Compression of Large Language Models and Its Effect on Bias

Owen Finucan
University of Pittsburgh
otf6@pitt.edu

Jonah Smith
University of Pittsburgh
jws130@pitt.edu

Abstract

Large Language Models (LLMs) are often trained on a vast corpus of text from the internet. These LLMs pick up on any social bias exhibited throughout the training material. We have developed multiple tools and datasets to demonstrate the level of bias in a given LLM. Additionally, due to the sheer volume of data trained on, the models are often quite large in size. We perform a study on the impact of LLM compression (via quantization and knowledge distillation) on social biases.

1 Introduction

Large Language models have become both ubiquitous and extremely important across a wide range of applications. However, current training methods expose models to the embedded bias of the training data. As a potential means of both bias mitigation and model efficiency, compressing models might serve as a valuable tool. Our replication of Gonçalves and Strubell (2023) aims to replicate their results regarding how compression affects bias and modeling performance across different models.

2 Summary of Gonçalves and Strubell (2023)

Gonçalves and Strubell (2023) investigated how model compression techniques (quantization and knowledge distillation) affect social bias in LLMs. The study addressed bias mitigation, model efficiency, and whether smaller/faster models exhibit altered bias profiles.

Methodology

The authors used several encoder and decoder LLMs, including BERT, RoBERTa, and Pythia, across various parameter scales (70M-6.9B). Compression was applied using:

- Dynamic post-training quantization (PTQ): converting 32-bit floating point weights to 8-bit integers after training.
- Knowledge distillation: transferring knowledge from large "teacher" models to smaller "student versions.

Measuring Bias

Bias was measured using Bias Bench, which unifies three datasets:

- CrowS-Pairs - minimal sentence pairs testing stereotypical vs. anti-stereotypical preferences.
- StereoSet - masked-language prediction tasks measuring bias and overal language modeling ability.
- SEAT - sentence-level tests assessing embedding-based associations for Gender, Race, and Religion.

Findings

- Quantization and distillation both reduced bias.
- Quantization acted as a regularizer, lowering bias while maintaining reasonable modeling capability (LM score).
- Larger and longer-trained models exhibited higher social bias.
- Quantized models reached their optimal LM-bias trade-off earlier in pretraining.

3 Project Implementation

The project is more than underway. We have broken the project into five different parts, labeled A-E respectively.

3.1 Section A - Setup

Section A is titled "Setup". Here, we initialize everything that could be needed for the project. This includes connecting to Hugging Face (for model access), seed setting (for reproducibility), and relevant helper methods. These helper methods are utilized to help load models (and their respective tokenizers), along with helping in dynamic post-training of the linear layer. Section A also features a function to re-pull the CrowS and StereoSet datasets from Hugging Face. This section is fully completed as described.

3.2 Section B - Evaluation Methods

Section B is titled "Replication Pipeline". Here, we initialize the functions used throughout the rest of the paper. These functions include the masked-LM pseudo-log-likelihood (PLL) scorer and robust evaluators for the CrowS and StereoSet datasets. We also run smoke testing on these functions with their respective scoring functions to ensure proper functionality. This section is fully completed as described.

3.3 Section C - Quantization Results

Section C is titled "Quantization Results". Here, we perform the quantization of the model. We call the appropriate functions to perform scoring. These results are then piped to a .CSV and .JSON for storage. This section, while fully coded, is not complete. As of now, only partial testing has been completed. All that needs to be done is run each of the models (BERT Base Uncased, RoBERTa Base, DistilBERT Base Uncased, DistilRoBERTa Base, BERT Large Uncased, and RoBERTa Large) over the full datasets.

3.4 Section D - Quantization Results

Section D is titled "Analysis & Reporting". Here, we perform all of the statistical analyses of the paper. As of now, the only piece of this section done is the Stereoset LM Score.
In the original paper, the analysis of bias effects was evaluated using point estimates derived from benchmark percentages on CrowS-Pairs and StereoSet. For a more robust analysis, we implemented a set of inferential statistical tests to formally test whether bias measuring how compression affects bias in specific models.

For the initial experiments, we introduced the following statistical tests:

- **Ordinary Least Squares (OLS) regression:** Models the continuous bias scores as a function of compression status.
- **Independent-samples *t*-test:** Compares mean bias levels between compressed and uncompressed models to test for significant differences in central tendency.
- **Bootstrap confidence intervals:** Quantify the uncertainty in the point estimates without assuming normality, providing a non-parametric measure of precision.

This statistical extension transforms the experiment from a purely descriptive replication into an inferential analysis framework, enabling formal hypothesis testing and the derivation of statistically sound conclusions rather than reliance on point estimates alone. While the foundational statistical tests are fully implemented, they are not yet producing conclusive results due to the current inclusion of only a single model in the pipeline (resulting in zero degrees of freedom). As additional models are incorporated, this framework will produce the statistical outputs we need to decide whether compression is conclusively affecting bias.

3.5 Section E - Extension

Currently, our results suggest that overall bias levels decrease following compression (based on the two models analyzed thus far), and that this downward trend persists even when stratifying across individual dimensions of bias such as *gender*, *race*, and *religion*. We plan to extend this analysis to additional social categories to more robustly evaluate how category type influences the extent of bias reduction.

Furthermore, we are considering experiments with varying levels of compression to investigate how the *magnitude* of compression relates to measured bias. This will allow us to test whether there exists a monotonic or nonlinear relationship between model size reduction and bias level. We intend to pursue this line of inquiry provided that sufficient computational resources are available.

4 Current Results

Currently, we have results from 2 different models: BERT Base Uncased and RoBERTa Base. Table 1 has our results while Table 2 contains results from the original paper from Gonçalves and Strubell

(2023). The results are consistent in outcome with that of the paper.

Given the completed compressed and uncompressed models, we can conduct a more detailed analysis of the results. Bootstrap confidence intervals for CrowS bias show that uncompressed FP32 models (95% CI $\approx$ [60.9, 61.3]) exhibit higher bias than their quantized INT8 counterparts (95% CI $\approx$ [56.4, 57.2]). This implies that, even when accounting for the inherent variance associated with such a small sample size ($n = 2$), bias is consistently and substantially reduced following compression.

Furthermore, we applied two additional formal hypothesis tests to confirm these findings.

A two-sample *t*-test was performed to determine whether the mean bias levels differ between compressed and uncompressed models. The test produced a p-value of 0.0325, suggesting that the observed difference in mean bias between FP32 and INT8 models is statistically significant. In other words, we reject the null hypothesis that the two means are equal, providing corroborating evidence for the bootstrap results.

We then conducted an Ordinary Least Squares (OLS) regression, modeling the continuous bias scores (Y) as a function of compression status (X), where $X = 0$ represents uncompressed FP32 models and $X = 1$ represents quantized INT8 models. This regression tests whether the slope coefficient (β_1) differs significantly from zero—that is, whether compression has a measurable effect on bias. The regression yielded a p-value of 0.01, indicating a significant linear relationship between compression and bias. This further reinforces the conclusion that quantization leads to a reduction in measured social bias.

In addition to the overall bias reduction, we also examined category-specific effects by stratifying the CrowS bias into *gender*, *race*, and *religion* dimensions. Table A7 summarizes the bootstrap confidence intervals and *t*-test results for each category.

Across all three dimensions, the mean bias decreased under compression, with INT8 models consistently showing lower bias values than their FP32 counterparts. For gender, the difference was statistically significant ($t = 8.50$, $p = 0.0316$), indicating that quantization notably reduces gender-related bias. The race and religion categories also exhibited reductions in mean bias ($\Delta_{\text{race}} = 3.39$, $\Delta_{\text{religion}} = 13.81$), though these differences were not statistically significant at the 0.05 level ($p = 0.11$ and $p = 0.15$, respectively), likely due to the small sample size ($n = 2$). Nevertheless, the direction of change is consistent across categories, supporting the conclusion that compression generally reduces model bias. We predict that the tests will show statistical significance when we add more models.

5 Teamwork Description

Owen was responsible for the development and deployment of the compression model pipeline. Jonah was responsible for everything involving the statistical analysis of the results.

References

G. Gonçalves and E. Strubell. 2023. Understanding the effect of model compression on social bias in large language models. *arXiv preprint arXiv:2312.05662*.

Appendix

Table A1: Our Results Uncompressed

Model	Size (MB)	Gender	Race	Religion
BERT Base	417.827	57.252	59.690	73.333
RoBERTa Base	475.747	57.634	58.140	69.524

Table A2: Our Results Compressed

Model	Size (MB)	Gender	Race	Religion
BERT Base	195.545	53.435	55.814	61.904
RoBERTa Base	267.925	54.198	55.232	53.333

Table A3: Gonçalves and Strubell (2023) Uncompressed Results

Model	Size (MB)	Gender	Race	Religion
BERT Base	438	57.25	62.33	62.86
RoBERTa Base	498	60.15	63.57	60.95

Table A4: Gonçalves and Strubell (2023) Compressed Results

Model	Size (MB)	Gender	Race	Religion
BERT Base	181	57.25	62.14	46.67
RoBERTa Base	242	53.64	58.53	49.52

Table A5: Our reproduction vs. Gonçalves and Strubell (2023). Compression ratio is FP32/INT8 (higher is better). CrowS "Difference" values are percentage point change from FP32 to INT8 (negative = improved)

Metric	Ours	Paper	Δ Abs	$\Delta 50$ Ours	$\Delta 50$ Paper
Compression ratio (Higher Better)					
`bert-base-uncased`	2.137	2.420	-0.283	—	—
`roberta-base`	1.776	2.057	-0.281	—	—
CrowS-Pairs: Δ stereotype (Lower Better)					
BERT Gender	-3.817	0.000	-3.817	3.435	7.25
RoBERTa Gender	-3.436	-6.510	3.074	4.198	3.64
BERT Race	-3.876	-0.190	-3.686	5.814	12.14
RoBERTa Race	-2.908	-5.040	2.132	5.232	8.53
BERT Religion	-11.429	-16.190	4.761	11.904	3.33
RoBERTa Religion	-16.191	-11.430	-4.761	3.333	0.48

Δ Abs = Ours $-$ Paper; $\Delta 50$ = score $-$ 50 (percentage points).

Table A6: Summary of Statistical Tests on CrowS Bias

Test	Statistic	p-value	Conclusion
Two-sample *t*-test	$t = 9.923$	0.0325	Significant difference in means
OLS Regression (slope β_1)	$t = -9.923$	0.010	Significant negative relationship
Bootstrap 95% CI (FP32)	–	–	[60.94, 61.27]
Bootstrap 95% CI (INT8)	–	–	[56.43, 57.23]

Statistical tests applied to aggregate CrowS-Pairs bias across two models (BERT Base and RoBERTa Base).

Table A7: Category-Specific Bias Comparison (Bootstrap and *t*-Tests)

Category	FP32 Mean	INT8 Mean	Mean Difference	*t*, p-value
Gender	57.443	53.817	3.626	$t = 8.50, p = 0.0316$
Race	58.915	55.523	3.392	$t = 4.10, p = 0.1100$
Religion	71.429	57.619	13.810	$t = 3.21, p = 0.1500$

Values reflect average CrowS-Pairs stereotype scores (lower = less bias) across BERT Base and RoBERTa Base.
All categories show a reduction in mean bias after quantization (INT8).