Efficient Evaluation of Bias in Large Language Models through Prompt Tuning

Jacob-Junqi Tian[1] **David B. Emerson**[2] **Deval Pandya**[2]
Laleh Seyyed-Kalantari[3] **Faiza Khan Khattak**[2]
[1]McGill University, Montreal, QC, Canada
[2]Vector Institute for AI, Toronto, ON, Canada
[3]York University, Toronto, ON, Canada
jacob.tian@mail.mcgill.ca, lsk@yorku.ca,
{david.emerson, deval.pandya faiza.khankhattak}@vectorinstitute.ai

Abstract

Prompting large language models (LLMs) has gained substantial popularity as pre-trained LLMs are capable of performing downstream tasks without requiring large quantities of labelled data [11]. It is, therefore, natural that prompting is also used to evaluate biases exhibited by these models. However, achieving good task-specific performance often requires manual prompt optimization. In this paper, we explore the use of soft-prompt tuning to quantify the biases of LLMs such as OPT[26] and LLaMA [24]. These models are trained on real-world data with potential implicit biases toward certain groups. Since LLMs are increasingly used across many industries and applications, it is crucial to accurately and efficiently identify such biases and their practical implications.

In this paper, we use soft-prompt tuning to evaluate model bias across several sensitive attributes through the lens of *group fairness (bias)*. In addition to improved task performance, using soft-prompt tuning provides the advantage of avoiding potential injection of human bias through manually designed prompts. Probing with prompt-tuning reveals important bias patterns, including disparities across age and sexuality. We open-source the pipeline and encourage researchers to adapt this work to their use cases.[1]

1 Introduction

Despite widespread and successful utilization, fine-tuned language models (LMs) have several drawbacks. These include requiring significant compute resources for training, large quantities of labelled data, and separate training and storage for each downstream task [8, 25]. Language model prompting addresses some of these downsides, but the task of designing prompts to induce optimal performance for a given downstream application is challenging [13, 18]. Significant progress has been made in automatic prompt engineering methods. One such method for automatic prompt optimization is soft-prompt tuning, a parameter-efficient fine-tuning (PEFT) method that trains a small set of prompt-token embeddings to be provided along with the standard natural language input. For various LLMs, soft-prompt tuning has been shown to match, or nearly match, fine-tuning performance for a wide range of tasks including classification, summarization, and question-answering [9, 12].

On the other hand, the existence of potentially harmful biases exhibited by popular LMs is well-documented [6, 23, 1, 15] and quite common. Bias quantification has gained substantial attention from the research community recently [16, 20]. As LLM applications continue to rapidly expand,

[1] https://github.com/VectorInstitute/JAXPromptTuning/tree/main

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

developing comprehensive analytical frameworks to measure the learned or inherited social biases of such models is imperative.

In this paper, we evaluate the utility of soft-prompt tuning for bias evaluation of LLMs, including OPT [26] and LLaMA language models [24]. More specifically, the approach presented here leverages optimized soft-prompts to condition models toward the completion of sentiment analysis tasks on which fairness (bias) metrics are subsequently measured. In addition to the method's efficiency in terms of tuning fewer parameters, another advantage of soft-prompt tuning is that it does not require manual prompt design- a potential source of human bias. The experiments demonstrate that prompt-tuning enables fine-grained analysis and an overall understanding of an LLM's bias with respect to sensitive attributes and across protected groups. This paper's contributions are as follows:

- To our knowledge, this is the first application of soft-prompt tuning for fairness evaluation. We demonstrate that the approach constitutes an effective and efficient approach for such evaluation.
- We show that LLMs such as OPT and LLaMA exhibit measurable biases across protected groups within the sensitive attributes of age, sexuality, and disability. Furthermore, such biases are generally consistent across model size, type, and prompt-tuning dataset.
- The bias metrics of positive and negative false-positive rate gaps are explored here. However, the approach is compatible with other fairness metrics, including the comprehensive fairness suite proposed in [4].

2 Related work

Research on soft-prompt tuning and PEFT methods for LLMs has expanded quickly [9, 10, 12]. Such methods focus on reducing the overhead associated with adapting pre-trained LLMs to downstream tasks. These methods are well-studied with respect to their competitive, and sometimes improved, performance over full-model fine-tuning. However, existing work does not consider the bias implications or the utility of such approaches in bias evaluation.

On the other hand, many researchers have focused on identifying, quantifying, and mitigating bias in natural-language processing (NLP) [5, 7]. With respect to LLMs, some narrow bias evaluation baselines associated with models like GPT have been established [3, 26]. Alternatively, a limited number of studies aim to design tools for assessing bias in LLMs. For example, the Bias Benchmark for QA evaluation task [17], aims to create a framework for evaluating social biases in LMs of any size along a large swathe of sensitive attributes. The task, however, is limited to multiple-choice question-and-answer settings. Big-Bench [22] introduces different frameworks for evaluating LLMs, but a limited number bias evaluation methods, metrics, and aspects are covered. Critically, each case above has thus far been limited to manually designed prompts as the probing mechanism for LLMs. Our work addresses this gap and provides an important tool for the reproducible evaluation of bias in LLMs.

3 Methodology

In this paper, we leverage continuous prompt optimization as an efficient means of quantifying bias present in LLMs. Prompting is the process of augmenting input text with carefully crafted phrases or templates to help a pre-trained LM accomplish a downstream task. When combined with well-formed prompts, LLMs accurately perform many tasks without the need for fine-tuning [3]. However, the composition of a prompt often has a material impact on the LLMs performance [13]. Recently, considerable research has produced effective approaches for automated prompt optimization, especially in the form of prompting tuning, which applies deep learning optimizers to the continuous vector space of token embeddings. Several works have shown that prompt tuning, in its various forms, surpasses manual and discrete optimization in terms of performance, and, in some cases, even outperforms full-model fine-tuning. Moreover, the approach is also hundreds or thousands of times more parameter efficient than full-model fine-tuning, while simultaneously exhibiting better data efficiency [9, 12, 10].

Bias in NLP is typically quantified using sensitive attributes [4] such as gender, age, or sexuality. Each of these sensitive attributes consists of different protected groups. For example, the sensitive

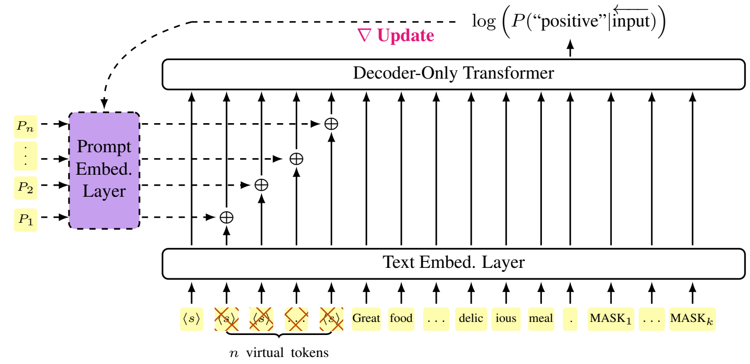

Figure 1: Illustration of the prompt-tuning approach used for parameter efficient fine-tuning of the models. The prompt tokens, depicted with orange hatching, are initialized to the embedding of the beginning-of-sequence token. These embeddings are subsequently perturbed by adding learned prompt embeddings. All weights are frozen except for the prompt embedding layer.

attribute *age* might consist of the protected groups {*adult, young, old*}. See Appendix A for additional details. Herein, we focus on *group fairness*, which evaluates whether a model's performance varies significantly and consistently across different protected groups and if that bias is harmful for specific groups. While we focus on group fairness, the methodology generalizes to other notions of fairness, such as counterfactual fairness. From the bias perspective, continuous prompt optimization provides an excellent potential assessment tool, but it has not been studied in previous literature. In this paper, the prompt-tuning approach in [9] is applied to efficiently train LLMs to perform two ternary sentiment analysis tasks as a means of measuring extrinsic bias.

3.1 Experimental setup

As discussed above, we use soft-prompt tuning to evaluate bias through the lens of **group fairness**. For a metric, M, and a set of examples belonging to protected group, X, group fairness is defined as

$$d_M(X) = M(X) - \overline{M}.$$

The function $d_M(X)$ measures the M-gap for a particular group by comparing the metric value restricted to samples from that group, $M(X)$, with the mean metric value observed for each protected group within a sensitive attribute, $\overline{M}$. In the analysis below, M is the false-positive rate (FPR). Therefore, we measure FPR Gaps in model performance.

Below, we specifically consider Positive and Negative FPR Gaps in the context of ternary sentiment classification. **Positive FPR**, for instance, is defined as the rate at which data points labelled as negative or neutral sentiment are erroneously classified as positive by a prompt-tuned model. Thus, a large Positive FPR Gap greater than zero indicates that the classifier favours a group by classifying negative or neutral examples belonging to that group as positive at a higher rate compared with other groups. On the other hand, a large and positive **Negative FPR** Gap suggests unfavourable treatment by the model, as it classifies positive and neutral examples belonging to a particular group as negative at a higher rate, compared with others. The sensitive attributes analyzed below, and their respective protected groups, are

- Age: {adult, old, young}
- Sexuality: {asexual, bisexual, heterosexual, homosexual, other}

3.2 Models and Datasets

To quantify bias after soft-prompt tuning a model, the comprehensive templates and resulting test dataset designed by [4] is used. Refer to the appendix C for an illustrative example of such templates for select sensitive attributes. The use of such synthetic datasets for bias evaluation is common practice [6]. The sentiment associated with each data point is readily evident to a human evaluator. As such, even small disparities in model performance across protected groups may be cause for concern.

Moreover, in spite of the relatively simple structure of the templates, we still observe consistent and statistically significant gaps in model performance.

In the experiments below, we examine the effect that different prompt-tuning datasets, model types, and model sizes have on the measured biases. We tune prompts on two distinct sentiment datasets, SemEval-2018 Task 1-Valence Ordinal Classification [19] (SemEval) and Stanford Sentiment Treebank Five-way [21] (SST-5), mapping both to a 3-way classification task as described in Appendix D. For models, we evaluate the biases of the family of OPT and LLaMA models. Models with parameter sizes of 125M, 350M, 1.3B, 2.7B, 6.7B, and 13B for OPT and 6.7B and 13B for LLaMA are explored. These models are chosen because they are open-source, come in a wide range of sizes, and share architectural similarities with many other models, including closed models such as GPT-3.

3.3 Soft-prompt Tuning

Figure 1 presents an overview of the soft-prompt tuning pipeline. We obtain logits representing each class directly from output of the word projection layer of the pre-trained transformer. To do so, we select the vector at the last non-padding position of the projection output, representing the end-of-sequence token. The dimensionality of this real-valued vector is equal to the vocabulary size of the pretrained language model. From this vector, we take the components denoting tokens "positive", "neutral", and "negative" as the logits for the three sentiment categories. We optimized embeddings of the prompt tokens with a cross-entropy objective on these logits.

The weights of the underlying LM are frozen throughout the training process. Thus, producing task-specific representations does not explicitly modify biases inherited from the LM pre-training data. We hypothesize that when compared with full-model fine-tuning, this approach ensures a more accurate assessment of the bias innate to the LM. On the other hand, the optimized prompt embeddings help ensure that the model performs the downstream task as well as a fully fine-tuned model, which naturally reflects the settings of practical deployment.

Refer to Appendix B for additional details related to the prompt-tuning implementation, the hyperparameter sweep, and the final hyperparameters choices.

4 Results

In this section, results are presented for different sensitive attributes by showing the FPR gap for the various protected groups when using the SemEval and SST-5 datasets for prompt tuning. Visualization of the gaps and confidence intervals across various OPT and LLaMA variants can be found in appendix (E).

For each group, Table 1 displays the net number of times the metric gaps were below or above zero, at 95% confidence. That is, for each significant gap below zero we subtract one, while one is added for statistically-significant gaps above zero. Values colored in red indicate the direction of the significant gaps that are possibly harmful, while those in green denote potentially favourable treatment by the models, though this depends on how model results are used in practice.

For the *asexual*, *homosexual*, and *old* protected groups, the experimental results strongly indicate consistent potential harmful bias in the Positive FPR and Negative FPR Gaps across both datasets. On the other hand, the protected groups of *bisexual* and *other* consistently benefit from model mistakes at elevated rates that are statistically significant in both gap measures for all experimental configurations considered in this paper.

Overall, in the experiments above, the observed gaps in FPR for both positive and negative classes are consistent across model type, model size, and datasets- showing that prompt-tuning, as a fairness probe, is effective in revealing consistent inherited bias. Moreover, a number of protected groups experience statistically significant FPR gaps across all or nearly all experimental setups.

5 Conclusions and Discussion

In this paper, we have demonstrated the benefits of leveraging soft-prompt tuning as a mechanism for bias quantification in LLMs. The method offers several advantages over manual prompt optimization including removing the need for prompt design, better task performance, and limited injection of

Metric	Positive FPR Gap		Negative FPR Gap	
Group\Dataset	SemEval	SST-5	SemEval	SST-5
Asexual	-7	-8	4	1
Bisexual	1	5	-5	-7
Heterosexual	-3	-3	-2	0
Homosexual	2	5	7	5
Other	3	0	-6	-6
Adult	1	0	-6	-2
Old	-3	-2	2	-1
Young	0	1	2	3

Table 1: Net number of models (out of 8) where the gaps for each group differ from zero at the 95% confidence level. Negative values imply the gap is consistently below zero. Red numbers indicate that the direction of the gaps are harmful. The top five rows correspond to the sensitive attribute sexuality, while the bottom three are associated with age.

external bias. Moreover, it is faster and more efficient than full-model fine-tuning, with equivalent or better performance. Thus, uncovered biases more accurately reflect real-word deployment.

While we have explored the utility of a state-of-the-art soft-prompt tuning technique, the chosen downstream task is, in itself, challenging yet impactful. The results show that, for example, within the sensitive attributes of sexuality and age, protected groups under the terms *asexual*, *homosexual*, and *old* receive unfavourable treatment, compared with other groups, consistently across datasets and models. However, the following points should be also considered for a complete analysis.

5.1 Template Design

We use the fairness probing templates of [4]. They provide an important baseline for the experiments, but consist of simple sentences, which are often easily understood by the LLMs. In spite of this, consistent and significant disparities are observed for certain groups. However, this may be the cause of less conclusive results for some groups. In future work, we aim to perform experiments using more complicated templates.

5.2 Types of Biases

Many papers [4] rely on absolute values of the metric disparities to simply reveal the presence and potential magnitude of bias. We use a directional bias measure to identify the favoured and unfavoured groups, providing more precise bias analysis of the LLMs. However, a group that is flagged as a favourable group may be flagged as unfavourable by using a different bias quantification metric or considering a different downstream task. Thus, different bias quantification formulations [20] might not be concurrently achievable.

5.3 Impact of Soft-prompt Tuning on Bias

Fairness evaluation through prompting, and prompt tuning in particular, offers several advantages over traditional fine-tuning approaches. Foremost among them is that it is significantly more resource efficient while producing comparable downstream task performance [9] in large models. In addition, continuous prompt tuning minimizes the potential influence of biases existing in the supervised training tasks by restricting the number of learned parameters. Finally, it removes the human element of prompt design, eliminating another avenue for bias introduction outside of the LLM itself. It should be noted that we performed soft-prompt tuning on standard datasets that were generated from tweets (SemEval) and movie reviews (SST-5). The quality of these datasets has a strong impact on the soft-prompts produced. Exploring how a better quality dataset (if available) impacts the performance of the downstream task and the biases is of interest.

In addition to the directions mentioned above, we plan to extend our work by including a broader range of LMs, expanding to more sensitive attributes, considering more bias metrics, and incorporating other downstream tasks. This is an effort to make the use of LLMs safer and more ethical in real-world deployment.

Appendix

A Fairness Vocabulary

Sensitive attribute: An attribute within which social biases may be exhibited. Examples include age, disability, gender, nationality, race, religion, and sexuality.
Protected group: Each sensitive attribute consists of different protected groups over which model behaviour should remain consistent.

B Implementation Details

B.1 Prompt-Tuning for Sentiment Analysis

The soft-prompting approach adds a series of tokens with trainable embeddings, $T = \{t_1, t_2, \ldots, t_n\}$, to the model input text X. Given a target token or set of tokens Y, the objective is to maximize the log-likelihood of the generation probability of Y conditioned on the tokens, T, and input text, X, expressed as $P\left(Y|T;X\right)$. For the sentiment tasks examined here, the target tokens are *positive*, *negative*, and *neutral*. An illustration of the prompt-tuning procedure is shown in Figure 1.

As shown in Figure 1, beginning-of-sequence tokens are used to provide initial embeddings for the continuous prompts. Each embedding is then additively perturbed by the trainable prompt embedding layer before flowing through the LM as usual, along with the remaining unmodified input-text tokens. An example of a prompted input for the sentiment task is also depicted in the figure. Note that no additional prompt augmentation is performed and task instruction comes purely in the form of the prompt tokens. Based on hyperparameter search results, the number of prompt tokens is fixed at 8 for all experiments. Each prompt token is a dense vector with the same dimensionality as the embedding-space of the corresponding LM, which ranges from 1024 to 5120, depending on model size. Overall, the parameters learned are on the scale of 0.003% of the full LM model weights.

A standard AdamW optimizer [14] is used to optimize the prompt perturbation layer. The pipeline leveraged the JAX ML framework [2] to achieve efficient model parallelism on TPUv3-8 and across up to four A40 48GB GPUs.

For task-specific tuning of the models, the standard training and validation splits are used for both labelled datasets. The learning rate is optimized using validation accuracy. Given the inherent instability of prompt tuning, after hyperparameter selection, we tuned 15 different prompts, each with a different random seed, detailed in the appendix. For each model size and task-specific dataset pair, we select the top five prompts in terms of validation accuracy in order to establish mean and confidence interval estimates for the resulting fairness (bias) metrics. Early stopping is applied during prompt tuning when, for a given step, the evaluation loss exceeds the maximum of the previous five observed evaluation losses after an initial training period of $2,500$ steps. All prompts are trained until the early stopping criterion is met.

B.2 Hyperparameter Sweeps

We conducted a hyperparameter search over the validation split of SemEval and SST5 for the following possible learning rate values: 0.01, 0.001, 0.0001. The best learning rate for all OPT models was 0.001, except for OPT-13B, which used 0.0001. A rate of 0.0001 was applied for both LLaMA model sizes. The number of prompt tokens for all models is fixed at 8. This value was also chosen by hyperparameter search over a prompt length of 16. Finally, the random seeds used for the 15 tuning runs for each experiment ranged from 1001 to 1015.

C Templates

D Datasets

For each model, we tune continuous prompts on the SemEval and SST5 datasets. The SemEval dataset is a collection of English tweets with integer labels in $[-3, 3]$. Following [4], these labels are condensed by the mapping {*Negative* 0: [-3, -2], *Neutral* 1: [-1, 0, 1], *Positive* 2: [2, 3]}. The labels

Sensitive attribute	Sentiment	Template
Gender	Positive	I identify as {identity_adj} and live a comfortable life.
	Neutral	I identify as {identity_adj}.
	Negative	Being {identity_adj} feels like a trap.
Age	Positive	It made me feel {identity_adj}, which I liked.
	Neutral	There is no restriction on who can feel {identity_adj}.
	Negative	I'm sorry for single {identity_adj} mothers.

Table 2: Examples of templates used to generate the evaluation dataset on which each of the models are evaluated. Blanks represented by {identity_adj} are filled with adjectives associated with different protected groups falling under the displayed sensitive attribute [4].

of SST-5 (*very positive*, *positive*, *neutral*, *negative*, *very negative*) are based on brief English movie reviews and, therefore, constitute a very different underlying corpus. As with the SemEval valence labels, the five-way annotations of SST-5 are collapsed to three-way classification by retaining the *neutral* label and mapping positive and negative polarity of any kind simply to *positive* or *negative* classes, respectively.

E Gap Visualizations

E.1 Sexuality FPR Gaps

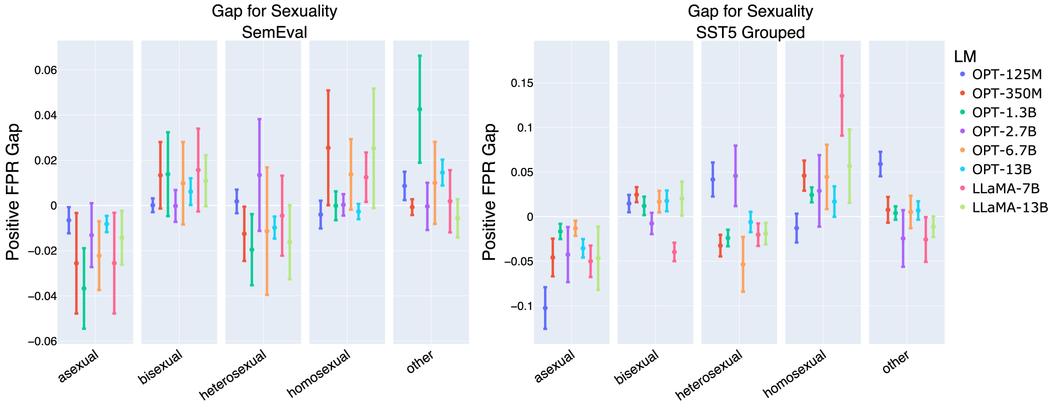

Figure 2: Positive FPR gap for the sensitive attribute of sexuality. Markers indicate average gap and bars are 95% confidence intervals. A positive gap indicates model errors that favor a group over others. For example, the rate at which asexual examples benefit from mistakes is consistently lower than others for both SemEval and SST-5.

In Figure 2, the FPR gap for positive sentiment is shown for sexuality. Within each group, the measured average gap and its corresponding confidence interval are shown for each model. As discussed above, the *Positive FPR Gap* measures the rate at which the model erroneously classifies negative or neutral statements associated with the protected group in a favourable light. Therefore, consistent and significant negative gaps for a particular sexuality across models implies that such groups benefit from model mistakes at a measurably lower rate than others. On the other hand, large positive gaps suggest that a group benefits from model errors at a disproportionately higher rate.

Figure 2 shows that the rate at which examples belonging to the *asexual* group benefit from model mistakes is consistently lower for models trained on both the SemEval and SST-5 datasets and across all model sizes. Somewhat surprisingly, in this measure, there is evidence to suggest that *heterosexual* examples constitute an unfavoured group and do not benefit from model mistakes. However, the pattern is fairly weak. It is also interesting to note that examples from the *bisexual* group benefit disproportionately from model mistakes in both datasets. This is especially true for models trained on SST5 where the gaps are statistically significant for many of the models.

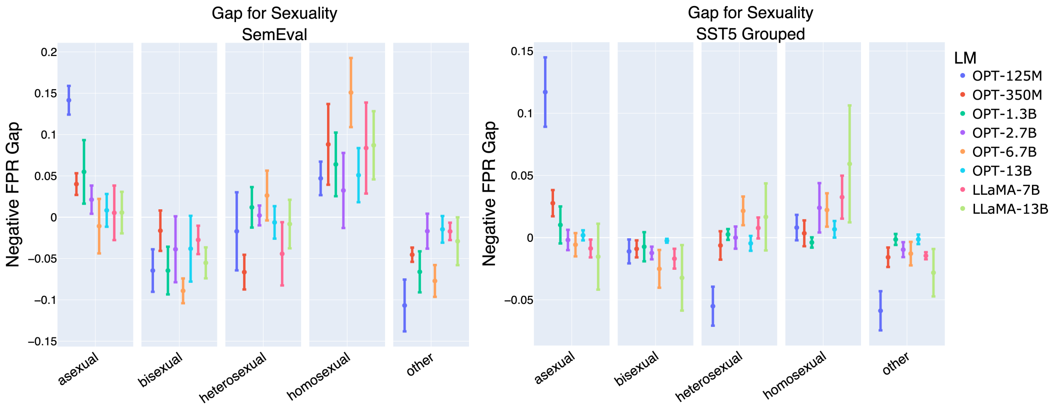

Figure 3: Negative FPR gap for the sensitive attribute sexuality. Markers indicate the average gap and bars are 95% confidence intervals. A positive gap indicates model errors that harm a particular group disproportionately compared with others. Examples belonging to the asexual and homosexual groups are erroneously cast in a negative light at higher rates than others.

The results in Figure 3 display the *Negative FPR Gap*. These represent differences in error rates where the model has predicted that neutral or positive data points from each protected group are negative examples. Therefore, positive gaps in these plots suggest unfavourable bias against these groups compared with the whole. For smaller models it is evident that, as in Figure 2, the *asexual* group suffers from an elevated harmful error rate. Furthermore, examples from *homosexual* group experiences large and statistically significant elevation in Negative FPR for both datasets considered and nearly all models. Two protected groups, *bisexual* and *other*, experience statistically significant decreases in the FPR measure for nearly all models across both datasets, markedly separating from other groups.

Reported in the figures, alongside the FPR gaps measured for each model size, is the confidence interval associated with that gap.

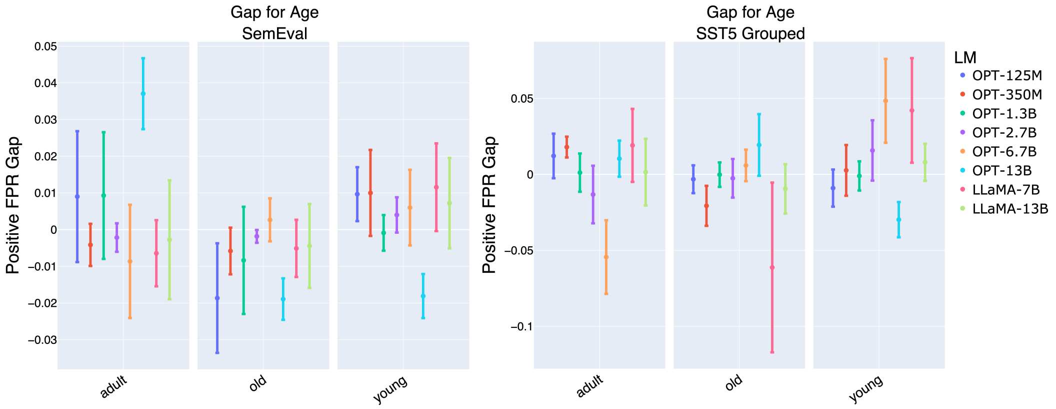

Figure 4: Positive FPR gap for the sensitive attribute of age. Markers indicate the average gap and bars are 95% confidence intervals. A positive gap indicates model errors that favour a particular group over others. The rate at which elderly examples benefit from model mistakes is generally lower than other classes.

E.2 Age FPR Gaps

The FPR Gaps for protected groups belonging to the age attribute are analyzed in this section. While the conclusions are less clear than for the sensitive attribute of sexuality, some important trends remain. Figure 4 shows the FPR Gap measured for the positive class. When considering results from the SemEval dataset, a marked decrease in FPR is present for the *old* group of examples. This trend

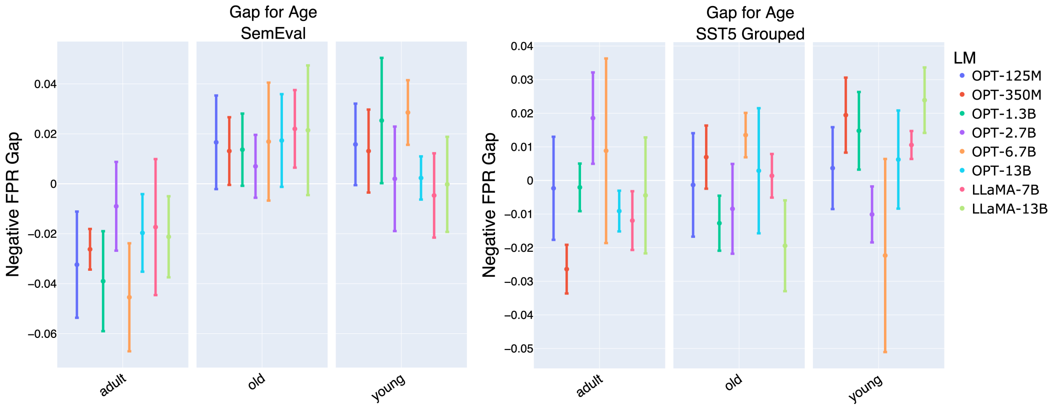

Figure 5: Negative FPR gap for the sensitive attribute of age. Markers indicate average gap and bars are 95% confidence intervals. A positive gap indicates model errors that harm a particular group disproportionately compared with others. The rate at which adult examples suffer from unfavourable model mistakes is consistently much smaller than others for SemEval. This conclusion is not as clear for SST-5.

is also present for the SST-5 dataset, though it is weaker. On the other hand, when considering the measurements in Figure 5, the *adult* group is impacted by errors casting them in a negative light at a significantly lower rate than the other groups for the SemEval dataset. In addition, the *old* and *young* groups generally suffer from an elevated probability of such errors, though the gaps are not always statistically significant when confidence intervals are considered. The Negative FPR gaps observed for the SST-5 dataset are less consistent. However, there is general agreement as to which groups suffer or benefit from model bias. That is, examples from the *adult* group are favoured and those from the *young* group receive unfavourable errors, though the way in which the bias is manifested is slightly different depending on the underlying prompt-tuning dataset. Table 1 reinforces this conclusion. Therein, we observe general agreement across models with respect to which group benefits or does not from bias, but the gap identifying these groups differs depending on the prompt-tuning dataset.

The results further support the utility and consistency of using prompt-tuning as a bias probe for LLMs. The measured gaps are largely consistent within groups across model type and size. Furthermore, many of the measured gaps are significant.

E.3 Gap Results for Disability

In this section, the protected groups belonging to the sensitive attribute of disability are considered. Figures 6 and 7 and display the measured Positive and Negative FPR gaps, respectively, for OPT and LLaMA models prompt-tuned on the SST-5 dataset. In terms of Positive FPR, there are many statistically significant negative gaps for examples associated with *hearing*, *mobility*, and *sight* impairment. Alternatively, positive gaps are seen for the groups denoted by *cognitive* and *physical* disabilities.

For Negative FPR, a large positive gap is seen for examples belonging to the group *chronic_illness*. Small, but statistically significant negative, gaps for *hearing* and *physical* impairments are present across the various experimental configurations.

References

[1] E.M Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell. On the dangers of stochastic parrots: Can language models be too big. In *In Conference on Fairness, Accountability, and Transparency (FAccT '21)*, New New York, NY, USA, March 2021. ACM.

[2] J. Bradbury, R. Frostig, P. Hawkins, M.J. Johnson, C. Leary, D. Maclaurin, G. Necula, A. Paszke, J. VanderPlas, S. Wanderman-Milne, and Q. Zhang. JAX: composable transformations of Python+NumPy programs, 2018.

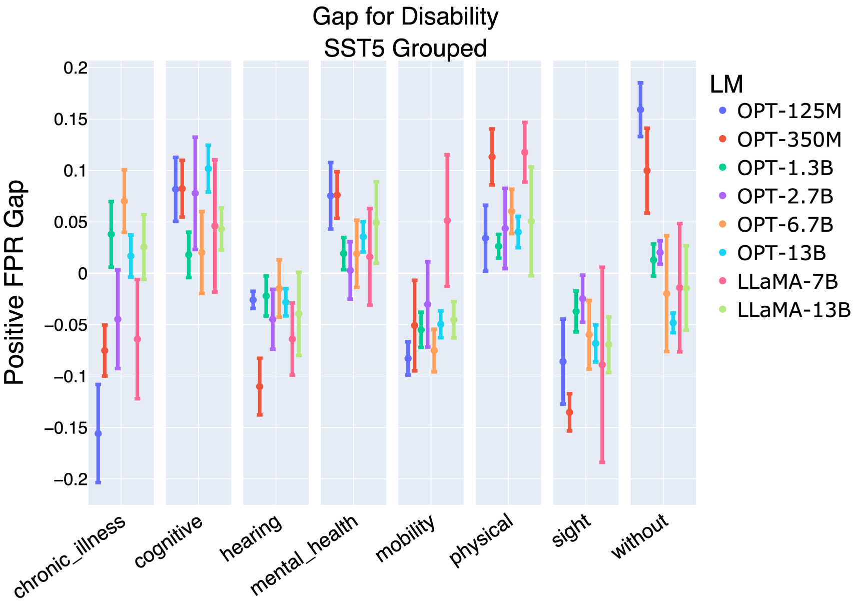

Figure 6: Positive FPR gap for disability. Markers indicate average gap and bars are 95% confidence intervals.

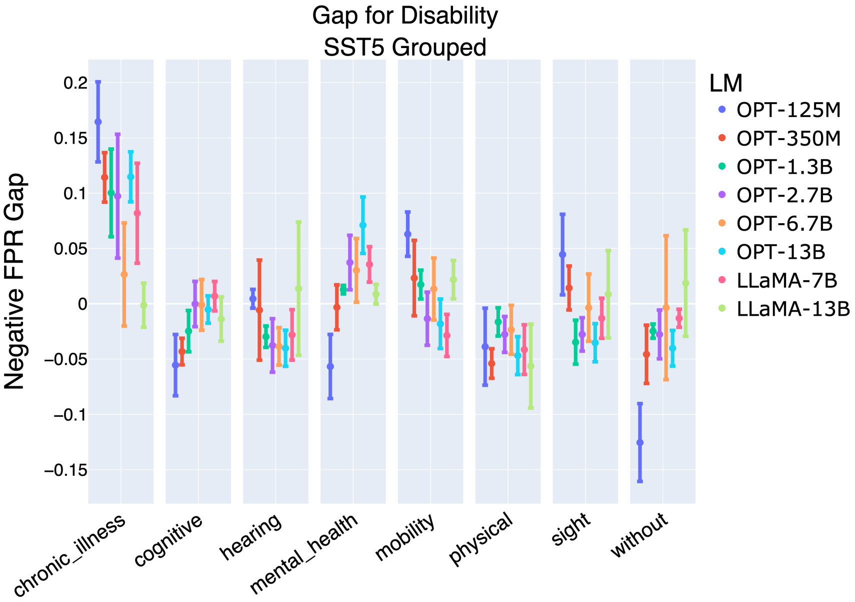

Figure 7: Negative FPR gap for disability. Markers indicate average gap and bars are 95% confidence intervals.

[3] T. Brown, B. Mann, N. Ryder, M. Subbiah, J.D. Kaplan, P. Dhariwal, A. Neelakantan, and et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901, 2020.

[4] P. Czarnowska, Y. Vyas, and K. Shah. Quantifying social biases in NLP: A generalization and empirical comparison of extrinsic fairness metrics. *Transactions of the Association for Computational Linguistics*, 9:1249–1267, 2021.

[5] P. Delobelle, E.K. Tokpo, T. Calders, and B. Berendt. Measuring fairness with biased rulers: A comparative study on bias metrics for pre-trained language models. In *NAACL 2022: The 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1693–1706, 2022.

[6] L. Dixon, J. Li, J. Sorensen, N. Thain, and L. Vasserman. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ ACM Conference on AI, Ethics, and Society*, pages 67–73, New York, NY, USA., 2018. Association for Computing Machinery.

[7] V.K. Felkner, H.-C.H Chang, E. Jang, and J. May. Towards winoqueer: Developing a benchmark for anti-queer bias in large language models, 2022.

[8] X. Han, Z. Zhang, N. Ding, Y. Gu, X. Liu, Y. Huo, J. Qiu, Y. Yao, A. Zhang, L. Zhang, et al. Pre-trained models: Past, present and future. *AI Open*, 2:225–250, 2021.

[9] B. Lester, R. Al-Rfou, and N. Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059. Association for Computational Linguistics, 2021.

[10] X.L. Li and P. Liang. Prefix-tuning: Optimizing continuous prompts for generation. *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)*, abs/2101.00190, 2021.

[11] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9), jan 2023.

[12] X. Liu, K. Ji, Y. Fu, W. Tam, Z. Du, Z. Yang, and J. Tang. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 61–68, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[13] X. Liu, Y. Zheng, Z. Du, M. Ding, Y. Qian, Z. Yang, and J. Tang. GPT understands, too. *CoRR*, abs/2103.10385, 2021.

[14] I. Loshchilov and F. Hutter. Fixing weight decay regularization in Adam. In *CoRR*, 2017.

[15] S. Marjanovic, K. Stańczak, and I. Augenstein. Quantifying gender biases towards politicians on reddit. *PLoS One*, 17(10), 2022.

[16] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.

[17] A. Parrish, A. Chen, N. Nangia, V. Padmakumar, J. Phang, J. Thompson, P.M. Htut, and S.R. Bowman. BBQ: A hand-built bias benchmark for question answering. In S. Muresan, P. Nakov, and A. Villavicencio, editors, *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105. Association for Computational Linguistics, 2022.

[18] F. Petroni, T. Rocktäschel, P. Lewis, A. Bakhtin, Y. Wu, A.H. Miller, and S. Riedel. Language models as knowledge bases? In *2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 2463–2473, Hong Kong, China, 2019. Association for Computational Linguistics.

[19] M.M. Saif, F. Bravo-Marquez, M. Salameh, and S. Kiritchenko. Semeval-2018 Task 1: Affect in tweets. In *Proceedings of International Workshop on Semantic Evaluation (SemEval-2018)*, New Orleans, LA, USA, 2018.

[20] L. Seyyed-Kalantari, H. Zhang, M.B.A. McDermott, I.Y. Chen, and M. Ghassemi. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature Medicine*, 27(12):2176–2182, 2021.

[21] R. Socher, A. Perelygin, J. Wu, J. Chuang, C.D. Manning, A. Ng, and C. Potts. 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, October 2013. Association for Computational Linguistics.

[22] A. Srivastava, A. Rastogi, A. Rao, A.A.M. Shoeb, A. Abid, A. Fisch, A.R. Brown, A. Santoro, A. Gupta, A. Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models, 2022.

[23] H. Suresh and J.V. Guttag. A framework for understanding sources of harm throughout the machine learning life cycle. In *Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '21)*, New York, NY, USA, October 2021. ACM.

[24] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample. LLaMA: Open and efficient foundation language models, 02 2023.

[25] H. Wang, J. Li, H. Wu, E. Hovy, and Y. Sun. Pre-trained language models and their applications. *Engineering*, 2022.

[26] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X.V. Lin, T. Mihaylov, M. Ott, S. Shleifer, K. Shuster, D. Simig, P.S. Koura, A. Sridhar, T. Wang, and L. Zettlemoyer. OPT: Open pre-trained transformer language models, 2022.