Overview of CCL25-Eval Task 10: Fine-grained Chinese Hate Speech Identification Evaluation Task

Junyu Lu[1]*, Zewen Bai[1]*, Shengdi Yin[1]*, Liang Yang[1,2], Hongfei Lin[1,2]†
[1] School of Computer Science and Technology,
[2] Key Laboratory of Social Computing and Cognitive Intelligence,
Dalian University of Technology, China
liang,hflin@dlut.edu.cn

Abstract

This paper provides an overview of the CCL25-Eval Task 10, i.e., Fine-grained Chinese Hate Speech Identification Evaluation. The primary objective of this task is to perform a fine-grained analysis of hateful samples. In addition to binary classification, systems are required to identify and extract the comment target, argument span, and the associated targeted group within each sample, thereby enhancing the model's capability in fine-grained detection and improving the interpretability of its decisions. In total, more than 300 teams registered for the task, with 100 teams submitting valid results. We present the submitted results and provide a comprehensive analysis of the technical approaches adopted by the top-performing teams. The dataset used in this task has been available[1].

1 Introduction

With the widespread adoption of social media, user-generated content has seen explosive growth, accompanied by the increasing spread of hate speech. Hate speech refers to harmful expressions that incite hatred or violence against individuals or groups based on characteristics such as race, religion, gender, region, sexual orientation, or physical condition (Nockleyby, 2000; Davidson et al., 2017). Compared to other forms of harmful content, hate speech is often more coercive, bullying, and inflammatory in nature, posing serious threats to individuals and society at large (Maarouf et al., 2024). Multiple laws and regulations in China, including *the Law of the People's Republic of China on Administrative Penalty* and *the Administrative Measures for Internet Information Services*, explicitly prohibit the dissemination of hate speech. As a result, how to effectively detect hate speech has attracted widespread attention from both academia and industry.

In recent years, researchers have leveraged natural language processing techniques for the detection of Chinese hate speech, resulting in the development of multiple open-source datasets and detection models. Specifically, (Chung and Lin, 2021) constructed a Traditional Chinese hate speech dataset and evaluated the detection performance of various machine learning methods. (Jiang et al., 2022) focused on the sexism content on Simplified Chinese social media and compared the effectiveness of machine learning, deep learning, and pretrained models. (Deng et al., 2022) introduced the large-scale COLDataset and proposed COLDetector, a prompt-based detection framework. (Zhou et al., 2022) addressed implicit hate speech with biased expressions and presented the Cdial-Bias dataset. (Lu et al., 2023) proposed ToxiCN, a fine-grained dataset towards multiple groups, along with a knowledge-enhanced method, TKE, which incorporates insult-related embeddings into the model to improve its understanding. (Wang et al., 2024b) built PCL, a bilingual Chinese-English prejudice dataset, and analyzed the effects of instruction tuning on large language models. (Xiao et al., 2024) studied Chinese lexical variants such as homophones and character decomposition, highlighting challenges for hate speech detection in Chinese. (Bai et al., 2025) further extended ToxiCN with richer annotations, including explicit attack targets and toxic

*Equal contribution.
†Corresponding author
[1]https://github.com/DUTIR-Emotion-Group/CCL2025-Chinese-Hate-Speech-Detection

Proceedings of the 24th China National Conference on Computational Linguistics, pages 387-394, Jinan, China, August 11-14, 2025.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Work	Platforms	Language	#Posts	Span	Tar.	Arg.	Group
TOCAB (Chung and Lin, 2021)	PTT	TC	121,344				✓
COLD (Deng et al., 2022)	Zhihu, Weibo	CN	37,480				
SWSR (Jiang et al., 2022)	Weibo	CN	8,969				✓
Cdial-Bias-Utt (Zhou et al., 2022)	Zhihu	CN	13,394				✓
Cdial-Bias-Ctx (Zhou et al., 2022)	Zhihu	CN	15,013				✓
ToxiCN (Lu et al., 2023)	Zhihu, Tieba	CN	12,011				✓
PCL (Wang et al., 2024b)	Zhihu, Tieba	CN, EN	94,324				✓
ToxiCloakCN (Xiao et al., 2024)	Zhihu, Tieba	CN	4,582				✓
STATE ToxiCN (Bai et al., 2025)	Zhihu, Tieba	CN	8,029	✓	✓	✓	✓

Table 1: Comparison of existing Chinese hate speech datasets based on *Platforms*, *Language*, number of *#Posts*, span-level annotations (*Span*), inclusion of Target (*Tar.*), Argument (*Arg.*), and *Group*. TC means Traditional Chinese, and CN means Simplified Chinese.

argument spans, to facilitate deeper analysis of model decision-making. The comparison of these studies is illustrated in Table 1.

In this evaluation, we designed a fine-grained Chinese hate speech detection task, aiming to advance the development of detection techniques, strengthen the regulation of harmful online behavior, and attract broader attention from the research community. The task requires participating systems not only to perform binary classification, but also to accurately extract spans corresponding to the comment target, argument, and the referenced targeted group. These elements together form a structured hate speech quadruple, which enhances the model's detection capability in fine-grained scenarios and improves the interpretability of its decisions. We adopt the STATE ToxiCN (Bai et al., 2025) as the evaluation dataset, due to its rich annotations across multiple granularities. During the evaluation, a total of 300 teams registered for the task, among which 100 teams successfully submitted valid results.

In this paper, we present the task description in Section 2. We briefly describe the STATE ToxiCN dataset in Section 3. In Section 4, we discuss the metrics used to rank participant submissions. We list the leaderboard of submissions and provide a more in-depth discussion in Section 5. In Section 6, we introduce the technical system of the excellent teams. We finally conclude the paper in Section 7.

2 Task Description

The evaluation task focuses on fine-grained span-level Chinese hate speech detection. Given an input social media text, the system should output hate speech quadruples in the following order: *Target*, *Argument*, *Group*, and *Hateful*. Specifically,

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

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

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

Additionally, for non-hateful samples and general offensive language that do not target specific groups, it is still necessary to extract the target group and argument, labeling them as "non-hate". Since a single sample may contain multiple targets, multiple quadruples are allowed. The sample annotations are shown in Table 2.

To enhance the organizational efficiency and fairness of the evaluation, the competition is divided into two phases: a preliminary round and a final round. The top fifteen teams based on preliminary scores will advance to the final round.

Proceedings of the 24th China National Conference on Computational Linguistics, pages 387-394, Jinan, China, August 11-14, 2025.

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

Table 2: Examples of annotated posts from STATE ToxiCN dataset with corresponding annotations of Target-Argument-Hateful-Group quadruples.

3 Dataset

In this section, we provide a brief overview of the STATE ToxiCN dataset used in the task. We first describe the process of its construction, followed by a presentation of its statistics across different levels of granularity.

3.1 Data Collection and Preprocessing

The STATE ToxiCN dataset is derived from ToxiCN (Lu et al., 2023), a post-level Chinese hate speech corpus. Through rigorous filtering and annotation, researchers refine this resource into a high-precision span-level hate speech dataset. The preprocessing stage involves eliminating noise from the raw data, including non-relevant content such as promotional material, nonsensical strings, and text segments that are excessively brief (under 5 characters) or lengthy (exceeding 500 characters).

Additionally, researchers meticulously verify existing hate speech labels, focusing on the explicitness of targeted entities and associated derogatory context. Samples with vague or undefined targets—particularly those lacking a coherent Target-Argument relationship—are discarded to maintain structural integrity. This ensures the dataset aligns with fine-grained hate speech analysis requirements.

3.2 Data Annotation

To guarantee annotation reliability, researchers implement a structured labeling protocol with multi-tier quality checks. The objective is to annotate each post with Target-Argument-Hateful-Group quadruples. The process begins with comprehensive guideline formulation, covering extraction protocols for Target-Argument pairs, criteria for hateful content identification, and taxonomy for group classification.

To ensure annotation quality, all texts undergo independent labeling by a minimum of two annotators adhering to standardized guidelines for identifying Target-Argument pairs, assessing hate speech intensity, and categorizing targeted groups. The quality control process includes random cross-validation of 20% samples, where different annotators re-examine selected cases to verify annotation consistency and maintain uniform interpretation of labeling criteria. This systematic review mechanism enables early detection and correction of potential annotation inconsistencies.

Cases exhibiting substantial inter-annotator disagreement are escalated to a panel of subject matter experts for resolution. The arbitration team conducts comprehensive evaluations by examining textual contexts and applying annotation protocols to reach consensus on final labels, thereby ensuring the reliability and objectivity of contentious annotations.

3.3 Data Description

The STATE ToxiCN dataset comprises 8,029 annotated posts, with hateful content appearing in 4,942 instances (61.55% of the total). The annotation process yielded 9,532 quadruples, among which 6,034 (63.30%) were identified as containing hateful content. As detailed in Table 3, the dataset reveals that Gender, Region, and Race emerge as the three most frequently targeted group categories. Notably, the dataset also includes 854 cases (8.96%) classified as "multi-group", representing instances where hateful expressions simultaneously target multiple distinct groups within a single Target-Argument pair. In the

Proceedings of the 24th China National Conference on Computational Linguistics, pages 387-394, Jinan, China, August 11-14, 2025.

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

Table 3: Statistics of annotated posts from the STATE ToxiCN dataset, including Target, Argument, Group, and Hatefulness classifications.

evaluation task, the data is randomly divided into three subsets: a training set, a preliminary test set, and a final test set, with a sample ratio of 2:1:1.

4 Evaluation Metrics

In this task, we adopt two evaluation metrics—hard match score and soft match score—drawing inspiration from prior work in aspect-level sentiment analysis. The final score, which determines the ranking index, is calculated as the average of the F1 scores from both matching strategies. Specifically, the calculation method is consistent with the scikit-learn (sklearn) machine learning library. The specific calculation formulas are as follows:

$$F_1 = 2 \cdot \frac{P \cdot R}{P + R} \tag{1}$$

In this formula, P is Precision and R is Recall.

Hard Match Score: A predicted quadruple is judged as correctly extracted if and only if every element of the predicted quadruple exactly matches the corresponding element in the ground truth answer.

Soft Match Score: A predicted quadruple is judged as correctly extracted if and only if the Targeted Group and Hateful elements of the predicted quadruple exactly match the corresponding two elements in the ground truth answer; The string matching similarity between the Target element of the predicted quadruple and its corresponding element in the ground truth answer exceeds 50%; The string matching similarity between the Argument element of the predicted quadruple and its corresponding element in the ground truth answer also exceeds 50%. The similarity calculation method is consistent with the `SequenceMatcher` function from Python's standard `difflib` module. The specific calculation is as follows:

$$\text{Similarity} = \frac{2 \cdot M}{\text{len}_{\text{pred}} + \text{len}_{\text{gold}}} \tag{2}$$

In the formula:

- len_{pred} is the length of the `Target` and `Argument` in the model's predicted quadruple.
- len_{gold} is the length of the corresponding element in the ground truth answer.
- M is the length of the Longest Common Subsequence (LCS) between the predicted quadruple's elements and the ground truth answer.

Additionally, in the calculation process of the soft match metric, the Longest Common Subsequence has a requirement for the order of characters in the text; only characters that are correct and whose character order is correct will be counted as the Longest Common Subsequence.

Proceedings of the 24th China National Conference on Computational Linguistics, pages 387-394, Jinan, China, August 11-14, 2025.

5 Results and Analysis

The preliminary and final results of this evaluation are shown in Table 4 and 5, respectively[1]. The overall performance in the preliminary round was slightly higher than that in the final round. This discrepancy can be attributed to a potential data leakage issue identified after the preliminary round. Specifically, we noticed that some teams had used the ToxiCN dataset during model training, which serves as the source data for the STATE ToxiCN dataset used in our evaluation. As a result, there was a risk that models may have been indirectly exposed to the test set. To ensure fairness in the final round, we explicitly prohibited the use of the ToxiCN dataset for training prior to the final evaluation.

Index	Team Name	Organization	Hard Score	Soft Score	Final Score
1	Nova-Z	ShanghaiTech University	0.2692	0.5133	0.3913
2	星辰之力	China Telecom Chongqing Branch	0.2675	0.5053	0.3864
3	珠科智能2班	Zhuhai College of Science and Technology	0.2630	0.4786	0.3708
4	zutNLP	Zhongyuan University of Technology	0.2488	0.4899	0.3693
5	青年先疯	Guangdong University of Foreign Studies	0.2432	0.4865	0.3648
6	没名字取了	Zhengzhou University	0.2429	0.4840	0.3634
7	地铺稀客队	Beijing Institute of Technology	0.2463	0.4654	0.3559
8	过来水个比赛队	East China Normal University	0.2437	0.4671	0.3554
9	BMI	Dalian University of Technology	0.2504	0.4602	0.3553
10	可唐	Yunnan University	0.2421	0.4614	0.3518
11	AIDM	Wuhan University	0.2479	0.4550	0.3515
12	AREA34	Individual	0.2472	0.4455	0.3463
13	_KING_	Harbin Institute of Technology, Shenzhen	0.2348	0.4431	0.3389
14	爆米蛙	Dalian University of Technology	0.2287	0.4372	0.3329
15	秒速五厘米组织	The Australian National University	0.2270	0.4357	0.3313

Table 4: Preliminary round results.

Index	Team Name	Organization	Hard Score	Soft Score	Final Score
1	过来水个比赛队	East China Normal University	0.2541	0.4740	0.3641
2	_KING_	Harbin Institute of Technology, Shenzhen	0.2558	0.4714	0.3636
3	星辰之力	China Telecom Chongqing Branch	0.2468	0.4715	0.3591
4	珠科智能2班	Zhuhai College of Science and Technology	0.2441	0.4692	0.3566
5	BMI	Dalian University of Technology	0.2498	0.4612	0.3555
6	zutNLP	Zhongyuan University of Technology	0.2342	0.4747	0.3545
7	Nova-Z	ShanghaiTech University	0.2423	0.4569	0.3496
8	地铺稀客队	Beijing Institute of Technology	0.2434	0.4547	0.3490
9	可唐	Yunnan University	0.2412	0.4538	0.3475
10	AIDM	Wuhan University	0.2388	0.4500	0.3444
11	爆米蛙	Dalian University of Technology	0.2260	0.4444	0.3352
12	青年先疯	Guangdong University of Foreign Studies	0.2190	0.4421	0.3306
13	没名字取了	Zhengzhou University	0.2207	0.4309	0.3258

Table 5: Final round results.

6 Participant Systems

The participating teams employed a variety of strategies to tackle the task of fine-grained Chinese hate speech identification. This section provides an overview of the technical approaches adopted by the top-performing teams. Their diverse methodologies highlight the richness of possible solutions in this field and offer valuable insights into future research directions.

[1]According to the earlier requirements, teams that did not indicate an intention to submit the registration form have been removed. Due to space limitations, only the top 15 teams that advanced to the final round are listed here. In addition, only 13 teams submitted their final results in the final round. The complete leaderboard can be found at https://tianchi.aliyun.com/competition/entrance/532298/rankingList.

Proceedings of the 24th China National Conference on Computational Linguistics, pages 387-394, Jinan, China, August 11-14, 2025.

- **[过来水个比赛队]**: This study details a prompt-optimized framework for fine-grained Chinese hate speech detection using the GLM4-9B-Chat model. Key optimizations include structured output formatting with explicit labels, rule-driven prompt refinement (e.g., strict text replication, defined hate categories), iterative Low-Rank Adaptation (LoRA) tuning with a train-validation split to prevent overfitting, and custom examples for enhanced few-shot learning. A four-step reasoning process ensures localized context analysis. The framework reportedly achieves stable performance on CCL2025 benchmarks.

- **[_KING_]**: The researchers introduce an innovative approach that integrates Task Reformulation (TR), Self-Retrieval-Augmented Generation (SRAG), and Multi-Round Accumulative Voting (MAV). Their method enhances output stability and performance, based on the Qwen2.5-7B-Instruct model, by reformulating the quadruplet extraction task into a triplet extraction task, leveraging dynamic retrieval from the training set for contextual prompts, and employing multi-round inference.

- **[星辰之力]**: The researchers developed a dynamic clue-augmented SFT and multi-stage optimization method for fine-grained Chinese hate speech detection with Qwen2.5-7B. They segmented quadruple extraction into three stages, enhanced by dynamic clue-augmented SFT. Classification benefited from label mapping, optimized training, and RL for loss, while extraction used a model-distilled CoT set. Their framework uniquely ranked in the TOP3 on two test sets, showing strong stability, generalization, and advancing the field.

- **[珠科智能2班]**: The authors propose a framework integrating the GLM-4-9B model with parameter-efficient LoRA fine-tuning for fine-grained Chinese hate speech detection. This model extracts structured "Target-Argument-Group-Hateful" tuples, reportedly outperforming baselines and improving accuracy on subtle, implicit hateful expressions. It demonstrates potential for real-world applications like content moderation, bias detection, and social media monitoring, while offering computational efficiency and interpretability.

- **[BMI]**: Addressing the challenge of hate speech on Chinese social media, where traditional systems struggle with context and slang, the authors propose a novel three-stage LLM-based framework. This framework involves prompt engineering for implicit hate patterns, supervised fine-tuning for domain adaptation, and LLM merging for robustness against out-of-distribution cases. Evaluations on the STATE-ToxiCN benchmark reportedly validate its effectiveness, demonstrating superior performance over baseline methods in detecting fine-grained hate speech.

- **[zutNLP]**: This study aims to identify fine-grained Chinese hate speech by extracting structured quadruples, enhancing detection accuracy and interpretability. To address data limitation and model optimization challenges, the authors propose an iterative multi-model collaborative fine-tuning framework. This framework efficiently optimizes the GLM4-9B model using LoRA, generating high-quality pseudo-labels via multi-model consensus to iteratively expand training data and improve performance. Preliminary results reportedly show significant outperformance over the baseline on key metrics.

7 Conclusions and Future Work

In this paper, we present an overview of CCL25-Eval Task 10: Fine-grained Chinese Hate Speech Identification Evaluation. This task focuses on the fine-grained analysis of online posts by extracting the comment target, argument span, and the associated targeted group, as well as determining whether the content conveys hate speech. We report the leaderboard of submitted results and provide a comprehensive analysis of the technical approaches adopted by the top-performing teams, offering valuable insights for future research in this area.

For future work, we plan to continue exploring effective methods for improving the identification of fine-grained Chinese hate speech. In addition, we aim to extend this task to multimodal data, such as memes (Lu et al., 2024; Gu et al., 2024) and videos (Wang et al., 2024a; Wang et al., 2025), to further study Chinese hate speech in more complex and diverse media formats.

Ethical Statement

This task aims to promote the fine-grained detection of Chinese hate speech. We are aware of the potential risk that malicious actors may attempt to reverse-engineer or misuse the comments contained in the dataset. We firmly denounce such behavior and emphasize the importance of human moderation to prevent any harmful use. The ownership of the dataset belongs to the Information Retrieval Laboratory at Dalian University of Technology. All resources are intended solely for scientific research and are prohibited from commercial use. During the dataset construction process, we strictly adhered to the data usage policies of each public online social platform. The opinions and findings contained in the dataset samples should not be interpreted as representing the views expressed or implied by the authors.

Acknowledgements

This research is supported by the Natural Science Foundation of China (No. 62376051) and the Fundamental Research Funds for the Central Universities (DUT24LAB123). We sincerely thank the Organizing Committee of CCL 2025 for their strong support of this evaluation.

References

Zewen Bai, Yuanyuan Sun, Shengdi Yin, Junyu Lu, Jingjie Zeng, Haohao Zhu, Liang Yang, and Hongfei Lin. 2025. State toxicn: A benchmark for span-level target-aware toxicity extraction in chinese hate speech detection. arXiv preprint arXiv:2501.15451.

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

Thomas Davidson, Dana Warmsley, Michael W. Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In Proceedings of the Eleventh International Conference on Web and Social Media, ICWSM 2017, Montréal, Québec, Canada, May 15-18, 2017, pages 512–515. AAAI Press.

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.

Tianlong Gu, Mingfeng Feng, Xuan Feng, and Xuemin Wang. 2024. Scare: A novel framework to enhance chinese harmful memes detection. IEEE Transactions on Affective Computing.

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.

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, July. Association for Computational Linguistics.

Junyu Lu, Bo Xu, Xiaokun Zhang, Hongbo Wang, Haohao Zhu, Dongyu Zhang, Liang Yang, and Hongfei Lin. 2024. Towards comprehensive detection of chinese harmful memes. In Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024.

Abdurahman Maarouf, Nicolas Pröllochs, and Stefan Feuerriegel. 2024. The virality of hate speech on social media. Proc. ACM Hum. Comput. Interact., 8(CSCW1):1–22.

J Nockleyby. 2000. 'hate speech in encyclopedia of the american constitution. Electronic Journal of Academic and Special librarianship.

Han Wang, Tan Rui Yang, Usman Naseem, and Roy Ka-Wei Lee. 2024a. Multihateclip: A multilingual benchmark dataset for hateful video detection on youtube and bilibili. In Proceedings of the 32nd ACM International Conference on Multimedia, MM 2024, Melbourne, VIC, Australia, 28 October 2024 - 1 November 2024, pages 7493–7502. ACM.

Hongbo Wang, Mingda Li, Junyu Lu, Hebin Xia, Liang Yang, Bo Xu, Ruizhu Liu, and Hongfei Lin. 2024b. Pclgpt: A large language model for patronizing and condescending language detection. In Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024, pages 6913–6928. Association for Computational Linguistics.

Hongbo Wang, Junyu Lu, Yan Han, Kai Ma, Liang Yang, and Hongfei Lin. 2025. Towards patronizing and condescending language in chinese videos: A multimodal dataset and detector. In ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 1–5.

Yunze Xiao, Yujia Hu, Kenny Tsu Wei Choo, and Roy Ka-Wei Lee. 2024. ToxiCloakCN: Evaluating robustness of offensive language detection in Chinese with cloaking perturbations. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 6012–6025, Miami, Florida, USA, November. Association for Computational Linguistics.

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

