

Fine-Grained Analysis. Fine-grained analysis of toxicity categories is crucial for understanding implicit toxic sentiments (Tang et al., 2019). Our Chinese CPCL dataset divides PCL into five subcategories. We split the CPCL dataset into five subsets based on these categories to test the sensitivity of PclGPT-CN to different toxicity types. We compared PclGPT-CN with Chinese-BERT and ChatGLM in these experiments.

4 Result and Analysis

4.1 Baselines and Settings

To validate the performance of PclGPT, we extensively tested various PLMs and LLMs with our PclGPT model group on three public datasets (two in English and one in Chinese). To ensure our model demonstrates the best performance on cross-language PCL detection, we used PclGPT-EN to detect the English datasets and PclGPT-CN for Chinese.

PLMs. Pre-trained language models have consistently been the most important types of models in traditional toxicity detection tasks. We employed BERT and its relevant variants within the PLM category, such as RoBERTa (Liu et al., 2019), Chinese-BERT (Sun et al., 2021), and Multilingual-BERT (M-BERT) (Pires et al., 2019). To ensure the optimal performance of PLMs on the test set, we used the standard training and fine-tuning workflow. The predicted probability results are ultimately mapped to polarity labels through a classification layer. The training portions of three public datasets were used for training the PLMs. Additionally, both PLMs and LLMs were evaluated using the same test set to ensure comparability. Detailed parameters are shown in Appendix A, providing comprehensive insights into our experimental setup.

Base-LLMs. The use of large language models is divided into two parts. For advanced but non-open-source LLMs, such as ChatGPT and Claude-3 (Anthropic, 2024), we accessed them via API calls. Meanwhile, we used the original versions of LLaMA-2-7B and ChatGLM-3-6B without any parameter fine-tuning as part of the PclGPT ablation study to evaluate the performance improvements. To ensure experimental consistency, we used the same instruction format for other LLMs as used for PclGPT.

For the results of both PLMs and LLMs, we evaluated the models using macro-average precision (P), recall (R), and F1-score (F1).

4.2 Overall Performance

Table 3 compares the performance of PclGPT with PLMs and other LLMs on three test sets.

LM	Model	DPM			TD			CCPC (CN)		
		P	R	F1	P	R	F1	P	R	F1
PLMs	RoBERTa	76.3	78.7	77.4	88.4	86.7	86.5	61.2	61.3	61.3
	RoBERTa-L	80.2	74.9	77.2	88.1	86.0	85.9	62.5	61.6	62.0
	Chinese-BERT	71.2	63.5	66.2	76.7	74.7	74.2	66.6	71.0	67.3
	M-BERT	69.2	76.0	71.8	87.6	87.4	87.4	65.8	67.8	66.6
Base-LLMs	ChatGPT	50.8	52.3	46.9	59.2	58.1	56.7	53.1	54.2	53.6
	GPT-4.0	51.5	57.5	54.3	60.8	60.3	60.5	55.4	56.3	55.7
	Claude-3	52.3	52.5	52.3	61.6	64.1	63.2	57.2	57.7	57.3
	LLaMA-2-7B	50.9	52.6	51.4	49.9	49.9	49.7	45.2	47.5	46.3
	ChatGLM-3-6B	N/A	N/A	N/A	N/A	N/A	N/A	51.9	50.2	51.0
LLMs(Ours)	**PclGPT-EN**	**80.4**	**81.8**	**81.1**	**89.9**	**89.0**	**88.9**	N/A	N/A	N/A
	PclGPT-CN	N/A	N/A	N/A	N/A	N/A	N/A	**69.1**	**72.0**	**70.2**

Table 3: The results indicate the macro-average precision (P), recall (R), and F1-score. The F1-score is calculated by weighting the F1 of positive and negative samples. Optimal and suboptimal scores are denoted in **bold** and underlined, respectively. (CN) indicates Chinese corpus. For optimal performance, we used the model test data for the respective language, with "N/A" for non-applicable segments.

- PLM still holds significant importance in the field of toxicity detection, but the disadvantages are apparent. From the perspective of subjective ambiguity, PLM performs well on the Talkdown (English) dataset, which has a uniform data distribution and clear definitions. However, it performs poorly on the DPM (English) and CCPC (Chinese) datasets, where the definition of condescension is more ambiguous.

- PclGPT has achieved SOTA results in both English and Chinese domains, with particularly noticeable improvements in detecting ambiguous data. Specifically, PclGPT improved by 3.7% on the DPM dataset compared to the best RoBERTa model, and by 2.9% on the CCPC dataset compared to the best Chinese-BERT model.

- Base-LLMs, without parameter adjustments, have not realized their potential in subjective toxicity detection. Due to insufficient emphasis on toxic texts, unadjusted LLMs show low performance in detecting implicit toxic texts like PCL. Compared to PLMs, LLMs' average performance drops by about 20.49% in precision, 18.87% in recall, and 19.66% in F1 score. This drop is intriguing as PCL samples often contain positive expressions and goodwill, interfering with LLMs' pre-trained features. PclGPT effectively guides LLMs in understanding PCL toxicity definitions and subcategories, providing essential guidelines for future LLM safety regulations.

Category	Chat-GLM	Chinese-BERT	PclGPT-CN
Unb.	52.1	66.5	**69.4** ↑ **2.9**
Spectators	44.3	71.3	**72.1** ↑ **0.8**
Prejudice	49.7	64.3	**67.5** ↑ **3.2**
Appeal	24.5	59.0	**65.0** ↑ **6.0**
Compassion	44.2	52.3	**57.4** ↑ **5.1**

Table 4: Experimental results for fine-grained PCL Detection. We evaluated our model using the macro-average F1-score (F1) as the metric.

4.3 Result for PCL Group Detection

In Figure 4, we compared the performance of PclGPT-EN and other models in detecting PCL across different vulnerable groups. The test set had an even distribution of various vulnerable groups and positive samples. However, the models showed a clear preference for identifying poor-families and homeless individuals, indicating that these groups exhibit more identifiable semantic features. Expressions of sympathy and pity towards these groups are more likely to be perceived as condescending. PclGPT further enhanced the detection capability

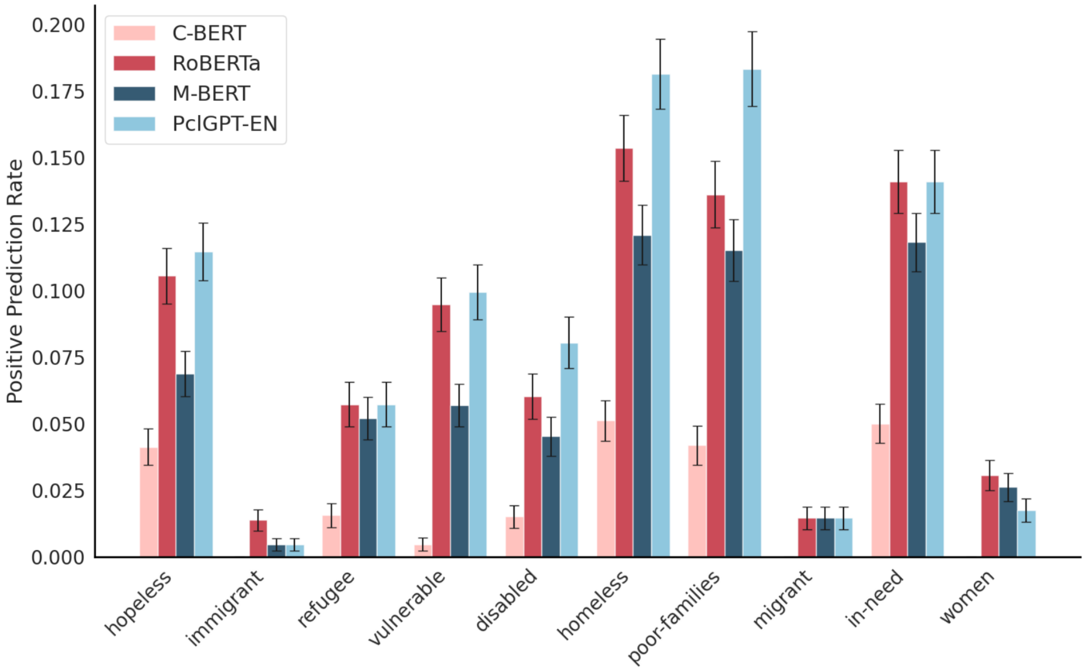

Figure 4: Group detection for different models. The test group consists of 10 different disadvantaged communities.

for these groups. In contrast, ambiguous discriminatory attitudes towards migrants and immigrants remain challenging to identify, suggesting that additional measures are necessary to protect these groups.

4.4 Result for Fine-grained PCL Detection

Table 4 presents the results of our fine-grained PCL testing. Our experiment indicated that models still exhibit varying degrees of bias in detecting different subcategories of PCL. In the "Appeal" and "Compassion" subcategories, subjective and ambiguous expressions effectively evade the recognizer's correct functioning. Notably, our PclGPT-CN showed improved performance across all subcategories, with the most significant improvement in the ambiguous "Appeal" subcategory.

5 Conclusion

This paper introduces PclGPT, a large-scale language model designed to detect patronizing and condescending language (PCL) targeting vulnerable groups. PCL, a subset of toxic speech, harms vulnerable groups through discriminatory language. Traditional pre-trained language models (PLMs) struggle with PCL detection due to its implicit harmful features. PclGPT significantly improves detection performance by leveraging the emotional semantic capabilities of LLMs. We collect, annotate, and merge the Pcl-PT/SFT dataset, and establish a bilingual PclGPT model through comprehensive pre-training and supervised fine-tuning process to detect PCL in both Chinese and English communities. PclGPT outperforms existing state-of-the-art models on three public datasets, showcasing its potential in handling implicit harmful language. Additionally, group detection and fine-grained toxicity analysis reveal significant bias differences against various vulnerable groups, highlighting the urgent need for societal protection. PclGPT's development enhances PCL recognition and provides new directions and tools for future toxic language detection research.

6 Limitation

PCL is an implicit and subjective classification of toxic language. Due to minimal existing research, further linguistic foundations are necessary to refine the standardized definition of this speech type. Our current research lacks an examination of "false positive" cases, such as insincere acts of kindness and disingenuous praise towards marginalized communities. Additionally, the subjectivity and morality of toxic speech make the use of reinforcement learning from human feedback (RLHF) for value alignment highly controversial.

References

Anthropic. 2024. Claude 3. Large Language Model developed by Anthropic.

Tom Bourgeade, Patricia Chiril, Farah Benamara, and Véronique Moriceau. 2023. What did you learn to hate? a topic-oriented analysis of generalization in hate speech detection. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3495–3508.

Rui Cao and Roy Ka-Wei Lee. 2020. Hategan: Adversarial generative-based data augmentation for hate speech detection. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6327–6338.

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

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*.

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

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

Ahmed El-Sayed and Omar Nasr. 2024. AAST-NLP at multimodal hate speech event detection 2024: A multimodal approach for classification of text-embedded images based on CLIP and BERT-based models. In *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024)*, pages 139–144, St. Julians, Malta. 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. *arXiv preprint arXiv:2109.05322*.

Paula Fortuna, Juan Soler, and Leo Wanner. 2020. Toxic, hateful, offensive or abusive? what are we really classifying? an empirical analysis of hate speech datasets. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6786–6794.

Fan Huang, Haewoon Kwak, and Jisun An. 2023. Is chatgpt better than human annotators? potential and limitations of chatgpt in explaining implicit hate speech. In *Companion proceedings of the ACM web conference 2023*, pages 294–297.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Junyu Lu, Hao Zhang, Tongyue Zhang, Hongbo Wang, Haohao Zhu, Bo Xu, and Hongfei Lin. 2022. Guts at semeval-2022 task 4: Adversarial training and balancing methods for patronizing and condescending language detection. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 432–437.

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

Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, et al. 2017. Mixed precision training. *arXiv preprint arXiv:1710.03740*.

Nicolas Benjamin 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. Association for Computational Linguistics.

OpenAI. 2022. Introducing chatgpt.

R OpenAI. 2023. Gpt-4 technical report. arxiv 2303.08774. *View in Article*, 2:13.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Carla Pérez-Almendros, Luis Espinosa Anke, and Steven Schockaert. 2022. Pre-training language models for identifying patronizing and condescending language: an analysis. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3902–3911.

Carla Pérez-Almendros, Luis Espinosa-Anke, and Steven Schockaert. 2020. Don't patronize me! an annotated dataset with patronizing and condescending language towards vulnerable communities. *arXiv preprint arXiv:2011.08320*.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Sarthak Roy, Ashish Harshavardhan, Animesh Mukherjee, and Punyajoy Saha. 2023. Probing llms for hate speech detection: strengths and vulnerabilities. *arXiv preprint arXiv:2310.12860*.

Omar Shaikh, Hongxin Zhang, William Held, Michael Bernstein, and Diyi Yang. 2022. On second thought, let's not think step by step! bias and toxicity in zero-shot reasoning. *arXiv preprint arXiv:2212.08061*.

Zijun Sun, Xiaoya Li, Xiaofei Sun, Yuxian Meng, Xiang Ao, Qing He, Fei Wu, and Jiwei Li. 2021. Chinesebert: Chinese pretraining enhanced by glyph and pinyin information. *arXiv preprint arXiv:2106.16038*.

Feilong Tang, Luoyi Fu, Bin Yao, and Wenchao Xu. 2019. Aspect based fine-grained sentiment analysis for online reviews. *Information Sciences*, 488:190–204.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Stanford alpaca: An instruction-following llama model.

Serra Sinem Tekiroglu, Yi-Ling Chung, and Marco Guerini. 2020. Generating counter narratives against online hate speech: Data and strategies. *arXiv preprint arXiv:2004.04216*.

Yuanhe Tian, Ruyi Gan, Yan Song, Jiaxing Zhang, and Yongdong Zhang. 2023. Chimed-gpt: A chinese medical large language model with full training regime and better alignment to human preferences.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Hongbo Wang, Mingda Li, Junyu Lu, Liang Yang, Hebin Xia, and Hongfei Lin. 2023. Ccpc: A hierarchical chinese corpus for patronizing and condescending language detection. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 640–652. Springer.

Zijian Wang and Christopher Potts. 2019. Talkdown: A corpus for condescension detection in context. *arXiv preprint arXiv:1909.11272*.

Jiaxin Wen, Pei Ke, Hao Sun, Zhexin Zhang, Chengfei Li, Jinfeng Bai, and Minlie Huang. 2023. Unveiling the implicit toxicity in large language models. *arXiv preprint arXiv:2311.17391*.

Gloria Wong, Annie O Derthick, EJR David, Anne Saw, and Sumie Okazaki. 2014. The what, the why, and the how: A review of racial microaggressions research in psychology. *Race and social problems*, 6:181–200.

Jinghua Xu. 2022. Xu at semeval-2022 task 4: Pre-bert neural network methods vs post-bert roberta approach for patronizing and condescending language detection. *arXiv preprint arXiv:2211.06874*.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). *arXiv preprint arXiv:1903.08983*.

Boyu Zhang, Hongyang Yang, and Xiao-Yang Liu. 2023. Instruct-fingpt: Financial sentiment analysis by instruction tuning of general-purpose large language models. *arXiv preprint arXiv:2306.12659*.

Yiming Zhu, Peixian Zhang, Ehsan-Ul Haq, Pan Hui, and Gareth Tyson. 2023. Can chatgpt reproduce human-generated labels? a study of social computing tasks. *arXiv preprint arXiv:2304.10145*.

A Parameter Settings

In this section, we provided a detailed description of the experimental parameter settings. This included the parameters for pre-training and supervised fine-tuning of the PclGPT.

A.1 PLM Settings

To compare our PclGPT, we fine-tuned our PLMs using the same size training and test sets as those used for PclGPT. Specifically, we conducted fine-tuning experiments for 10 epochs on 2 A100 GPUs and used the best epoch model weights for test set predictions. We tested RoBERTa, Chinese-BERT, and M-BERT models on three datasets. The specific parameters are as shown in Table 5.

Parameter_for_PLM	Value
Lr	1e-2
Max_len	1024
Batchsize	16
Training Epochs	5
warmip_steps	500
GPUs	A100_PCIE*2 (40G)

Table 5: Detailed parameter settings for the fine-tuning and testing phases of PLMs.

Parameter_for_PT	Value	Parameter_for_SFT	Value
Lr	2e-4	Lr	2e-5
Batchsize	32	Batchsize	16
Training Epochs	10	Training Epochs	5
Max Source Len	512	Block Size	1024
Max Target Len	512	-	-
GPUs	RTX 4090*8 (24G)	GPUs	RTX 4090*8 (24G)
-	-	GPUs_inference	A100_PCIE*2 (40G)

Table 6: Detailed configuration parameters for the pre-training and supervised fine-tuning phases of PclGPT. The inference phase uses the same GPU configuration as the PLM test.

A.2 PclGPT Settings

For PclGPT, due to the scale effect of the pre-training corpus, we set a higher learning rate and batch size than supervised fine-tuning. Both the pre-training and supervised fine-tuning were conducted on 8 RTX 4090 GPUs. We accomplished this procedure and guaranteed the consistency of the pertinent training parameters in both Chinese and English models. During the inference phase, to control for a single variable, we used the same configuration of 2 A100 GPUs as in the PLM fine-tuning, as shown in Table 6. This inference setup is also applicable to the zero-shot inference process for non-API Base-LLMs, like LLaMA-2-7B and ChatGLM-3-6B.

B Detailed Construction of the Pcl-PT Dataset

RAL-P. In the process of transforming RAL-E, we used LLM to construct a patronizing language dictionary. Specifically, we had the LLM generate 300 words that best reflect patronizing semantics based on confidence levels, which were then manually verified. Part of the word cloud information sorted by confidence levels is shown in Figure 5. For sentences in RAL-E that did not contain any dictionary information, we retained only 30% as non-patronizing corpus, while all sentences containing any dictionary information were retained. The original text corpus consisted of 1,476,472 sentences, and the filtered corpus contained 1,091,945 sentences, which were used as RAL-P pre-training data.

WEB-C. We uniformly collected data from various vulnerable community groups on the Weibo platform as our WEB-C Chinese pre-training corpus. Detailed information on community categories can be found in Table 9. For filtering, we removed duplicate and irrelevant samples (including common fixed tags on Weibo such as "# 话题内容" and "# 评论日期"), and we replaced user information with #USER to comply with the community privacy agreement. We retained the emojis in the samples and converted them to the corresponding Chinese text specified by the platform to preserve as much of the emotional semantic information conveyed by the emojis as possible.

Figure 5: Word cloud statistics of the condescending dictionary.

C Detailed Construction of the Pcl-SFT Dataset

CPCL. We adopted the same method as WEB-C described in Appendix B for data selection and filtering, and manually annotated the high-quality texts. This section provides a detailed description of the annotation and statistics of our constructed CPCL dataset. Due to the subjective nature of PCL speech, we abandoned the automatic annotation method by LLM and continued to use manual annotation. We recruited four annotators with diverse gender, age, and educational backgrounds (two pri-

	Disabled	Women	Elderly	Children	Single-parents	Ordinary.	Disadv. groups	Total
zhihu	1208	1147	1131	1619	1113	1093	1959	9270
zhihu_p	338	248	294	374	264	263	354	2135
prop.(%)	28.0	21.6	26.0	23.1	23.7	24.1	18.1	23.0
weibo	1102	974	1247	1588	1077	944	2051	8983
weibo_p	310	241	267	592	389	123	533	2455
prop.(%)	28.1	24.7	21.4	37.3	36.1	13.0	26.0	27.3
Total	2310	2121	2378	3207	2190	2037	4010	18253

Table 7: Statistical Results of CPCL from different Platforms. Platform_p represents samples marked as PCL, whereas prop.(%) represents a percentage.

Binary-classification	*Kappa IAA*
All labels	0.62
Remove Weak level	0.67
Multi-classification	*Kappa IAA*
Unbalanced Power Rel.	0.65
Spectators	0.54
Prejudice	0.61
Appeal	0.48
Sympathy	0.71

Table 8: Kappa IAA scores of CPCL binary and multi-class annotations.

Community	*Total*
# Disabled	38981
# Women	40256
# Elderly	39385
# Children	38475
# Single-parent	40689
# Ordinary People	37589
# Disadvantaged	40324
# Others	39375

Table 9: The collection status of different PCL communities. *Total* is the total number of sentences collected for each community.

mary annotators and two proofreaders) (50% female, 50% male; age 25±5 years; two master's degree holders, two PhD holders). We adopted the standard proposed by Wang et al. (2023) and conducted detailed training on test samples before annotation to ensure that annotators understood the subtle toxicity differences of PCL. The annotation was uniformly conducted using the annotation template as shown in Figure 8. To ensure annotation consistency, we calculated the Kappa inter-annotator agreement (IAA) for binary and multi-class annotations. The IAA results are shown in Table 8. If we ignore all annotations marked as low toxicity intensity by at least one annotator, the IAA improves. This indicates that PCL with weak toxicity intensity has higher ambiguity. Detailed statistics of the CPCL dataset are shown in Table 7.

D Case Study for PclGPT

To further illustrate the rationales of PclGPT, and to determine whether the model can effectively identify the fuzzy subcategory of PCL. We selected samples from the Chinese and English test results respectively for case testing. The results are detailed in Table 10. Regarding the English part, we selected BERT-multi, RoBERTa, GPT-4.0, Claude-3, LLaMA-2-7B and PclGPT-EN for comparative analysis. For Chinese data, we choose Chinese pre-trained Chinese-BERT, ChatGLM-3-6B and PclGPT-CN for comparison.

- Case A generally selects cases with "Unbalanced Power Relations" and "Prejudice" labels in PCL. In these examples, advantaged groups place themselves in a higher social status and display strong discriminatory characteristics against disadvantaged groups. For example, "so-called" in A(i) satirizes that poor communities should not receive subsidies, a severe expression of prejudice. A(ii) expresses the stereotype that "children from single-parent families are difficult to get along with". The toxicity of this type of speech is apparent. Although there is no precise attack vocabulary, the models can detect it effectively. In A(i), most models can effectively identify

EN	Case A(i)	Case B(i)
Text	*After already receiving relief funds, what else do these so-called 'poor' families think they deserve?*	*For some of these male prostitutes, the 'clients' they picked up on this corner were their only means of survival.*
Category	"Unbalanced Power Relations","Prejudice"	"Spectator","Elicit Compassion"
Explanation	The phrase "so-called 'poor' families" suggests a condescending attitude towards impoverished households, reflecting an unbalanced power relationship, where those with more resources view those with less through a biased perspective. The tone is dismissive and judgmental.	The phrasing of this sentence suggests a spectator's indifferent attitude towards male prostitutes. It implies that these men have no other choice but to engage in sex work for survival. Spectators elicit compassion for their plight while maintaining a superior stance. The toxicity of such descriptive statements is often complex to detect.
Recognition Difficulty	Middle	High
Prediction	BERT-multi:✔, RoBERTa:✔, GPT-4.0:✗, Claude-3:✔, LLaMA-2:✗, PclGPT-EN:✔	BERT-multi:✗, RoBERTa:✗, GPT-4.0:✗, Claude-3:✔, LLaMA-2:✗, PclGPT-EN:✔

CN	Case A(ii)	Case B(ii)
Text	单亲的小孩大概率很难相处。 **Translation:** *Children from single-parent families often face difficulties in getting along with others.*	农民工挣钱不容易的，确保工资该发就发吧。 **Translation:** *Making a living as a migrant worker is no easy task, let's make sure they receive their rightful wages.*
Category	"Unbalanced Power Relations","Prejudice"	"Appeal","Elicit Compassion"
Explanation	This statement reflects an unbalanced power relation and prejudice against single-parent families. It assumes that children from such backgrounds inherently face social difficulties, ignoring the complexity of individual experiences and the diverse support systems that may exist.	This superficial appeal for fairness to migrant workers hides implicit bias. It simplifies their fight and focuses solely on the wage situation. Due to the lack of offensive intent, this condescending attitude is difficult to detect without deeper analysis.
Recognition Difficulty	Middle	High
Prediction	RoBERTa:✗, Chinese-BERT:✔, GPT-4.0:✗, Claude-3:✔, ChatGLM-3:✔, PclGPT-CN:✔	RoBERTa:✗, Chinese-BERT:✗, GPT-4.0:✗, Claude-3:✗, ChatGLM-3:✔, PclGPT-CN:✔

Table 10: Illustration of case study. We selected typical PCL samples from the English and Chinese test sets respectively. "Category" represents the fine-grained toxicity category of PCL, "Explanation" is a manual annotation analysis, and the key information is marked in red. ✔ indicates that the model has made a correct judgment, ✗ indicates a wrong judgment.

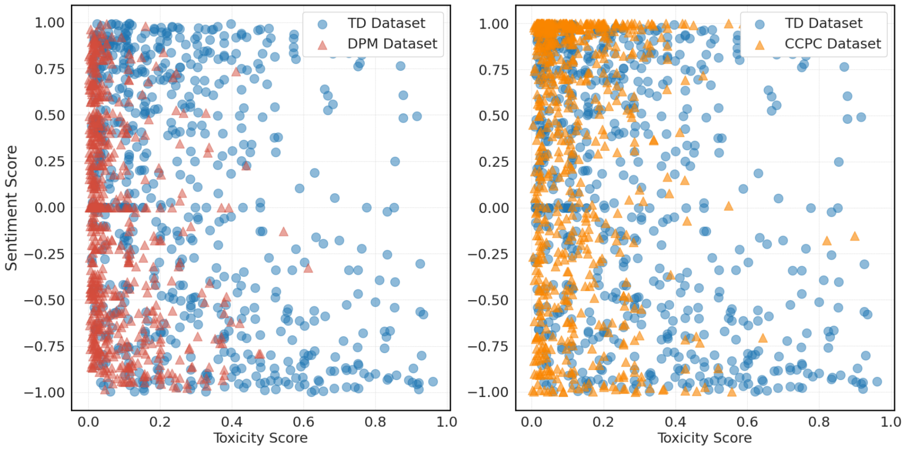

Figure 6: Toxicity score scatter plots for three PCL datasets.

the result. Similar results were obtained in A(ii), indicating that the Chinese domain also uses the semantic information of PCL.

- The cases selected in Case B are mostly subcategories of "Spectator" and "Elicit Compassion". These categories place advantaged groups as bystanders, offering superficial opinions to solve problems or expressing sympathy for disadvantaged groups. In B(i), people's sympathy for the "client" is aroused through descriptive sentences, and in B(ii), people's concern for the "migrant worker" is aroused, and people are called for guaranteed wages. The PCL toxicity of these remarks is hidden in vague expressions, and it is difficult for the model to detect the implicit toxicity. For B(i), only Claude-3 and PclGPT-EN correctly identified the result, while for B(ii), only ChatGLM-3 and PclGPT-CN correctly identified the result. This demonstrates the importance of PclGPT for implicit toxicity detection.

E Add Implicit Interference Samples

We conducted additional experiments to assess PclGPT's detection capabilities for implicit toxicity. As a subjective sentiment, the ambiguous part of PCL's semantic information often results in interference samples during annotation. These samples have more marginal condescending attributes, hindering the model's ability to distinguish positive samples effectively. We experimented with three dataset scenarios: without any interference samples, with a limited number of interference samples, and with all interference samples included.

Result. Identifying interference samples encom-

Model	*S-None*	*S-Few*	*S-All*
BERT	67.1 (0)	67.2 (+0.1)	67.1 (-0.6)
ChatGLM	48.1 (0)	48.8 (+0.7)	48.3 (-0.5)
ChatGPT	64.3 (0)	61.3 (-3.0)	52.4 (-8.9)
GPT-4.0	65.5 (0)	63.2 (-2.3)	54.5 (-8.7)
PclGPT	**67.7 (0)**	**71.5 (+3.8)**	**72.8 (+1.3)**

Table 11: The test results of each model after gradually adding fuzzy samples. The percentage in parentheses indicates the change after addition compared with before addition.

passing complex and implicit emotions is a difficult objective in toxicity analysis. Table 11 displays the following test results. It is evident that when the number of interference intermediate examples increases, both the BERT model and the GPT baseline model experience a decrease in performance. Notably, ChatGPT and GPT-4 decline over 8%, suggesting that they inadequately capture the condescending traits of these fuzzy cases. PclGPT is the only model that can effectively detect these interference samples in the S-Few and S-All datasets, which fully demonstrates the robust testing capabilities of our model.

F Toxicity Scores and Implicit Features

Figure 6 uses a scatter plot to show the toxicity scores of the three PCL test sets. The TD dataset has a smooth distribution across the entire range, while the DPM and CCPC datasets have lower average toxicity scores, with samples concentrated in low or zero-score regions. This correlates with the weaker F1 scores in the DPM and CCPC data, indicating that lower toxicity scores often align with higher implicit features, suggesting more exploration is needed for implicit toxicity. The scatter plot also shows that sentiment scores (vertical axis) have a limited impact on PCL detection, as the sentiment scores do not exhibit distinct distribution patterns.

```
{
 "conversations": [
        {
            "from": "human",
            "value": "Suppose you are a linguist and you are asked to judge whether a
                      text is patronizing and condescending. <definition of PCL>
                      # Main Subcategories and Criteria:
                      **Unbalanced Power Relations: Language that conveys a
                      power differential, positioning the speaker as superior...
                      **<definition of Subcategories2>...
                      If any one of above subcategories is met, the text is patronizing
                      and condescending. If none of above subcategories apply, the
                      text is considered non-condescending.
                      The toxicity intensity of this sentence is severe because the text
                      using offensive words to express prejudice.
                      # Your return: Based on the following conversation, make a
                      decision and return your choice. FALSE means the text is not a
                      condescending language,and TRUE means the text is a
                      condescending language.
                      Here is the text->
                     (Yes, everyone does. Humans should not be doing manual labour,
                      and the idiots who fight tooth and nail against any form of
                      education are just holding automation back.)
        },
        {
              "from": "gpt",
              "value": "TRUE"
         }
              ]
    },
```

Figure 7: Pcl-SFT data sample in JSON format.

Patronizing and Condescending Language (PCL) is a form of implicitly toxic speech aimed at vulnerable groups with the potential to cause them long-term harm. Please determine if the following text is PCL. *If it is, further assess the toxicity level and classify it into the appropriate categories.*

Tips:

(1) The PCL text itself is less aggressive, and a clear characteristic is that the speaker is expressing their views from a position evidently different from that of the disadvantaged group.

(2) Statements with clear insulting vocabulary and hate/offensive language targeting specific individuals are not considered PCL; they are categorized as non-PCL.

(3) To reduce subjective errors, please indicate the toxicity level when annotating PCL: Weak, Middle, or Strong. No further labeling is required for non-PCL statements.

Text:

You can't always blame your incompatibility on her being from a single-parent family.

1. Is this text patronizing or condescending? *(Skip (2) and (3) if 'No' is selected)*

○ Yes ○ No

2. Please determine the subcategory of PCL. *(multiple choices)*

☐ Unbalanced Power Relations ☐ Spectators ☐ Prejudice Impression

☐ Appeal ☐ Elicit Sympathy

3. Please further assess the toxicity level of PCL.

○ Weak ○ Middle ○ Strong

Figure 8: We used a web-based layered annotation questionnaire, which includes the definitions of annotations, annotation tips, and input texts. Every time we changed the text, we performed batch annotation.