CRetor: A Mandarin Counter-Speech Dataset Annotated for Rhetorical Strategy

Michael Bennie[1], Yajie Xu[2], Chenfeng Su[1], Yichuan Shen[1], Xiaoyu Chen[1], Rosa Weng[3], Bushi Xiao[1],

[1]University of Florida, [2]University of South Florida,
[3]Shandong Institute of Petroleum and Chemical Technology
Correspondence: michaelbennie@ufl.edu, xiaobushi@ufl.edu

Abstract

We introduce CRetor, the first Mandarin counter-speech corpus with annotations of rhetorical strategies. We extend existing taxonomies for English to reflect culturally specific nuances in Chinese online discourse by identifying six rhetorical strategies employed in anti-bias responses and analyzing their distributions across platforms. Following our guidelines, human annotators curated and annotated corpus consisting of 824 gold standard question-answer pairs. To enable large-scale analysis, we develop a three-stage annotation pipeline: embedding-based classification, GEPA-optimized LLM topic extraction, and rhetorical strategy labeling that lead to the creation of a machine annotated corpus consisting of 114,479 items. Our cross-platform analysis reveals that Chinese counter-speech tends to be more succinct and favors structural explanations or evidential rebuttals, in contrast to English corpora, where redirecting the conversation toward other positive qualities of a group is more prevalent. This indicates that culturally aware models must adapt to localized rhetorical norms. Our dataset establishes a critical foundation for future research comparing human and machine-generated counter-speech in Mandarin and beyond.

1 Introduction

Many young people are exposed to online hate-speech (Hawdon et al., 2017). Exposure to hate speech can weaken the brain's capacity to empathize and cause adverse psychological effects in adolescents (Obermaier and Schmuck, 2022; Wachs et al., 2022). Research analyzing online discourse in Western countries shows that organized counter-speech can be more effective than individual efforts in curbing hate speech, but inappropriate counter-speech may also backfire (Garland et al., 2022; Schäfer et al., 2024). Previous experiments on social media, Hangartner et al. (2021) found that empathy-based counter-speech is particularly effective in English-language contexts, but the role counter-speech plays in other languages is still up for debate.

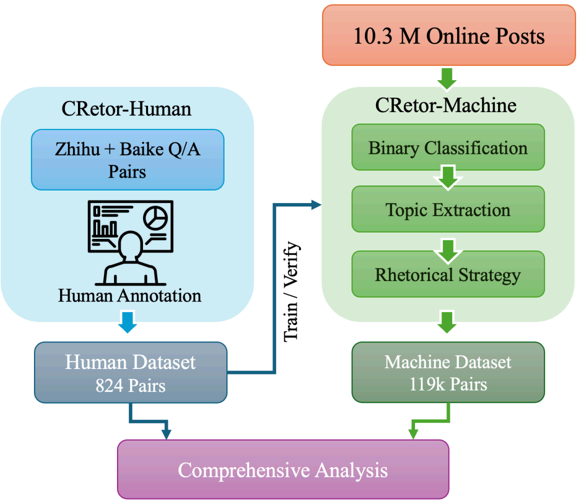


Figure 1: Generation process of CRetor dataset.

To counter hate speech online and reduce its negative emotional impact, researchers have focused on Large-Language-Models (LLMs)'s counter-speech generation ability (Deng et al., 2022; Bennie et al., 2025a; Hong et al., 2024). However, current approaches still have limitations. Firstly, LLMs suffer from inaccuracy while evaluating the hate speech target, as they tend to misinterpret the original hateful content to a notable degree (Zheng et al., 2023). Second, LLMs employ different rhetorical strategies than humans when constructing counter-speech. It has been shown that hey often fail to address the nuanced, implied stereotypes that human responders typically refute in English(Mun et al., 2024). As such, more human written and annotated data from other languages, like Mandarin, is needed before larger, cross linguistic language models can be evaluated.

While various research has focused on counter-speech rhetoric in English, other languages remain understudied. This linguistic imbalance stems from the predominance of English content on major Western platforms (Das et al., 2024). Unlike many other understudied, non-Western languages, Mandarin Chinese is spoken by nearly a billion people; yet, research on counter-speech is still limited. Existing Mandarin hate-speech corpora (Zhou et al., 2022; Deng et al., 2022; Zhao et al., 2023) have mapped what biases appear in online discourse, but lack annotations for argumentative structure or rhetorical strategies. As such, it is an open question as to what linguistic qualities constitute effective counter speech in Mandarin Chinese.

Understanding how rhetorical strategies are used in Mandarin counter-speech is critical, as prior research shows that strategy choice significantly impacts counter-speech effectiveness; appropriate approaches reduce the spread of hate speech, while other strategies show minimal effect (Hangartner et al., 2021; Bär et al., 2024). To address this gap, we introduce CRetor, the first Mandarin counter-speech corpus systematically annotated for rhetorical strategies. Figure 1 illustrates the construction process of our dataset with two subsets: CRetor-Human and CRetor-Machine. Native Mandarin speakers manually annotated 824 high-quality question-answer pairs from two major Chinese online platforms. Then we developed a three-stage machine annotation pipeline and applied it to 10.3 million online posts from different platforms: (i) binary classification to identify bias-related content, (ii) LLM-based topic extraction to label bias categories, and (iii) rhetorical strategy annotation. The CRetor-Human data serves as a golden standard for training the classifier and verifying the final result for CRetor-Machine. This yields 119,181 high-quality machine-annotated hate-speech counter-speech pairs. Together, these datasets enable comprehensive analysis of six rhetorical strategies across multiple platforms.

Our work makes three key contributions:

1. **High-quality benchmark** We extend existing taxonomies to identify rhetorical strategies, which capture Mandarin-specific argumentative patterns. Through a rigorous annotation process and uncertainty-driven active learning loop, we achieve a high inter-annotator agreement.

2. **Machine-annotated method** We develop a three-stage machine annotation pipeline and apply it to 10.3M Mandarin-language posts, which finally yields 62,916 hate-speech counter-speech pairs with high-precision annotations.

3. **Cross-linguistic insights** Our analyses reveal divergences between Mandarin and English counter-speech, providing a theoretical foundation that culturally aware counter-speech models must adapt to localized rhetorical norms.

2 Related Work

Early Chinese resources (CDial-Bias(Zhou et al., 2022) and its follow-up variants) mapped what biases appear in Zhihu exchanges, but did not include annotations for argumentative structure. Other corpora, such as CHBias (Zhao et al., 2023) and COLD (Deng et al., 2022), focus on labeling toxicity spans rather than recording reparative dialogue. As a result, we still lack data on how Mandarin speakers counter stereotypes in everyday question–and–answer (Q/A) threads. A detailed comparison of previous Chinese hate-speech/counter-speech corpora can be found in Table 1. Notably, our work fills this gap by providing a dataset that contains contextualized, multi-platform data on hate speech and counter-speech, while serving as the first Mandarin corpus to include rhetorical device annotations.

A rich literature in English suggests that rebuttals do more than merely “call out” hate: strategies that supply counter-examples, alternative qualities, or structural explanations measurably shift bystanders’ attitudes (Fanton, Margherita and Bonaldi, Helena and Tekiroğlu, Serra Sinem and Guerini, Marco, 2021). Yet those findings stem almost entirely from Western or cross-platform data

Name	Counter / Total Instances	Contextualized	Multi-platform	Bias Type	Rhetorical Devices
PANDA (Bennie et al., 2025b)	785 / 785	y	y		
COLD (Deng et al., 2022)	668 / 37,480		y		
CDial (Zhou et al., 2022)	454 / 28,343	y		y	
Political (Wang et al., 2022)	0 / 315,795				
CHSD (Zhu and Bhat, 2021)	0 / 17,430				
ToxiCN (Lu et al., 2023)	0 / 12,011		y	y	
SWSR (Jiang et al., 2022)	0 / 8,969				
Human Annotation (Ours)	414 / 824	y	y	y	y
Machine Annotations (Ours)	62,916 / 114,479	y	y	y	y

Table 1: Dataset statistics for previous and current Mandarin hate speech and counter speech corpora. Columns indicate instance counts and characteristics such as format, platform, and presence of bias or rhetorical features.

such as CONAN, Change-My-View, and Twitter crawls, where communicative norms differ starkly from Chinese fora.

From a rhetorical perspective, previous research has found that American English tend to use more heterogenetic engagement resources compared to Mandarin prose, while Mandarin speakers tended to avoid outright rejecting other viewpoints as much (Pinying, 2018). Likewise, Chinese projection clauses are markedly different from their English counterparts (Xuan and Chen, 2020). Given the different social and linguistic background of Mandarin Chinese speakers compared to the previous studies English and German speakers, it is not self-evident that Mandarin speakers will have a similar distribution of argumentative strategies when compared to previous corpora. Existing cross-cultural rhetoric studies suggest that Chinese argumentation tends to be more indirect and context-oriented, prioritizing harmony over direct confrontation (Yang and Cahill, 2008). Based off the previous research, we would infer that counter speech in Mandarin may be more euphemistic, such as using “external factors,” rather than directly challenging the other party’s claims with “general denouncement” claims. Yet, the paper will later show that Mandarin speakers are more likely to directly criticize hate-speech on online platforms.

3 Labeling

Our annotation scheme extends the context–sensitive taxonomy introduced in Zhou et al. (2022) and the rhetorical-strategy typology of Mun et al. (2023). It comprises two main layers: bias labeling, which operates at the question level, and rhetorical labeling, which operates at the response level.

Question-level Annotation: For every question, we apply two labels:

- **Bias Type**: 偏见表达 (Bias-Expression); 偏见讨论 (Bias-Discussion); and 不相关 (Irrelevant).
- **Topic**: Targeted social groups, separating gender and sexuality into distinct classes.

Response-level Annotation: Each response is labeled for its **attitude** toward the implied bias: 反偏见/正面 (Anti-Bias), 中立无偏 (Neutral), 偏见/负面 (Prejudiced), and 不相关 (Irrelevant). If the response is labeled as Anti-Bias, we additionally annotate the specific rhetorical strategies it uses to counter the bias:

1. **Alternate Groups** (替代群体) –Highlight that other groups exhibit the same behavior.
2. **Alternate Qualities** (替代特质) –Attribute contrasting traits to the target group.
3. **Counterexamples** (反例) –Cite concrete individuals or sub-groups who disprove the blanket claim.
4. **External Factors** (外部因素) –Explain the behavior by external reasons rather than group character.
5. **Expansion** (扩展) –Emphasize human commonality or intra-group diversity.
6. **General Denouncing** (普遍谴责) –Simply condemn the hate speech.

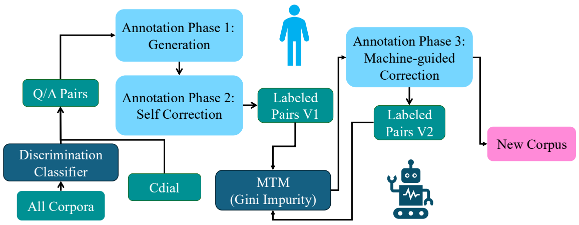

Annotation Phase 1: Generation
Q/A Pairs
Annotation Phase 2: Self Correction
Labeled Pairs V1
Annotation Phase 3: Machine-guided Correction
Labeled Pairs V2
New Corpus
Discrimination Classifier
All Corpora
Cdial
MTM (Gini Impurity)

Figure 2: Generation process of the CRetor-Human Dataset.

4 CRetor-Human

4.1 Data Sources

The data used for CRetor-Human was primarily sourced from two copora: Zhihu (Zhou et al., 2022) and Baike2018QA, a general web scrape of Q/A data, (Xu, 2019).

To filter contents about social bias, we trained a binary logistic regression classifier using 1,951 labeled questions from the CDial, COLD, PANDA, and SWSR datasets, as well as 800 hand-labeled examples from Baike. Input texts were converted into semantic embedding vectors using a Chinese GTE model (gte-base-zh) (Li et al., 2023).

The classifier achieved a mean F1 score of 0.93 and was then applied to 100,000 randomly sampled Baike entries. From this, 1,730 lines were identified as potentially containing bias.

We then use the Llama 4-Maverick model to remove irrelevant or sexually explicit questions. This model was used due to previous comparisons of the Llama-4 Maverick model (rank in lmarena.ai against other LLM models in Mandarin Chinese text generation) shows that it has similar performance to other Chinese focused models like Qwen2.5-plus-1127 and Qwen2.5-72b-instruct. Finally we combined 589 items from Baike with 235 items from Zhihu, yielding a dataset of 824 question–answer pairs.

4.2 Human Annotation

All eleven annotators are native Mandarin Chinese speakers who. Each annotator completed a one-hour, synchronous training session before the annotation process. In addition to the rhetorical devices, topics, and stances, our dataset also contains the scores of how well each response was from 1 to 10. 100% of CRetor-Human received annotations from at least two independent annotators.

To resolve the issue of annotator disagreements, we utilize the multi-task model (MTM) from Davani et al. (2022) to generate a probability distribution for each label and simulate a majority vote for each item. We use the Gini impurity function to calculate uncertainty in annotations. As shown in figure Figure 2, the uncertainty scores were then used in an iterative process of reviewing the top 20 Q/A pairs with the highest scores over 8 rounds of updating. The median of each annotation's scores was used as the gold standard, as more complex models performed similarly on numerical/ordinal annotator aggregation tasks (Braylan et al., 2023).

We measure the inter-annotator agreement using Krippendorff's alpha (κ). This ranges from 0.77 to 0.99 for response attitude and rhetorical strategies, comparable to other standard hate-speech datasets (Baheti et al., 2021) and exceeded previous agreement scores from other Chinese datasets who had κ ranging from 0.4 to 0.74 (Zhou et al., 2022).

4.3 Platform-specific strategies

Data from Zhihu (CDial) heavily favor Counterexamples (60.3%) and General Denouncing (43.8%), whereas Baike authors utilize a more balanced range of strategies, notably External Factors (38.7%) and General Denouncing (39.3%). A chi-square test confirms a significant difference between these two platforms in terms of rhetorical strategy distribution ($\chi^2(5) = 86.08$, $p < 0.001$). A plausible explanation for this difference is response length, as the median Baike response length (100 characters) is significantly longer than Zhihu's (26 characters), allowing for more comprehensive discussions of external contexts.

Combination	% of CS
Counterexamples	21.0
General Denouncing	15.5
Counterexamples + Alt. Qual.	7.7
Counterexamples + Gen. Denouncing	7.5
External Factors	6.9
Ext. Factors + Gen. Denouncing	6.4

Table 2: Most used strategies and combinations across both datasets (anti-prejudice responses only).

Figure 11 illustrates these distributions clearly, indicating both platforms frequently use multiple rhetorical tactics within a single response. A Mann-Whitney U test indicates this small difference is not statistically significant ($U = 17,182.5$, $p = 0.324$), suggesting similar inclinations toward combining rhetorical strategies in both Zhihu and Baike anti-prejudice discussions.

5 CRetor-Machine

While we already have gold-standard dataset CRetor-Human for Mandarin rhetorical strategy analysis, its limited size and high amount of time required to manually annotate additional data constrains broader comparative analyses.

To overcome these constraints,

a three-stage annotation pipeline is adopted to ensure high-precision identification of HS and CS:

1. An embedding-based binary classifier that filters questions into *non-biased* vs. *biased / bias-discussing* content.
2. An LLM-based topic category model that assigns the bias topics to each candidate question and maps phrases related to bias to named entities of targeted groups occurring in the text.
3. An LLM-based counter speech recognition model that predicts (a) the bias type of the original content, (b) the respondent' s attitude toward that bias, and (c) the rhetorical strategies used.

Table 3 summarizes the scale and main usage location of each resource. The Zhihu-KOL dataset is composed of question-response pairs from Zhihu, a social media platform for question answering. Importantly, it is the only corpus in the list that provides extensive response metadata, including post date and number of stars.

Datasets from three other social media platforms, Tieba, PTT, and Weibo, were also chosen for machine annotation, as they represent several different types of text-based social media. Alongside the social media corpora, the Baike2018QA dataset was also used, as it contains a general web scrape of question-answer data drawn from multiple platforms on Chinese internet.

5.1 Embedding-Based Binary Classification

We trained a binary classifier to distinguish bias-related content (BIAS-EXPRESSION or BIAS-DISCUSSION) from non-bias text (NON-BIAS). At this stage, all instances that neither express nor discuss social bias are treated as a single NON-BIAS category.

Each question was embedded using three different promptsinto a vector space using a sentence-embedding model. For models that support Matryoshka Representation Learning (Kusupati et al., 2022), each prompt's vector was truncated to 512 dimensions and concatenated. This was used to reduce the chances of overfitting to the relatively small 22k item dataset.

We evaluated seven different models with eight different binary classification heads, and measure their performance with a 20-fold cross-validation. Results shows the Qwen-8B model consistently outperforms other models across multiple classification heads (precision = 0.97 and F1 = 0.94).

Active Learning Loop To improve classification on edge cases, we applied pool-based active learning over six rounds, prioritizing uncertain instances where $|p(\text{bias} \mid x) - 0.5|$ was smallest.

Labeling Results After retraining on the full 24,839-item set, the classifier was applied to datasets from Table 3 to identify bias-related content. Table 4 shows the distribution of predicted probabilities: 7.39% of 10.2M items exceeded 0.5, and 3.36% exceeded 0.75. Most datasets showed distributions heavily skewed toward zero, except Tieba where 9.9% exceeded 0.75, possibly due to platform-specific discourse characteristics or collection strategies.

5.2 Topic and Relation Extraction

After flagged as bias-related, an LLM will assign topic categories and extract relational tuples of the form ⟨topic category, targeted phrase, evaluative phrase⟩ (Figure 3). This structured output grounds predictions in specific text spans, and reduces spurious labels (Sap et al., 2020; Mun et al., 2023).

The topic-labeling prompt was optimized us-

Corpus	Source platform	# entries	Main usage region	License	Citation
Zhihu-KOL	Zhihu (2011–2023)	1,006,218	China	MIT	Wang (2023)
PTT-Gossiping	PTT (2015–2017)	418,201	Taiwan	Apache-2.0	Yang (2019)
Tieba	Baidu Tieba forum (pre-2017)	3,059,172	China	Apache-2.0	codemayq (2018)
Weibo	Weibo microblog (pre-2017)	4,435,959	China	Apache-2.0	codemayq (2018)
Baike2018QA	General forum webscrape (pre-2018)	1,470,142	China	MIT	Xu (2019)
Total		**10,389,692**			

Table 3: Chinese corpora used for CRetor-Machine.

data source	%>0.5	%>0.75
Baike	1.09% (15,995)	0.42% (6,179)
PTT	6.58% (27,509)	2.46% (10,269)
Tieba	20.93% (640,315)	9.91% (303,044)
weibo	0.57% (25,450)	0.15% (6,649)
zhihu	5.42% (44,007)	2.01% (16,280)

Table 4: Distribution of predicted bias probabilities by data source.

ing GEPA (Agrawal et al., 2025) algorithm via the `dspy` framework. The CRetor-Human data is utilized for training (40%), validation (30%), and testing (30%). The objective balanced correct topic identification and tuple consistency, without penalizing `Other` labels due to low inter-annotator agreement ($\kappa = 0.22$). We used the state-of-the-art model Qwen3-Next-80B-A3B-Thinking for this process (Team, 2025).

Table 5 shows `Qwen3-Next-80B-A3B-Instruct` achieved the best non-thinking model performance (macro F1: 0.7001), while `Qwen3-VL-30B-A3B-Thinking` scored highest overall (0.7075). It is also 40 times faster (2.03 minutes vs. 1 hour 25 minutes on 8 NVIDIA B200 GPUs). We therefore adopt `Qwen3-Next-80B-A3B-Instruct` with the optimized prompt as our primary topic category model, and exam it's further ability on the test set.

Table 6 shows the model performs well on most categories, but the lowest F1 scores appear for *vocation/education*,*religion* and *other*. Error inspection reveals the model frequently labeled phrases like athletic or military roles as professions for the vocation/education category, while human annotators did not consistently treat these as vocational bias targets.

Although the embedding-based classifier exhibits high recall, it can over-trigger on bias-related keywords like ” 黑人” (Black people) and ” 贤惠” (virtuous or submissive), without any social bias being expressed or discussed.

Topic Labeling Example

Input (Title):
亚洲男人在欧美会很不受女人欢迎吗?

(Translation: Are Asian men very unpopular with women in Europe and America?)

Output (Extracted Tuples):

- **Groups:** [种族文化, 性别]
  (Race, Gender)
- **Relations:**
  - [种族文化, 亚洲男人, 很不受女人欢迎]
    (Race, Asian men, very unpopular)
  - [性别, 亚洲男人, 很不受女人欢迎]
    (Gender, Asian men, very unpopular)

Figure 3: An example of the relation extraction task illustrating the mapping from a question to group-quality tuples. The group “亚洲男人” (asian men) refers to a specific gender and race, so it was repeated twice in the Relations section.

In contrast, the LLM produces explicit group–relation tuples, providing structured evidence of social group references and evaluations. We access three scoring variants: (1) baseline using only embedding probability $p_{\text{embed}}(x)$; (2) tuple-scaled multiplying $p_{\text{embed}}(x)$ by weight $w(k) = 1 - \alpha \exp(-\beta k)$ based on tuple count k; (3) binary-switch clamping probability to zero when no tuples exist.

We use Qwen3 model for category tuple generation and the same Qwen3-Embedding-8B SVM classifier for generating probability scores. Cross-validation (Table 13) showed the binary-switch model achieved 0.99 precision at threshold 0.75, though with lower recall. Since the downstream goal requires high-precision

Model	Params	Prompt	Macro F1 ↑	Macro Accuracy ↑	Hamming ↓	Subset ↑
Qwen3-VL-2B-Instruction	2B	basic	0.4762	0.8278	0.1722	0.3035
		optim.	0.4180	0.8997	0.1003	0.3187
Qwen3-VL-2B-Thinking*	2B	basic	0.5495	0.9298	0.0702	0.5171
		optim.	0.3855	0.8997	0.1003	0.3187
Qwen3-VL-30B-Instruction	30B	basic	0.5280	0.9366	0.0634	0.5535
		optim.	0.6501	0.9330	0.0670	0.6345
Qwen3-VL-30B-Thinking*	30B	basic	0.6549	0.9466	0.0534	0.6151
		optim.	0.7075	0.9553	0.0447	0.6715
Kimi-Linear-48B-Instruction	48B	basic	0.5197	0.9264	0.0736	0.5556
		optim.	0.5460	0.9284	0.0716	0.5782
Qwen3-Next-80B-Instruction	80B	basic	0.5871	0.9482	0.0518	0.5958
		optim.	**0.7001**	**0.9539**	**0.0461**	**0.6765**
Qwen3-Next-80B-Thinking*	80B	basic	0.6748	0.9413	0.0587	0.6164
		optim.	0.6833	0.9454	0.0546	0.6490
Ring-flash-2.0*	100B	basic	0.6323	0.9486	0.0514	0.6089
		optim.	0.6970	0.9488	0.0512	0.6282
Ling-flash-2.0*	100B	basic	0.4097	0.8603	0.1397	0.3280
		optim.	0.4762	0.9204	0.0796	0.4202
GLM-4.6-FP8†	358B	basic	0.7009	0.9500	0.0500	0.6405
		optim.	0.6895	0.9527	0.0473	0.6422

Table 5: Topic-label prediction performance for different LLMs with basic vs. GEPA-optimised prompts. Thinking models are marked with $\star$, and the hybrid model is labeled with a †. **Macro F1** (↑) and **Macro Accuracy** (↑) are macro-averaged over labels. For each metric column, **bold** numbers indicate the best-performing non-thinking model, and underlined numbers indicate the best-performing thinking model.

Topic	F1	Precision	Recall
Chinese Minorities	0.9891	1.0000	0.9783
Disability	0.9286	1.0000	0.8667
Sexual Orientation	0.8333	1.0000	0.7143
Gender	0.8000	1.0000	0.6667
Race	0.7826	0.7750	0.7905
Region	0.7181	0.9375	0.5819
Vocation/Education	0.3676	0.6667	0.2404
Religion	0.4000	1.0000	0.2500
Other	0.1875	0.3333	0.1304

Table 6: Per-topic F1, precision, and recall for `Qwen3-Next-80B-A3B-Instruct` (optimized prompt), sorted by F1.

candidates, we only label items with probability > 0.75 and containing non-"other" tuples.

5.3 Rhetorical Strategies

The final step assigns counter-speech attributes to each post-response pair. Given bias-labeled posts and extracted tuples, an LLM determines the post type (expressing or discussing bias), response stance, and rhetorical strategies. Each strategy is encoded as a 4-tuple that incudes the strategy name, targeted group, biased phrase from the original post, and the counter speech phrase found in the response. This was done to entourage strategies to be anchored to specific bias content and can be later used for co-occurrence analysis of key words.

The prompt was tuned using the GEPA algorithm using `Qwen3-Next-80B-A3B-Thinking`. After prompt optimization, a series of non-thinking models were tested on the optimized and and non-optimized prompts and the results can be found in Table 7 shows non-thinking models performed well on bias-expressing posts but struggled with bias-discussing posts and comprehensive rhetorical labeling. Overall `Qwen3-Next-80B-A3B-Instruct` achieved best and was used to label all the items in the dataset that were previously marked as possibly containing bias during embedding-based binary classification and topic extraction.

5.4 Results

Table 8 summarizes key corpus-level statistics. Machine generation pipeline produced 118,605 labeled pairs; 114,479 were predicted to be bias-related posts (Bias-Expression or Bias-Discussion) and form the analysis pool. Within this pool, 62,916 responses were classified as Counter-speech. Platform contributions were highly imbalanced: Tieba accounts for 98,023 bias-related pairs, while Baike and Zhihu contribute smaller but substantially longer-form Q/A exchanges (median CS response lengths: 99 and 232 characters,

Model	Params	Prompt	Post macro F1 ↑	Bias expression F1 ↑	Resp. Macro F1 ↑	Resp. CS F1 ↑	Rhetorical Devices Macro F1 ↑
Qwen3-VL-2B-Instruct	2B	basic	0.41	0.58	0.21	0.14	0.22
		optim.	0.40	0.76	0.30	0.55	0.29
Qwen3-VL-30B-A3B-Instruct	30B	basic	0.45	0.57	0.35	0.25	0.27
		optim.	0.46	0.75	0.35	0.56	0.29
Kimi-Linear-48B-A3B-Instruct	48B	basic	0.42	0.55	**0.38**	0.20	0.29
		optim.	0.45	0.73	0.35	0.52	0.30
Qwen3-Next-80B-A3B-Instruct	80B	basic	0.50	0.59	0.33	0.26	0.31
		optim.	0.45	**0.79**	0.36	**0.66**	**0.37**
Ling-flash-2.0	100B	basic	0.50	0.57	0.35	0.42	0.33
		optim.	**0.51**	0.71	0.32	0.57	0.35

Table 7: Performance by model and prompt (basic vs. optimized). **Post macro F1** is the macro-averaged F1 over post-level bias-tuple labels. **Bias expression F1** is the F1 for the post-level bias-expression label. **Response Macro F1** is the macro-averaged F1 over response-level bias-tuple labels. **Response CS F1** is the F1 for the response counter-speech label. **Rhetorical Devices Macro F1** is the macro-averaged F1 over eligible rhetorical-device labels.

Corpus	#Pairs	%CS	Med. CS len	Mean #strat.
Baike	6,086	61.6	99	1.74
Zhihu	5,792	74.2	232	1.99
PTT	1,367	56.9	35	1.05
Weibo	3,211	36.6	24	1.02
Tieba	98,023	54.0	24	1.05
Combined	**114,479**	54.9	–	1.15

Table 8: Composition of the CRetor-Machine analysis pool (bias-related posts only). **#Pairs** denotes the number of post–response pairs retained after bias filtering. **%CS** is the proportion of responses predicted as counter-speech. **Med. CS len** is the median character length of counter-speech responses. **Mean #strat.** is the mean number of unique rhetorical strategies per counter-speech response (Stage 3).

respectively), compared to microblog/forum corpora like Weibo and Tieba (median CS response lengths: 24 and 24 characters).

Bias type and response stance varied by platform. Platform was strongly associated with whether a post was predicted as Bias-Expression versus Bias-Discussion ($\chi^2(4) = 14513.96$, $p < 0.01$, Cramér's $V = 0.36$). As shown in Figure 5, Zhihu contained a substantially higher share of Bias-Discussion posts (52.2%) than Tieba (13.4%), which may be due to the more long form, Q/A style of Zhihu encouraging meta-level discussion (e.g., "why do people say X?") rather than direct stereotyping.

Targeted groups varied by Mandarin social media platforms. Before comparing topic distributions, we normalized raw topic strings generated by the LLM into a fixed set of canonical topics . After normalization, topic distributions differed strongly by platform ($\chi^2(28) = 14385.60$, $p < 10^{-300}$, Cramér's $V = 0.18$). Figure 9 shows clear platform signatures: Weibo and Baike were dominated by gender-related tuples, while Tieba contained substantially more region- and vocation/education-related tuples, and PTT showed relatively higher representation of region and sexuality topics.

Pairwise co-occurrence statistics showed that the most common tuple pairs involve *region*, *gender*, and *vocation/education*. . The high co-occurrence of gender and regional bias across all categories suggests that there is high *intersectionality* between gender/regional bias and other sensitive categories across online forums. For example, in the sentence "为什么上海本地人，尤其是老太太特别有优越感?" (Why do native Shanghainese people, especially older women, have such a sense of superiority over other people?), the first clause's target (Shanghainese people) was later intersected with the group "older women." As such, a model that would generate effective Counter-Speech (CS) to counter these biases would need to address both components simultaneously.

Rhetorical strategy use was partially platform-specific and correlated with response length. We next analyzed rhetorical strategies among responses classified as Counter-speech. Because a single response may employ multiple strategies, we compared corpora using strategy-instance counts. Strategy distributions differed substantially across platforms ($\chi^2(20) = 4569.49$, $p < 10^{-300}$, Cramér's $V = 0.14$). Figure 7 highlights two consistent patterns. First, General Denouncing appeared broadly across all corpora (about 61% of CS responses overall), suggesting that brief normative condemnation is a cross-platform default. Second, Evidence- or explanation-heavy strategies were much more platform-dependent. For example, COUNTEREXAMPLES and EXTERNAL FACTORS were most prevalent on long-form Q/A platforms (Zhihu: 35.4% and 12.1%; Baike: 10.0% and 11.6%), but were far rarer on Tieba and Weibo (e.g., COUNTEREXAMPLES 6.5% and 5.5%).

6 Conclusion

This study presents the first systematic study of rhetorical strategies in Mandarin counter-speech, combining a human-annotated corpus with a machine-annotated hate speech-counter speech pairs from five major Chinese platforms. We extend existing English taxonomies to identify six rhetorical strategies and analyzing their distributions across platforms.

Our findings reveal a markedly different rhetorical landscape for Mandarin counter-speech than the one for English in Mun et al. (2023). While prior English-language research found that counter-speech overwhelmingly favored ALTERNATE QUALITIES and treated EXTERNAL FACTORS as rare and less persuasive, our Mandarin data presents a more nuanced platform-dependent pattern.

We also observe that Chinese counterspeakers typically keep their tactical repertoire small. This differs from finding from Mun et al. (2023) that persuasive English comments cluster around two to three tactics. This remains true even when responses had relatively high ratings of a rating of 8 or higher or higher (*mean* = 1.66). This may suggest that Chinese counter-speech argumentative strategies may naturally be more succinct.

Limitations

Moreover, although 71% of items were encoded by 3 or more annotators, the remaining double-coded (28%) and could still be strongly biased by a single annotator's latent biases. The use of a aggregation model that models each speaker's labeling preferences was used to simulate having all eleven annotators and reduce individual bias as described in other papers (Davani et al., 2022). that While we trained annotators carefully, personal biases or misunderstandings could still influence the labels, especially for tricky categories like deciding whether a question expresses or just discusses bias. Additionally, our annotators were mostly young, educated, urban Mandarin speakers, which may limit how well the dataset represents the views of older, rural, or more linguistically diverse communities.

While we aim to enhance the diversity of the current HS/CS corpus, we recognize several potential risks. First, it can be exploited to train models for generating more sophisticated toxic dialogues (Zhou et al., 2022). In addition, stereotypes may be reinforced if counter-speech is not thoughtfully crafted (Jia and Schumann, 2025). Furthermore, the dataset might reflect specific cultural and linguistic biases, given the fact that the dialogues are extracted only from online platforms, potentially limiting the diversity and fairness of Chinese-speaking communities. To mitigate these concerns, we suggest mitigation strategies, such as including a broader range of representatives and conducting ethical reviews involving community stakeholders (Eder, 2023).

Another important point is that the labels in our dataset show how people responded to biased questions, but they do not tell us whether those responses were actually effective in changing opinions or reducing harm. Just because a comment uses a certain rhetorical strategy does not mean it works better. Future work should connect these strategies to actual outcomes, like how people respond or whether harmful behavior decreases.

When selecting answers, machine-assisted filtering may have favored stereotypical content that fits existing Western taxonomies as the majority of the data used to train LLaMA 4 was English. As such, the exact distribution of targeted groups may not actually be that representative. Finally, training the MTM took about four hours on an A100 GPU. This contributed some energy use and emissions, though relatively small compared to larger AI models. Training and annotating the models from Experiment 2 required 8 B100 GPUs and took 80 hours in total.

References

Lakshya A Agrawal, Shangyin Tan, Dilara Soylu, Noah Ziems, Rishi Khare, Krista Opsahl-Ong, Arnav Singhvi, Herumb Shandilya, Michael J Ryan, Meng Jiang, and 1 others. 2025. Gepa: Reflective prompt evolution can outperform reinforcement learning. *arXiv preprint arXiv:2507.19457*.

Ashutosh Baheti, Maarten Sap, Alan Ritter, and Mark Riedl. 2021. Just say no: Analyzing the stance of neural dialogue generation in offensive contexts. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4846–4862, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Michael Bennie, Bushi Xiao, Chryseis Xinyi Liu, Demi Zhang, Jian Meng, and Alayo Tripp. 2025a. Codeofconduct at multilingual counterspeech generation: A context-aware model for robust counterspeech generation in low-resource languages. In *Proceedings of the First Workshop on Multilingual Counterspeech Generation*.

Michael Bennie, Demi Zhang, Bushi Xiao, Jing Cao, Chryseis Xinyi Liu, Jian Meng, and Alayo Tripp. 2025b. Panda - paired anti-hate narratives dataset from asia: Using an llm-as-a-judge to. In *Proceedings of the First Workshop on Multilingual Counterspeech Generation*.

Alexander Braylan, Madalyn Marabella, Omar Alonso, and Matthew Lease. 2023. A general model for aggregating annotations across simple, complex, and multi-object annotation tasks. *Journal of Artificial Intelligence Research*, 78:901–973.

Dominik Bär, Abdurahman Maarouf, and Stefan Feuerriegel. 2024. Generative ai may backfire for counterspeech. *Preprint*, arXiv:2411.14986.

codemayq. 2018. Chinese chatbot corpus. https://github.com/codemayq/chinese_chatbot_corpus.

Mithun Das, Saurabh Pandey, Shivansh Sethi, Punyajoy Saha, and Animesh Mukherjee. 2024. Low-resource counterspeech generation for Indic languages: The case of Bengali and Hindi. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1601–1614, St. Julian's, Malta. Association for Computational Linguistics.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

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.

Milton Eder. 2023. Aligning clinical research ethics with community-engaged and participatory research in the united states. *Frontiers in Public Health*, 11:1122479.

Fanton, Margherita and Bonaldi, Helena and Tekiroğlu, Serra Sinem and Guerini, Marco. 2021. Human-in-the-Loop for Data Collection: a Multi-Target Counter Narrative Dataset to Fight Online Hate Speech. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Joshua Garland, Keyan Ghazi-Zahedi, Jean-Gabriel Young, Laurent Hébert-Dufresne, and Mirta Galesic. 2022. Impact and dynamics of hate and counter speech online. *EPJ data science*, 11(1):3.

Dominik Hangartner, Gloria Gennaro, Sary Alasiri, Nicholas Bahrich, Alexandra Bornhoft, Joseph Boucher, Buket Buse Demirci, Laurenz Derksen, Aldo Hall, Matthias Jochum, and 1 others. 2021. Empathy-based counterspeech can reduce racist hate speech in a social media field experiment. *Proceedings of the National Academy of Sciences*, 118(50):e2116310118.

James Hawdon, Atte Oksanen, and Pekka Räsänen. 2017. Exposure to online hate in four nations: A cross-national consideration. *Deviant behavior*, 38(3):254–266.

Lingzi Hong, Pengcheng Luo, Eduardo Blanco, and Xiaoying Song. 2024. Outcome-constrained large language models for countering hate speech. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4523–4536, Miami, Florida, USA. Association for Computational Linguistics.

Yue Jia and Sandy Schumann. 2025. Tackling hate speech online: The effect of counter-speech on subsequent bystander behavioral intentions. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, 19(1).

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.

Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, and 1 others. 2022. Matryoshka representation learning. *Advances in Neural Information Processing Systems*, 35:30233–30249.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.

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.

Jimin Mun, Emily Allaway, Akhila Yerukola, Laura Vianna, Sarah-Jane Leslie, and Maarten Sap. 2023. Beyond denouncing hate: Strategies for countering implied biases and stereotypes in language. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9759–9777, Singapore. Association for Computational Linguistics.

Jimin Mun, Cathy Buerger, Jenny T Liang, Joshua Garland, and Maarten Sap. 2024. Counterspeakers' perspectives: Unveiling barriers and ai needs in the fight against online hate. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Magdalena Obermaier and Desirée Schmuck. 2022. Youths as targets: factors of online hate speech victimization among adolescents and young adults. *Journal of Computer-Mediated Communication*, 27(4):zmac012.

Chen Pinying. 2018. A comparative study on engagement resources in american and chinese csr reports. *English Language Teaching*, 11(11):122–135.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Svenja Schäfer, Isabella Rebasso, Ming Manuel Boyer, and Anna Maria Planitzer. 2024. Can we counteract hate? effects of online hate speech and counter speech on the perception of social groups. *Communication Research*, 51(5):553–579.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Sebastian Wachs, Manuel Gámez-Guadix, and Michelle F Wright. 2022. Online hate speech victimization and depressive symptoms among adolescents: The protective role of resilience. *Cyberpsychology, behavior and social networking*, 25(7):416—423.

Chih-Chien Wang, Min-Yuh Day, and Chun-Lian Wu. 2022. Political hate speech detection and lexicon building: A study in taiwan. *IEEE Access*, 10:44337–44346.

Rui Wang. 2023. Data Scraping Project for Zhihu Dataset. `https://github.com/wangrui6/Zhihu-KOL`. Accessed: December 7, 2025.

Bright Xu. 2019. Nlp chinese corpus: Large scale chinese corpus for nlp.

Winfred Wenhui Xuan and Shukun Chen. 2020. Taking stock of accumulated knowledge in projection studies from systemic functional linguistics: a research synthesis. *Functional Linguistics*, 7:1–19.

Kai-Chou Yang. 2019. Ptt-gossiping-corpus.

Lin Yang and David Cahill. 2008. The rhetorical organization of chinese and american students' expository essays: A contrastive rhetoric study. *International Journal of English Studies (IJES)*, 8.

Jiaxu Zhao, Meng Fang, Zijing Shi, Yitong Li, Ling Chen, and Mykola Pechenizkiy. 2023. CHBias: Bias evaluation and mitigation of Chinese conversational language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13538–13556, Toronto, Canada. Association for Computational Linguistics.

Yi Zheng, Björn Ross, and Walid Magdy. 2023. What makes good counterspeech? a comparison of generation approaches and evaluation metrics. In *Proceedings of the 1st Workshop on CounterSpeech for Online Abuse (CS4OA)*, pages 62–71, Prague, Czechia. 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, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Wanzheng Zhu and Suma Bhat. 2021. Generate, prune, select: A pipeline for counterspeech generation against online hate speech. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 134–149, Online. Association for Computational Linguistics.

A Appendix

B Semantic Classifiers

Age range	Annotators (n)	Share (%
15–19	3	27
20–24	4	36
25–29	4	36
Current residence		
Mainland China	4	36
United States	6	54
United Kingdom	1	10

Table 9: Aggregate demographic profile of the annotation team.

Rating	Avg. Strategies
1	1.000
2	1.159
3	1.214
4	1.318
5	1.533
6	1.378
7	1.513
8	1.667
9	2.000
10	1.632

Table 10: Average number of argumentative strategies per rating across both datasets for CS responses.

C Embedding Results

General Categories	Accuracy (%)	κ
Question Type	52.9	0.75
Response Attitude	78.3	0.68
Topics		
Region	89.4	0.91
Chinese Minorities	98.6	0.93
Race	94.5	0.92
Gender	77.3	0.81
Sexual Orientation	96.3	0.88
Disability	98.4	0.88
Vocation/Education	96.5	0.89
Religion	99.7	1.00
Other	83.3	0.22
Response Strategies		
Alternative Group	92.8	0.71
Alternative Trait	82.3	0.71
Counterexample	76.8	0.78
External Factors	86.2	0.67
Expansion	89.4	0.64
General Denouncing	79.3	0.72

Table 11: Inter-annotator agreement: Accuracy and Krippendorff's alpha (κ) for each item type after 4 rounds of uncertainty corrections.

C.1 Prompt Engineering References

The following lists the three prompts used to generate the combined embedding vector. In all prompts, the placeholder `{text}` is replaced with the input sentence to be embedded. In the few-shot prompt, the labels `True` and `False` indicate whether the example text involves social bias. Specifically, `True` denotes content that either expresses or discusses bias, while `False` represents content not expressing bias to a social group.

1. **Raw Input:**
   `{text}`

2. **Direct Interrogation:**
   "Does the input '`{text}`' talk about race, ethnicity, gender, sex, occupations, or sexuality and hint at discrimination?"

3. **Few-Shot Examples (Chinese):**
   "下面是一些示例：'那你有哪里来的样本说明贵州人素质低了？' (True)，'你喜欢美国人歧视亚裔的文化吗？你喜欢中国人歧视越南人的文化吗？你不喜欢，仅仅是受益人不是你罢了，人家没违法没犯罪，

Embedding Model	AdaBoost	Dec. Tree	KNN (k=10)	LogReg	Rand. Forest	SVM (C=1)	SVM (Lin.)	SVM (Sig.)
Qwen3-0.6B	0.91	0.87	0.90	0.94	0.92	0.95	0.95	0.94
GTE-Base-ZH	0.91	0.87	0.90	0.94	0.92	0.95	0.95	0.94
Xiaobu-v2	0.91	0.87	0.90	0.94	0.92	0.95	0.95	0.94
E5-Large-Instruct	0.91	0.87	0.90	0.94	0.92	0.95	0.95	0.94
RITrieve-v1	0.91	0.87	0.90	0.94	0.92	0.95	0.95	0.94
Qwen3-4B	0.91	0.87	0.90	0.94	0.92	0.95	0.95	0.94
Qwen3-8B	**0.96**	**0.96**	**0.97**	**0.97**	**0.97**	**0.97**	**0.97**	**0.97**

Table 12: Semantic classification precision of different embedding models, reported to two decimal places.

Model	Threshold	Accuracy	Precision	Recall	F1
Baseline (embed)	0.50	**0.952**	0.949	**0.857**	**0.901**
Baseline (embed)	0.75	0.923	0.984	0.705	0.822
Tuple-scaled	0.50	**0.952**	0.949	**0.857**	**0.901**
Tuple-scaled	0.75	0.923	0.984	0.705	0.822
Binary-switch	0.50	0.918	0.976	0.692	0.810
Binary-switch	0.75	0.897	**0.991**	0.594	0.743
AUC (all Baseline / Tuple-scaled)			**0.988**		
AUC (Binary-switch)			0.867		

Table 13: Ablation of embedding-only vs. embedding + LLM tuple models for predicting bias. The tuple-scaled model's optimal parameters are $\alpha = \beta = 0$, making it equivalent to the baseline.

> 你为什么讨厌？’ (True)，‘hub 么？我在国外都不用翻墙 ’ (False)，‘你搁发廊作市场调研是吧，剩下的一个是不是扫黄的警察？说实话是不是集体蹲墙角刚被放出来？’ (False)，‘作者你有父母吗？来把匿名打开，你在那装你妈’ (False)。请问以上示例是否涉及敏感话题或含有歧视性内容？`{text}`(”

C.2 Binary labeling Distributions

Table 14: Binary classification performance using different embedding models, measured by F1 score to two decimal places.

Model	**AdaBoost**	**Dec. Tree**	**KNN (k=10)**	**LogReg**	**Rand. Forest**	**SVM (C=1)**	**SVM (Lin)**	**SVM (Sig)**
Qwen-0.6B	0.92	0.87	0.86	**0.94**	**0.91**	**0.95**	**0.95**	**0.94**
GTE-Base-ZH	0.92	0.87	0.86	**0.94**	**0.91**	**0.95**	**0.95**	**0.94**
Xiaobu-v2	0.92	0.87	0.86	**0.94**	**0.91**	**0.95**	**0.95**	**0.94**
E5-Large-Instruct	0.92	0.87	0.86	**0.94**	**0.91**	**0.95**	**0.95**	**0.94**
RITrieve-v1	0.92	0.87	0.86	**0.94**	**0.91**	**0.95**	**0.95**	**0.94**
Qwen-4B	0.92	0.87	0.86	**0.94**	**0.91**	**0.95**	**0.95**	**0.94**
Qwen-8B	**0.93**	**0.89**	**0.90**	0.93	**0.91**	0.94	0.94	**0.94**

Table 15: Semantic Classification Performance using different embedding models, measured by accuracy to two decimal places.

Embedding Model	**AdaBoost**	**Dec. Tree**	**KNN (k=10)**	**LogReg**	**Rand. Forest**	**SVM (C=1)**	**SVM (Lin)**	**SVM (Sig)**
Qwen-0.6B	**0.92**	**0.87**	0.88	**0.95**	**0.92**	**0.95**	**0.95**	0.94
GTE-Base-ZH	**0.92**	**0.87**	0.88	**0.95**	**0.92**	**0.95**	**0.95**	**0.95**
Xiaobu-v2	**0.92**	**0.87**	0.88	**0.95**	**0.92**	**0.95**	**0.95**	**0.95**
E5-Large-Instruct	**0.92**	**0.87**	0.88	**0.95**	**0.92**	**0.95**	**0.95**	**0.95**
RITrieve-v1	**0.92**	**0.87**	0.88	**0.95**	**0.92**	**0.95**	**0.95**	**0.95**
Qwen-4B	**0.92**	**0.87**	0.88	**0.95**	**0.92**	**0.95**	**0.95**	**0.95**
Qwen-8B	0.91	0.86	**0.90**	0.93	0.91	0.93	0.93	0.93

Counterspeech Labeling Example

Input (Post):
为什么有人说“女生不适合学理工”?
(Why do some people say “women are not suited for STEM”?)

Input (Response):
这种说法不对。我认识的女生做算法很强，而且问题不在性别，是教育资源和机会分配不均。
(This claim is wrong. I know women who are great at algorithms, and the issue isn't gender; it's unequal access to educational resources and opportunities.)

Stage 2 evidence (Extracted tuple):
[性别, 女生, 不适合学理工] [*gender, women, not suited for stem*]

Stage 3 outputs:

- **Post type:** 偏见讨论
  (Bias-Discussion)
- **Response stance:** 反偏见/正面
  (Anti-Bias)
- **Rhetorical strategy tuples:**
  - [普遍谴责, 性别, 不适合学理工, 说法不对]
    [*General Denouncement, gender, not suited for STEM, that's incorrect to say*]
  - [替代特質, 性别, 不适合学理工, 女生做算法很强]
    [*Alternate Qualities, gender, not suited for STEM, women are very good at algorithms*]
  - [外部因素, 性别, 不适合学理工, 教育资源和机会分配不均]
    [*External Factors, gender, not suited for STEM, that's due to differences in educational opportunities*]

Figure 4: Stage 3 assigns response stance and rhetorical strategies. Strategy labels are encoded as tuples grounded in the Stage 2 group–evaluation evidence, enabling scalable aggregation and co-occurrence analysis across topics and platforms.

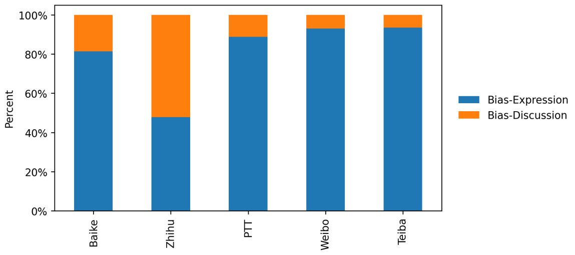


Figure 5: Model-predicted post type distribution by corpus (Bias-Expression vs. Bias-Discussion).

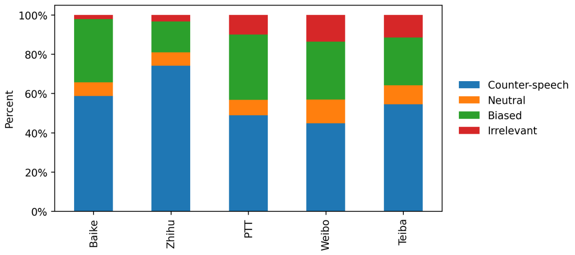


Figure 6: Model-predicted response stance distribution by corpus.

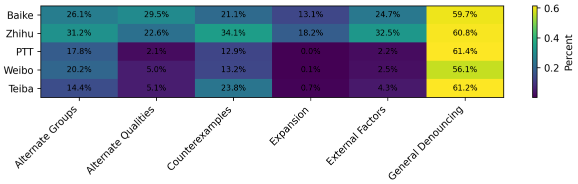


Figure 7: Rhetorical strategy prevalence by corpus among Counter-speech responses. Each cell shows the percentage of CS responses in that corpus that contain the strategy.

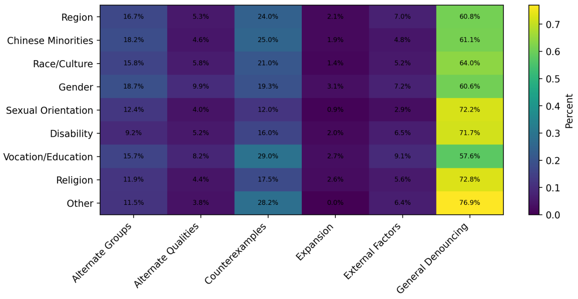


Figure 8: Topic–strategy heatmap. Values show the conditional distribution of strategies within each topic (strategy-instance proportions).

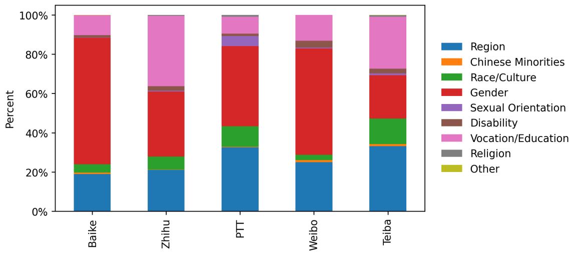


Figure 9: Distribution of normalized topic tuples by corpus (counts aggregated over Stage 2 group–evaluation tuples).

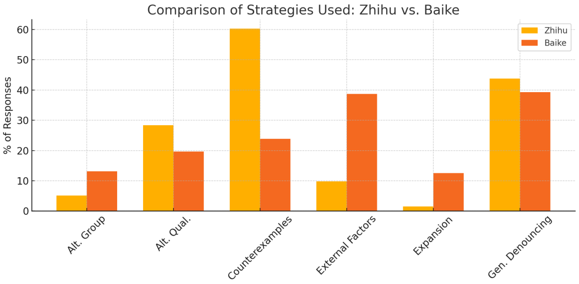


Figure 10: Comparison of strategies used between the Q/A platforms Zhihu and Baike. Stars above a category indicate a statistically significant difference between the two platforms ($p < 0.05$).

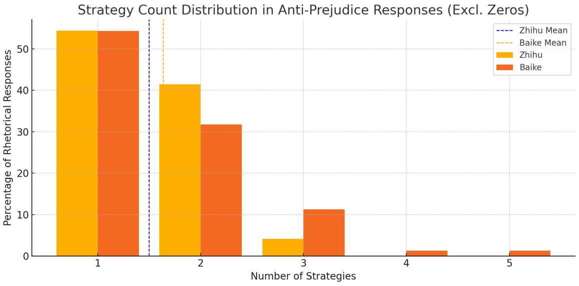
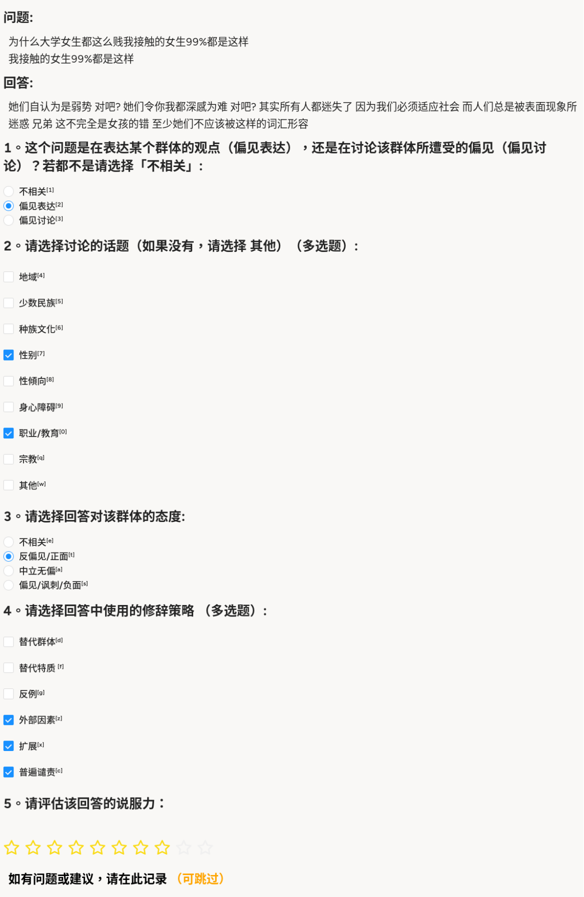


Figure 11: The relative frequencies of rhetorical strategy counts per item from all annotators. Each bar reflects the percentage of rhetorical items in each dataset used that number of strategies. The dashed vertical lines mark the mean number of strategies per item for each dataset.



问题:

为什么大学女生都这么贱我接触的女生99%都是这样
我接触的女生99%都是这样

回答:

她们自认为是弱势 对吧? 她们令你我都深感为难 对吧? 其实所有人都迷失了 因为我们必须适应社会 而人们总是被表面现象所迷惑 兄弟 这不完全是女孩的错 至少她们不应该被这样的词汇形容

1。这个问题是在表达某个群体的观点（偏见表达），还是在讨论该群体所遭受的偏见（偏见讨论）？若都不是请选择「不相关」：

○ 不相关[1]
● 偏见表达[2]
○ 偏见讨论[3]

2。请选择讨论的话题（如果没有，请选择 其他）（多选题）：

□ 地域[4]
□ 少数民族[5]
□ 种族文化[6]
☑ 性别[7]
□ 性倾向[8]
□ 身心障碍[9]
☑ 职业/教育[0]
□ 宗教[q]
□ 其他[w]

3。请选择回答对该群体的态度:

○ 不相关[e]
● 反偏见/正面[t]
○ 中立无偏[a]
○ 偏见/讽刺/负面[s]

4。请选择回答中使用的修辞策略（多选题）：

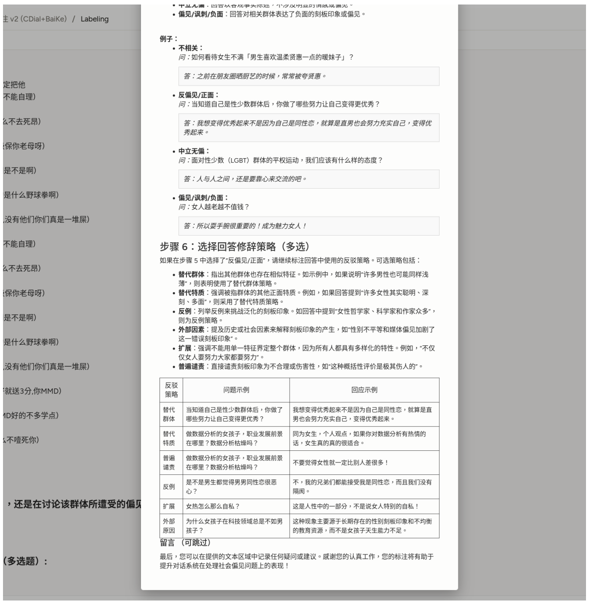
□ 替代群体[d]
□ 替代特质 [f]
□ 反例[g]
☑ 外部因素[z]
☑ 扩展[x]
☑ 普遍谴责[c]

5。请评估该回答的说服力：

★★★★★★★★☆☆

如有问题或建议，请在此记录（可跳过）

Figure 12: Screenshot of the browser-based interface shown to annotators. The prompt contains a biased question from Baike asking why college girls are "so cheap," which expresses a strong negative stereotype. The response attempts to mediate by attributing confusion to broader societal factors and discouraging derogatory language. Annotators classified the question as " 偏见表达" (bias expression), with the topics of " 性别" (gender) and " 职业/教育" (education). The response was labeled as having a " 反偏见/正面" (anti-bias/positive) attitude, using rhetorical strategies such as " 外部因素" (external factors), " 扩展" (generalization), and " 普遍谴责" (general denouncing).



- **中立无偏**：回答以客观事实陈述，不涉及明显的情感或偏见。
- **偏见/讽刺/负面**：回答对相关群体表达了负面的刻板印象或偏见。

例子：

- **不相关：**
  *问：*如何看待女生不满「男生喜欢温柔贤惠一点的暖妹子」？

  答：之前在朋友圈晒厨艺的时候，常常被夸贤惠。

- **反偏见/正面：**
  *问：*当知道自己是性少数群体后，你做了哪些努力让自己变得更优秀？

  答：我想变得优秀起来不是因为自己是同性恋，就算是直男也会努力充实自己，变得优秀起来。

- **中立无偏：**
  *问：*面对性少数（LGBT）群体的平权运动，我们应该有什么样的态度？

  答：人与人之间，还是要靠心来交流的吧。

- **偏见/讽刺/负面：**
  *问：*女人越老越不值钱？

  答：所以要手腕很重要的！成为魅力女人！

步骤 6：选择回答修辞策略（多选）

如果在步骤 5 中选择了"反偏见/正面"，请继续标注回答中使用的反驳策略。可选策略包括：

- **替代群体**：指出其他群体也存在相似特征。如示例中，如果说明"许多男性也可能同样浅薄"，则表明使用了替代群体策略。
- **替代特质**：强调被指群体的其他正面特质。例如，如果回答提到"许多女性其实聪明、深刻、多面"，则采用了替代特质策略。
- **反例**：列举反例来挑战泛化的刻板印象。如回答中提到"女性哲学家、科学家和作家众多"，则为反例策略。
- **外部因素**：提及历史或社会因素来解释刻板印象的产生，如"性别不平等和媒体偏见加剧了这一错误刻板印象"。
- **扩展**：强调不能用单一特征界定整个群体，因为所有人都具有多样化的特性。例如，"不仅仅女人要努力大家都要努力"。
- **普遍谴责**：直接谴责刻板印象为不合理或伤害性，如"这种概括性评价是极其伤人的"。

反驳策略	问题示例	回应示例
替代群体	当知道自己是性少数群体后，你做了哪些努力让自己变得更优秀？	我想变得优秀起来不是因为自己是同性恋，就算是直男也会努力充实自己，变得优秀起来。
替代特质	做数据分析的女孩子，职业发展前景在哪里？数据分析枯燥吗？	同为女生，个人观点，如果你对数据分析有热情的话，女生真的真的很适合。
普遍谴责	做数据分析的女孩子，职业发展前景在哪里？数据分析枯燥吗？	不要觉得女性就一定比别人差很多！
反例	是不是男生都觉得男男同性恋很恶心？	不，我的兄弟们都能接受我是同性恋，而且我们没有隔阂。
扩展	女热怎么那么自私？	这是人性中的一部分，不是说女人特别的自私！
外部原因	为什么女孩子在科技领域总是不如男孩子？	这种现象主要源于长期存在的性别刻板印象和不均衡的教育资源，而不是女孩子天生能力不足。

留言 （可跳过）

最后，您可以在提供的文本区域中记录任何疑问或建议。感谢您的认真工作，您的标注将有助于提升对话系统在处理社会偏见问题上的表现！

Figure 13: A screenshot of the annotation interface with a pop-up of the instructions.

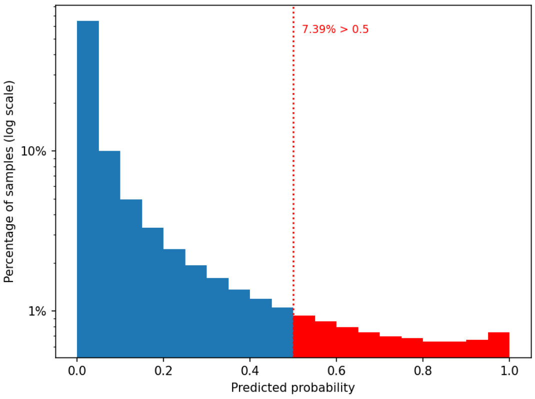


Figure 14: combined distribution of predicted bias probabilities across all data sources. The combined curve smooths over the heterogeneity of individual sources, but the influence of Tieba's high-probability peak remains visible as a long right tail.

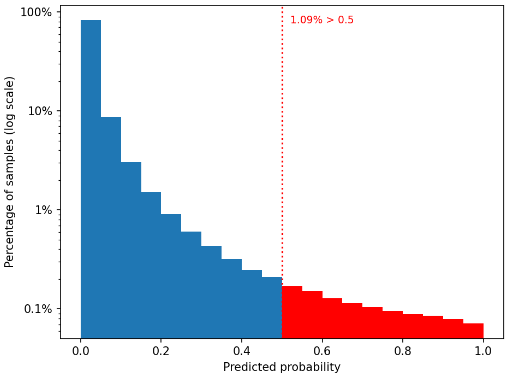


Figure 15: Distribution of predicted bias probabilities for the Baike data source. The distribution is heavily concentrated near zero, with only a very small proportion of items exceeding the 0.5 threshold (1.09%; 15,995 items). This indicates that Baike contains comparatively little bias-related content relative to other data sources.

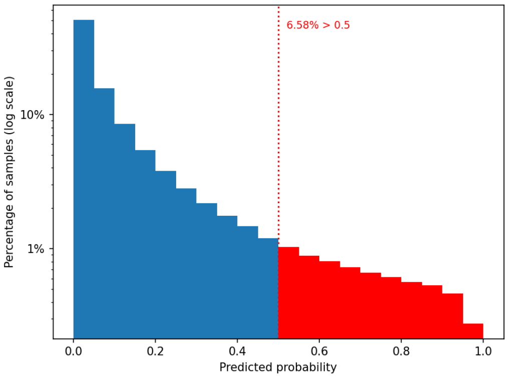


Figure 16: Distribution of predicted bias probabilities for the PTT data source. The distribution remains skewed toward low probabilities but shows a noticeable tail of moderately high-probability items, with 6.58% of samples exceeding 0.5 (27,509 items) and 2.46% exceeding 0.75 (10,269 items).

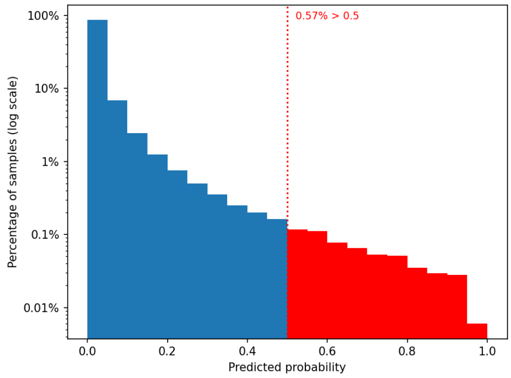


Figure 17: Distribution of predicted bias probabilities for the Weibo data source. The distribution is sharply concentrated near zero, with almost no mass above 0.5 (0.57%; 25,450 items) or 0.75 (0.15%; 6,649 items).

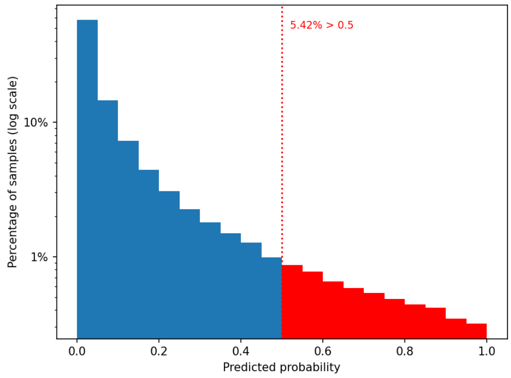


Figure 18: Distribution of predicted bias probabilities for the Zhihu data source. Zhihu shows a moderate right tail, with 5.42% (44,007 items) exceeding 0.5 and 2.01% (16,280 items) exceeding 0.75.

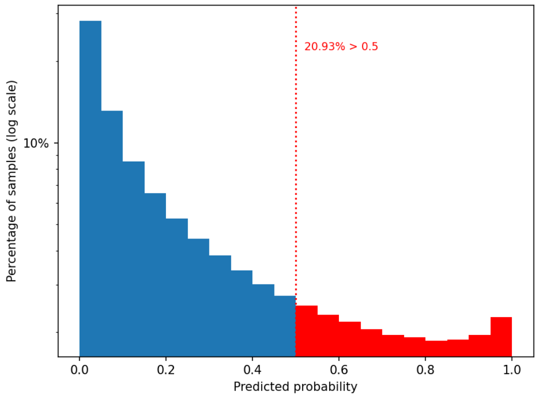


Figure 19: Distribution of predicted bias probabilities for the Tieba data source. Tieba is unique among all data sources in exhibiting a clear bimodal distribution, with a pronounced second peak near 1.0.

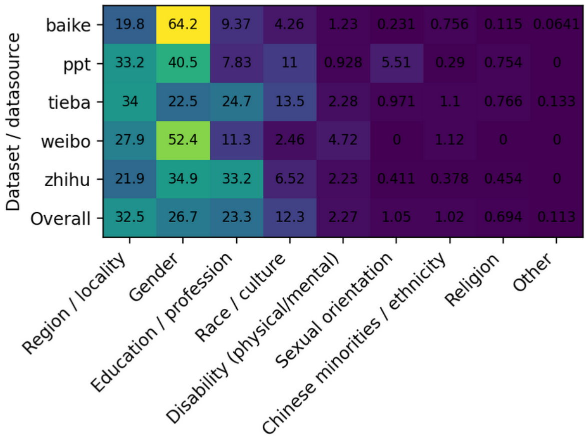


Figure 20: Distribution of category labels for HS across the 5 corpora.