A study of distilled pre-trained models on StereoSet

Mehrnaz Moslemi and **The Minh Nguyen**
University of Montreal
Mehrnaz.moslemi@umontreal.ca and the.minh.nguyen@umontreal.ca

Abstract

Machine learning models can develop biases based on the patterns they see in the training data. Popular pre-trained language models such as BERT, RoBERTa, ALBERT, and GPT-2 also learn bias patterns from their training corpus. Since language models often reflect society's inequalities, we may find that the datasets for language models contain these biases, especially between genders, races, professions, and religions. Many recent papers have been trying to debias these large language models. In this project, we measure the bias and their language modeling performance on a dataset called StereoSet and its metric, icat. Our goal is to find methods that can debias models efficiently or at least convey certain properties of models' bias. Firstly, we have reproduced the experiments of BERT, RoBERTa, GPT-2, and ALBERT models on StereoSet dataset from our baseline paper Meade et al. (2022). After reproducing the result, we found a mistake in the language modeling scores of models, and we address it through our experiments. Then we have finetuned these models under our low settings using Wikitext with or without dropout configuration to see its effect on bias. We found that this configuration improves icat scores in all models. Distillation is a good practice for language model tasks to perform as well as the teacher model but in a more efficient way, so we experiment on the distilled models of BERT, RoBERTa, GPT-2 to monitor the icat behavior of these models before and after dropout configuration. We have found that using distillation decreases the performance of BERT family models (BERT and RoBERTa) while improving the GPT-2 model's icat in all target terms. We have also proposed a hypothesis that encoders-only and decoders-only transformers react differently to distillation and finetuning in terms of icat, which could be explored for deeper understanding of language models.

1 Introduction

Machine Learning systems embrace many aspects of human life. The use of these systems is increasing in many fields, including movie recommendations, abusive language detection, and who to date with. Although AI algorithms bring multiple benefits compared to humans, their decision-making could contain biases the same as human minds (Mehrabi et al., 2021). However, much work has been done to identify and mitigate bias in the different tasks and models. Still, there are inequities in the different downstream tasks, for example, coreference resolution (Zhao et al., 2018), sentiment analysis (Bhardwaj et al., 2021). Pre-trained language models could reproduce bias from the pre-training corpus (Abid et al., 2021). BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), ALBERT (Lan et al., 2020), and GPT-2 (Radford et al., 2019) are used to do many NLP tasks. So, identifying and mitigating bias in these pre-trained models is a crucial task. There are datasets with their evaluation metrics specifically designed for identifying bias in the pre-trained language models, such as StereoSet, a large-scale natural dataset in English to measure stereotypical biases in four domains: gender, profession, race, and religion, with the Context Association Tests (CATs) proposed by Nadeem et al. (2021). Their experiments were conducted on popular models to measure and compare their performance along with their stereotypical bias. On the other hand, drop-out regularization is commonly used to prevent overfitting, but it is also used to mitigate gender bias in the coreference resolution task (Webster et al., 2020). This paper Meade et al. (2022) also implemented different bias mitigation methods, including dropout regularization on the state-of-the-art pre-trained language models. This paper Meade et al. (2022) also used StereoSet and its corresponding scores (ss, lms) to evaluate bias in the pre-trained language models (BERT, RoBERTa, GPT-2 and ALBERT) in three target domains: Religion, gender, and race before and after implementing dropout regularization. Our experiments are divided into four sections; the first section reproduces the baseline results Meade et al. (2022) to verify our configuration. In this step, we reproduced the results on the StereoSet dataset and its corresponding scores on the test dataset for all pre-trained models. We compared our reproduced results' scores with the baseline paper (Meade et al., 2022), and we found that the authors mistakenly considered the same lms for all target terms of the same pre-trained models. After that, we fine-tuned these models with our low settings. We explored our settings with and without the drop-out regularization method and found that the drop-out configuration could improve the icat score. On the other hand, distillation (Hinton et al., 2015) is a good practice to efficiently use pre-trained models while having pretty much the same model performance; After observing the distilled model icat scores, we found that distillation decreases the BERT family models icat while improving GPT-2 model's score in all target terms. This paper aims to explore bias in the state-of-the-art pre-trained models in different settings and see the icat's behavior and opens a direction to researchers to find the reason behind these trends.

2 Related Works

Studies show that pre-trained language models could induce bias in the downstream tasks (Zhao et al., 2018). To identify the bias, specifically stereotypes, a new dataset named StereoSet was introduced by Nadeem et al. (2021). The novelty of the paper comes from the extensive size of the dataset compared to other previously limited sets of sentences (May et al., 2019). Nadeem et al. (2021) developed the Context Association Tests (CATs) measuring bias and language model ability at the sentence level (intrasentence) and at the discourse level (intersentence). The latter test is a new contribution since earlier work only tackled the issue at sentence level (May et al., 2019). CATs are also an intrinsic approach to evaluate bias in pre-trained models since it does not require training and does not mix the bias of task-specific training data and pre-trained representations (Kiritchenko and Mohammad, 2018). The other paper we want to focus on used drop-out regularization to mitigate gender bias in BERT in the coreference resolution task (Webster et al., 2020). There is another study that used the drop-out regularization method to mitigate bias in state-of-the-art pre-trained models, such as BERT (Meade et al., 2022). This paper only evaluated three stereotypes such as gender, race, and religion, in the intrasentence section of the Stereotype dataset with and without the dropout regularization method. In this project, we reproduced the results of BERT-Uncased on intrasentence sections of the Stereoset dataset and compared our results with the baseline model Meade et al. (2022). We also used the profession target term in our experiment while it has not been experimented with in the baseline model. For our other contribution, we fine-tuned all pre-trained models with our low-settings configuration with and without drop-out regularization which gives us low training time while the same performance. We also used the distillation of these models (BERT, RoBERTa, GPT-2) to see if this method could help with bias or not. We observed that while distillation decreased the BERT family icat score, it improves GPT-2 icat score in all target terms. This observation could open a new direction for bias detection to see why distillation results in better performance in GPT-2 models while worsening performance in BERT-family models.

3 Methods

For this experiment, we have used this repository for our experiment with several modifications `https://github.com/McGill-NLP/bias-bench`. To evaluate the scores on StereoSet, the first step is to obtain the probability for each of the options of intrasentence examples and then use the evaluation code to calculate the related scores (ss, lms, icat) for each target term (gender, race, religion, and profession). We had some modifications to the code that we will explain in the experiments' section. At the beginning, we reproduced the results for all the base models in bias-bench (Meade et al., 2021) including BERT, ALBERT, RoBERTa, and GPT-2. Then we finetuned all models under our low settings with or without dropout configuration and measured their scores on StereoSet. Towards the end of this project, we are more curious about how all the distilled models perform, so we include DistilBERT, DistilRoBERTa, and DistilGPT2 as our contributions. We carried out our experiments using files from bias-bench's GitHub website (their evaluation on StereoSet was adopted from

Nadeem et al. (2021) while their finetuning files are from Hugging Face with added debiasing functions). We have integrated the distilled models into their code with some adjustments as there was incompatibility. For finetuning, as we use a smaller dataset with less training, which we call "low settings", compared to bias-bench's settings, we redo the intrasentence test for the finetuned version of all their base models (ALBERT, BERT, RoBERTa, GPT-2). We exclude intersentence tests because some models were not pretrained on NSP. We finetune all seven models over three seeds with and without specific dropout configurations. Our first goal is to have a comparison between our finetuned base models and bias-bench's to be more confident with our low settings. Then we want to measure the effect of dropout configuration on all fine-tuned models. Finally, we want to monitor all model's trends of lms, ss, and icat after fine-tuning.

4 Experiments

4.1 Dataset

Nadeem et al. (2021) crowd-source StereoSet, a dataset for associative contexts in English containing 4 target domains (gender, profession, race, and religion), 321 target terms, and 16,995 test instances (triplets). This dataset is split into two sets: 25% of target terms for the development set and 75% for the test set. With the intrasentence CAT, if the domain is "Gender", the target is "Girl", and the context is "Girls tend to be more [blank] than boys", the model will have to choose from 3 options "soft" (stereotype), "determined" (anti-stereotype), "fish" (unrelated). Similarly, with the intersentence CAT, if the domain is "Race", the target is "Arab", and the context is "He is an Arab from the Middle East", the model will have to choose from 3 options "He is probably a terrorist with bombs" (stereotype), "He is a pacifist" (anti-stereotype), "My dog wants a walk" (unrelated). In both examples, we want the pre-trained model to show its language modeling ability by not choosing the unrelated option while having an equal chance of choosing the stereotype and anti-stereotype options. For finetuning, we use wikitext-2-raw-v1 (Merity et al., 2016) as our dataset, which is from Hugging Face and is a popular choice for training language models.

4.2 Baselines

We used the original metrics for CATs (Nadeem et al., 2021): Language Modeling Score (lms), Stereotype Score (ss), and Idealized CAT Score ($icat$). lms is defined as the percentage of instances in which a language model prefers the meaningful (either stereotype or anti-stereotype option) over meaningless (unrelated option) association, while the overall lms is the average lms. ss is defined as the percentage of examples in which the model prefers a stereotypical association over an anti-stereotypical one, while the overall ss is the average ss. Combining both of these scores, we have $icat = lms * \frac{min(ss,100-ss)}{50}$, indicating the language modeling ability of a model to behave in an unbiased manner. An ideal model would have $lms = 100,\ ss = 50,\ icat = 100$. Our baselines are all pre-trained model (BERT, RoBERTa, GPT-2, ALBERT) scores with and without dropout regularization from this paper (Meade et al., 2022). This paper does not calculate the icat score directly, so in our experiment, we calculated the icat score for all baselines. The scores on the test set are divided by the domains. For example, for BERT model without dropout regularization and gender target terms, we have $ss = 60.28$, $lms = 84.17$. For the race target terms: $ss = 57.03$, $lms = 84.17$, and for the religion target terms: $ss = 59.70$, $lms = 84.17$.

4.3 Evaluation Methods

We have used the original metrics of this paper (Nadeem et al., 2021): Language modeling score (lms), Stereotype score (ss), and Idealized CAT score (ICAT). lms defines as the percentage of the examples in which the language model prefers stereotype or anti-stereotype examples over meaningless examples or model accuracy. The other metric is designed for measuring the stereotypical bias in the language models, which is the ss. ss is defined as the percentage of examples in which the model prefers stereotype association over anti-stereotypes. Combining both of these scores, we have $icat = lms * \frac{min(ss,100-ss)}{50}$, indicating the language modeling ability of a model to behave in an unbiased manner. An ideal model would have $lms = 100,\ ss = 50,\ icat = 100$.

4.4 Experimental details

With a change of plan from the proposal and midway report, we conducted the experiments from (Meade et al., 2021), with some modifications and novelties; this is the related repository: https://github.com/McGill-NLP/bias-bench. Firstly, we reproduced the baselines of four base models: BERT, ALBERT, RoBERTa, GPT-2. Next, we added DistilBERT, DistilRoBERTa, and DistilGPT2 to evaluate the effect of distillation on bias. We integrated these models into bias-bench's programs and solved any incompatibility such as the lack of token_type_ids in DistilBERT. Finally, we finetuned all seven models on their pretrained tasks, with and without specific dropout configurations, using dataset wikitext-2-raw-v1 (Merity et al., 2016) for 3 epochs. The batch size ranges from 2 to 8 to fit into our gpu depending on how large the models are. These settings require less training time than bias-bench's finetuning settings. In this report, we will compare our finetuned models to bias-bench's to see if our settings are reasonable. The specific dropout configurations with hidden and attention probabilities, respectively, are 0.20 and 0.15 (for BERT, DistilBERT, RoBERTa, DistilRoBERTa), 0.05 and 0.05 (for ALBERT) while the resid_pdrop, embd_pdrop, attn_pdrop are all 0.15 (for GPT-2, DistilGPT2). We did 3 seeds for each finetuned version for more reliable results. We also added $fp16$ training argument to avoid the CUDA memory problem. The files that we used for finetuning are bias-bench's $run\_mlm.py$ and $run\_clm.py$, which are from Hugging Face, but with added functions for debiasing methods. We have adjusted the dropout function in these two files for the distilled models. For clarity, we add the suffix "baseline" for the models from the bias-bench, "finetuned" for the finetuned models with default dropout probabilities, and "dropout" for the finetuned models with dropout configuration.

Model	ss	lms	icat
Gender			
BERT-baseline	60.3	84.2	66.9
BERT	60.3	85.7	68.1
ALBERT-baseline	59.9	89.8	73.1
ALBERT	59.9	89.4	71.6
RoBERTa-baseline	66.3	88.9	59.9
RoBERTa	66.3	89.8	60.5
GPT-2-baseline	62.7	91.0	68.0
GPT-2	62.7	92.0	68.7
Race			
BERT-baseline	57.0	84.2	72.3
BERT	57.0	84.0	72.2
ALBERT-baseline	57.5	89.8	76.3
ALBERT	57.5	90.3	76.7
RoBERTa-baseline	61.7	88.9	68.2
RoBERTa	61.7	89.9	68.9
GPT-2-baseline	58.9	91.0	74.8
GPT-2	58.9	90.3	74.8
Religion			
BERT-baseline	59.7	84.2	67.8
BERT	59.7	84.2	67.9
ALBERT-baseline	60.3	89.8	71.3
ALBERT	60.3	92.7	73.6
RoBERTa-baseline	64.3	88.9	63.5
RoBERTa	64.3	88.0	62.9
GPT-2-baseline	63.3	91.0	66.9
GPT-2	63.3	91.2	67.0
Profession			
BERT	58.9	83.9	68.9
ALBERT	60.4	89.0	70.5
RoBERTa	61.5	87.5	67.4
GPT-2	61.3	90.7	70.2

Table 1: Reproduction of the baselines (Meade et al., 2022)

4.5 Results and Analysis

4.5.1 Baselines reproduction

Regarding to our baseline Meade et al. (2022); first, we reproduced the experiments with the four popular pre-trained models (RoBERTa, BERT, AlBERT, GPT-2) on the stereo set dataset. We compare our results with the baseline to see if our reproducing results are reliable or not. In this stage, we could also find which pre-trained model is more biased.

In Table 1, we reproduced the experiments on the four popular pre-trained models (BERT, RoBERTa, ALBERT, GPT-2) and compared the results with the baseline models (Meade et al., 2022). All replicated scores are almost the same except for the accuracy scores (lms) for all pre-trained models (by a small deviation). We think the authors might mistakenly use the average lms to replace the lms for each target. Intuitively, the accuracy of all target terms could not be the same. Since Meade et al. (2022) did not mention icat, we calculated icat based on the respective ss and lms from their results. Totally we could rely on our reproducing results to do the rest of the experiments. After a quick look, we could find that ALBERT has the best performance (icat score) compared to other pre-trained models in all target terms. BERT, GPT-2, and RoBERTa have their best performance in the Race examples and ALBERT has the best performance in the Religion examples. The baseline paper Meade et al. (2022) does not consider "profession target terms" in the Stereoset dataset, but in our settings, we include this dataset as well to see the models' performance on this target term. In the following, we fine-tune these models and measure the previous scores to see the models' behavior after fine-tuning.

4.5.2 Low settings vs bias-bench's settings

For fine-tuning, we have used wikitext-2-raw-v1 instead 10% of Wikipedia dump-like bias-bench. Our dataset is smaller, which requires less training time with 3 epochs. This makes it feasible for us to finetune 2 versions (with and without dropout configuration) of each model and 3 seeds for each version which leads to a total of 42 finetuned models.

Model (dropout)	ss	lms	icat
Gender			
BERT-baseline	60.3	83.1	66.1
BERT	59.7	85.4	68.9
ALBERT-baseline	57.4	77.5	66.1
ALBERT	56.5	83.7	72.8
RoBERTa-baseline	66.2	88.8	60.0
RoBERTa	65.8	89.7	61.3
GPT-2-baseline	63.1	90.4	68.8
GPT-2	64.1	90.7	65.1
Race			
BERT-baseline	57.0	83.1	71.6
BERT	56.3	83.4	72.8
ALBERT-baseline	51.6	77.5	75.0
ALBERT	51.5	82.0	79.5
RoBERTa-baseline	60.5	88.8	70.2
RoBERTa	59.6	90.2	72.9
GPT-2-baseline	57.5	90.4	76.9
GPT-2	56.5	88.8	77.3
Religion			
BERT-baseline	59.7	83.1	67.0
BERT	59.2	82.6	67.5
ALBERT-baseline	54.7	77.5	70.2
ALBERT	55.3	84.7	75.7
RoBERTa-baseline	62.5	88.8	66.6
RoBERTa	63.0	87.4	64.8
GPT-2-baseline	64.3	90.4	64.6
GPT-2	61.7	89.8	68.9
Profession			
BERT	58.2	83.4	69.7
ALBERT	58.5	81.9	68.0
RoBERTa	61.6	87.3	67.0
GPT-2	61.3	89.2	69.0

Table 2: Reproduction of the baselines, models implemented with drop out configuration (Meade et al., 2022)

In Table 2, we want to compare our results of dropout models to bias-bench's to be more confident in our low settings. As we could see in the table, all the scores with our low settings exceed those of the baseline models in all target terms. The exception is for GPT-2 in the gender target term and RoBERTa in the religion target where we have lower icat. However, in the grand scheme, we can say that our settings are reasonable to conduct further experiments.

4.5.3 Finetune with and without dropout

In this stage, we have used the drop-out regularization method on the four pre-trained models to see the effect of this method on the icat score. This method previously is used to prevent over-fitting in machine learning models, and it is proven that it could mitigate bias in language models Webster et al. (2020). We used this method in our low settings and fine-tuned models to see the behavior of our models in terms of icat score.

Model	ss	lms	icat
ALBERT	59.0	89.8	73.6
ALBERT-finetuned	56.0	82.3	72.5
ALBERT-dropout	54.9	82.3	74.3
BERT	58.2	84.2	70.3
BERT-finetuned	57.8	83.5	70.4
BERT-dropout	57.6	83.6	71.0
RoBERTa	62.3	88.9	67.1
RoBERTa-finetuned	61.3	88.8	68.7
RoBERTa-dropout	61.2	89.0	69.0
GPT-2	60.4	91.0	72.0
GPT-2-finetuned	59.5	89.3	72.3
GPT-2-dropout	59.4	89.2	72.4

Table 3: Comparison between base models, finetuned models (finetuned without dropout) and dropout models (finetuned with dropout), respectively.

As we can see in Table 3, drop-out configuration improved ss for all fine-tuned models while lowering their lms. Fortunately, the improvement of ss is enough to outweigh the decrease in lms, which results in a better icat for all models. Dropout models tend to have slightly better icat as their ss is closer to 50. Overall, dropout configuration provides a tradeoff between ss and lms for a small icat's improvement. One exception is RoBERTa where its lms is higher or almost the same after finetuning. This could be an effect of its robustness.

4.5.4 Consider distilled models

Knowledge distillation is a process of training smaller models to mimic the performance of larger and more complex models. These models use fewer resources. In this project, we are curious whether these models could perform better in icat scores or not. Table 4 shows our results for the three base models and their corresponding distilled variants. We also include "finetuned" and "dropout" models to analyze the trends of ss, lms, and icat.

Model	ss	lms	icat
BERT	58.2	84.2	70.3
BERT-finetuned	57.8	83.5	70.4
BERT-dropout	57.6	83.6	71.0
DistilBERT	59.2	85.1	69.5
DistilBERT-finetuned	58.5	83.7	69.4
DistilBERT-dropout	58.4	83.6	69.5
RoBERTa	62.3	88.9	67.1
RoBERTa-finetuned	61.3	88.8	68.7
RoBERTa-dropout	61.2	89.0	69.0
DistilRoBERTa	61.7	89.2	68.4
DistilRoBERTa-finetuned	59.0	83.2	68.3
DistilRoBERTa-dropout	59.1	83.7	68.4
GPT-2	60.4	91.0	72.0
GPT-2-finetuned	59.5	89.3	72.3
GPT-2-dropout	59.4	89.2	72.4
DistilGPT2	58.9	89.2	73.3
DistilGPT2-finetuned	57.9	88.4	74.5
DistilGPT2-dropout	57.6	88.3	74.8

Table 4: Comparison between teacher and student models and their corresponding finetuned versions.

Without any finetuning, DistilBERT has more bias and is also better at language modeling than its teacher. The worse ss agrees with the results from Ahn et al. (2022) as they observe the amplified bias in DistilBERT, specifically. The better lms can be explained by the absence of dynamic masking in BERT's pretraining technique. After finetuning (with and without dropout configuration), DistilBERT's results follow the same trend that we have seen earlier with better ss and worse lms. However, the drop in lms is bigger and keeps the icat the same or even slightly worse.

For DistilRoBERTa and RoBERTa, the gap of their lms is smaller than that of DistilBERT and BERT, at the checkpoint. This can be explained by the fact they were both pretrained with dynamic masking and the hypothesis that a smaller model can generalize slightly better. It is unexpected to see DistilRoBERTa's ss is lower than its teacher's at the checkpoint. Again, DistilRoBERTa's icat has little to no development like DistilBERT's after finetuning, with a significant drop in lms.

As being from another family of models, DistilGPT2 is different with lower ss and lms compared to its teacher at the checkpoint. This contradicts the conclusion from Gupta et al. (2022) as they confirm DistilGPT2 has more bias than GPT-2. We can argue that their tests are specific to a type of bias, a combination of profession and gender, while we measure bias for all 4 targets. We observe somewhat of a reverse trend in icat after finetuning. For the BERT family, the teachers' icat improved while the students' remain consistent. For the GPT family, the teacher has a relatively small increase compared to its students.

In addition, using distillation as a debiasing method seems to be ineffective for BERT family models as the teachers have higher icat than their corresponding students while being effective for GPT family models as the student has higher icat.

4.6 Discussion

To sum up our results, DistilGPT2-dropout achieves the best icat (74.8) out of all models that we have tested with the second place goes to ALBERT-dropout at (74.3). At checkpoint, the ranking is reversed as ALBERT has an icat of 73.6 while DistilGPT2 has an icat of 73.3. Because StereoSet has a limited size as a dataset and its metrics (ss, lms, icat) can be unreliable, our tests cannot be treated as an absolute way to measure bias and language modeling performance. Therefore, DistilGPT2-dropout and ALBERT-dropout should be seen as a good suggestion for other work.

However, we are interested in explaining the less obvious results in our experiments, the trend of icat between the checkpoint and finetuned versions. Δ icat calculated in Table 5 are between dropout and checkpoint models. On the one hand, we observe that the teachers in the BERT family always have a moderate amount of increase in icat while their students have no improvement. On the other hand, in the GPT family, GPT-2 only has a relatively small boost while DistilGPT2 has the biggest jump in

Model	Δ icat
ALBERT	+ 0.7
BERT	+ 0.7
DistilBERT	+ 0
RoBERTa	+ 1.9
DistilRoBERTa	+ 0
GPT-2	+ 0.4
DistilGPT2	+ 1.5

Table 5: icat's trend after finetuning with dropout configuration.

icat. Although we are simplifying some factors such as the different reactions of each model to finetuning, we believe that this phenomenon is the outcome of the difference in each family's special structure. We propose a bold hypothesis for their properties that if both bias and language modeling performance is considered, encoders-only transformers tend to have better behaviors when they are full-scale, whereas decoders-only transformers tend to have better behaviors when they are compressed, which is fascinating for us to verify or debunk in further research. This also links to the results that distillation seems to work better for decoders-only models like GPT-2. As far as we know, there has been no research that covers bias in ALBERT, BERT, RoBERTa, GPT-2, and their respective distilled variants taking into account their behavior after finetuning. Since all models are assessed on the same settings and environment over three seeds, we believe that any pattern we can find may be worth looking into.

There are some limitations in our project that we wish to address in our future work. One of them is the potential difference in each model's pattern after finetuning. In our project, we make an implicit assumption that each model behaves similarly given the same settings. We are curious to see if in other settings, for example, bias-bench's, the earlier mentioned phenomenon still exists. In addition, we would also need to analyze our data more to reconcile with other work on the bias since there could be some contradicting beliefs at the moment. For example, Gupta et al. (2022) says DistilGPT2 has more bias than GPT-2, and Ahn et al. (2022) says DistilBERT has more bias than BERT. We disagree with the former while agreeing with the latter. Our opinion is compressed models may not have the same bias pattern if we also consider their architecture (encoders-only versus decoders-only).

4.7 Conclusion

In conclusion, we have presented the comprehensive bias experiment on the most prominent pretrained models (BERT, RoBERTa, GPT-2, and ALBERT) to measure bias trends with different configurations in four main stereotypical biases (gender, race, profession, and religion) using StereoSet. We have found that there is a mistake in the lms scores of (Meade et al., 2022) paper which consider the same scores for all target terms. After that, we implemented our fine-tuned low-settings model with efficient training time while having the same performance as the baseline model. As distillation is a useful practice to have efficient training with the same performance, we explored some distilled models to measure their absolute performance in icat and observe the icat's trend in these models. We realize that dropout configuration is beneficial for all models' icat while distillation is only effective on GPT family models. In terms of absolute scores, we found that finetuned DistilGPT2 with dropout configuration achieves the best icat out of all experiments, including checkpoints and finetuned models. Finally, our project discovers a bizarre phenomenon about the different patterns of the icat trends between BERT family and GPT family. At the moment, we makes a proposal that encoders-only and decoders-only transformers could be the factor that explains this observation, which would be our future research direction.

References

Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent anti-muslim bias in large language models. New York, NY, USA. Association for Computing Machinery.

Jaimeen Ahn, Hwaran Lee, Jinhwa Kim, and Alice Oh. 2022. Why knowledge distillation amplifies gender bias and how to mitigate from the perspective of DistilBERT. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 266–272, Seattle, Washington. Association for Computational Linguistics.

Rishabh Bhardwaj, Navonil Majumder, and Soujanya Poria. 2021. Investigating gender bias in bert. *Cognitive Computation*, 13(4):1008–1018.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Umang Gupta, Jwala Dhamala, Varun Kumar, Apurv Verma, Yada Pruksachatkun, Satyapriya Krishna, Rahul Gupta, Kai-Wei Chang, Greg Ver Steeg, and Aram Galstyan. 2022. Mitigating gender bias in distilled language models via counterfactual role reversal.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network.

Svetlana Kiritchenko and Saif Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 43–53, New Orleans, Louisiana. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations.

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.

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. *arXiv preprint arXiv:2110.08527*.

Nicholas Meade, Elinor Poole-Dayan, and Siva Reddy. 2022. An empirical survey of the effectiveness of debiasing techniques for pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1878–1898, Dublin, Ireland. Association for Computational Linguistics.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models.

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.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, Ed Chi, and Slav Petrov. 2020. Measuring and reducing gendered correlations in pre-trained models. *arXiv preprint arXiv:2010.06032*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. *arXiv preprint arXiv:1804.06876*.