LLM Sensitivity Challenges in Abusive Language Detection: Instruction-Tuned vs. Human Feedback

Yaqi Zhang,[1] **Viktor Hangya**[2,*] **and Alexander Fraser**[1,3]
[1]School of Computation, Information and Technology, Technical University of Munich
[2]Fraunhofer IIS, Erlangen, Germany
[3]Munich Center for Machine Learning
{yaqi.zhang, alexander.fraser}@tum.de viktor.hangya@iis.fraunhofer.de

Abstract

The capacity of large language models (LLMs) to understand and distinguish socially unacceptable texts enables them to play a promising role in abusive language detection. However, various factors can affect their sensitivity. In this work, we test whether LLMs have an unintended bias in abusive language detection, i.e., whether they predict more or less of a given abusive class than expected in zero-shot settings. Our results show that instruction-tuned LLMs tend to under-predict positive classes, since datasets used for tuning are dominated by the negative class. On the contrary, models fine-tuned with human feedback tend to be overly sensitive. In an exploratory approach to mitigate these issues, we show that label frequency in the prompt helps with the significant over-prediction.

1 Introduction

The rapid development of social media facilitates a surging amount of user-generated content, which inevitably includes abusive language,[1] making automatic detection crucial. Recent developments of large language models (LLMs) enable their application for different NLP tasks (Radford et al., 2019; Brown et al., 2020), including abusive language detection (Li et al., 2024). LLMs can even be applied for abuse detection without fine-tuning, i.e., zero-shot, making them invaluable for communities who do not have the resources[2] to annotate datasets for their specific needs (Plaza-del arco et al., 2023).

Various LLMs are available. Instruction-tuned models, such as Flan-T5 (Chung et al., 2024), are fine-tuned on different datasets and tasks, including abuse detection, in order to perform well over a range of tasks. Abuse datasets, some of which are included in LLM instruction tuning, suffer from label imbalance, leading to low recall in supervised classifiers (Steimel et al., 2019; Rizos et al., 2019; Al-Azzawi et al., 2023). Thus, the question arises: Do instruction-tuned LLMs have the same problem? Other models, such as LLaMA 2-Chat (Touvron et al., 2023), are fine-tuned using reinforcement learning with human feedback (RLHF) to align the model to human preferences for helpfulness and safety. RLHF can make LLMs conservative and sensitive to unsafe contexts. As shown by Touvron et al. (2023), with more safety data mixed in the tuning process, the LLaMA 2-Chat model exhibits a higher false refusal rate by refusing to answer the actual non-adversarial prompts out of safety considerations. This might affect LLMs' fairness in the abusive language detection task.[3] It's crucial to figure out these issues when using LLMs as annotators for abuse detection.

In this work, we evaluate the performance of four LLMs (Flan T5, OPT-IML, LLaMA 2-Chat, and GPT 3.5) in zero-shot settings with prompting techniques. We consider an LLM to be biased when it predicts more (over-prediction) or less (under-prediction) of a given class than it should. To measure this deviation we calculate the ratio of the predicted and expected frequency of a given label, using 4 binary and 3 multi-class English datasets covering a wide range of annotation methodologies to ensure representativeness and robustness against dataset biases. Results show that instruction-tuned models predict less positive abusive labels and even suffer from under-prediction, while RLHF tuning leads to over-prediction of positive labels. Additionally, we present exploratory

[*]This work was done while the author was affiliated with LMU Munich.

[1]We use the term *abusive language* as any type of socially unacceptable content.

[2]Although LLMs need financial resources, they can eliminate the need for experts to annotate or define guidelines.

[3]In this work, we hypothesize subtle annotation differences do not significantly impact biases in abusive language detection. As noted by Touvron et al. (2023), while annotations may vary, skilled annotators can provide consistent evaluations that models can reliably learn from. Humans are adept at distinguishing response quality, adding to RLHF robustness.

Proceedings of the 31st International Conference on Computational Linguistics, pages 2765–2780

	#Params	Instruction-tuned	RLHF	Open Source	$(temp, top_p)$
Flan-T5	3B	✓		✓	(0.7, 0.7)
OPT-IML	1.3B	✓			(0.9, 0.7)
GPT 3.5		✓	✓		(0.9, 0.3)
Llama 2-Chat	7B	✓	✓	✓	(0.1, 1.0)

Table 1: Summary of the examined models in our experiments. Our setups of $temperature$ and top_p for each model are listed. Our choice of models in this work covers the spectrum of LLMs and we made sure not to have data contamination with our considered test datasets.

experiments aiming at mitigating such biases by informing LLMs about the desired label distributions that they should output. Our experiments show that LLMs having significant over-prediction can be positively steered, however, the opposite or no effect can be achieved in the case of mild over-prediction or under-prediction.

2 Methods

Our goal is to test the abusive language bias in LLMs caused by fine-tuning procedures and data. We test off-the-shelf LLMs using zero-shot prompting and no abusive language-specific fine-tuning.

Base Prompt We employ a base prompt and ask LLMs to classify if a given text belongs to a specific abusive or non-abusive class. For example, we input: `Text: {text} Is this NORMAL, OFFENSIVE, or HATESPEECH? Answer in one word with NORMAL, OFFENSIVE, or HATESPEECH only.` where `{text}` is the input example, and we take the generated texts as the label.

Adding Label Distribution As a preliminary set of experiments motivated by traditional imbalance learning techniques (Zhang et al., 2024), especially *thresholding* methods which compensate for the prior class probabilities (Buda et al., 2018), we test whether the output label distribution of LLMs can be steered with information about label distributions. In our method denoted by **numeric**, we specify label distribution in numbers, e.g., `Consider that the post originates from a dataset where 16.8% labels are NORMAL, 77.4% labels are OFFENSIVE, and 5.8% labels are HATESPEECH.` Additionally, considering that some models may lack the ability to process numerical information, in the **word** method, we specify the relative frequency of labels, e.g., `Consider that the post originates from a dataset where OFFENSIVE occurs more frequently than NORMAL, NORMAL occurs more frequently than HATESPEECH, and OFFENSIVE occurs more frequently than HATESPEECH.` Note that we relied on training set distributions in this exploratory method. Appendices C and D show that if no training dataset is available giving feedback to the model after manually investigating a few samples, or instructing the model about a balanced distribution can reach competitive results.

3 Experiments

Experimental Setup We focus on Flan-T5-XL, OPT-IML-1.3B, LLaMA 2-Chat 7B, and GPT 3.5[4] in this work. A summary of models examined in our work is listed in Table 1. As it is known from previous work that LLMs are sensitive to prompts (Zhu et al., 2023; Pezeshkpour and Hruschka, 2023), we experimented with different prompt variants, including permuting labels (Shu et al., 2024), to eliminate bias from the prompt itself. For each LLM we selected the prompt that performed best on average based on the development splits of our datasets using macro F_1 score. The final prompts are listed in Table 7. On top of different prompt formats, we adopted grid search on $temperature$ and top_p and compared the average macro F_1 scores across all datasets for each combination of parameter values using the development split of each dataset. Regarding the prediction, we take the first token of a generated text as the label. When the first token is not among the valid labels, we exclude this sample when measuring the model performance. Our results were the average of three seeds (0, 21, 42). The code and experiment outputs are available in github.[5]

Datasets To have robust results and minimize the negative effects of dataset biases, we experimented on datasets with various abusive task types and label sets (toxic, hate speech, etc.) from different social media platforms (Twitter, Wikipedia comments), which are thus representative of a wide

[4]We adopted the gpt-3.5-turbo-0125 variant.

[5]https://github.com/zhangyaqi20/llm_sensitivity_challenges

	Macro F_1	non-abusive		abusive		bias_{agg}
		bias	$\mathbf{F_1}$	bias	$\mathbf{F_1}$	
Flan-T5 XL	**73.03**	**6.33**	**87.17**	**−23.55**	**58.90**	**15.14**
Civil-Comments	71.15±0.76	2.81±0.07	96.14±0.09	−32.41±0.76	46.16±1.44	17.61±0.42
HASOC-2019 task1	74.65±0.37	7.94±0.47	88.84±0.17	−23.84±1.40	60.45±0.61	15.89±0.93
HatEval	70.07±0.48	−0.29±0.44	74.40±0.38	0.39±0.59	65.76±0.59	0.48±0.32
OLID	76.26±0.55	14.84±0.58	89.29±0.24	−38.33±1.50	63.23±0.87	26.59±1.04
OPT-IML 1.3B	**59.15**	**16.81**	**83.24**	**−11.76**	**35.06**	**44.72**
Civil-Comments	56.44±0.56	−9.71±0.35	89.15±0.26	112.03±4.03	23.73±0.89	60.87±2.19
HASOC-2019 task1	57.52±1.56	20.00±1.50	85.16±0.14	−60.07±4.51	29.88±3.04	40.03±3.01
HatEval	59.31±0.96	38.80±1.58	74.30±0.55	−52.07±2.12	44.32±1.52	45.43±1.86
OLID	63.31±2.67	18.17±0.80	84.33±1.16	−46.94±2.06	42.29±4.20	32.56±1.42
LLaMA 2-Chat 7B	**65.07**	**−16.88**	**77.21**	**138.76**	**52.92**	**77.82**
Civil-Comments	47.22±0.11	−44.08±0.18	70.16±0.14	508.36±2.01	24.27±0.07	276.22±1.09
HASOC-2019 task1	73.26±0.27	−0.69±0.12	86.50±0.11	2.08±0.35	60.02±0.43	1.39±0.23
HatEval	66.72±0.11	−11.40±0.20	69.40±0.12	15.30±0.27	64.03±0.11	13.35±0.24
OLID	73.06±0.19	−11.35±0.73	82.76±0.03	29.31±1.88	63.35±0.40	20.32±1.30
GPT 3.5	**70.79**	**0.06**	**84.82**	**20.59**	**56.76**	**17.32**
Civil-Comments	65.78±0.20	−7.49±0.35	92.18±0.08	86.38±4.08	39.38±0.45	46.94±2.22
HASOC-2019 task1	75.48±0.31	−2.43±0.41	87.30±0.24	7.29±1.25	63.65±0.39	4.86±0.83
HatEval	67.18±0.42	12.04±0.80	74.24±0.26	−16.16±1.07	60.13±0.62	14.10±0.94
OLID	74.70±0.21	−1.88±0.67	85.54±0.24	4.86±1.73	63.87±0.19	3.37±1.20

(a) Binary datasets.

	Macro F_1	normal		offensive		hate speech		bias_{agg}
		bias	$\mathbf{F_1}$	bias	$\mathbf{F_1}$	bias	$\mathbf{F_1}$	
Flan-T5 XL	**61.81**	**−10.29**	**73.89**	**−18.77**	**63.51**	**61.68**	**48.03**	**31.70**
Davidson-2017	70.22±0.11	4.36±0.45	85.69±0.68	−5.49±0.29	91.73±0.09	60.96±3.86	33.23±0.68	23.60±1.36
HateXplain	53.40±0.45	−24.94±0.26	62.09±0.63	−32.05±1.42	35.28±0.56	62.40±1.56	62.83±0.51	39.80±1.06
OPT-IML 1.3B	**37.09**	**110.52**	**46.13**	**−52.72**	**36.67**	**114.61**	**28.45**	**97.92**
Davidson-2017	35.79±0.41	174.51±1.39	39.73±0.70	−56.14±0.33	52.40±0.68	245.11±2.29	15.24±1.32	158.59±0.93
HateXplain	38.38±1.02	46.52±0.93	52.53±0.55	−49.30±0.37	20.94±1.70	−15.90±0.85	41.66±0.96	37.24±0.71
LLaMA 2-Chat 7B	**55.10**	**−34.88**	**52.89**	**22.91**	**64.48**	**56.48**	**47.91**	**41.31**
Davidson-2017	62.27±0.19	4.08±0.87	70.31±0.16	−5.59±0.35	86.81±0.05	63.17±2.13	29.67±0.66	24.28±1.11
HateXplain	47.92±0.36	−73.83±0.27	35.47±0.16	51.40±1.53	42.15±0.57	49.78±1.57	66.14±0.40	58.33±0.20
GPT 3.5	**52.39**	**−48.75**	**51.81**	**79.88**	**67.45**	**16.33**	**37.92**	**67.41**
Davidson-2017	61.36±0.65	−28.29±1.08	62.41±1.14	−0.52±0.16	86.78±0.30	89.39±1.80	34.88±0.77	39.40±0.88
HateXplain	43.42±0.57	−69.22±0.20	41.20±0.24	160.28±2.39	48.12±0.41	−56.73±2.43	40.95±1.45	95.41±1.55

(b) Multi-class datasets.

Table 2: Results with the base prompt. Averaged results follow the model names. ± indicates standard deviation.

range of abusive tasks and domains. Details about the used datasets can be found in Appendix B. We argue that by averaging results over multiple datasets their differences and biases cancel out.

Evaluation Metrics We follow the common practice of evaluating an abusive language classifier using F_1 and macro F_1 scores. Additionally, we measure a model's prediction bias as the difference between the label distributions of the gold (test set) and model output. Note that all occurrences of **bias** in our work refer to the prediction bias. Consider a dataset with 5 non-abusive and 5 abusive texts. A model M_1 classifies all the samples as non-abusive, whereas M_2 correctly classifies 3 of 5 non-abusive and 4 of 5 abusive samples. Both models have 67% as the F_1 score for the non-abusive class. However, M_1 is clearly biased towards predicting the negative class. This demonstrates that the F_1 score alone cannot provide a complete picture of the prediction distributions. Inspired by Dixon et al. (2018), for a given class $c \in C$ we define $bias_c$ to measure how much the model predicts c more or less than it should as:

$$\text{bias}_c = \frac{(\text{TP}_c + \text{FP}_c) - (\text{TP}_c + \text{FN}_c)}{\text{TP}_c + \text{FN}_c} \quad (1)$$

where TP_c counts the number of samples which are correctly classified as c, FN_c counts samples that are wrongly classified as non-c, while FP_c is the number of non-c samples that are wrongly classified as c. A classifier with no bias towards label c should have $\text{bias}_c = 0$. A $\text{bias}_c > 0$ indicates that label c is over-predicted, e.g., $\text{bias}_c = 0.5$ means 50% more, while $\text{bias}_c < 0$ shows under-prediction. Finally, we define the overall bias of a

LLaMA 2-Chat 7B	Macro F_1	non-abusive		abusive		bias_{agg}
		bias	$\mathbf{F_1}$	bias	$\mathbf{F_1}$	
Average	65.07	−16.88	77.21	138.76	52.92	77.82
+ numeric (train)	67.57	−12.44	80.41	82.32	54.74	47.38
+ word (train)	63.75	10.51	84.15	−9.25	43.35	35.39
Civil-Comments	47.22±0.11	−44.08±0.18	70.16±0.14	508.36±2.01	24.27±0.07	276.22±1.09
+ numeric (train)	55.83±0.01	−24.70±0.10	82.91±0.04	284.79±1.13	28.74±0.04	154.75±0.61
+ word (train)	60.46±0.19	−8.14±0.19	90.73±0.11	93.90±2.10	30.19±0.28	51.02±1.14
OLID	73.06±0.19	−11.35±0.73	82.76±0.03	29.31±1.88	63.35±0.40	20.32±1.30
+ numeric (test)	76.35±0.62	−4.25±0.25	86.08±0.35	10.97±0.64	66.62±0.90	7.61±0.44
+ numeric (train)	75.85±0.39	−8.01±0.34	85.13±0.21	20.70±0.87	66.58±0.58	14.35±0.60
+ word (train)	67.11±0.55	21.13±0.28	86.65±0.22	−54.59±0.72	47.56±0.89	37.86±0.50

Table 3: LLaMA 2-Chat 7B results on binary datasets with label distribution in the prompt. The first three rows present average results across four binary datasets. "+ numeric (train/test)" is the prediction with train/test set distribution specified by percentage numbers. "+ word (train)" is the prediction with training set distribution specified by textual descriptions of label frequency. *bias* in green indicates the corresponding method mitigates the prediction bias.

HateXplain	Macro F_1	**normal** (28.5%)		**offensive** (40.6%)		**hate speech** (30.9%)		bias_{agg}
		bias	$\mathbf{F_1}$	bias	$\mathbf{F_1}$	bias	$\mathbf{F_1}$	
OPT-IML 1.3B	38.38±1.02	46.52±0.93	52.53±0.55	−49.30±0.37	20.94±1.70	−15.90±0.85	41.66±0.96	37.24±0.71
+ numeric (train)	35.60±0.18	−5.54±1.60	45.50±0.79	98.12±1.53	40.97±0.49	−83.36±0.57	20.34±0.82	62.34±0.85
+ word (train)	42.82±1.61	29.25±2.97	54.61±1.23	−25.73±4.47	29.76±2.32	−14.77±2.38	44.09±1.45	23.25±2.38
LLaMA 2-Chat 7B	47.92±0.36	−73.83±0.27	35.47±0.16	51.40±1.53	42.15±0.57	49.78±1.57	66.14±0.40	58.33±0.20
+ numeric (train)	50.87±0.39	36.96±0.59	67.85±0.23	−59.06±1.34	22.27±0.61	5.84±0.70	62.49±0.51	33.95±0.81
+ word (train)	53.79±0.48	−34.57±0.58	57.56±0.58	56.57±1.42	44.76±0.42	−6.68±0.76	59.04±0.74	32.61±0.86
GPT 3.5	43.42±0.57	−69.22±0.20	41.20±0.24	160.28±2.39	48.12±0.41	−56.73±2.43	40.95±1.45	95.41±1.55
+ numeric (train)	42.52±0.29	−83.88±0.66	25.26±0.95	143.31±2.28	46.39±0.30	−21.77±1.40	55.92±1.08	82.99±1.40
+ word (train)	46.09±0.24	−76.64±0.27	33.59±0.35	86.19±0.94	45.31±0.31	21.38±1.18	59.37±0.69	61.40±0.18

Table 4: Results on HateXplain with label distribution in the prompt. The percentage number after each label denotes the label distribution in the training set.

given model as:

$$\text{bias}_{\text{agg}} = \frac{1}{|C|} \sum_{c \in C} |\text{bias}_c| \tag{2}$$

Importantly, $bias_c$ only measures the amount of predicted labels compared to the gold value and ignores whether the individual instances are correctly classified. An approximately zero bias score does not guarantee a high F_1 score, and vice-versa. Thus we take both F_1 and the bias scores into account when discussing the model performance.

4 Results and Analysis

Base Results Table 2a presents the results on the binary datasets. We found that GPT 3.5 and LLaMA 2-Chat over-predicted the abusive class, while Flan-T5 and OPT-IML under-predicted it. Although there is a small variance in the amount of bias of the models across the different datasets, the direction (+/-) of the bias is constant for all models, with only two exceptions, indicating a general model tendency. We conjecture that the similar behavior of GPT 3.5 and LLaMA 2-Chat is rooted in RLHF, especially the fine-tuning towards the *safety* metrics (Touvron et al., 2023), so that they rather label any suspicious sentence as abusive rather than leave them unfiltered. However, Flan-T5 and OPT-IML are instruction-tuned, including abusive datasets which suffer from label imbalance, leading to the negative prediction bias. Table 2b presents the results on two fine-grained datasets. As above, the RLHF models tend to over-predict the positive labels and under-predict the negative one. In contrast, Flat-T5 and OPT-IML have mixed results. They over-predict normal class with one exception, under-predict the offensive (positive) class, but over-predict the hate class. Still, compared to the RLHF models they predict more of the negative classes and less of the positive classes.

Overall, Flan-T5 shows the lowest amount of bias_{agg} compared with the other three models. However, they are comparable to Flan-T5 on many of the datasets when considered individually.

Prompting with Label Distribution Adding label distribution information leads to a smaller non-abusive and abusive bias on average in case of

LLaMA 2-Chat on the binary datasets (Table 3). The word variant wins over the numeric variant on average, although there are exceptions. It can also be seen that mitigating the bias also brings improvements to the overall classification performance with a higher macro F_1 score. We find similar results on the fine-grained HateXplain dataset in Table 4.

Overall, we conclude that including label distribution in the prompt can alleviate the abusive bias in RLHF models on datasets with a larger degree of prediction bias from LLMs. In contrast, we also found that this approach is ineffective for Flan-T5 and OPT-IML, and can hurt the model performance on unbiasedly predicted datasets with the base prompt. We perform further analyses in Appendices C, D and E, to highlight some of the negative results that did not fit the main paper.

5 Conclusion

In this work, we analyzed the abusive language bias of four popular LLMs. Our results show that instruction-tuned models tend to under-predict the abusive labels, while RLHF models have the opposite tendency. Our work raises awareness about potential pitfalls in current LLM fine-tuning strategies when imbalanced training data is used for instruction-tuning as well as about achieving a better trade-off in the helpfulness and safety of LLMs in RLHF when abusive language is included. We experimented with a preliminary method to mitigate model bias and showed promising results in the case of biased models.

Limitations

Although we tested on 7 datasets, our experiments are limited to English corpora. Our current concentration on the English-speaking community is to avoid involving potential cross-cultural biases. We think, however, it is also worth extending this task in subsequent studies to include other cultural backgrounds and communities. We believe that our findings on the prediction bias in LLMs also hold for other languages, but verification is needed.

Furthermore, the methods to mitigate the prediction bias in LLMs in our work are just exploratory, including negative findings. There is a large room for further improvement in addressing the label bias in LLMs, starting from the root cause of the bias. It is thus worth exploring more methods and strategies in future work.

Acknowledgement

The work was supported by the European Research Council (ERC) under the European Union's Horizon Europe research and innovation programme (grant agreement No. 101113091) and by the German Research Foundation (DFG; grant FR 2829/7-1).

References

Sana Al-Azzawi, György Kovács, Filip Nilsson, Tosin Adewumi, and Marcus Liwicki. 2023. NLP-LTU at SemEval-2023 task 10: The impact of data augmentation and semi-supervised learning techniques on text classification performance on an imbalanced dataset. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1421–1427, Toronto, Canada. Association for Computational Linguistics.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Mateusz Buda, Atsuto Maki, and Maciej A. Mazurowski. 2018. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Thomas Davidson, Dana Warmsley, Michael W. Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *International Conference on Web and Social Media*.

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.

Lingyao Li, Lizhou Fan, Shubham Atreja, and Libby Hemphill. 2024. "hot" chatgpt: The promise of chatgpt in detecting and discriminating hateful, offensive, and toxic comments on social media. *ACM Transactions on the Web*, 18(2):1–36.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel. 2019. Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. In *Proceedings of the 11th Annual Meeting of the Forum for Information Retrieval Evaluation*, FIRE '19, page 14–17, New York, NY, USA. Association for Computing Machinery.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2020. Hatexplain: A benchmark dataset for explainable hate speech detection. *CoRR*, abs/2012.10289.

Pouya Pezeshkpour and Estevam Hruschka. 2023. Large language models sensitivity to the order of options in multiple-choice questions. *arXiv preprint arXiv:2308.11483*.

Flor Miriam Plaza-del arco, Debora Nozza, and Dirk Hovy. 2023. Respectful or toxic? using zero-shot learning with language models to detect hate speech. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 60–68, Toronto, Canada. Association for Computational Linguistics.

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

Georgios Rizos, Konstantin Hemker, and Björn Schuller. 2019. Augment to Prevent: Short-Text Data Augmentation in Deep Learning for Hate-Speech Classification. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, page 991–1000.

Sarthak Roy, Ashish Harshvardhan, Animesh Mukherjee, and Punyajoy Saha. 2023. Probing LLMs for hate speech detection: strengths and vulnerabilities. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6116–6128, Singapore. Association for Computational Linguistics.

Bangzhao Shu, Lechen Zhang, Minje Choi, Lavinia Dunagan, Lajanugen Logeswaran, Moontae Lee, Dallas Card, and David Jurgens. 2024. You don't need a personality test to know these models are unreliable: Assessing the reliability of large language models on psychometric instruments. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5263–5281, Mexico City, Mexico. Association for Computational Linguistics.

Kenneth Steimel, Daniel Dakota, Yue Chen, and Sandra Kübler. 2019. Investigating multilingual abusive language detection: A cautionary tale. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1151–1160, Varna, Bulgaria. INCOMA Ltd.

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*.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1415–1420, Minneapolis, Minnesota. Association for Computational Linguistics.

Yaqi Zhang, Viktor Hangya, and Alexander Fraser. 2024. A study of the class imbalance problem in abusive language detection. In *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)*, pages 38–51, Mexico City, Mexico. Association for Computational Linguistics.

Kaijie Zhu, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, Wei Ye, Yue Zhang, Neil Zhenqiang Gong, et al. 2023. Promptbench: Towards evaluating the robustness of large language models on adversarial prompts. *arXiv preprint arXiv:2306.04528*.

A Prompts

In Table 7 we listed prompts used in our experiments. We experimented with various prompt variants. It is worth mentioning that different LLMs are more adequate for different formats of input prompts. So the listed prompts are different in terms of wording and structure. Further, as shown by (Shu et al., 2024), some LLMs might generate inconsistent responses under perturbations targeting the question content. To eliminate the potential bias from the prompts as much as possible, we also experimented with permutations of label orders, for example, instead of `Is this text NORMAL, OFFENSIVE, or HATESPEECH?`, we ask `Is this text OFFENSIVE, NORMAL, or HATESPEECH?` or `Is this text HATESPEECH, NORMAL, or OFFENSIVE?` Then we chose the prompt variant resulting in the highest macro F_1 score for each method on each dataset. Additionally to the *numeric* and *word* prompt variants, we also examine if an LLM can learn from its own mistakes in our **feedback** method as presented in Appendix C. We point out how much a given model deviates from the expected distribution, e.g., `You wrongly predicted less NORMAL, less HATESPEECH, but much more OFFENSIVE than what is actually present in the dataset.`

B Datasets

We experimented on the datasets below which involve various abusive types (toxic, hate speech, etc.) and different social media platforms (Twitter, Facebook, Wikipedia comments), representing a wide range of abusive language detection tasks.

Civil-Comments is a multi-label dataset[6] used to identify and classify various types of toxic Wikipedia comments. We utilized its binary label. We sub-sampled 5,000 instances due to limited computational resources.

Davidson-2017 is collected to better differentiate between serious hate speech and commonplace offensive language (Davidson et al., 2017). We used its fine-grained labels.

HASOC-2019 contains Twitter and Facebook posts for the identification of hate speech and offensive content in Indo-European languages (Mandl et al., 2019). We experimented on the English set with the binary task (Task 1) to detect hate and offensive language (HOF), as well as the fine-grained task (Task 2) to differentiate between three subtypes of HOF: hate, offensive, and profane.

HatEval is the dataset used in SemEval 2019 Task 5 (Basile et al., 2019) to detect hate speech against immigrants and women in tweets. We used its English dataset with the binary label.

HateXplain is a benchmark dataset by Mathew et al. (2020) to capture human rationales for hate speech labeling. We used its fine-grained set to classify a text into hate, offensive, or normal.

OLID is a dataset compiled by Zampieri et al. (2019) for offensive content using a fine-grained three-layer annotation scheme. We used its binary labels.

[6] `https://www.kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge/data`

GPT 3.5	non-abusive	abusive	bias_{agg}
base	0.06	20.59	17.32
+ numeric (train)	-8.15	34.37	21.25
+ word (train)	14.15	-37.12	25.63
+ feedback	2.41	-4.76	4.42

Table 5: Prediction bias in GPT 3.5 on binary datasets with various prompting strategies.

C Prompting by Learning from Base Results

Given a development set of the target domain, we can measure and tell the model where it is wrong. We can prompt the model that it predicted more or less of a particular class than needed. We call this variant of prompting, *feedback* (more details in Appendix A). Although neither the *numeric* nor the *word* helped in mitigating the bias of GPT 3.5, the feedback format decreased it significantly. With extra experiments, we achieved results as shown in Table 5. As can be seen from the table, the feedback significantly alleviates the prediction bias in GPT 3.5. We conjecture the effectiveness is due to stating that the model has made an error, similarly as in self-correct approaches (Madaan et al., 2023).

D Prompting with Made Up Information

In addition to prompting models with the training set distribution, we would also like to observe model performance when made-up information is given, i.e., to test model sensitivity to distribution information variations and eliminate the need for gold label distribution information. In Table 6 we test the model bias, averaged over the binary dataset, when the distribution of training data and an equal distribution are given. With a higher proportion of abusive labels, GPT 3.5 and LLaMA 2-Chat tend to predict more abusive samples, and the **numeric** variant has a greater influence than the word one. In comparison, Flan-T5 and OPT-IML predict less abusive samples when increasing the portion of the abusive class in the prompt. When the model is informed about a balanced distribution (+numeric(50%) and +word(same)) the models perform competitively with the prompts using the train distribution. This indicates that the balanced setup is a good choice in case of no information about the true distribution. Further, as shown by Table 6 and detailed results on all datasets in Appendix F, we found that results on the Flan-T5 model are

	Flan-T5 XL			OPT-IML 1.3B		
	non-abusive	abusive	bias_{agg}	non-abusive	abusive	bias_{agg}
base	6.33	−23.55	15.14	16.81	−11.76	44.72
+ numeric (train)	6.77	−28.32	17.54	25.12	−69.37	47.25
+ numeric (50%)	6.43	−26.53	16.48	22.34	−62.32	42.33
+ numeric (75%)	6.07	−25.81	15.94	22.43	−63.57	43.00
+ word (train)	5.67	−23.34	14.58	29.09	−75.38	52.23
+ word (same)	5.99	−22.02	14.00	25.05	−64.89	44.97

	GPT 3.5			LLaMA 2-Chat 7B		
	non-abusive	abusive	bias_{agg}	non-abusive	abusive	bias_{agg}
base	0.06	20.59	17.32	−16.88	138.76	77.82
+ numeric (train)	−8.15	34.37	21.25	−12.44	82.32	47.38
+ numeric (50%)	−22.91	116.92	69.92	−24.21	149.97	87.09
+ word (train)	14.15	−37.12	25.63	10.51	−9.25	35.39
+ word (same)	−7.56	39.77	24.14	2.81	16.69	24.57

Table 6: Average bias on binary datasets with various prompting strategies. Values in the parentheses specify the made-up information added to the prompt. For example, + numeric (75%) means we indicate that the abusive class is 75%, while the non-abusive one is 25%, while + word(same) means we state that all labels occur with the same frequency.

close to each other, indicating that Flan-T5 is hard to steer with numeric or word descriptions. This aligns with the experimental results shown in (Roy et al., 2023) when applying different prompting strategies.

E Bias towards Various Abusive Types

Our experiments discussed above relate to datasets with both abusive and non-abusive samples. In this section, we examine the prediction bias in LLMs when classifying different abusive types. Thus, we conducted experiments on the HASOC-2019 task2 with labels: hate speech, offensive, and profane. From the results in Table 8g, we conclude that for all models there is no clear tendency towards any of the labels when the base prompt is used. We also prompt the models with label distribution information. On LLaMA 2-Chat 7B we achieved better aggregated results with both **numeric** and **word** variants. But on GPT 3.5, we found that, although the prompt with label distribution successfully flipped the prediction to output more hate speech and less offensive/profane, the magnitude of the change exceeds much more than needed, leading to an over-prediction in hate speech and under-prediction in the other two labels. Even with the test data distribution, the situation was not improved. We then experimented further with **feedback** prompt and obtained a better distribution on two of the labels as well as a lower aggregated bias value. We see that on both instruction-tuned models, adding label information even harms the model performance. From these results we conclude that the models can be steered, however they are sensitive to various factors which we leave to future work to be explored.

F Full Experimental Results

We presented the detailed results on all of our datasets in Table 8.

LLaMA 2-Chat 7B	
Base	Text: ``` {text}``` Classify this text with "NORMAL" / "OFFENSIVE" / "HATESPEECH". Answer 0 for "NORMAL", 1 for "OFFENSIVE", and 2 for "HATESPEECH". The answer is
+ Numeric	The following text originates from a dataset where 16.8% labels are "NORMAL", 77.4% labels are "OFFENSIVE", and 5.8% labels are "HATESPEECH". Text: ``` {text}``` Classify this text with "NORMAL" / "OFFENSIVE" / "HATESPEECH". Answer 0 for "NORMAL", 1 for "OFFENSIVE", and 2 for "HATESPEECH". The answer is
+ Word	The following text originates from a dataset where "OFFENSIVE" occurs more frequently than "NORMAL", "NORMAL" occurs more frequently than "HATESPEECH", and "OFFENSIVE" occurs more frequently than "HATESPEECH". Text: ``` {text}``` Classify this text with "NORMAL" / "OFFENSIVE" / "HATESPEECH". Answer 0 for "NORMAL", 1 for "OFFENSIVE", and 2 for "HATESPEECH". The answer is
GPT 3.5	
Base	Text: ``` {text}``` Is this text NORMAL, OFFENSIVE, or HATESPEECH? Answer in one word with NORMAL, OFFENSIVE, or HATESPEECH only.
+ Numeric	Text: ``` {text}``` Consider that the post originates from a dataset where 16.8% labels are NORMAL, 77.4% labels are OFFENSIVE, and 5.8% labels are HATESPEECH. Is this text NORMAL, OFFENSIVE, or HATESPEECH? Answer in one word with NORMAL, OFFENSIVE, or HATESPEECH only.
+ Word	Text: ``` {text}``` Consider that the post originates from a dataset where OFFENSIVE occurs more frequently than NORMAL, NORMAL occurs more frequently than HATESPEECH, and OFFENSIVE occurs more frequently than HATESPEECH. Is this text NORMAL, OFFENSIVE, or HATESPEECH? Answer in one word with NORMAL, OFFENSIVE, or HATESPEECH only.
+ Feedback	Text: ``` {text}``` You wrongly predicted less NORMAL, less HATESPEECH, but much more OFFENSIVE than what is actually present in the dataset. Is this text OFFENSIVE, HATESPEECH, or NORMAL? Answer in one word with OFFENSIVE, HATESPEECH, or NORMAL only.

(a) Prompts for RLHF models

	Flan-T5-XL
Base	Classify the post delimited by three backticks as normal, offensive, or hatespeech. Provide the answer as either normal, offensive, or hatespeech only. Example output for normal : normal Example output for offensive : offensive Example output for hatespeech : hatespeech ``` {text}```
+ Numeric	Consider that the text originates from a dataset where 16.8% labels are normal, 77.4% labels are offensive, and 5.8% labels are hatespeech. Based on this classify the post delimited by three backticks as normal, offensive, or hatespeech. Provide the answer as either normal, offensive, or hatespeech only. Example output for normal : normal Example output for offensive : offensive Example output for hatespeech : hatespeech ``` {text}```
+ Word	Consider that the text originates from a dataset where offensive occurs more frequently than normal, normal occurs more frequently than hatespeech, and offensive occurs more frequently than hatespeech. Based on this classify the post delimited by three backticks as normal, offensive, or hatespeech. Provide the answer as either normal, offensive, or hatespeech only. Example output for normal : normal Example output for offensive : offensive Example output for hatespeech : hatespeech ``` {text}```
	OPT-IML 1.3B
Base	Is this text hatespeech, normal, or offensive? " text " OPTIONS: - hatespeech - normal - offensive Answer:
+ Numeric	Consider that the post originates from a dataset where 30.9% labels are hatespeech, 40.6% labels are normal, and 28.5% labels are offensive. Is this text hatespeech, normal, or offensive? " text " OPTIONS: - hatespeech - normal - offensive Answer:
+ Word	Consider that the post originates from a dataset where hatespeech occurs more frequently than offensive, normal occurs more frequently than offensive, and normal occurs more frequently than hatespeech. Is this text hatespeech, normal, or offensive? " text " OPTIONS: - hatespeech - normal - offensive Answer:

(b) Prompts for instruction-tuned models.

Table 7: Prompts used in our experiments. The prompts for Flan-T5 XL are adapted from the work by Roy et al. (2023). The label options are adapted to each dataset (the table depicts the labels of HateXplain).

Civil-Comments	Macro F_1	non-toxic (92.0%) bias	non-toxic (92.0%) $\mathbf{F_1}$	toxic (8.0%) bias	toxic (8.0%) $\mathbf{F_1}$	$\mathbf{bias_{agg}}$
Flan-T5 XL						
base	$71.15_{\pm 0.76}$	$2.81_{\pm 0.07}$	$96.14_{\pm 0.09}$	$-32.41_{\pm 0.76}$	$46.16_{\pm 1.44}$	$17.61_{\pm 0.42}$
+ numeric (train)	$66.08_{\pm 0.79}$	$4.77_{\pm 0.10}$	$96.08_{\pm 0.07}$	$-55.06_{\pm 1.18}$	$36.07_{\pm 1.52}$	$29.91_{\pm 0.64}$
+ numeric (50%)	$66.97_{\pm 0.04}$	$4.45_{\pm 0.06}$	$96.08_{\pm 0.02}$	$-51.30_{\pm 0.63}$	$37.86_{\pm 0.08}$	$27.87_{\pm 0.34}$
+ numeric (75%)	$66.90_{\pm 0.38}$	$4.41_{\pm 0.14}$	$96.06_{\pm 0.04}$	$-50.79_{\pm 1.67}$	$37.74_{\pm 0.73}$	$27.60_{\pm 0.90}$
+ word (train)	$68.51_{\pm 1.12}$	$3.76_{\pm 0.10}$	$96.06_{\pm 0.11}$	$-43.36_{\pm 1.15}$	$40.95_{\pm 2.11}$	$23.56_{\pm 0.62}$
+ word (equal)	$69.97_{\pm 1.19}$	$3.12_{\pm 0.22}$	$96.07_{\pm 0.11}$	$-35.93_{\pm 2.56}$	$43.87_{\pm 2.28}$	$19.52_{\pm 1.39}$
OPT-IML 1.3B						
base	$56.44_{\pm 0.56}$	$-9.71_{\pm 0.35}$	$89.15_{\pm 0.26}$	$112.03_{\pm 4.03}$	$23.73_{\pm 0.89}$	$60.87_{\pm 2.19}$
+ numeric (train)	$54.40_{\pm 0.58}$	$6.17_{\pm 0.17}$	$95.31_{\pm 0.03}$	$-71.18_{\pm 2.05}$	$13.48_{\pm 1.16}$	$38.68_{\pm 1.12}$
+ numeric (50%)	$56.53_{\pm 0.79}$	$5.47_{\pm 0.05}$	$95.25_{\pm 0.10}$	$-63.08_{\pm 0.58}$	$17.82_{\pm 1.50}$	$34.27_{\pm 0.31}$
+ numeric (75%)	$54.41_{\pm 1.33}$	$5.70_{\pm 0.26}$	$95.11_{\pm 0.23}$	$-65.66_{\pm 3.04}$	$13.70_{\pm 2.46}$	$35.68_{\pm 1.65}$
+ word (train)	$55.31_{\pm 0.72}$	$6.14_{\pm 0.34}$	$95.39_{\pm 0.07}$	$-70.76_{\pm 3.90}$	$15.22_{\pm 1.51}$	$38.45_{\pm 2.12}$
+ word (equal)	$56.42_{\pm 0.63}$	$5.17_{\pm 0.11}$	$95.11_{\pm 0.11}$	$-59.57_{\pm 1.28}$	$17.73_{\pm 1.16}$	$32.37_{\pm 0.70}$
LLaMA 2-Chat 7B						
base	$47.22_{\pm 0.11}$	$-44.08_{\pm 0.18}$	$70.16_{\pm 0.14}$	$508.36_{\pm 2.01}$	$24.27_{\pm 0.07}$	$276.22_{\pm 1.09}$
+ numeric (train)	$55.83_{\pm 0.01}$	$-24.70_{\pm 0.10}$	$82.91_{\pm 0.04}$	$284.79_{\pm 1.13}$	$28.74_{\pm 0.04}$	$154.75_{\pm 0.61}$
+ numeric (50%)	$49.19_{\pm 0.13}$	$-39.77_{\pm 0.24}$	$73.30_{\pm 0.17}$	$458.65_{\pm 2.71}$	$25.09_{\pm 0.15}$	$249.21_{\pm 1.47}$
+ word (train)	$60.46_{\pm 0.19}$	$-8.14_{\pm 0.19}$	$90.73_{\pm 0.11}$	$93.90_{\pm 2.10}$	$30.19_{\pm 0.28}$	$51.02_{\pm 1.14}$
+ word (equal)	$61.72_{\pm 0.10}$	$-10.05_{\pm 0.13}$	$90.35_{\pm 0.04}$	$115.96_{\pm 1.53}$	$33.10_{\pm 0.20}$	$63.01_{\pm 0.83}$
GPT 3.5						
base	$65.78_{\pm 0.20}$	$-7.49_{\pm 0.35}$	$92.18_{\pm 0.08}$	$86.38_{\pm 4.08}$	$39.38_{\pm 0.45}$	$46.94_{\pm 2.22}$
+ feedback	$62.43_{\pm 0.68}$	$0.22_{\pm 0.64}$	$94.09_{\pm 0.11}$	$-2.59_{\pm 7.33}$	$30.77_{\pm 1.48}$	$3.41_{\pm 1.21}$
+ numeric (train)	$65.75_{\pm 0.50}$	$-6.47_{\pm 0.10}$	$92.50_{\pm 0.14}$	$74.60_{\pm 1.24}$	$39.00_{\pm 0.87}$	$40.53_{\pm 0.67}$
+ numeric (50%)	$56.80_{\pm 0.09}$	$-27.94_{\pm 0.42}$	$82.00_{\pm 0.19}$	$322.22_{\pm 4.86}$	$31.61_{\pm 0.15}$	$175.08_{\pm 2.64}$
+ word (train)	$60.44_{\pm 0.48}$	$4.97_{\pm 0.01}$	$95.49_{\pm 0.06}$	$-57.22_{\pm 0.14}$	$25.39_{\pm 0.91}$	$31.10_{\pm 0.08}$
+ word (equal)	$64.99_{\pm 0.05}$	$-8.56_{\pm 0.24}$	$91.65_{\pm 0.07}$	$98.66_{\pm 2.77}$	$38.32_{\pm 0.16}$	$53.61_{\pm 1.51}$

(a) Civil-Comments

HASOC-2019 task1	Macro F_1	**non-hof** (61.4%/75.0%) bias	$\mathbf{F_1}$	**hof** (38.6%/25.0%) bias	$\mathbf{F_1}$	$\mathbf{bias_{agg}}$
Flan-T5 XL						
base	$74.65_{\pm 0.37}$	$7.94_{\pm 0.47}$	$88.84_{\pm 0.17}$	$-23.84_{\pm 1.40}$	$60.45_{\pm 0.61}$	$15.89_{\pm 0.93}$
+ numeric (test)	$75.09_{\pm 0.14}$	$6.90_{\pm 0.74}$	$88.84_{\pm 0.18}$	$-20.71_{\pm 2.21}$	$61.33_{\pm 0.20}$	$13.81_{\pm 1.47}$
+ numeric (train)	$75.12_{\pm 0.59}$	$6.63_{\pm 0.48}$	$88.81_{\pm 0.31}$	$-19.91_{\pm 1.45}$	$61.44_{\pm 0.89}$	$13.27_{\pm 0.96}$
+ numeric (50%)	$74.96_{\pm 0.37}$	$5.78_{\pm 0.60}$	$88.58_{\pm 0.20}$	$-17.36_{\pm 1.80}$	$61.34_{\pm 0.59}$	$11.57_{\pm 1.20}$
+ numeric (75%)	$75.28_{\pm 0.53}$	$5.66_{\pm 0.53}$	$88.70_{\pm 0.30}$	$-17.01_{\pm 1.59}$	$61.86_{\pm 0.78}$	$11.34_{\pm 1.06}$
+ word (train)	$75.37_{\pm 0.53}$	$5.05_{\pm 0.44}$	$88.63_{\pm 0.28}$	$-15.16_{\pm 1.32}$	$62.12_{\pm 0.81}$	$10.11_{\pm 0.88}$
+ word (equal)	$75.46_{\pm 0.58}$	$4.43_{\pm 0.58}$	$88.56_{\pm 0.22}$	$-13.31_{\pm 1.75}$	$62.36_{\pm 0.95}$	$8.87_{\pm 1.16}$
OPT-IML 1.3B						
base	$57.52_{\pm 1.56}$	$20.00_{\pm 1.50}$	$85.16_{\pm 0.14}$	$-60.07_{\pm 4.51}$	$29.88_{\pm 3.04}$	$40.03_{\pm 3.01}$
+ numeric (test)	$58.25_{\pm 1.66}$	$20.69_{\pm 0.81}$	$85.63_{\pm 0.32}$	$-62.15_{\pm 2.43}$	$30.86_{\pm 3.00}$	$41.42_{\pm 1.62}$
+ numeric (train)	$58.34_{\pm 0.74}$	$20.65_{\pm 0.37}$	$85.64_{\pm 0.17}$	$-62.04_{\pm 1.12}$	$31.04_{\pm 1.32}$	$41.35_{\pm 0.74}$
+ numeric (50%)	$60.07_{\pm 0.92}$	$17.80_{\pm 0.70}$	$85.38_{\pm 0.40}$	$-53.47_{\pm 2.09}$	$34.75_{\pm 1.52}$	$35.64_{\pm 1.39}$
+ numeric (75%)	$59.01_{\pm 0.49}$	$19.11_{\pm 1.33}$	$85.40_{\pm 0.25}$	$-57.41_{\pm 4.01}$	$32.61_{\pm 1.20}$	$38.26_{\pm 2.67}$
+ word (train)	$55.45_{\pm 0.32}$	$23.20_{\pm 0.66}$	$85.50_{\pm 0.25}$	$-69.67_{\pm 2.00}$	$25.40_{\pm 0.58}$	$46.44_{\pm 1.34}$
+ word (equal)	$61.31_{\pm 0.40}$	$18.23_{\pm 0.18}$	$85.96_{\pm 0.14}$	$-54.75_{\pm 0.53}$	$36.65_{\pm 0.68}$	$36.48_{\pm 0.35}$
LLaMA 2-Chat 7B						
base	$73.26_{\pm 0.27}$	$-0.69_{\pm 0.12}$	$86.50_{\pm 0.11}$	$2.08_{\pm 0.35}$	$60.02_{\pm 0.43}$	$1.39_{\pm 0.23}$
+ numeric (test)	$73.48_{\pm 0.24}$	$1.81_{\pm 0.18}$	$87.11_{\pm 0.09}$	$-5.44_{\pm 0.53}$	$59.84_{\pm 0.40}$	$3.62_{\pm 0.35}$
+ numeric (train)	$73.37_{\pm 0.36}$	$-0.54_{\pm 0.07}$	$86.59_{\pm 0.18}$	$1.62_{\pm 0.20}$	$60.16_{\pm 0.54}$	$1.08_{\pm 0.13}$
+ numeric (50%)	$74.37_{\pm 0.32}$	$-14.80_{\pm 0.20}$	$84.35_{\pm 0.21}$	$44.45_{\pm 0.60}$	$64.39_{\pm 0.43}$	$29.62_{\pm 0.40}$
+ word (train)	$61.96_{\pm 0.31}$	$22.46_{\pm 0.27}$	$87.41_{\pm 0.13}$	$-67.47_{\pm 0.8}$	$36.51_{\pm 0.53}$	$44.97_{\pm 0.54}$
+ word (equal)	$74.07_{\pm 0.82}$	$0.81_{\pm 0.20}$	$87.20_{\pm 0.44}$	$-2.43_{\pm 0.60}$	$60.93_{\pm 1.20}$	$1.62_{\pm 0.40}$
GPT 3.5						
base	$75.48_{\pm 0.31}$	$-2.43_{\pm 0.41}$	$87.30_{\pm 0.24}$	$7.29_{\pm 1.25}$	$63.65_{\pm 0.39}$	$4.86_{\pm 0.83}$
+ feedback	$76.08_{\pm 0.36}$	$2.54_{\pm 0.76}$	$88.51_{\pm 0.26}$	$-7.64_{\pm 2.28}$	$63.66_{\pm 0.52}$	$5.09_{\pm 1.52}$
+ numeric (test)	$76.46_{\pm 0.34}$	$-9.36_{\pm 0.31}$	$86.60_{\pm 0.23}$	$28.16_{\pm 0.88}$	$66.33_{\pm 0.46}$	$18.76_{\pm 0.60}$
+ numeric (train)	$75.96_{\pm 0.34}$	$-11.71_{\pm 0.82}$	$85.88_{\pm 0.32}$	$35.19_{\pm 2.46}$	$66.05_{\pm 0.40}$	$23.45_{\pm 1.64}$
+ numeric (50%)	$74.85_{\pm 0.02}$	$-16.41_{\pm 0.42}$	$84.34_{\pm 0.07}$	$49.31_{\pm 1.25}$	$65.37_{\pm 0.10}$	$32.86_{\pm 0.83}$
+ word (train)	$75.66_{\pm 0.46}$	$4.47_{\pm 0.55}$	$88.66_{\pm 0.24}$	$-13.43_{\pm 1.64}$	$62.66_{\pm 0.70}$	$8.95_{\pm 1.09}$
+ word (equal)	$76.53_{\pm 0.15}$	$-8.13_{\pm 0.24}$	$86.84_{\pm 0.10}$	$24.42_{\pm 0.72}$	$66.22_{\pm 0.21}$	$16.28_{\pm 0.48}$

(b) HASOC-2019 task1

HatEval	Macro F_1	non-hate speech ($58.0\%/57.3\%$) bias	$\mathbf{F_1}$	hate speech ($42.0\%/42.7\%$) bias	$\mathbf{F_1}$	$\mathbf{bias_{agg}}$
Flan-T5 XL						
base	$70.07_{\pm 0.48}$	$-0.29_{\pm 0.44}$	$74.40_{\pm 0.38}$	$0.39_{\pm 0.59}$	$65.76_{\pm 0.59}$	$0.48_{\pm 0.32}$
+ numeric (train)	$69.53_{\pm 0.67}$	$1.75_{\pm 0.46}$	$74.28_{\pm 0.60}$	$-2.34_{\pm 0.62}$	$64.77_{\pm 0.75}$	$2.04_{\pm 0.54}$
+ numeric (50%)	$69.27_{\pm 0.56}$	$2.04_{\pm 0.66}$	$74.11_{\pm 0.49}$	$-2.73_{\pm 0.89}$	$64.43_{\pm 0.66}$	$2.38_{\pm 0.77}$
+ numeric (75%)	$69.43_{\pm 0.75}$	$1.05_{\pm 0.46}$	$74.07_{\pm 0.62}$	$-1.40_{\pm 0.62}$	$64.78_{\pm 0.89}$	$1.22_{\pm 0.54}$
+ word (train)	$69.10_{\pm 0.71}$	$0.81_{\pm 1.02}$	$73.76_{\pm 0.49}$	$-1.09_{\pm 1.37}$	$64.44_{\pm 0.95}$	$1.23_{\pm 0.74}$
+ word (equal)	$68.74_{\pm 0.53}$	$2.85_{\pm 0.72}$	$73.82_{\pm 0.41}$	$-3.82_{\pm 0.98}$	$63.67_{\pm 0.68}$	$3.34_{\pm 0.85}$
OPT-IML 1.3B						
+ base	$59.31_{\pm 0.96}$	$38.80_{\pm 1.58}$	$74.30_{\pm 0.55}$	$-52.07_{\pm 2.12}$	$44.32_{\pm 1.52}$	$45.43_{\pm 1.86}$
+ numeric (train)	$58.15_{\pm 1.39}$	$37.11_{\pm 0.79}$	$73.16_{\pm 0.88}$	$-49.80_{\pm 1.06}$	$43.13_{\pm 1.93}$	$43.46_{\pm 0.92}$
+ numeric (50%)	$59.85_{\pm 1.08}$	$30.60_{\pm 1.40}$	$72.75_{\pm 0.89}$	$-41.06_{\pm 1.88}$	$46.96_{\pm 1.35}$	$35.83_{\pm 1.64}$
+ numeric (75%)	$59.57_{\pm 0.89}$	$29.38_{\pm 0.53}$	$72.28_{\pm 0.66}$	$-39.42_{\pm 0.72}$	$46.87_{\pm 1.14}$	$34.40_{\pm 0.62}$
+ word (train)	$54.44_{\pm 1.33}$	$51.31_{\pm 0.97}$	$74.49_{\pm 0.55}$	$-68.85_{\pm 1.31}$	$34.39_{\pm 2.12}$	$60.08_{\pm 1.14}$
+ word (equal)	$56.86_{\pm 1.18}$	$42.81_{\pm 1.40}$	$73.74_{\pm 0.74}$	$-57.45_{\pm 1.88}$	$39.97_{\pm 1.76}$	$50.13_{\pm 1.64}$
LLaMA 2-Chat 7B						
base	$66.72_{\pm 0.11}$	$-11.40_{\pm 0.20}$	$69.40_{\pm 0.12}$	$15.30_{\pm 0.27}$	$64.03_{\pm 0.11}$	$13.35_{\pm 0.24}$
+ numeric (train)	$65.24_{\pm 0.05}$	$-16.52_{\pm 0.99}$	$67.02_{\pm 0.22}$	$22.17_{\pm 1.33}$	$63.46_{\pm 0.18}$	$19.35_{\pm 1.16}$
+ numeric (50%)	$64.46_{\pm 0.52}$	$-10.01_{\pm 0.51}$	$67.61_{\pm 0.46}$	$13.43_{\pm 0.68}$	$61.30_{\pm 0.60}$	$11.72_{\pm 0.59}$
+ word (train)	$65.47_{\pm 0.29}$	$6.57_{\pm 0.44}$	$71.81_{\pm 0.15}$	$-8.82_{\pm 0.59}$	$59.12_{\pm 0.43}$	$7.70_{\pm 0.51}$
+ word (equal)	$65.18_{\pm 0.21}$	$4.95_{\pm 0.27}$	$71.25_{\pm 0.20}$	$-6.64_{\pm 0.36}$	$59.1_{\pm 0.23}$	$5.79_{\pm 0.31}$
GPT 3.5						
base	$67.18_{\pm 0.42}$	$12.04_{\pm 0.80}$	$74.24_{\pm 0.26}$	$-16.16_{\pm 1.07}$	$60.13_{\pm 0.62}$	$14.10_{\pm 0.94}$
+ feedback	$65.14_{\pm 0.51}$	$7.27_{\pm 1.41}$	$71.68_{\pm 0.28}$	$-9.76_{\pm 1.89}$	$58.59_{\pm 0.84}$	$8.51_{\pm 1.65}$
+ numeric (25%)	$66.31_{\pm 0.57}$	$-1.75_{\pm 0.18}$	$70.89_{\pm 0.48}$	$2.34_{\pm 0.24}$	$61.73_{\pm 0.66}$	$2.04_{\pm 0.21}$
+ numeric (train)	$66.59_{\pm 0.10}$	$-7.68_{\pm 1.36}$	$70.00_{\pm 0.19}$	$10.31_{\pm 1.83}$	$63.17_{\pm 0.35}$	$8.99_{\pm 1.60}$
+ numeric (50%)	$66.98_{\pm 0.53}$	$-20.94_{\pm 0.35}$	$67.84_{\pm 0.47}$	$28.10_{\pm 0.47}$	$66.12_{\pm 0.60}$	$24.52_{\pm 0.41}$
+ word (train)	$58.44_{\pm 0.41}$	$35.43_{\pm 0.70}$	$72.94_{\pm 0.27}$	$-47.54_{\pm 0.94}$	$43.93_{\pm 0.64}$	$41.48_{\pm 0.82}$
+ word (equal)	$66.64_{\pm 0.62}$	$0.81_{\pm 0.83}$	$71.67_{\pm 0.61}$	$-1.09_{\pm 1.11}$	$61.62_{\pm 0.66}$	$0.95_{\pm 0.97}$

(c) HatEval

OLID	Macro F_1	non-offensive ($66.8\%/72.1\%$) bias	non-offensive $\mathbf{F_1}$	offensive ($33.2\%/27.9\%$) bias	offensive $\mathbf{F_1}$	$\mathbf{bias_{agg}}$
Flan-T5 XL						
base	$76.26_{\pm 0.55}$	$14.84_{\pm 0.58}$	$89.29_{\pm 0.24}$	$-38.33_{\pm 1.50}$	$63.23_{\pm 0.87}$	$26.59_{\pm 1.04}$
+ numeric (test)	$76.85_{\pm 1.06}$	$13.60_{\pm 1.21}$	$89.36_{\pm 0.29}$	$-35.14_{\pm 3.13}$	$64.34_{\pm 1.83}$	$24.37_{\pm 2.17}$
+ numeric (train)	$76.76_{\pm 1.06}$	$13.92_{\pm 1.21}$	$89.37_{\pm 0.28}$	$-35.97_{\pm 3.13}$	$64.16_{\pm 1.85}$	$24.95_{\pm 2.17}$
+ numeric (50%)	$76.72_{\pm 0.93}$	$13.44_{\pm 1.37}$	$89.27_{\pm 0.20}$	$-34.72_{\pm 3.54}$	$64.18_{\pm 1.68}$	$24.08_{\pm 2.46}$
+ numeric (75%)	$76.95_{\pm 1.19}$	$13.17_{\pm 1.46}$	$89.33_{\pm 0.34}$	$-34.03_{\pm 3.76}$	$64.57_{\pm 2.05}$	$23.60_{\pm 2.61}$
+ word (train)	$77.31_{\pm 1.38}$	$13.07_{\pm 0.84}$	$89.48_{\pm 0.57}$	$-33.75_{\pm 2.17}$	$65.15_{\pm 2.20}$	$23.40_{\pm 1.50}$
+ word (equal)	$76.91_{\pm 1.31}$	$13.55_{\pm 1.16}$	$89.38_{\pm 0.44}$	$-35.00_{\pm 3.00}$	$64.45_{\pm 2.19}$	$24.27_{\pm 2.09}$
OPT-IML 1.3B						
base	$63.31_{\pm 2.67}$	$18.17_{\pm 0.80}$	$84.33_{\pm 1.16}$	$-46.94_{\pm 2.06}$	$42.29_{\pm 4.20}$	$32.56_{\pm 1.42}$
+ numeric (test)	$46.49_{\pm 1.05}$	$36.67_{\pm 0.49}$	$84.28_{\pm 0.13}$	$-94.72_{\pm 1.27}$	$8.69_{\pm 1.98}$	$65.70_{\pm 0.89}$
+ numeric (train)	$46.91_{\pm 1.16}$	$36.56_{\pm 0.65}$	$84.36_{\pm 0.10}$	$-94.45_{\pm 1.68}$	$9.45_{\pm 2.22}$	$65.50_{\pm 1.17}$
+ numeric (50%)	$48.19_{\pm 0.66}$	$35.48_{\pm 0.33}$	$84.34_{\pm 0.22}$	$-91.67_{\pm 0.84}$	$12.05_{\pm 1.15}$	$63.58_{\pm 0.58}$
+ numeric (75%)	$48.97_{\pm 0.91}$	$35.54_{\pm 0.10}$	$84.59_{\pm 0.24}$	$-91.81_{\pm 0.24}$	$13.35_{\pm 1.58}$	$63.67_{\pm 0.16}$
+ word (train)	$48.73_{\pm 0.40}$	$35.70_{\pm 0.34}$	$84.58_{\pm 0.04}$	$-92.22_{\pm 0.87}$	$12.88_{\pm 0.79}$	$63.96_{\pm 0.60}$
+ word (equal)	$51.13_{\pm 0.05}$	$33.98_{\pm 0.47}$	$84.70_{\pm 0.15}$	$-87.78_{\pm 1.21}$	$17.57_{\pm 0.24}$	$60.88_{\pm 0.83}$
LLaMA 2-Chat 7B						
base	$73.06_{\pm 0.19}$	$-11.35_{\pm 0.73}$	$82.76_{\pm 0.03}$	$29.31_{\pm 1.88}$	$63.35_{\pm 0.40}$	$20.32_{\pm 1.30}$
+ numeric (test)	$76.35_{\pm 0.62}$	$-4.25_{\pm 0.25}$	$86.08_{\pm 0.35}$	$10.97_{\pm 0.64}$	$66.62_{\pm 0.90}$	$7.61_{\pm 0.44}$
+ numeric (train)	$75.85_{\pm 0.39}$	$-8.01_{\pm 0.34}$	$85.13_{\pm 0.21}$	$20.70_{\pm 0.87}$	$66.58_{\pm 0.58}$	$14.35_{\pm 0.60}$
+ numeric (50%)	$69.36_{\pm 0.46}$	$-32.26_{\pm 0.81}$	$75.77_{\pm 0.53}$	$83.33_{\pm 2.09}$	$62.94_{\pm 0.41}$	$57.80_{\pm 1.45}$
+ word (train)	$67.11_{\pm 0.55}$	$21.13_{\pm 0.28}$	$86.65_{\pm 0.22}$	$-54.59_{\pm 0.72}$	$47.56_{\pm 0.89}$	$37.86_{\pm 0.50}$
+ word (equal)	$70.20_{\pm 0.13}$	$15.54_{\pm 0.65}$	$86.71_{\pm 0.20}$	$-40.14_{\pm 1.69}$	$53.69_{\pm 0.09}$	$27.84_{\pm 1.17}$
GPT 3.5						
base	$74.70_{\pm 0.21}$	$-1.88_{\pm 0.67}$	$85.54_{\pm 0.24}$	$4.86_{\pm 1.73}$	$63.87_{\pm 0.19}$	$3.37_{\pm 1.20}$
+ feedback	$75.37_{\pm 0.45}$	$-0.38_{\pm 0.65}$	$86.18_{\pm 0.24}$	$0.97_{\pm 1.69}$	$64.55_{\pm 0.69}$	$0.67_{\pm 1.17}$
+ numeric (test)	$73.63_{\pm 1.16}$	$-4.68_{\pm 0.85}$	$84.39_{\pm 0.53}$	$12.08_{\pm 2.21}$	$62.86_{\pm 1.78}$	$8.38_{\pm 1.53}$
+ numeric (train)	$74.00_{\pm 1.02}$	$-6.72_{\pm 0.89}$	$84.23_{\pm 0.78}$	$17.36_{\pm 2.29}$	$63.78_{\pm 1.26}$	$12.04_{\pm 1.59}$
+ numeric (50%)	$71.86_{\pm 0.45}$	$-26.34_{\pm 1.22}$	$78.94_{\pm 0.49}$	$68.06_{\pm 3.16}$	$64.76_{\pm 0.53}$	$47.20_{\pm 2.18}$
+ word (train)	$71.16_{\pm 0.52}$	$11.72_{\pm 0.94}$	$86.34_{\pm 0.40}$	$-30.28_{\pm 2.44}$	$55.98_{\pm 0.70}$	$21.00_{\pm 1.69}$
+ word (equal)	$73.87_{\pm 0.17}$	$-14.35_{\pm 0.74}$	$82.71_{\pm 0.20}$	$37.09_{\pm 1.91}$	$65.03_{\pm 0.21}$	$25.72_{\pm 1.32}$

(d) OLID

Davidson-2017	Macro F_1	offensive (77.4%)		neither (16.8%)		hate speech (5.8%)		bias_{agg}
		bias	$\mathbf{F_1}$	bias	$\mathbf{F_1}$	bias	$\mathbf{F_1}$	
Flan-T5 XL								
base	70.22±0.11	−5.49±0.29	91.73±0.09	4.36±0.45	85.69±0.68	60.96±3.86	33.23±0.68	23.60±1.36
+ numeric (train)	69.16±0.54	−1.89±0.27	92.51±0.01	10.84±0.62	85.41±0.19	−6.18±2.05	29.58±1.82	6.31±0.43
+ numeric (33.3%)	71.23±0.06	−2.63±0.38	92.80±0.21	10.20±0.55	84.94±0.37	5.59±3.65	35.94±0.42	6.14±1.49
+ word (train)	71.47±0.85	−2.12±0.36	92.88±0.14	6.00±1.20	86.09±0.16	10.96±2.58	35.45±2.52	6.36±1.13
+ word (equal)	71.20±0.54	−3.50±0.10	92.54±0.12	7.36±0.18	86.45±0.26	25.52±1.81	34.62±1.24	12.13±0.58
OPT-IML 1.3B								
base	35.79±0.41	−56.14±0.33	52.40±0.68	174.51±1.39	39.73±0.70	245.11±2.29	15.24±1.32	158.59±0.93
+ numeric (train)	42.87±0.42	21.65±0.03	88.19±0.06	−68.83±0.54	34.29±1.04	−90.20±1.52	6.14±2.31	60.23±0.34
+ numeric (33.3%)	46.99±0.31	18.21±0.43	88.14±0.24	−57.54±1.57	40.89±1.49	−76.90±2.49	11.93±1.94	50.89±1.26
+ word (train)	46.64±0.27	8.95±0.54	86.47±0.22	−12.09±2.61	47.15±0.56	−85.05±1.94	6.29±0.84	35.36±1.01
+ word (equal)	38.66±1.14	24.14±0.29	87.91±0.23	−78.67±1.28	25.18±2.24	−94.76±0.7	2.88±1.02	65.85±0.59
LLaMA 2-Chat 7B								
base	62.27±0.19	−5.59±0.35	86.81±0.05	4.08±0.87	70.31±0.16	63.17±2.13	29.67±0.66	24.28±1.11
+ numeric (train)	54.15±0.12	9.87±0.16	91.00±0.08	−11.16±0.67	71.46±0.30	−99.88±0.20	0.00±0.00	40.31±0.34
+ numeric (33.3%)	51.11±0.37	−2.28±0.25	87.18±0.08	44.42±1.07	65.00±0.26	−98.72±0.40	1.15±0.79	48.47±0.36
+ word (train)	53.92±0.12	14.91±0.04	90.93±0.02	−36.61±0.24	66.03±0.13	−93.48±0.20	4.82±0.38	48.33±0.05
+ word (equal)	38.50±0.13	−34.16±0.21	69.48±0.21	191.52±1.02	46.01±0.19	−99.42±0.20	0.00±0.00	108.36±0.47
GPT 3.5								
base	61.36±0.65	−0.52±0.16	86.78±0.30	−28.29±1.08	62.41±1.14	89.39±1.80	34.88±0.77	39.40±0.88
+ numeric (train)	56.03±0.26	15.83±0.08	89.03±0.19	−66.95±0.66	47.51±0.87	−17.48±2.86	31.54±0.31	33.42±0.76
+ numeric (33.3%)	38.81±0.21	−72.32±0.61	38.12±0.53	−36.61±0.67	63.78±0.24	1077.16±10.05	14.53±0.14	395.36±3.77
+ word (train)	58.82±0.43	2.28±0.20	86.68±0.10	−48.62±0.36	60.64±0.47	111.07±1.93	29.15±0.75	53.99±0.53
+ word (equal)	31.48±0.46	−89.54±0.52	16.29±0.97	−32.89±0.83	65.09±0.41	1297.44±4.72	13.04±0.09	473.29±1.51

(e) Davidson-2017

HateXplain	Macro F_1	offensive (40.6%)		hate speech (30.9%)		normal (28.5%)		bias_{agg}
		bias	$\mathbf{F_1}$	bias	$\mathbf{F_1}$	bias	$\mathbf{F_1}$	
Flan-T5 XL								
base	53.40±0.45	−32.05±1.42	35.28±0.56	62.40±1.56	62.83±0.51	−24.94±0.26	62.09±0.63	39.80±1.06
+ numeric (train)	56.73±0.79	−18.49±0.82	41.15±0.89	40.46±0.54	63.71±1.10	−17.78±0.78	65.31±0.55	25.57±0.31
+ numeric (33.3%)	57.16±0.75	−15.09±1.52	42.04±1.12	37.60±0.59	64.05±1.02	−17.99±1.09	65.38±0.70	23.56±0.40
+ word (train)	54.03±0.50	−35.77±1.46	34.82±0.74	55.56±0.68	62.89±0.37	−17.14±1.36	64.38±0.53	36.15±0.35
+ word (equal)	55.54±0.41	−22.75±2.13	39.06±0.52	44.56±2.17	63.10±0.21	−17.90±0.67	64.46±0.55	28.40±1.47
OPT-IML 1.3B								
base	38.38±1.02	−49.30±0.37	20.94±1.70	−15.90±0.85	41.66±0.96	46.52±0.93	52.53±0.55	37.24±0.71
+ numeric (train)	35.60±0.18	98.12±1.53	40.97±0.49	−83.36±0.57	20.34±0.82	−5.54±1.60	45.50±0.79	62.34±0.85
+ numeric (33.3%)	37.67±1.58	99.63±2.07	42.38±1.02	−79.48±0.49	24.25±2.53	−9.55±1.62	46.36±2.12	62.89±1.13
+ word (train)	42.82±1.61	−25.73±4.47	29.76±2.32	−14.77±2.38	44.09±1.45	29.25±2.97	54.61±1.23	23.25±2.38
+ word (equal)	41.46±0.66	−19.59±4.04	29.33±1.00	−40.14±2.36	40.15±1.19	44.18±2.98	54.91±0.47	34.63±2.38
LLaMA 2-Chat 7B								
base	47.92±0.36	51.40±1.53	42.15±0.57	49.78±1.57	66.14±0.40	−73.83±0.27	35.47±0.16	58.33±0.20
+ numeric (train)	50.87±0.39	−59.06±1.34	22.27±0.61	5.84±0.70	62.49±0.51	36.96±0.59	67.85±0.23	33.95±0.81
+ numeric (33.3%)	51.87±0.09	−52.49±1.49	26.31±0.46	31.76±0.68	63.44±0.54	12.66±0.56	65.86±0.04	32.31±0.90
+ word (train)	53.79±0.48	56.57±1.42	44.76±0.42	−6.68±0.76	59.04±0.74	−34.57±0.58	57.56±0.58	32.61±0.86
+ word (equal)	38.90±0.20	143.86±0.28	45.22±0.26	−91.13±0.76	12.57±1.16	−31.59±0.39	58.92±0.42	88.86±0.22
GPT 3.5								
base	43.42±0.57	160.28±2.39	48.12±0.41	−56.73±2.43	40.95±1.45	−69.22±0.20	41.20±0.24	95.41±1.55
+ feedback	38.17±0.73	124.69±2.11	44.78±0.9	7.24±1.75	56.75±1.31	−92.88±0.15	12.97±0.12	74.94±0.17
+ numeric (train)	42.52±0.29	143.31±2.28	46.39±0.30	−21.77±1.40	55.92±1.08	−83.88±0.66	25.26±0.95	82.99±1.40
+ numeric (33.3%)	39.05±0.56	173.53±1.92	46.90±0.24	−45.28±2.15	49.17±1.68	−87.13±0.27	21.07±0.56	101.98±1.27
+ word (train)	46.09±0.24	86.19±0.94	45.31±0.31	21.38±1.18	59.37±0.69	−76.64±0.27	33.59±0.35	61.40±0.18
+ word (equal)	41.15±0.44	165.51±1.58	47.66±0.19	−59.48±2.24	37.36±1.50	−70.80±0.6	38.40±0.37	98.60±1.07

(f) HateXplain

HASOC-2019 task2	Macro F_1	**hate speech** (50.5%/43.1%) bias	**hate speech** (50.5%/43.1%) $\mathbf{F_1}$	**offensive** (19.9%/24.7%) bias	**offensive** (19.9%/24.7%) $\mathbf{F_1}$	**profane** (29.5%/32.3%) bias	**profane** (29.5%/32.3%) $\mathbf{F_1}$	$\mathbf{bias_{agg}}$
Flan-T5 XL								
base	55.97±1.11	1.61±2.13	67.46±1.85	−45.07±2.82	26.08±1.49	32.26±1.86	74.38±0.06	26.31±1.48
+ numeric (train)	50.84±1.07	10.75±1.68	65.55±1.42	−78.40±0.81	15.43±2.56	45.52±2.48	71.54±0.64	44.89±0.56
+ numeric (33.3%)	50.53±0.13	−3.23±2.14	63.92±3.17	−69.48±3.54	17.21±3.37	57.34±1.24	70.47±0.62	43.35±0.93
+ word (train)	56.40±1.37	−12.63±3.06	61.97±1.88	−32.86±2.93	32.56±3.06	41.94±1.86	74.68±2.10	29.14±0.69
+ word (equal)	55.94±1.47	4.84±1.61	66.66±2.24	−55.40±2.15	27.29±2.35	35.84±0.62	73.86±0.94	32.03±1.09
OPT-IML 1.3B								
base	26.44±0.58	1.35±1.68	46.46±0.59	128.64±2.94	32.85±1.21	−100.00±0.00	0.00±0.00	76.66±0.42
+ numeric (train)	13.18±0.00	−100.00±0.00	0.00±0.00	305.63±0.00	39.55±0.00	−100.00±0.00	0.00±0.00	168.54±0.00
+ numeric (33.3%)	13.18±0.00	−100.00±0.00	0.00±0.00	305.63±0.00	39.55±0.00	−100.00±0.00	0.00±0.00	168.54±0.00
+ word (train)	25.78±0.16	−23.12±3.05	41.92±1.70	170.89±5.86	35.43±1.64	−100.00±0.00	0.00±0.00	98.00±2.97
+ word (equal)	14.23±0.86	−97.04±1.23	3.11±2.70	300.47±2.15	39.59±0.54	−100.00±0.00	0.00±0.00	165.84±1.13
LLaMA 2-Chat 7B								
base	32.20±1.95	−87.63±1.23	10.04±1.36	43.66±6.45	28.10±1.95	83.51±6.57	58.44±2.65	71.60±0.45
+ numeric (train)	38.04±2.14	−70.97±2.42	24.97±2.03	110.80±8.61	38.99±1.70	10.04±3.46	50.15±3.07	63.93±2.55
+ numeric (33.3%)	31.19±1.38	−63.71±2.13	30.37±1.56	184.04±5.69	38.87±0.76	−55.55±2.24	24.34±1.99	101.10±3.14
+ word (train)	40.02±1.41	−60.48±1.62	30.82±1.42	7.04±6.46	30.39±1.78	75.27±5.99	58.85±1.13	47.60±1.19
+ word (equal)	44.63±0.75	54.60±2.71	64.12±0.39	−50.23±4.53	18.79±1.48	−34.05±3.45	50.97±1.18	46.30±2.24
GPT 3.5								
base	49.29±0.92	−58.29±2.39	40.73±0.42	156.96±6.32	44.64±1.38	−43.52±3.05	62.51±2.07	86.26±3.91
+ feedback	33.71±0.61	68.92±1.95	63.92±0.60	0.47±4.53	21.98±2.41	−91.76±0.62	15.23±1.06	54.65±0.36
+ numeric (test)	35.24±0.56	105.69±2.82	65.25±0.80	−84.04±3.54	6.46±1.27	−75.63±1.64	34.00±0.60	88.45±2.44
+ numeric (train)	26.35±1.05	122.58±1.40	61.83±0.28	−95.77±1.41	0.89±1.54	−90.32±2.15	16.32±2.02	102.89±1.09
+ numeric (33.3%)	35.40±1.00	89.99±2.08	65.05±0.46	−54.93±1.41	8.41±1.13	−77.42±2.15	32.72±2.13	74.12±1.88
+ word (train)	24.85±0.25	112.64±1.68	57.09±0.72	−74.65±2.82	6.73±2.16	−93.19±0.62	10.73±1.10	93.49±1.52
+ word (equal)	35.83±1.75	−16.53±4.09	50.78±1.97	143.66±8.45	36.91±1.48	−87.82±1.24	19.79±2.00	82.67±4.53

(g) HASOC-2019 task 2

Table 8: Experimental Results. The percentage numbers after each label denote the label distribution in the training set (optionally the test as well where different). For example, in HASOC-2019 task2 dataset, **hate speech** (50.5%/43.1%) indicates that there are 50.5% hate speech samples in the training set, while 43.1% in the test set.