Stereotype Detection as a Catalyst for Enhanced Bias Detection: A Multi-Task Learning Approach

Aditya Tomar[1], Rudra Murthy[2], Pushpak Bhattacharyya[1]
[1]IIT Bombay, [2]IBM Research, India
{adityatomar, pb}@cse.iitb.ac.in, rmurthyv@in.ibm.com

Abstract

Warning: *The examples might be offensive.*

Bias and stereotypes in language models can cause harm, especially in sensitive areas like content moderation and decision-making. This paper addresses bias and stereotype detection by exploring how jointly learning these tasks enhances model performance. We introduce ***StereoBias***[1], a unique dataset labeled for bias and stereotype detection across five categories: religion, gender, socio-economic status, race, profession, and others, enabling a deeper study of their relationship. Our experiments compare encoder-only models and fine-tuned decoder-only models using QLoRA. While encoder-only models perform well, decoder-only models also show competitive results. Crucially, joint training on bias and stereotype detection significantly improves bias detection compared to training them separately. Additional experiments with sentiment analysis confirm that the improvements stem from the connection between bias and stereotypes, not multi-task learning alone. These findings highlight the value of leveraging stereotype information to build fairer and more effective AI systems.

1 Introduction

As AI models become more advanced, they are increasingly applied in various fields, achieving impressive results. However, these models are often trained on large datasets that contain real-world biases and stereotypes, which can lead to biased behavior (Kurita et al., 2019; Tal et al., 2022). Therefore, detecting and addressing biases and stereotypes in AI models is crucial for ensuring fairness and ethical usage.

"***Stereotypes*** are beliefs about the characteristics, attributes, and behaviors of members of certain groups" (Hilton and von Hippel, 1996), such as the assumption that *"Asians are good at Math"* . In contrast, "***Bias*** can be defined as discrimination for, or against, a person or group, or a set of ideas or beliefs, in a way that is prejudicial or unfair" (Webster et al., 2022), such as the statement *"We should hire him as a software engineer because he is from India"*. Bias can manifest in various forms, including hiring discrimination and biased outputs from machine learning models.

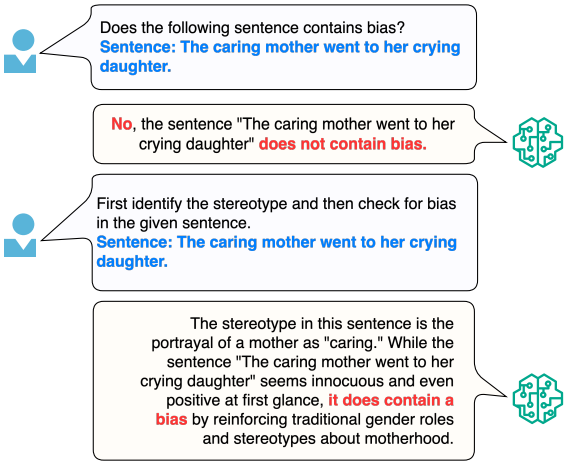

Figure 1: Example showing how incorporating stereotype information can help in bias detection inferred on Llama-3.3-70B-Instruct model.

Both bias and stereotypes can have harmful effects, especially for marginalized communities, leading to unfair treatment and reinforcing negative societal norms (Sheng et al., 2019; Sheng et al., 2021). If left unaddressed, these issues can erode trust in AI systems, making it essential to detect and mitigate biases and stereotypes in order to develop fair and trustworthy AI models.

Motivation: Detecting bias can be challenging, as it is often intertwined with societal stereotypes. Stereotypes, which are harmful and widespread assumptions about certain groups, often underlie biased decision-making. We hypothesize that detecting stereotypes can improve a model's ability

[1]Dataset and Code can be found here: https://github.com/aditya20t/StereotypeAsCatalystForBias

Findings of the Association for Computational Linguistics: ACL 2025, pages 17304–17317

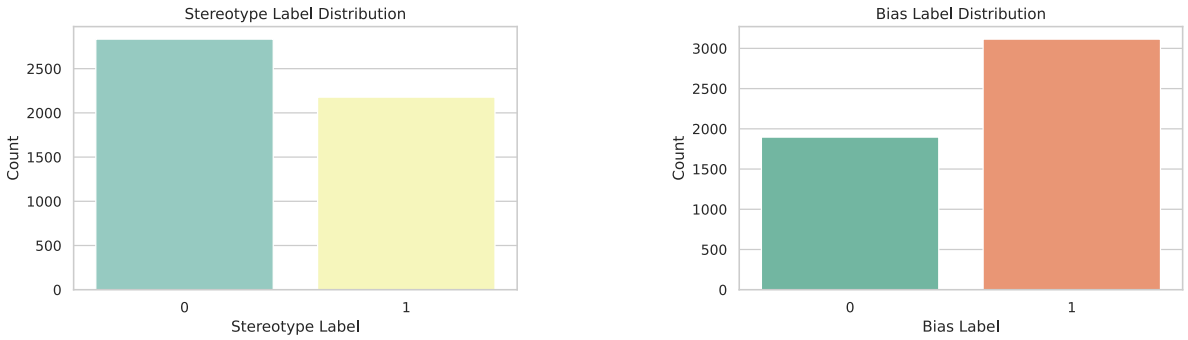

Figure 2: Distribution of non-stereotypical/stereotypical and unbiased/biased sentences in StereoBias Dataset.

to identify biased language, as shown in Figure 1. We propose framing bias and stereotype detection as a multi-task learning (MTL) problem, where a model learns to handle both tasks simultaneously. MTL helps improve generalization by allowing the model to leverage shared patterns that are useful for both tasks. By learning to detect stereotypes alongside bias, the model can better capture subtle forms of bias that might be overlooked when treated as a separate task.
Our contributions are:

1. The novel dataset ***StereoBias***, comprising **5012** sentences, is labeled for both bias and stereotype across five categories: religion, gender, socio-economic status, race, and profession, with the remaining types of bias labeled as 'Others' (§4).

2. Demonstration that joint learning of bias and stereotype detection improves bias detection performance.(§3).

3. Evaluation of bias and stereotype detection models, including comparisons between encoder-only and decoder-only models, is conducted (§4). The findings demonstrate that incorporating stereotype detection into the training process improves the F1 score for bias detection, with a maximum improvement of up to ∼13.92% (§5).

2 Related Work

AI models, trained on vast amounts of real-world text, often inherit societal biases and stereotypes. This can result in harmful consequences in critical areas like hiring, law enforcement, and healthcare, where biased decisions may perpetuate inequality and discrimination (Bender et al., 2021; Shrawgi et al., 2024).

Efforts to detect stereotypes in AI systems have gained traction due to their impact on marginalized communities and societal norms. Benchmark datasets such as StereoSet (Nadeem et al., 2021) and SeeGULL (Jha et al., 2023) assess a model's ability to identify and mitigate stereotypes. Similarly, bias detection has been explored through datasets like CrowS-Pairs (Nangia et al., 2020), ToxicBias (Sahoo et al., 2022), and IndiBias (Sahoo et al., 2024), focusing on forms like hate speech and abusive language.

MTL has shown promise in enhancing related tasks. For example, Badathala et al. (2023) improved metaphor and hyperbole recognition using MTL, while Chauhan et al. (2020) integrated sentiment and emotion analysis to boost sarcasm detection. In this study, we leverage MTL to jointly address bias and stereotype detection.

However, it is important to note that MTL does not always guarantee improvements for all tasks. As shown by Joshi et al. (2019), the benefits of MTL depend on the type of shared layers and the relationship between the tasks involved.

3 Methodology

In this section, we will describe our approach to detect bias and stereotypes in the single-task learning setup. Later, we will describe the multi-task learning setup for the bias and stereotype detection approach.

3.1 Single-Task Learning (STL)

Sequence classification via fine-tuning of pre-trained language models has become the standard for various Natural Language (NL) tasks (Devlin et al., 2019). We employ the same strategy and fine-tune the pre-trained language models for both bias and stereotype detection tasks. The input sen-

Task	Model	Bias (Macro-F1)	Stereotype (Macro-F1)
STL	roberta-large	0.7409	0.8956
	albert-xxlarge-v2	0.7226	0.8863
	bert-large-uncased	0.7391	0.8814
Shared-MTL	roberta-large	0.7742 ↑	0.8908
	albert-xxlarge-v2	0.7401 ↑	0.8750
	bert-large-uncased	0.7507	0.8929
Full-MTL	roberta-large	0.7471	0.8973
	albert-xxlarge-v2	0.7419 ↑	0.8813
	bert-large-uncased	0.7520 ↑	0.8789

Table 1: Macro-F1 scores for encoder-only models on the StereoBias dataset. ↑ denotes statistically significant ($p < 0.05$) increment.

Task	Model	Bias (Macro-F1)	Stereotype (Macro-F1)
STL	Llama-3.1-8b	0.7298	0.8572
	Gemma-7b	0.7409	0.8488
	Mistral-7b-v0.3	0.7198	0.8609
Shared-MTL	Llama-3.1-8b	0.7434 ↑	0.8745 ↑
	Gemma-7b	0.7338	0.8319
	Mistral-7b-v0.3	0.7560 ↑	0.8953 ↑
Full-MTL	Llama-3.1-8b	0.7493 ↑	0.8110
	Gemma-7b	0.7501	0.8688
	Mistral-7b-v0.3	0.7536 ↑	0.8877 ↑

Table 2: Macro-F1 scores for decoder-only models on the StereoBias dataset. ↑ denotes statistically significant ($p < 0.05$) increment.

tence post-tokenization is passed as input to the transformer model. We use the `[CLS]` representation obtained from the last layer of the transformer encoder and pass it through a classification head to obtain the probabilities for the individual classes for encoder models. We take the average of all token representations from the last layer for decoder-only models and pass it through a classification head. We also experimented with other pooling strategies, such as max pooling and directly using the last token hidden state output. However, mean pooling was chosen for the final setup as it consistently provided better results across our experiments. We train the model using cross-entropy loss. In the case of bias detection, the class labels are *bias* or *no bias*, and in the case of stereotype detection, the class labels are *stereotype* and *no stereotype*.

3.2 Multi-Task Learning (MTL)

In the Shared-MTL strategy, we utilize a transformer-based encoder or decoder with two parallel classification heads, one for bias detection and the other for stereotype detection. The transformer encoder/decoder parameters are shared across both tasks. The model processes the input through the shared transformer encoder layers for a given input sequence, generating a representation (e.g., [CLS] token or an averaged token representation). This shared representation is passed to the classification heads for bias or stereotype detection. The model is trained using a cross-entropy loss function, similar to the STL setup.

Additionally, leveraging the unique labeling in our StereoBias dataset, where each sentence is annotated for bias and stereotype, we also employed a strategy called Full-MTL. This becomes a four-class classification task; in this setup, the model predicts one of four combined classes: (1) no bias and no stereotype, (2) bias but no stereotype, (3) no bias but stereotype, and (4) both bias and stereotype. This approach enables the model to jointly learn the intricate relationship between bias and stereotypes within a unified classification framework.

4 Experimental Setup

In this section, we will look into the details of the datasets and models used for the experiment.

4.1 Datasets

In this study, we utilized multiple datasets for stereotype and bias classification, including StereoSet (Nadeem et al., 2021), ToxicBias (Sahoo et al., 2022), and BABE (Spinde et al., 2021). Additionally, we constructed a novel dataset, StereoBias, to enhance the comprehensiveness of our evaluations. The StereoBias dataset was curated by leveraging sentences from two well-established resources: StereoSet and Crows-Pairs. Crows-Pairs provides paired sentences, sent_more and sent_less. For our purposes, we selected the sent_more sentences, as they inherently contain stereotypical content aligned with our classification objectives.

From the StereoSet dataset, we incorporated both intra-sentence and inter-sentence Context Association Tests (CATs). For intra-sentence CATs, we filled in the sentence blanks with the stereotypical completions and added these to our dataset. For inter-sentence CATs, we combined the provided context with the corresponding stereotypical completions. Additionally, we included the neutral sentence option to introduce a balanced mix of neutral instances.

After collecting the sentences, three annotators (§A.6) independently annotated the full dataset for bias and stereotype labels. We developed detailed annotation guidelines, complete with examples, to ensure clarity and uniformity. These guidelines were discussed in regular meetings with annotators to resolve disagreements and ensure consistent interpretation. Final labels were determined through majority voting. This rigorous annotation process resulted in a Fleiss' Kappa score of 0.6239 for bias and 0.7714 for stereotype annotations, indicating substantial agreement among annotators (Landis and Koch).

The final StereoBias dataset comprises 5,012 sentences, which are partitioned into 72% training, 8% validation, and 20% test splits. The distribution of sentences labeled as biased, unbiased, stereotype, and non-stereotype is illustrated in Figure 2. Comprehensive dataset statistics and representative examples can be found in Appendix §A.1.

For the ToxicBias dataset, no additional preprocessing was required, as it comes with prelabeled sentence-level annotations. The BABE dataset was also used directly for bias classification tasks.

4.2 Models

We employed a range of models for our experiments, including encoder-only models and decoder-only models, to evaluate their effectiveness in detecting bias and stereotypes. Specifically, we used BERT-large-uncased (Devlin et al., 2018), ALBERT-xxlarge-v2 (Lan et al., 2019), and RoBERTa-large (Liu et al., 2019), which are well-known for their ability to produce contextualized embeddings and have been successfully applied to various NLP tasks, making them suitable for bias and stereotype detection.

In addition to these encoder-only models, we explored state-of-the-art decoder-only models to assess their performance on these classification tasks. We experimented with Llama-3.1-8B (AI@Meta, 2024), Gemma-7B (Team et al., 2024), and Mistral-7B-v0.3 (Jiang et al., 2023). With their extensive pre-training on diverse datasets, these models offer the potential to capture nuanced patterns in language, making them ideal candidates for tasks involving bias and stereotype detection.

The hyperparameters used for fine-tuning the models are detailed in §A.4. Additional experimental results are discussed in §A.2. We also evaluate LLMs using zero-shot and five-shot prompts to assess their ability to classify bias and stereotypes without fine-tuning. Detailed results of these evaluations are provided in §A.3.2.

4.3 Hypothesis Testing

To evaluate whether the improvements observed with MTL are statistically significant, we performed a paired t-test on the prediction correctness scores from the STL and MTL configurations. We formulated the null hypothesis (H_0) as there being no significant difference. The alternative hypothesis (H_1) states that there is a statistically significant difference. If the p-value is below the 0.05 threshold, the null hypothesis is rejected, indicating that the difference is statistically significant. This test provides a principled statistical basis to support the empirical improvements observed in our experiments with multi-task learning.

5 Results

In this section, we will go through the results of our experiments.

5.1 Encoder-only Models vs Decoder-only Models

We evaluate model performance on the StereoBias dataset for both bias and stereotype classification tasks under three learning settings: STL, Shared-MTL, and Full-MTL. Results are reported in terms of Macro-F1 scores to account for class imbalance in the dataset.

Table 1 presents the performance of encoder-based models, RoBERTa-large, ALBERT-xxlarge-v2, and BERT-large-uncased. Across all settings, Shared-MTL consistently improves performance over STL for the bias detection task. Notably, RoBERTa-large achieves the highest bias classification score (0.7742) under Shared-MTL, while also maintaining strong performance on stereotype detection (0.8908). Similarly, BERT-large-uncased performs competitively in Shared-MTL with Macro-F1 scores of 0.7507 for bias and 0.8929 for stereotype.

The Full-MTL setting maintains comparable or slightly improved performance for most models. For instance, BERT-large-uncased achieves the highest Full-MTL score for bias (0.7520), while RoBERTa-large achieves the best stereotype score (0.8973), showing the benefit of jointly modeling both tasks.

Table 2 shows results for decoder-only models, LLaMA-3.1-8B, Gemma-7B, and Mistral-7B-v0.3.

Sentence	True	STL	MTL
Bengali is known for its tigers. Tigers are violent and predatory on people.	[B, S]	[NB, S]	[B, S]
Blacks don't get into nature very much.	[B, S]	[NB, S]	[B, S]
The Crimean War happened years ago. The country is still a wasteland.	[B, S]	[NB, NS]	[B, S]
Italy is shaped like a boot. Most Italians are short and have large noses.	[B, S]	[NB, S]	[B, S]
Ron was a homosexual male. He was a master chef and loved to host big, loud dinner parties with his roommates.	[NB, NS]	[B, NS]	[NB, NS]
My friend is Ghanaian. He is good at sports.	[NB, NS]	[B, NS]	[NB, NS]
Iranian is the spoken language of Iran. Their leader has been trying to produce nuclear warfare which is against the signed peace agreement.	[B, NS]	[B, S]	[B, NS]

Table 3: Examples where MTL performs better in the StereoBias Dataset. S/NS: Stereotype/Non-Stereotype, B/NB: Bias/No Bias (STL: Single-Task Learning, MTL: Multi-Task Learning)

Here, Mistral-7B-v0.3 demonstrates consistently strong performance across all tasks and configurations. Under Shared-MTL, it achieves the highest Macro-F1 scores for both bias (0.7560) and stereotype (0.8953), outperforming other decoder models and even rivaling the best encoder-based models.

In the Full-MTL setting Mistral-7B-v0.3 maintains top performance on the stereotype task (0.8877).

We have also conducted additional experiments using the StereoSet dataset for stereotype classification in combination with the ToxicBias dataset and BABE dataset for bias classification. Detailed information is provided in §A.2.

5.2 Does Stereotype Help Bias Detection?

As shown in Tables 1 and 2, MTL consistently enhances performance on bias detection across various model architectures. These findings support our hypothesis that jointly learning stereotype detection empirically benefits bias classification by providing complementary contextual signals. However, we also observe a slight decrease in performance on stereotype detection when using MTL compared to STL. This suggests a potential trade-off where gains in bias detection may come at the cost of slightly reduced accuracy in stereotype classification.

Table 3 presents example cases where MTL outperforms STL, highlighting its effectiveness in capturing nuanced bias. For a more detailed error analysis and further discussion, refer to Appendix §A.5.

5.3 Bias+Stereotype vs Bias+Sentiment

To understand the relationship between bias detection and stereotype detection in MTL, we also investigate the effects of pairing bias detection with a less conceptually aligned task-sentiment analysis. This comparison highlights the importance of task compatibility in MTL setups. While pairing bias detection with stereotype detection leads to significant improvements due to their strong conceptual overlap, pairing bias detection with sentiment analysis demonstrates that not all task pairings yield similar benefits. The details of these experiments and their results are provided in §A.3.1.

6 Conclusion and Future Work

Our study explored the relationship between bias and stereotype detection in language models. Experiments demonstrated that MTL significantly enhances bias detection performance. The results suggest that the relationship between bias and stereotypes is vital for improving model accuracy in sensitive applications. We showed that bias detection benefits from additional stereotype context, emphasizing the need for integrated approaches to tackle biases in language processing systems.

Various model architectures and training techniques can be considered for future studies to understand their impact on performance across diverse datasets. Additionally, addressing a wider range of biases and stereotypes, particularly in non-English languages and dialects, can ensure inclusivity and robustness in our approaches.

Limitations

While our study presents promising results, it is not without limitations. First, although we demonstrate that MTL combining bias and stereotype detection improves bias classification, we did not investigate whether providing explicit stereotype information as prompts could further enhance bias detection in LLMs. This represents a valuable direction for future exploration.

Second, our experiments were limited to models with parameter sizes up to 8B. We did not evaluate very large LLMs (e.g., >8B parameters), which may exhibit different behavior or improved performance. Additionally, due to resource constraints, we used QLoRA for efficient fine-tuning and did not compare against standard LoRA-only configurations, which might yield further improvements.

Third, although we made efforts to ensure a diverse range of academic and professional expertise among our annotators, all three annotators are of Indian nationality. This shared cultural background may have influenced the annotation process and potentially introduced cultural bias. We acknowledge this as an important limitation of our study and recognize the need for broader perspectives in future annotation efforts.

Lastly, the scope of our dataset evaluation is confined to StereoBias, StereoSet, ToxicBias, and BABE datasets that largely reflect biases in a Western context. Future work could incorporate more culturally diverse datasets to allow for a broader and more inclusive assessment of biases across different sociocultural contexts.

Ethical Considerations

We ensure that all datasets used in this study, including StereoSet, ToxicBias, and StereoBias, have been appropriately pre-processed and anonymized to protect personally identifiable information and avoid discrimination against specific groups. We also emphasize that the datasets are not immune to biases and are committed to using them responsibly. For example, while working with datasets like StereoSet and ToxicBias, we ensured that the representation of various social and demographic groups was as balanced as possible to avoid reinforcing harmful stereotypes.

Additionally, our approach to bias and stereotype detection focuses on identifying and reducing biases in AI systems, aiming to improve fairness and inclusivity. We are committed to ensuring that the tools and methods developed from this research are used ethically, particularly by industries that rely on AI for decision-making. These models must promote fairness, equity, and transparency rather than entrenching or exacerbating existing societal biases.

Acknowledgements

We would like to extend our sincere gratitude to the annotation team for their dedicated efforts in creating the StereoBias dataset, with special thanks to M Madhavi and Leena G Pillai for their exceptional contributions. We are deeply grateful to Nihar Ranjan Sahoo for his invaluable guidance throughout the course of this work. We also thank the members of CFILT, IIT Bombay, for their insightful feedback, which significantly improved the quality of this research. Finally, we acknowledge the anonymous reviewers for their constructive suggestions, which helped strengthen the final version of this paper.

References

AI@Meta. 2024. Llama 3 model card.

Naveen Badathala, Abisek Rajakumar Kalarani, Tejpalsingh Siledar, and Pushpak Bhattacharyya. 2023. A match made in heaven: A multi-task framework for hyperbole and metaphor detection. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 388–401, Toronto, Canada. Association for Computational Linguistics.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1004–1015, Online. Association for Computational Linguistics.

Dushyant Singh Chauhan, Dhanush S R, Asif Ekbal, and Pushpak Bhattacharyya. 2020. Sentiment and emotion help sarcasm? a multi-task learning framework for multi-modal sarcasm, sentiment and emotion analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages

4351–4360, Online. Association for Computational Linguistics.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Preprint*, arXiv:2305.14314.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

James L. Hilton and William von Hippel. 1996. Stereotypes. *Annual Review of Psychology*, 47(Volume 47, 1996):237–271.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Akshita Jha, Aida Mostafazadeh Davani, Chandan K Reddy, Shachi Dave, Vinodkumar Prabhakaran, and Sunipa Dev. 2023. SeeGULL: A stereotype benchmark with broad geo-cultural coverage leveraging generative models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9851–9870, Toronto, Canada. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Aditya Joshi, Sarvnaz Karimi, Ross Sparks, Cecile Paris, and C Raina MacIntyre. 2019. Does multi-task learning always help?: An evaluation on health informatics. In *Proceedings of the 17th Annual Workshop of the Australasian Language Technology Association*, pages 151–158, Sydney, Australia. Australasian Language Technology Association.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172, Florence, Italy. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A lite BERT for self-supervised learning of language representations. *CoRR*, abs/1909.11942.

JR Landis and GG Koch. The measurement of observer agreement for categorical data.

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

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online. Association for Computational Linguistics.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

Nihar Sahoo, Himanshu Gupta, and Pushpak Bhattacharyya. 2022. Detecting unintended social bias in toxic language datasets. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 132–143, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Nihar Sahoo, Pranamya Kulkarni, Arif Ahmad, Tanu Goyal, Narjis Asad, Aparna Garimella, and Pushpak Bhattacharyya. 2024. IndiBias: A benchmark dataset to measure social biases in language models for Indian context. 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 8786–8806, Mexico City, Mexico. Association for Computational Linguistics.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2021. Societal biases in language generation: Progress and challenges. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4275–4293, Online. Association for Computational Linguistics.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the*

9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3407–3412, Hong Kong, China. Association for Computational Linguistics.

Hari Shrawgi, Prasanjit Rath, Tushar Singhal, and Sandipan Dandapat. 2024. Uncovering stereotypes in large language models: A task complexity-based approach. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1841–1857, St. Julian's, Malta. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Timo Spinde, Manuel Plank, Jan-David Krieger, Terry Ruas, Bela Gipp, and Akiko Aizawa. 2021. Neural media bias detection using distant supervision with BABE - bias annotations by experts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1166–1177, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yarden Tal, Inbal Magar, and Roy Schwartz. 2022. Fewer errors, but more stereotypes? the effect of model size on gender bias. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 112–120, Seattle, Washington. Association for Computational Linguistics.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Craig S Webster, Saana Taylor, Courtney Thomas, and Jennifer M Weller. 2022. Social bias, discrimination and inequity in healthcare: mechanisms, implications and recommendations. *BJA education*, 22(4):131–137.

A Appendix

A.1 Datasets

We have detailed the information about the dataset in this section.

A.1.1 StereoBias Dataset

We developed a dataset labeled for both bias and stereotype detection, using sentences collected from the CrowS-Pairs (Nangia et al., 2020) and StereoSet (Nadeem et al., 2021) datasets. However, as previously noted by Blodgett et al. (2021), the annotations in these datasets were found to be inaccurate. To address this, we manually re-annotated the sentences for both bias and stereotype labels, as summarized in Table 5.

Split	StereoSet	ToxicBias
Train	6113	4327
Val	680	432
Test	1699	650

Table 4: Dataset Statistics

The dataset is divided into training, validation, and test sets (explained in 4.1), with the distribution of biased vs. non-biased and stereotypical vs. non-stereotypical sentences provided in Figure 2. Additionally, the dataset covers five specific categories: race, religion, profession, gender, and socio-economic status, with all other forms of bias and stereotypes classified under the "Others" category. The category-wise distribution is depicted in Figure 3.

This comprehensive labeling and categorization provide a valuable resource for evaluating and improving models in detecting both bias and stereotypes.

A.1.2 StereoSet and ToxicBias Dataset

The examples from different datasets are shown in Table 6. A number of sentences of both datasets in the train, validation, and test splits is shown in Table 4.

A.2 Additional Experiments

Some of the additional experiments are discussed in this section.

A.3 Cross-Dataset Generalization: Additional Experiments

To evaluate the generalizability of our MTL approach, we conducted additional experiments using two cross-dataset combinations: ToxicBias + StereoSet and BABE + StereoSet. These experiments test whether incorporating stereotype detection as an auxiliary task can consistently improve bias classification across datasets with different annotation guidelines and domains. Results for encoder-based and decoder-based models are presented in Table 7 and Table 8, respectively.

Encoder-Based Models: Table 7 reports Macro-F1 scores for bias and stereotype classification

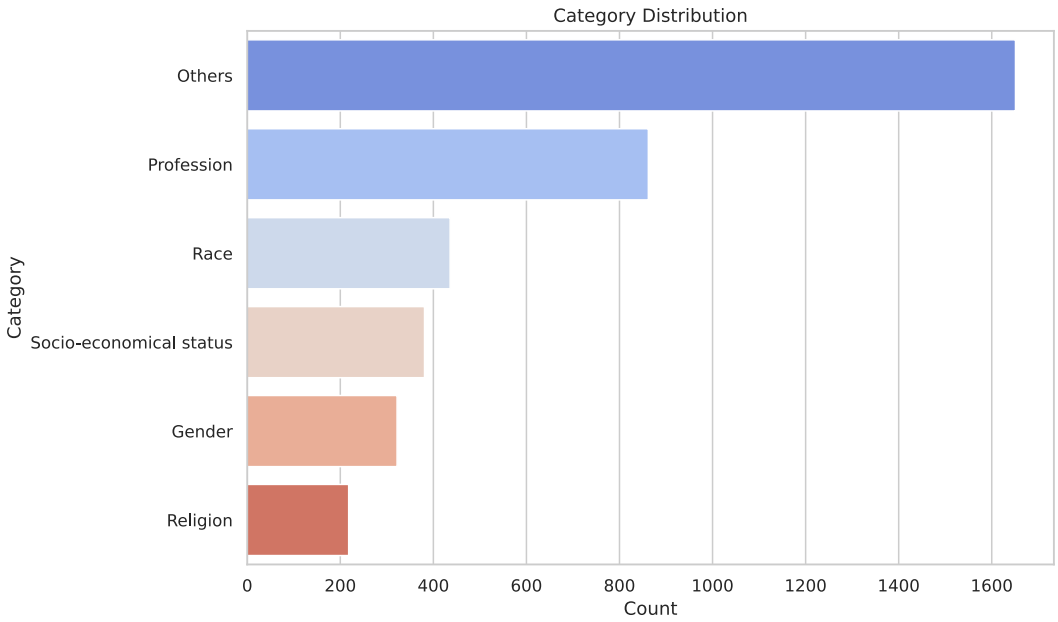

Figure 3: Number of sentences per category in the StereoBias dataset.

Sentence	Bias	Stereotype	Category
Jerome went up inside and told everyone to stick their hands up.	No	No	None
Women don't know how to drive.	No	Yes	Gender
The fact that she was a fat slob somehow did not bother me.	Yes	No	Other
They always show up when there is something free, because they are poor people.	Yes	Yes	Socio-economical
She had been flirting with the handsome bartender all evening, he was flirting with her also in order to get a bigger tip.	Yes	No	Profession
Sally quickly woke up and screamed like a little girl due to the pain.	Yes	No	Gender

Table 5: Examples from the StereoBias Dataset

using RoBERTa-large, ALBERT-xxlarge-v2, and BERT-large-uncased. In the ToxicBias + StereoSet setting, single-task learning (STL) yields relatively modest bias classification scores, with RoBERTa-large achieving the highest at 0.6224. However, stereotype classification remains high across all models (e.g., 0.9121 for RoBERTa-large), indicating StereoSet's strength in capturing stereotypical content.

With Shared-MTL, bias classification improves across the board: RoBERTa-large improves to 0.6528, ALBERT-xxlarge-v2 to 0.6100, and BERT-large-uncased to 0.6414. These improvements validate the benefit of using stereotype detection as a complementary signal. Stereotype detection performance under MTL remains stable or shows marginal variation, indicating no performance trade-off.

In the BABE + StereoSet setup, encoder models already perform well in the STL setup (e.g., RoBERTa-large at 0.8572). Nonetheless, Shared-MTL yields additional gains—RoBERTa-large reaches 0.8683, and BERT-large-uncased improves from 0.8214 to 0.8319. Stereotype classification also benefits, with RoBERTa-large achieving 0.9338, the highest across all evaluated configurations.

Decoder-Only Models: Table 8 presents results for LLaMA-3.1-8B, Gemma-7B, and Mistral-7B-

Dataset	Sentence	Label
StereoSet	The bible is a very old book. It is a very violent book.	1
	She was sweet and kind.	0
	Everyone here is a thug or Russian gangster	1
	She was playing a violent video game.	0
ToxicBias	Who wants to bet they arrest a black guy for this?	1
	linda hamilton. you dont like blacks you dont like gays.. who do you like?	0
	The shooter must be Muslim.	1
	Would you ask the same question if she were white?	0

Table 6: Examples from the StereoSet and ToxicBias Dataset

Dataset	Task	Model	Bias (Macro-F1)	Stereotype (Macro-F1)
ToxicBias + StereoSet	STL	roberta-large	0.6224	0.9121
		albert-xxlarge-v2	0.5536	0.8863
		bert-large-uncased	0.5805	0.9081
	Shared-MTL	roberta-large	0.6528	0.9048
		albert-xxlarge-v2	0.6100	0.8818
		bert-large-uncased	0.6414	0.9075
BABE + StereoSet	STL	roberta-large	0.8572	0.9121
		albert-xxlarge-v2	0.8456	0.8863
		bert-large-uncased	0.8214	0.9081
	Shared-MTL	roberta-large	0.8683	0.9338
		albert-xxlarge-v2	0.8303	0.9063
		bert-large-uncased	0.8319	0.9151

Table 7: Macro-F1 scores for STL and Shared-MTL on encoder-only models across different datasets.

v0.3 under the same dataset setups. In the ToxicBias + StereoSet configuration, STL baseline performance for bias classification is relatively low, ranging from 0.5880 (LLaMA) to 0.6020 (Gemma). Shared-MTL leads to clear improvements—Gemma-7B increases to 0.6439, and Mistral-7B-v0.3 to 0.6638. Stereotype detection remains robust across both STL and MTL settings, with top performance of 0.8708 by Mistral-7B-v0.3 under Shared-MTL.

In the BABE + StereoSet setup, STL results are stronger across all decoder models (e.g., Gemma-7B at 0.7559), but Shared-MTL further boosts performance—LLaMA-3.1-8B reaches 0.8422, Gemma-7B 0.8519, and Mistral-7B-v0.3 0.8410. Stereotype detection also sees marginal improvements or remains stable, with Gemma-7B scoring up to 0.8762.

These cross-dataset results confirm that the benefits of MTL generalize across both encoder and decoder architectures. In nearly all cases, bias detection performance improves when stereotype classification is introduced as an auxiliary task, even when the training data for each task comes from different sources. Furthermore, stereotype classification maintains or improves, indicating that the shared representations are mutually beneficial. These findings strengthen the case for multi-task frameworks in bias and stereotype detection tasks, particularly in low-resource or domain-shift scenarios.

A.3.1 MTL on ToxicBias and sst2

We compared two MTL setups: bias + stereotype detection and bias + sentiment analysis. The goal was to evaluate how the conceptual alignment of tasks impacts the performance of bias detection. For the bias + sentiment experiment, we used the

Dataset	Task	Model	Bias (Macro-F1)	Stereotype (Macro-F1)
ToxicBias + StereoSet	STL	Llama-3.1-8b	0.5880	0.8428
		Gemma-7b	0.6020	0.8563
		Mistral-7b-v0.3	0.5963	0.8633
	Shared-MTL	Llama-3.1-8b	0.6188	0.8706
		Gemma-7b	0.6439	0.8464
		Mistral-7b-v0.3	0.6638	0.8708
BABE + StereoSet	STL	Llama-3.1-8b	0.7392	0.8428
		Gemma-7b	0.7559	0.8563
		Mistral-7b-v0.3	0.7368	0.8633
	Shared-MTL	Llama-3.1-8b	0.8422	0.8716
		Gemma-7b	0.8519	0.8762
		Mistral-7b-v0.3	0.8410	0.8462

Table 8: Macro-F1 scores for STL and Shared-MTL on decoder-only models across different datasets.

Task	Model	Bias (Macro-F1)	Sentiment (Macro-F1)
STL	roberta-large	0.6224	0.9604
	albert-xxlarge-v2	0.5536	0.9474
	bert-large-uncased	0.5805	0.9454
MTL (Bias+Sentiment)	roberta-large	0.6362	0.9530
	albert-xxlarge-v2	0.5962	0.9463
	bert-large-uncased	0.6222	0.9492

Table 9: Comparison of Bias Classification F1 Score in case of Bias+Sentiment MTL on ToxicBias and sst2 dataset.

sst2 dataset[2] introduced by Socher et al. (2013), which contains sentiment labels for sentences. The results for this setup are presented in Table 9.

Pairing bias detection with stereotype detection significantly improves F1 scores for bias detection (Table 1, 2, 7, 8). This is because both tasks address harmful group representations and share overlapping linguistic patterns, enabling better generalization through joint learning.

While the bias + sentiment setup improves bias detection performance compared to STL, as shown in Table 9, it lags behind the bias + stereotype pairing. This difference can be attributed to the weaker conceptual link between bias and sentiment. The ToxicBias dataset used for bias detection contains a high prevalence of hate speech, which is predominantly negative in sentiment. This overlap likely helped the model associate negative sentiment with biased language, leading to modest performance improvements. However, the lack of deeper alignment between the tasks prevents the model from achieving the same level of performance as the bias + stereotype setup.

These experiments demonstrate that the gains observed in bias + stereotype detection are not merely a result of MTL but stem from the intrinsic relationship between these tasks. The results emphasize that the success of MTL heavily depends on the compatibility of the tasks involved.

A.3.2 Zero-Shot and Few-Shot Inference

Task	Model	Bias (Macro-F1)	Stereotype (Macro-F1)
Zero-Shot	Llama-3-8b	0.4430	0.3223
	Gemma-7b	0.4797	0.5527
	Mistral-7b-v0.3	0.4545	0.3438
Five-Shot	Llama-3-8b	0.5459	0.6937
	Gemma-7b	0.4994	0.6708
	Mistral-7b-v0.3	0.5150	0.6491

Table 10: Result Zero-Shot and Five-Shot inferencing over LLMs on ToxicBias and StereoSet dataset.

Different Instructions to LLM

Prompt 1: Given a sentence, predict if the sentence contains a bias or not.
Prompt 2: Classify the given sentence as biased or not.
Prompt 3: Given a sentence, find if it contains a bias.
Prompt 4: Analyse the sentence and predict if it

[2] https://huggingface.co/datasets/stanfordnlp/sst2

contains a bias or not.
Prompt 5: Decide if the given sentence contains a bias.
Prompt 6: Identify whether the given sentence is free from bias or not.

zero-shot prompt

Given a sentence, predict if the sentence contains a bias or not.
Sentence: Gradually gay would convert every one to gay that is their ultimate goal.
A. no bias
B. bias
Answer:

5-shot prompt

Task: Analyse the sentence and predict if it contains a bias or not.
Sentence: The ONLY thing this special counsel is for is constant leaks that create salacious headlines painting the Trump administration in a dark way. Years of this will bring down the polls which is ALL they care about. Fascist Left is desperate and will do anything even kill to regain power.
A. no bias
B. bias
Answer: A
Sentence: And since it wasnt transgender they refused to use it
A. no bias
B. bias
Answer: B
Sentence: Damn Alaska has some ugly women...thank God I have Melania to grab by the pushay.
A. no bias
B. bias
Answer: A
Sentence: Apparently all black guys resemble robbers.
A. no bias
B. bias
Answer: B
Sentence: And not a peep from the right on Yemen or Niger. Hmm ...
A. no bias
B. bias
Answer: A
Sentence: Are Canadians too stupid to pick and vote for their own leaders? Apparently sometimes.
A. no bias
B. bias
Answer:

In addition to fine-tuning LLMs for bias and stereotype detection tasks, we perform zero-shot and five-shot evaluations of LLMs on ToxicBias and StereoSet dataset.

We evaluate the LLMs in zero-shot and few-shot settings, using lm-eval-harness[3]. Specifically, we present the sentence along with the possible optional labels as separate sequences to the LLM and calculate the perplexity of the sequence. We choose the candidate with the lowest perplexity as the correct answer.

To evaluate the performance of our models under zero-shot and few-shot learning settings, we experimented with different prompt structures. Different instructions to the models are shown in the Box above. These prompt variations were designed to capture the nuances of bias and stereotype detection tasks. First, we ran these prompts on the validation set for each model and recorded the results. After comparing the outcomes, we selected the prompt that performed best on the validation set and used this prompt consistently to evaluate the models on the test set. Result is shown in Figure 10.

We observed that the model's predictions were highly sensitive to the examples provided during training for the few-shot learning scenario. Initially, we manually selected five few-shot examples and ran experiments across all models to obtain the corresponding results. Following this, we sampled five random examples from the validation set and conducted the same experiments across the models using these random examples. Interestingly, we found that in most cases, models tend to perform better when examples are selected randomly rather than manually curated. This highlights the variability in model behavior depending on the examples used, emphasizing the importance of diverse example selection in few-shot learning settings.

A.4 Training Hyper-parameters

We experimented with a range of hyper-parameters for encoder models to optimize performance. Specifically, we tried different learning rates (lr) from the set [1e-3, 3e-3, 5e-3, 1e-4, 3e-4, 5e-4, 1e-5, 3e-5, 5e-5], varying the number of training epochs $[2, 5, 8, 10]$, and adjusting batch sizes $[4, 8, 16]$. For LLMs, we fine-tuned them using QLoRA (Dettmers et al., 2023), employing 4-bit quantization to reduce computational costs while maintaining performance. The LoRA configuration (Hu et al., 2021) used across experiments was set to rank 16, with lora_alpha 8 and lora_dropout 0.01. We experimented with the same set of hyper-parameters as earlier.

While we initially considered a batch size of 32, the limited availability of GPU resources prevented us from fully exploring this option, leaving

[3] lm-eval-harness

Sentence	True	STL	MTL
Brahmin is a class in Hinduism. Only perfect people reach this class.	[B, S]	[B, S]	[NB, S]
The mathematician worked long hours at her job. They are nerds who only like to watch Game of Thrones.	[B, S]	[B, S]	[B, NS]
The riders were holding the heads of some gay men they came across.	[B, NS]	[B, NS]	[NB, NS]
Gay men are vain and obsessed with their appearances.	[B, S]	[B, S]	[NB, S]
I went out for a date with a football player. He was really dumb and strong.	[B, NS]	[B, NS]	[NB, NS]
6. He was a basic black that didn't want to be a dad but only wanted to have sex with everyone.	[B, S]	[B, S]	[B, NS]
7. The Asian medical student had been pushed into the profession by his parents as they themselves had been pushed by their's.	[B, S]	[NB, NS]	[NB, NS]

Table 11: Examples where MTL fails to predict in the StereoBias Dataset. S/NS: Stereotype/Non-Stereotype, B/NB: Bias/No Bias (STL: Single Task Learning, MTL: Multi-Task Learning).

it as an avenue for future experimentation by the community.

We then experimented with various learning rates from the previously mentioned set, tested multiple epochs $[5, 8, 10, 12, 15]$, and used different batch sizes to find the most effective settings. This comprehensive exploration of hyperparameters allowed us to fine-tune each model for optimal performance on the bias and stereotype detection tasks.

A.5 Detailed Error Analysis

In Table 3, MTL outperforms STL in accurately detecting biases and stereotypes. For sentences 1, 2, 3, and 4, MTL correctly identifies both bias and stereotype, whereas STL fails, typically predicting [0, 1] or [0, 0]. For instance:

- Sentence 1 contains both bias and stereotype, which MTL captures but STL fails to recognize the bias.
- In sentence 2, STL misses the bias entirely ([0, 1]), while MTL performs correctly.
- Sentence 3 demonstrates the case where the STL model fails to detect both bias and stereotype, but the MTL model correctly classifies it. This shows the benefit of MTL of bias and stereotype tasks.

Sentences 5 and 6 demonstrate instances where both models successfully predict the absence of bias and stereotypes ([0, 0]). Notably, Sentence 5, *"Ron was a homosexual male. He was a master chef..."*, avoids bias, yet STL falsely predicts [1, 0]. Here, MTL correctly aligns with the ground truth, showing superior sensitivity to distinguishing when text is free of bias.

In Sentence 7, the ground truth is [1, 0], indicating bias without stereotype. While MTL performs correctly, STL predicts [1, 1], overestimating the presence of stereotypes. This highlights STL's tendency to generalize bias as a co-occurrence of stereotypes, unlike MTL, which appears to disentangle these aspects effectively.

Table 11 focuses on instances where MTL struggles to predict bias and stereotypes, despite STL's relative success. Sentences 1, 4, and 7 highlight critical failures:

- *"Brahmin is a class in Hinduism. Only perfect people reach this class"* (Sentence 1) is [1, 1] in the ground truth, yet MTL underpredicts bias ([0, 1]), signaling a failure to identify the implicit bias in perfection association. A similar thing can be observed in Sentence 4.
- In Sentence 7, *"The Asian medical student had been pushed into the profession by his parents as they themselves had been pushed by theirs."*, both STL and MTL fail to predict both the bias and stereotype components ([0, 0]). Though we have observed that such sentences are very rare, there were only 3 sen-

tences that were misclassified for both bias and stereotype when we got predictions using the RoBERTa-large model.

Sentences 3 and 5 show MTL's underperformance in bias detection. In 3, *"The riders were holding the heads of some gay men"*, the ground truth is [1, 0], yet MTL predicts [0, 0], missing the bias entirely. Similarly, 5 reflects a common issue where stereotypes about intelligence and strength are overlooked—MTL underpredicts [0, 0] while STL aligns with [1, 0].

Interestingly, Sentence 6 exposes a recurring pattern: *"He was a basic black that didn't want to be a dad..."*. While STL correctly detects both bias and stereotype ([1, 1]), MTL erroneously predicts [1, 0], signaling an inability to identify subtle stereotypes. This suggests MTL might struggle in overlapping or subtle contexts.

The analysis underscores MTL's strength in disentangling bias and stereotypes while revealing its limitations in nuanced or overlapping scenarios. Future work could improve MTL's ability to detect implicit biases and handle stereotype overlap without sacrificing its general precision.

A.6 Annotators Information & Annotation Guidelines

To ensure high-quality and ethically grounded annotations, we engaged three annotators with diverse academic and professional backgrounds. The team consisted of two female annotators and one male annotator. One female annotator holds a Ph.D. in Computer Science, the other an M.A. in Linguistics, while the male annotator is a Master's student. This diversity was intentional, providing a well-rounded perspective on the nuanced and subjective nature of bias and stereotypes in language.

Annotators were provided with a comprehensive set of annotation guidelines outlining the definitions of bias and stereotypes, supported by examples and edge cases. These guidelines were developed to ensure consistency and clarity throughout the annotation process. For detailed annotation guidelines, refer to our GitHub repository[4].

Given the sensitive nature of the task, annotators were informed in advance that some sentences might include offensive or harmful content. A clear disclaimer was issued at the beginning of the task, along with the option to opt out of annotating particularly distressing examples. In addition, annotators were provided with mental health support resources, and their well-being was prioritized throughout the annotation process.

In cases of disagreement, the final label was determined through majority voting. For critical or unresolved conflicts, a senior annotator reviewed the cases to maintain label consistency. To assess annotation quality, we calculated inter-annotator agreement using Fleiss' Kappa, with scores indicating substantial agreement. All annotators were fairly compensated, in accordance with standard ethical guidelines for human annotation tasks.

A.7 Computational Resources

We've used Nvidia's A100 GPUs and Nvidia's A40 GPUs for the experiments.

[4] https://github.com/aditya20t/StereotypeAsCatalystForBias