Evaluating Scalability of Bias Mitigation Techniques on Large Language Models

Martin Blanckaert
UQAC
mblanckaer@etu.uqac.ca

Clément Delteil
UQAC
cdelteil@etu.uqac.ca

Valentin Porchet
UQAC
vporchet@etu.uqac.ca

Thomas Sirvent
UQAC
tsirvent@etu.uqac.ca

Abstract

The recent explosion in popularity of large language models (such as GPT-4, Falcon, PaLM, and LLaMA) has highlighted inherent biases against certain demographics or cultures induced by training data. There has been extensive work on modest-sized models such as BERT or GPT-2, fine-tuning these models on specialized datasets. These approaches have succeeded to some extent in mitigating the bias of certain models, but remain very limited. This work examines the possibility of applying such methods to larger models such as Bloomz and FLAN-T5. Using Hugging Face metrics and an adaptation of the StereoSet benchmark, we found that both models experienced significant performance degradation after fine-tuning.

1 Introduction

In the field of Natural Language Processing (NLP), the widespread adoption of Deep Learning models has become standard practice, owing to their remarkable ability to comprehend context and vast knowledge. Among these models, a particular type of network called Large Language Models (LLM) has gained significant popularity through applications such as ChatGPT (Brown et al., 2020), LLaMA (Touvron et al., 2023), and Bard (Thoppilan et al., 2022). This surge in popularity has attracted attention from the general public, investors, and major tech companies such as Google, Meta, OpenAI, etc. However, this exponential growth has also shed light on certain shortcomings in the behavior of these algorithms, particularly their inclination to perpetuate common occidental stereotypes induced by the training data (Bolukbasi et al., 2016) or the training process itself. They often appear in the form of harmful opinions or misrepresentations of a social group. These kinds of stereotypes are called biases. The biases present in current LLMs are problematic, both for the user experience at an individual level and for the general public acceptance of these emerging technologies, since people fear AIs that act against the good of the people.

Our motivation is therefore to contribute to existing research on the correction of these biases, using the methods and datasets they employ on new generations of LLMs. We will then be able to see whether advances in the field of LLMs can be adjusted with these methods.

Detecting and quantifying these biases is crucial for developing strategies to mitigate their negative effects. A popular approach in the literature is to use word embedding associations as a measure of bias on specific benchmarks. Extensive research has been conducted on language models of the BERT family, given their possibility to be fine-tuned with limited computational resources. However, to the best of our knowledge, no comparative study has been conducted to investigate the efficacy of bias mitigation techniques between language models and large language models.

In this work, our goal is to study the effectiveness of bias mitigation techniques on large language models with billions of parameters, employing benchmarks that have demonstrated some degree of effectiveness on language models. Our results will give directions for the development of bias mitigation techniques on large language models.

This paper is organized as follows. Firstly, we will define related works in the field of bias mitigation, and the emerging methods that come with it. Then, we will present the methodology used by our team to address these biases, taking inspiration from the related works. Finally, we will conduct our experiments and make conclusions on the effectiveness of our methods and the perspective of unbiased LLMs. We publicly release our notebooks on GitHub[1] as well as the fine-tuned

[1] https://github.com/Wazzabeee/Bias-Mitigation-In-LLM

models on Hugging Face[2][3][4].

2 Related Work

To effectively address the bias present in large language models, it is essential to understand the existing techniques employed for measuring such biases. Three principal ones emerged from our literature review. The first technique, Sentence Encoder Association Test (SEAT) (May et al., 2019), is an extension of another test, Word Embedding Association Test (WEAT) (Caliskan et al., 2017). SEAT evaluates the model's propensity to associate specific words with the beginnings of sentences. By comparing the model's completions of sentences with predetermined unbiased completions, a score can be assigned to this model. The second technique, StereoSet (Nadeem et al., 2021), is a crowdsourced dataset that captures four different types of biases. Each example sentence in the dataset has three possible completions, provided by the benchmark: a stereotyped one, a non-stereotyped one, and one unrelated to the beginning of the sentence. By calculating the percentage of times that the model favors the non-stereotyped answer over the stereotyped one, a stereotype score can be established. The third technique, Crowdsourced Stereotype Pairs (CrowS-Pairs) (Nangia et al., 2020), is also crowdsourced and offers pairs of stereotyped and non-stereotyped sentences. The pairs usually differ on a single word that completely reverses the expected stereotype. By examining the model's preference for the single word associated with the stereotype, it becomes possible to gauge the level of bias exhibited by the model. The higher the likelihood of the model choosing the word representing the stereotype, the stronger the indication of bias.

From these benchmarks, different methods were considered by the researchers to improve the scores of language models. With BERT and DistilBERT models, it was shown that only 4 epochs of model fine-tuning on non-stereotyped sentences from the StereoSet and CrowsPairs datasets were enough to drastically reduce the SEAT score of the models (Dolci, 2022). Nevertheless, fine-tuning all parameters of the model using inverted stereotyped sentences presents certain limitations. The extensive time and cost associated with retraining all the model's weights raise concerns of an economic and environmental nature. Furthermore, this complete fine-tuning has demonstrated a potential decline in the model's performance on its original task due to the "catastrophic forgetting" phenomenon (Kirkpatrick et al., 2017). Thus, alternative approaches that are more cost-effective have been investigated, such as unfreezing only some specific layers of the model. Research has shown that adjusting merely 1% of the GPT-2 parameters can yield scores comparable to those achieved through complete fine-tuning (Gira et al., 2022).

However, simply drawing attention to bias without a deliberate approach provides little benefit, especially considering the long-standing tendency of LLMs to propagate and even exacerbate stereotypes (Bolukbasi et al., 2016). To ensure the efficacy of the mentioned interventions, it is crucial to correctly identify, detect and measure these biases. Existing literature lacks a clear alignment between bias measures and the specific harms they address. To address this gap, researchers put forward a practical framework that establishes connections between biases and specific harms while offering a set of guiding documentation questions to guide the development of bias measures (Dev et al., 2022). Additionally, they present case studies illustrating how different measures align with distinct harms. The framework includes five types of harm: Stereotyping, Disparagement, Dehumanization, Erasure, and Quality of Service (QoS). By aligning bias measures with these harms, practitioners can better articulate the limitations, appropriate use cases, and implications of their measures. Furthermore, it is important to acknowledge that bias measures may inadvertently interpret mentions of social group denominations as bias occurrences (Davani et al., 2020). For example, mentioning the word "Muslim" in a text can be flagged as biased content due to the association of numerous stereotypes with that group. As a result, classifiers have learned to associate the presence of such words in a text with a significant probability of biased content. These false positives highlight the need for a carefully crafted methodology when measuring bias to avoid unintended consequences and ensure accuracy.

Along with those limitations, it is important to keep in mind that since measuring techniques cannot, and should not, target every type of bias, resulting scores must always be interpreted in context. A

[2] https://huggingface.co/Wazzzabeee/PoliteBloomz

[3] https://huggingface.co/Wazzzabeee/PoliteT5Base

[4] https://huggingface.co/Wazzzabeee/PoliteT5Small

low score of bias occurrence does not imply that our predictions are completely unbiased, it means progress has been made in mitigating the targeted bias. Attempting to address all types of biases is an unattainable goal due to their various forms. Gender bias, for example, can be expressed through the use of different pronouns in associations with specific words (jobs, occupations, etc), while bias concerning other demographics can be shown through preconceived opinions (usually negative).

Furthermore, it is crucial to acknowledge the limitations of the existing evaluation methods such as SEAT, StereoSet, and CrowS-Pairs, which may not provide reliable measures of bias in these models (Meade et al., 2021). Simply reducing stereotype scores does not necessarily indicate successful debiasing, as it could be achieved by compromising the overall language modeling ability of the model. This raises concerns about the effectiveness of certain debiasing techniques in truly mitigating bias. As exposed earlier, most debiasing techniques tend to worsen a model's language modeling ability. A comprehensive assessment of debiasing techniques should go beyond superficial measures and delve into their effects on the fundamental language modeling capabilities of large language models.

3 Methodology

The initial phase of our research is to assess the existing biases in the models selected for this work, namely FLAN-T5, and Bloomz.

FLAN-T5 represents a versatile family of LLM with a diverse range of sizes, encompassing models with parameter counts spanning from 80 million to 11 billion. On the other hand, Bloomz represents a multilingual family of models encompassing a parameter range of 560 million to 176 billion. Bloomz is a multitask fine-tuned version of the Bloom open-source model, capable of following task instructions zero-shot. We have chosen this specific version for this ability, which makes Bloomz particularly suitable for conducting the evaluations mentioned below.

The selection of both FLAN-T5 and Bloomz models was also driven by their capability to address a significant concern in our research: hardware limitations. These models include implementations with parameter counts below one billion, enabling us to fine-tune them using accessible resources without compromising the feasibility of our experiments. Additionally, the advantage of scaling our methodology across different sizes within the same model family is facilitated by the fact that these models perform the same tasks but with improved performance. As a result, we can confidently apply our methodology to various model sizes, knowing that the underlying tasks remain consistent and the models exhibit enhanced capabilities with more parameters.

To quantify the biases, we employ various metrics from Hugging Face's Evaluate library, along with other widely recognized metrics in the literature, such as the *Language Modeling Score* (LMS), the *Stereotype Score* (SS) and the *Idealized Context Association Test* (ICAT).

In the second phase of our research, we undertake the fine-tuning of the models using different combinations of datasets aiming to mitigate the previously measured bias. Ultimately, we evaluate the models once again to assess the impact of the fine-tuning process on reducing the pre-established biases.

3.1 Evaluation

3.1.1 HuggingFace

These metrics are simple to implement and offer some level of modularity. Some, like Toxicity and Regard, can analyze our generated completions using aggregations. This allows us to have multiple indicators such as average and maximum harmful content ratios. Moreover, metrics such as HONEST and Regard can be applied to specific groups, which makes their evaluations more focused.

HONEST The HONEST score is a metric for assessing hurtful sentence completions generated by language models. It leverages HurtLex, a multilingual lexicon containing hurtful language, to evaluate these completions. The primary objective of the HONEST score is to quantify the frequency with which sentences are completed using hurtful words and determine if there are disparities among different groups, such as genders, sexual orientations, and more. This metric is applied using the BOLD(Dhamala et al., 2021) dataset, a large-scale dataset that consists of over 20,000 English text generation prompts for bias benchmarking across five domains: profession, gender, race, religion, and political ideology.

Toxicity The objective of the toxicity measurement is to quantitatively assess the level of toxicity present in input texts utilizing a pre-trained hate

speech classification model. This metric is applied using the RealToxicityPrompts dataset provided by the Allen Institute for AI. The dataset consists of 100,000 sentence snippets sourced from the web, designed to assist researchers in examining and mitigating the potential for neural models to exhibit toxic behavior.

Regard The regard measurement provides an estimation of the language polarity and social perceptions associated with a particular demographic, such as gender, race, or sexual orientation. To calculate this measurement, a model trained on labeled data from the paper titled "The Woman Worked as a Babysitter: On Biases in Language Generation"(Sheng et al., 2019) is utilized. By leveraging this model, insights can be acquired regarding the regard and sentiment expressed towards different groups within the analyzed texts. Furthermore, the regard measurement can take into account two different sets of content, which allows it to compare how badly a specific group is treated compared to another. A classic example of this usage is to put side-by-side male and female-related content.

3.1.2 StereoSet

The remaining three metrics are based on the StereoSet benchmark, as discussed earlier. However, it should be noted that this benchmark was initially designed for models like Bert or GPT-2, which allow for easy extraction of model completion probabilities. In our case, with FLAN-T5 and Bloomz, this is not possible. Therefore, in this work, we propose an adapted evaluation approach by formulating queries to the models in the form of prompts with multiple choices.

The original benchmark consisted of two distinct tasks. The intrasentence task assessed model bias and language modeling ability at the sentence level. It involved presenting a context sentence with a *BLANK* word to be completed, offering three choices for potential word completions. The intersentence task evaluated bias and language modeling ability at the discourse level. Here, three possible sentences were provided as options, and the model was required to associate the most appropriate sentence with the given context.

Figure 1 illustrates the prompt format utilized to evaluate the models and assess their choices. The format has been refined to maximize the number of consistent responses from the models. By adapting 2123 questions for the first task and 2106 for the second, we capture the model choices necessary for calculating the same metrics as those used in the original benchmark.

Question: The context is missing a word defined as "blank". Select which sentence among the 3 choices below is the best way to replace it.
Context: The physicist had a reputation for being BLANK.
Choice1: The physicist had a reputation for being intelligent. (anti-stereotype)
Choice2: The physicist had a reputation for being powdery. (unrelated)
Choice3: The physicist had a reputation for being nerdy. (stereotype)
Possible answer: The physicist had a reputation for being nerdy.
(a) The Intrasentence Context Association Test as a Prompt

Question: Select the item from this list which relates the most to the context.
Context: Many people live in Ethiopia.
Choice1: The people are fat and unathletic. (anti-stereotype)
Choice2: The people are very thin and good at distance runing. (stereotype)
Choice3: Cats have sharp claws. (unrelated)
Possible answer: The people are fat and unathletic.
(b) The Intersentence Context Association Test as a Prompt

Figure 1: Context Association Tests (CATs) adapted in prompt form to measure the bias and the language modeling ability of Large-Language-Models.

Language Modeling Score (LMS) The LMS metric represents the percentage of times the model chose a "logical" response (stereotyped or non-stereotyped) compared to the unrelated one. The ideal LMS score for a model is 100, indicating that the model has never selected a sentence or word that has no direct link with the given context.

Stereotype Score (SS) The SS metric represents the percentage of times the model favors the stereotyped response to the non-stereotyped one. The ideal SS score for a model is 50, as this means that the model has shown no preference for stereotyped or non-stereotyped responses, and has selected them to the same extent.

Idealized Context Association Test (ICAT) Finally, the ICAT score is a combination of the two aforementioned metrics and is calculated as follows:

$$ICAT = LMS \times \frac{min(SS, 100 - SS)}{50}$$

This formula summarizes the two metrics presented above as the ICAT scores reaches 100 when its *lms* is 100 and *ss* is 50.

3.2 Fine-Tuning

To fine-tune the models, our primary focus was on utilizing the Crows-Pairs dataset. This dataset, as outlined earlier, provided us with a format that aligned perfectly with our research objectives, making it the ideal choice for our experimentation. Our approach involves fine-tuning these models by exposing them to debiased sentences and presenting

StereoSet Benchmark Scores						
Model	**Pre Fine-Tuning**			**Post Fine-Tuning**		
	LMS	**SS**	**ICAT**	**LMS**	**SS**	**ICAT**
	Intrasentence Task					
FLAN-T5 Small	79.2	54.02	**72.83**	1.19	48.0	1.14
FLAN-T5 Base	**87.46**	63.52	63.81	0.28	66.67	0.19
Bloomz-560m	47.72	**52.04**	45.77	0	0	0
	Intersentence Task					
FLAN-T5 Small	78.62	47.69	74.99	3.06	52.31	2.92
FLAN-T5 Base	**95.9**	**51.77**	**92.91**	0.24	40.0	0.19
Bloomz-560m	64.3	47.62	61.24	0	0	0
	Global					
FLAN-T5 Small	78.91	50.85	77.57	2.13	51.11	2.08
FLAN-T5 Base	**91.7**	57.35	**78.22**	0.26	54.55	0.24
Bloomz-560m	56.04	**49.49**	55.47	0	0	0

Table 1: Performance of our models on the Intersentence and Intrasentence tasks of the StereoSet benchmark. Each metric was measured using custom prompts given in Figure 1. Each experiment was run n=1 times as the results were always the same. The best score for each task is in bold characters.

the preferred responses as reference answers. As presented in the introduction, through this experiment we aim to find out whether this technique is still effective on models with several billion parameters.

More precisely, the question at hand is whether fine-tuning on relatively small datasets, in comparison to the extensive training datasets of these models, is sufficient to modify their word associations and reduce their harmfulness over time.

Considering the constraints imposed by hardware limitations, our primary focus revolved around three models: the Small (80M) and Base (250M) variants of the FLAN-T5 model, and the 560M parameter version of Bloomz. Both models being multi-tasking LLM, they are well-suited for both decision-making and text-generation tasks. In our research, we leverage their text generation capabilities, where the model directly produces text that is then evaluated using our metrics to assess the effectiveness of our methodology. Additionally, the multi-task nature of FLAN-T5 and Bloomz allows us to incorporate specific prompts into the input data if deemed relevant for our testing purposes.

4 Results and Discussion

4.1 Results

See Table 1, Table 3, and Table 4 for experimental results. Before fine-tuning, we can see that overall it's the FLAN-T5 family of models that fared best in the StereoSet benchmark, with FLAN-T5 Base's unbeatable performance on the intersentence task clearly demonstrating its superior ability to reason more globally over several sentences. Across the board, fine-tuning these models resulted in a global decrease in performance. We hypothesize that the overall decline in model language modeling capabilities may be due to two reasons.

The first concerns the way we fine-tune models. The FLAN-T5 and Bloomz models have been trained for certain specific tasks, and "rewriting a sentence in a less offensive way" is not one of them. Unfortunately, we haven't found any other way of adapting the CrowSPairs dataset for this work. Furthermore, we didn't want to use the StereoSet dataset for this fine-tuning task as the authors made it clear in their article that this was of little interest.

The second phenomenon we encountered is catastrophic forgetting, which is probably a consequence of the first one. It's not just that the models have lost some of their ability to model the language despite a drop in bias, it's as if they've completely forgotten what they were trained to do. This may be due to the hyperparameters used in our experiments, which left too much room for change, but we don't have the necessary hindsight on these methods to confirm or refute this hypothesis.

The HuggingFace metrics confirm the benchmark's hypothesis. The toxicity evaluation increased for all models after the fine-tuning, achiev-

ing the opposite of what was intended. The HONEST and Regard metric scores are harder to interpret, with no clear evolution across all models. The modifications are model-specific but have to be placed back in context. Indeed, after the fine-tuning, the models - especially Bloomz - suffered from catastrophic forgetting. The metric calculations are limited in that sense, with the models struggling to perform the basic tasks they were evaluated on.

4.2 Discussion

While our results may not be as conclusive as we anticipated, they offer valuable insights. Firstly, it is clear that using small-sized models had a significant impact on various aspects, from text generation to instruction comprehension. This often led to models generating new sentences instead of completing the given prompts as intended.

Secondly, our main approach of fine-tuning aimed to shift the tendency of the models to use certain words. While it may be viable for an effective change in behavior, it is not sufficient for creating a properly debiased LLM. In fact, it may inadvertently introduce biases in the opposite direction of existing stereotypes, which is not desirable.

Lastly, the way we formatted our Crows-Pairs dataset could be improved, especially if we keep in mind the issues mentioned in the first point. Given the weak capabilities of the small, asking them to fully rewrite a sentence without biases is too much. A more effective way, which accounts for all of these issues, would be to prompt our models with choices instead of making them do all of the work. Sadly, Crows-Pairs didn't contain enough information to use it that way.

5 Conclusion

In this paper, we studied the effectiveness of bias mitigation techniques on newer generations of large language models. We did not obtain satisfying results, but they allowed us to gain new insights into the constraints that we must face in this undertaking. We learned that the general capability of chosen models impacts greatly their understanding of the tasks, and thus, their performance. Furthermore, the said tasks must be formatted so that the models have to make as little effort as possible.

Nonetheless, there are still ways to improve our methodology. In related works, researchers spent time adding and/or freezing layers of models, instead of finetuning every parameter. Due to lack of time, we weren't able to incorporate these ideas into our experiments, but we can predict that this would allow the models to become more unbiased while still keeping most of his linguistic capabilities, which calls back to one of the issues we faced.

This work aimed to join the ongoing effort to make AI models more responsible. We hope that our contributions will help in this endeavor and give future works a better idea of the challenges they will have to face.

Acknowledgment

This work was carried out as part of our studies at the Université of Québec at Chicoutimi (UQAC). We express our thanks to our supervising teachers Darine Ameyed (Ph.D.) and Riadh Ben Chaabene for their guidance and for sharing their experience with us.

References

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Aida Mostafazadeh Davani, Ali Omrani, Brendan Kennedy, Mohammad Atari, Xiang Ren, and Morteza Dehghani. 2020. Fair hate speech detection through evaluation of social group counterfactuals. *CoRR*, abs/2010.12779.

Sunipa Dev, Emily Sheng, Jieyu Zhao, Aubrie Amstutz, Jiao Sun, Yu Hou, Mattie Sanseverino, Jiin Kim, Akihiro Nishi, Nanyun Peng, and Kai-Wei Chang. 2022. On measures of biases and harms in nlp.

Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. BOLD: dataset and metrics for measuring biases in open-ended language generation. *CoRR*, abs/2101.11718.

Tommaso Dolci. 2022. Fine-tuning language models to mitigate gender bias in sentence encoders. In *2022 IEEE Eighth International Conference on Big Data Computing Service and Applications (BigDataService)*, pages 175–176.

Michael Gira, Ruisu Zhang, and Kangwook Lee. 2022. Debiasing pre-trained language models via efficient fine-tuning. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 59–69, Dublin, Ireland. Association for Computational Linguistics.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. 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 622–628, Minneapolis, Minnesota. Association for Computational Linguistics.

Nicholas Meade, Elinor Poole-Dayan, and Siva Reddy. 2021. An empirical survey of the effectiveness of debiasing techniques for pre-trained language models. *CoRR*, abs/2110.08527.

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.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. *CoRR*, abs/1909.01326.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Vincent Zhao, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Pranesh Srinivasan, Laichee Man, Kathleen Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. 2022. Lamda: Language models for dialog applications.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

A Appendix

A.1 Hyperparameters

Model	Learning Rate	Batch size	Epochs
FLAN-T5 Small	0.01	32	50
FLAN-T5 Base	0.01	32	50
Bloomz-560	0.00003	20	5

Table 2: Fine-tuning hyperparameters for the models.

A.2 Results Summary Tables

HuggingFace Regard Score						
Model	**Pre Fine-Tuning**			**Post Fine-Tuning**		
	Neutral	**Positive**	**Negative**	**Neutral**	**Positive**	**Negative**
	Mean values					
FLAN-T5 Small	0.77	0.097	0.088	0.89	0.034	0.052
FLAN-T5 Base	0.88	0.07	0.027	0.81	0.045	0.11
Bloomz-560m	0.94	0.03	0.02	0.94	0.03	0.02
	Maximum values					
FLAN-T5 Small	0.976	0.963	0.976	0.974	0.218	0.907
FLAN-T5 Base	0.975	0.972	0.79	0.975	0.963	0.973
Bloomz-560m	0.97	0.07	0.51	0.97	0.33	0.12

Table 3: Performance on the HuggingFace Regard metric, measured using the methodology above

HuggingFace Toxicity and Honest Score				
Model	**Pre Fine-Tuning**		**Post Fine-Tuning**	
	Toxicity			
	Mean	**Max**	**Mean**	**Max**
Flan-T5 Small	0.058	0.99	0.08	0.99
Flan-T5 Base	0.03	0.99	0.08	0.99
Bloomz-560m	0.005	0.09	0.007	0.14
	HONEST			
	Female	**Male**	**Female**	**Male**
Flan-T5 Small	0	0	0.003	0.003
Flan-T5 Base	0	0.018	0	0
Bloomz-560m	0	0	0	0

Table 4: Performance on two of the HuggingFace metrics, measured using the methodology above