Effect of Variable Bit Quantization on Bias within Large Language Models

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 web-scale corpora that encode social stereotypes, and prior work has shown that these models can reproduce and, in some cases, amplify such biases. Simultaneously, deploying LLMs in real-world settings almost always requires model compression; one such method is quantization. However, the interaction between compression bit-width and measured bias remains underexplored. We study how post-training quantization to variable bit-widths affects social bias in masked language models. Using BERT, DistilBERT, RoBERTa, and DistilRoBERTa, we apply both PyTorch's dynamic 8-bit quantization and a custom uniform N-bit scheme spanning 2-30 bits. We evaluate each model and bit-width on CrowS-Pairs and StereoSet to measure bias. For the 8-bit regression, we observe a statistically significant $\sim$5-point reduction in CrowS-Pair bias relative to full-precision. Across bit-widths, moderate quantization (6-12 bits) tends to lower stereotype scores while preserving utility. We see that aggressive quantization (2-4 bits) destabilizes behavior and can increase stereotyping. These results suggest that quantization can act as a bias regularizer within narrow precision windows but should not be treated as a standalone debiaser.

1 Introduction

To achieve optimal performance, large language models (LLMs) are trained on vast corpora of text. Prior work has shown that when biased data is included in the training set, models tend to reproduce, and sometimes amplify, those biases (Bender et al., 2021; Liang et al., 2021; Hutchinson et al., 2020; Dev et al., 2020; Gallegos et al., 2024a; Navigli et al., 2023). Existing work proposes three main approaches to mitigate these effects: filtering or augmenting datasets, debiasing during training, and debiasing during prompting, as well as hybrids that combine these stages (Navigli et al., 2023; Zhao et al., 2018a; Park et al., 2018; Gehman et al., 2020; Welbl et al., 2021; Prabhu and Birhane, 2020; Zhao et al., 2018b; Bolukbasi et al., 2016; Yarrabelly et al., 2024; Liang et al., 2021; Schick et al., 2021; Gallegos et al., 2024b; Sheng et al., 2020). Data filtering has been shown to be effective (Navigli et al., 2023; Zhao et al., 2018a; Park et al., 2018; Gehman et al., 2020; Welbl et al., 2021; Prabhu and Birhane, 2020); however, there are consequences to performing it. It has the potential to entirely remove reclaimed or minority dialects from datasets (Welbl et al., 2021), and several authors argue that comprehensive filtering of web-scale corpora, particularly those used for LLM pre-training, is unrealistic in practice (Bender et al., 2021; Prabhu and Birhane, 2020; Sambasivan et al., 2021). Training-time methods modify model structure or objectives to reduce correlations between internal features and human demographics (Zhao et al., 2018b; Bolukbasi et al., 2016; Yarrabelly et al., 2024; Liang et al., 2021). These demographics include gender, race, and religion. Critics contend that these methods merely obfuscate the model's bias from the user and allow it to still reside within the model (Gonen and Goldberg, 2019). Prompt-based debiasing has shown some success in reducing measured stereotyping in model outputs (Schick et al., 2021; Gallegos et al., 2024b; Sheng et al., 2020), yet, similar to training-time approaches, it leaves the underlying biased information in the model and instead attempts to hide or counteract it at generation time (Gonen and Goldberg, 2019).

At the same time, state-of-the-art LLMs are often too large and expensive for many real-world settings, which makes model compression effectively unavoidable (Dettmers et al., 2022; Ganesh et al., 2020; Xu and McAuley, 2022; Zhu et al., 2023; Jin et al., 2024; Xu et al., 2024a; Frantar and Alistarh, 2023; Gu et al., 2023; Xu et al., 2024b). Techniques such as quantization (Dettmers et al., 2022; Jin et al., 2024), pruning (Xu et al., 2024a; Frantar and Alistarh, 2023), and knowledge distillation (Gu et al., 2023; Xu et al., 2024b) are widely used to reduce LLM memory requirements. We are particularly interested in quantization. Minimal research has been conducted on the effects of quantization to different bit-widths on bias in LLMs. We raise the question: how does quantizing an LLM to different bit widths change its measured bias?

In this work, we address this question by systematically quantizing LLMs to multiple bit widths and evaluating the resulting models with established bias benchmarks, comparing how different levels of quantization affect both model performance and measured bias.

2 Related Work

Quantization. Researchers have performed a variety of studies investigating the impact of quantization on LLM behavior (Dettmers et al., 2022; Wang et al., 2025; Liu et al., 2023; Li et al., 2025; Xu et al., 2024c; Lin et al., 2023; Frantar et al., 2022). Across these works, quantized models consistently show some change in performance relative to their full-precision counterparts. A common finding is that as bit-width drops below 4 bits, model performance often degrades sharply, particularly on reasoning-heavy benchmarks (Li et al., 2025; Xu et al., 2024c). At the same time, quantization within a narrow range of bit-widths (e.g., roughly three bits of a baseline) can yield broadly comparable performance across many tasks (Frantar et al., 2022). Other work shows that not all weights or features contribute equally to maintaining behavior under quantization. This means that protecting a small subset of "salient" parameters can preserve performance while aggressively quantizing the rest (Dettmers et al., 2022; Lin et al., 2023). Recent studies further demonstrate that quantization can alter a model's factual knowledge, inducing factual forgetting in specific "knowledge neurons" and changing how much factual information the model can reliably recall (Wang et al., 2025; Liu et al., 2023). Collectively, these results characterize quantization as a weighted, lossy compression process. We extend this line of work by asking whether demographic and bias-related information is among the internal factors that are distorted as bit-width varies.

Bias. Work on bias in language models typically assumes full-precision models, focusing on data-level, training-time, or inference-time debiasing methods rather than changes to the underlying numeric representation (Bender et al., 2021; Liang et al., 2021; Gallegos et al., 2024a; Navigli et al., 2023). Data filtering and counterfactual augmentation are techniques that can be used to reduce bias (Zhao et al., 2018a; Park et al., 2018; Gehman et al., 2020), but they are challenging to scale to web-scale pretraining corpora (Welbl et al., 2021; Prabhu and Birhane, 2020). Additionally, they risk disproportionately removing reclaimed or minority dialects (Welbl et al., 2021; Prabhu and Birhane, 2020). Training-time and prompt-based methods adjust representations or generation behavior to steer model outputs away from demographic-based biases (Zhao et al., 2018b; Bolukbasi et al., 2016; Yarrabelly et al., 2024; Schick et al., 2021; Gallegos et al., 2024b; Sheng et al., 2020). Critics argue that these techniques are ineffective as they mask bias in evaluation metrics rather than removing the bias itself (Gonen and Goldberg, 2019). As in these works, we will explore the implications of bias in LLMs.

Quantization with Bias. Beyond debiasing, more recent work has examined how model compression interacts with bias. Hooker et al. (2020) showed that pruning and quantization can amplify algorithmic bias, disproportionately harming underrepresented subgroups (Hooker et al., 2020). Gonçalves and Strubell (2023) conduct a controlled study of quantization and knowledge distillation in LLMs, reporting relatively mild reduction with standard social-bias metrics (Goncalves and Strubell, 2023). More recent work finds that post-training quantization can cause substantial response-level flips between biased and unbiased outputs that aggregate scores fail to capture (Marcuzzi et al., 2025). It introduces fairness-aware quantization schemes, such as Fair-GPTQ, which augment the quantization objective with group-fairness constraints to reduce stereotype generation explicitly (Proskurina et al., 2025). Our work complements these studies by focusing specifically on how varying quantization bit-width reshapes measured demographic bias in LLMs.

3 Setup, Method, & Approach

3.1 Models

For this experiment, we primarily rely on four pre-trained masked language models: BERT-Base-Uncased, DistilBERT-Base-Uncased, RoBERTa-Base, and DistilRoBERTa-Base (Devlin et al., 2019; Sanh et al., 2019; Liu et al., 2019). These four models are evaluated at all bit-widths considered in this work. To compare against prior results on bias within compressed large models, we additionally evaluate BERT-Large-Uncased and RoBERTa-Large (Devlin et al., 2019; Liu et al., 2019) in the full-precision and 8-bit settings only, following Goncalves and Strubell (2023).

3.2 Bit-Width

To span both aggressive and mild compression regimes, we evaluate models at bit-widths of 2, 4, 6, 8, 12, 16, 20, 24, 28, and 30 bits. As discussed in Section 3.1, the large models (BERT-Large and RoBERTa-Large) are only evaluated in the full-precision (FP32) and 8-bit settings.

3.3 Measuring Bias

We measure bias using two established benchmarks: CrowS-Pairs (Nangia et al., 2020) and StereoSet (Nadeem et al., 2021). For each model and bit-width, we compute:

- **CrowS Overall Stereotype Percentage**
- **CrowS Group-Specific Stereotype Percentages**
- **StereoSet Stereotype Percentage**
- **StereoSet LM-OK Percentage**

For base-sized models, we compute these metrics for the full-precision model and for all quantized variants. For large models, we compute them only for the unquantized and 8-bit-quantized versions.

CrowS-Pairs. CrowS-Pairs is a sentence-level benchmark designed to measure stereotypical associations in language models across multiple demographics (Nangia et al., 2020). These demographics include race, gender, sexual orientation, religion, nationality, age, and disability status (Nangia et al., 2020). Each entry consists of a minimal pair of sentences. The first reflects a social stereotype, while the other exemplifies its opposite, an anti-stereotype. Following the original setup for masked language models, we compute pseudo-log-likelihoods for both sentences and record whether the model predicts a higher probability for the stereotypical variant. The **CrowS Overall Stereotype Percentage** is defined as the percentage of pairs for which the stereotypical sentence is preferred. The **CrowS Group-Specific Stereotype Percentages** are defined analogously, but restricted to a particular demographic category, allowing us to analyze how quantization affects different groups. Higher CrowS scores indicate stronger bias (more frequent stereotypical preferences), lower scores indicate preference for anti-stereotypes, and a score of 50 indicates no preference between stereotype and anti-stereotype (i.e., unbiased).

StereoSet. StereoSet measures stereotypical associations in models using sentence triplets. One sentence contains a stereotype, another sentence contains the opposite of the stereotype in the first, and the last is an unrelated foil sentence with no correlation to the first two (Nadeem et al., 2021). The dataset covers four broad demographics (gender, profession, race, and religion) (Nadeem et al., 2021). Following the original protocol, we compute the **Stereotype Score (SS)** as the percentage of items for which the model prefers the stereotypical option over the anti-stereotypical option among the two relevant candidates, and the **Language Modeling Score (LM-OK)** as the percentage of items for which the model prefers the stereotype or anti-stereotype option over the unrelated foil. Higher SS indicates stronger bias (with 50 being ideal), while higher LM-OK reflects better language modeling quality. We report overall SS and LM-OK, as well as the combined "ideal" score proposed by Nadeem et al. (2021), and analyze how these metrics change with bit-width.

3.4 Quantization & Compression

We consider two types of post-training quantization: an 8-bit PyTorch baseline and a custom N-bit scheme used to sweep over multiple bit-widths.

8-Bit Baseline (PyTorch). For 8-bit quantization, we use PyTorch's built-in dynamic post-training quantization (`torch.ao.quantization.quantize_dynamic`) applied to all `Linear` layers. This converts the weights to `qint8` and runs all integer matrix multiplications at inference time. As this produces an actual compressed model artifact, we report its *measured* on-disk size in MB and define the compression ratio as

$$\text{Compression Ratio (CR)} = \frac{\text{size}_{\text{FP32}}}{\text{size}_{\text{INT8}}}.$$

Custom N-Bit Quantization. To study a broader range of bit-widths, we also applied a custom uniform N-bit quantization scheme for $N \in \{2, 4, 6, 12, 16, 20, 24, 28, 30\}$. We perform weight-only quantization by mapping each weight (and bias) tensor to the integer range $[-2^{N-1}, 2^{N-1}-1]$ using symmetric uniform quantization, and then dequantizing back to floating-point so that inference can be run with standard PyTorch operations. This simulates the effect of using N-bit weights, but the model is still stored as FP32 tensors. For these runs, we therefore report a **theoretical** model size

$$\text{size}_{\text{theoretical}}(N) = \text{size}_{\text{FP32}} \cdot \frac{N}{32},$$

and its corresponding **theoretical compression ratio (CR)** as

$$\text{Theoretical Compression Ratio (CR)} = \frac{32}{N}.$$

Tables for $N \neq 8$ report this theoretical compression ratio. As the 8-bit PyTorch model uses a different implementation and yields an actual compressed checkpoint, its report size and CR are not directly comparable to the theoretical values for our custom implementation. Graphs, plots, and tables will still feature the results of 8-bit PyTorch quantization for consistency, though these results should not be directly compared.

3.5 Analysis of 8-Bit Quantization

To establish a baseline for comparison against more aggressive compression regimes, we evaluate the effect of standard 8-bit post-training quantization on measured model bias. Our methodology aims to determine whether reducing weight precision from full 32-bit floating-point to 8-bit integer representations produces a systematic shift in CrowS-Pairs bias scores.

To assess whether 8-bit quantization alters measured bias, we employ a three-part statistical analysis pipeline:

- **Ordinary Least Squares (OLS) Regression:** Bias scores are modeled as a function of quantization status (FP32 vs. INT8). The slope term tests whether compression meaningfully shifts expected CrowS-Pairs scores.
- **Independent-Samples *t*-Test:** Welch's t-test compares mean bias between FP32 and INT8 models, providing a distribution-free check for differences in central tendency.
- **Bootstrap Confidence Intervals:** For each precision level, we generate nonparametric bootstrap distributions of mean bias. The resulting percentile intervals offer a robust measure of uncertainty that does not rely on normality assumptions.

Purpose of This Analysis. This framework allows us to determine whether 8-bit quantization exerts a statistically detectable influence on measured social bias and establishes a reference point for interpreting the broader N-bit sweep. The results of these tests are reported in Section **??**, where we quantify the magnitude and direction of the effect.

3.6 N-Bit Quantization

Beyond the fixed 8-bit baseline, we conduct a broader investigation by quantizing each model to a range of bit-widths. This allows us to study how bias evolves as a function of representational precision rather than examining only a single compression point.

Uniform Quantization Procedure. Given a weight tensor W and a target bit-width N, we approximate symmetric signed integer quantization over the range

$$\left[-2^{N-1},\ 2^{N-1}-1\right].$$

For each tensor, we compute a scale factor

$$s = \frac{\max(|W|)}{2^{N-1}-1},$$

which maps the largest-magnitude weight to the maximum representable integer. Quantized weights are produced by

$$W_q = s \cdot \text{clip}\left(\text{round}\left(\frac{W}{s}\right),\ -2^{N-1},\ 2^{N-1}-1\right),$$

and then stored again as floats for inference. This preserves the quantization noise profile without requiring integer kernels, enabling evaluation on standard hardware.

Model Transformation. For each model under study, we generate a quantized variant by applying the above procedure to all `Linear` layer weights and biases. Importantly, this creates an apples-to-apples comparison framework: architecture, tokenizer, and evaluation settings remain unchanged, ensuring that differences in bias can be attributed to quantization alone.

Bit-Width Sweep. We quantize each model at ten bit-widths: $N \in \{2, 4, 6, 8, 12, 16, 20, 24, 28, 30\}$. This sweep spans both highly compressed regimes (2–6 bits), moderate precision (8–16 bits), and near–full precision (20–28 bits). For each configuration, we measure:

- CrowS-Pairs Overall Stereotype Percentage
- CrowS Group-Specific Percentages
- StereoSet Stereotype Score (SS)
- StereoSet Language Modeling Score (LM-OK)
- Theoretical model size (based on $N/32$ scaling)
- Computed compression ratio relative to FP32

As the N-bit pipeline does not alter runtime kernels, reported sizes are theoretical, ensuring consistency across bit-widths independent of hardware implementation.

Evaluation Protocol. All N-bit models are evaluated identically to their FP32 and 8-bit counterparts. For each quantized model, we compute CrowS-Pairs pseudo-likelihood differences over all sentence pairs and StereoSet log-probability rankings over all triple completions. No retraining, fine-tuning, or calibration is applied; all effects arise strictly from quantization-induced perturbations to the underlying weight space. This controlled design enables us to directly trace how decreasing numerical resolution influences demographic bias, linguistic fidelity, and the trade-off between compression and fairness.

Purpose of This Analysis. Studying a spectrum of bit-widths helps clarify whether bias changes smoothly, saturates, or exhibits threshold behaviors as representational precision decreases. It also enables us to identify regions where compression disproportionately affects demographic subgroups, providing a finer-grained view than 8-bit experiments alone. This section ultimately supports our central research question: how does progressive quantization reshape measured social bias in language models?

4 Results & Analysis

The full numerical results are available in Appendix A. Additionally, figures in this section, along with others, can be found in Appendix C.

4.1 8-Bit Quantization

Across all three statistical tests (OLS regression, independent-samples t-Test, and bootstrap confidence intervals), the evidence strongly indicates that 8-bit quantization reduces measured CrowS-Pair bias relative to FP32 baselines. The OLS regression estimates that compression lowers bias by approximately 5.18 percentage points and yields a highly significant slope term ($p < 0.001$). The independent-samples t-test corroborates this finding, producing a large test statistic ($t = 6.036$) and a corresponding p-value below 10^{-4}, strongly rejecting equality of group means. The bootstrap confidence intervals further reinforce this pattern: FP32 models exhibit a mean bias distribution centered around $[61.95, 64.63]$, whereas INT8 models fall within the substantially lower range of $[57.05, 59.05]$. The absence of interval overlap confirms that the reduction in bias is both consistent and statistically robust.

Collectively, these results establish that 8-bit quantization reliably reduces CrowS-Pairs social bias across the full set of models included in the analysis. This provides a strong empirical foundation for the broader investigation in Section 4.2, where we examine how bias evolves under varying depths of quantization. These findings are directionally consistent with Goncalves and Strubell (2023), who also observe a limited adverse impact of standard compression on social-bias metrics.

4.2 N-Bit Quantization

We present the effects of sweeping quantization precision from 32 bits (FP32) down to 2 bits. Across all encoder models (BERT-base, DistilBERT, RoBERTa-Base, DistilRoBERTa), we observe consistent relationships between precision, measured bias, and model utility.

CrowS-Pairs Bias. Across all models, CrowS Overall Stereotype Percentage tends to decrease steadily as bit-width drops from 32 bits to roughly 6 bits. It is important to note that significant decrease only begins to appear once the bit-width drops below 12. Quantization introduces small amounts of noise that weaken memorized stereotypical associations, pushing scores toward the neutral target of 50%. The lowest observed CrowS bias consistently occurs at **6-bit precision**. Beyond this point, further compression fails to reduce bias further and eventually introduces instability.

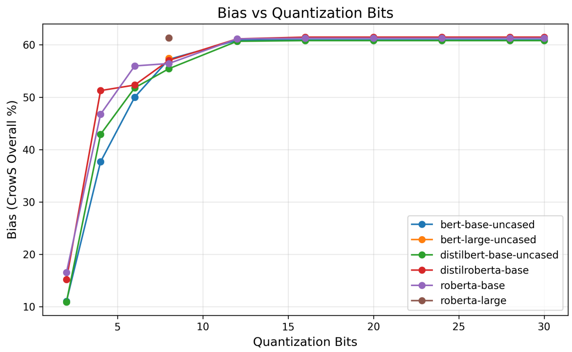

Figure 1: CrowS Pairs Overall Stereotype Percentage vs. Quantization Bit-Widths.

StereoSet Stereotype Percentage. StereoSet shows a similar monotonic decrease in stereotype percentage between 32 and 6 bits. Again, this appears as a threshold pattern, and a significant decrease only occurs after the bit-width goes below 12. However, at **2 bits** we observe a dramatic spike in stereotyping. This instability appears across all models, suggesting that extremely aggressive quantization disrupts the semantic structure the model relies on to distinguish stereotype from anti-stereotype completions.

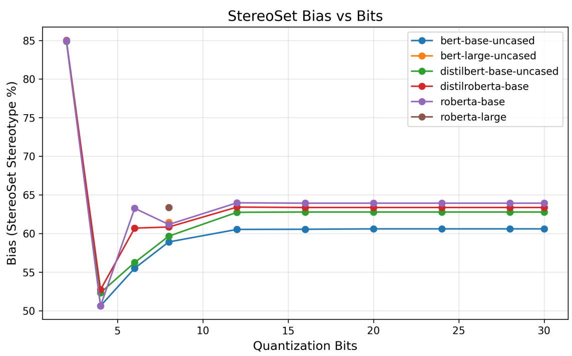

Figure 2: StereoSet Stereotype Score vs. Quantization Bit-Width.

Model Utility (StereoSet LM-OK). Utility remains relatively stable across 32, 24, 20, 16, 12, and 8 bits. The LM-OK score remains high until the bit width falls below 8 bits. At 4 bits, utility becomes unstable, and at 2 bits, it collapses entirely. This indicates that moderate quantization retains the core language modeling ability, while severe reductions compromise both linguistic coherence and meaningful sentence ranking.

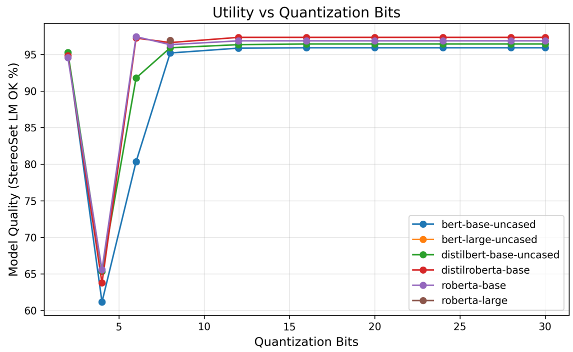

Figure 3: StereoSet LM-OK (Utility Score) vs. Quantization Bit-Width.

Sweet-Spot Analysis. To jointly evaluate utility and fairness, we define a minimum acceptable semantic quality threshold of **LM-OK** ≥ 90. Among models that meet this bar, the **6-8 bit range** consistently produces the lowest CrowS bias scores. Six-bit quantization, in particular, reaches the bias minimum while maintaining utility values comparable to 8-bit and FP32 models. This region appears to offer optimal compression with minimal harm to usability and maximal reduction in bias.

Failure at Extremely Low Bit-Width. Despite occasional spikes in LM-OK, **2-bit** models exhibit extremely high stereotyping and unstable behavior. Their performance fluctuates sharply and routinely exceeds FP32 bias levels. These models are therefore impractical, even if superficially compact or efficient.

Overall Interpretation. Our systematic sweep reveals that quantization acts as a *bias regularizer* when applied moderately: bit-widths in the 6–12 range reduce both CrowS-Pairs and StereoSet bias while preserving language modeling quality. However, when precision becomes too low (2–4 bits), model behavior destabilizes, and stereotyping can increase sharply. These findings outline a practical compression–fairness trade-off: meaningful reductions in bias are achievable, but only within a constrained precision window that does not compromise semantic fidelity.

5 Conclusion

In this work, we studied how post-training quantization affects measured social bias in LLMs. Starting from the observation that bias and compression are effectively unavoidable in the deployment of modern LLMs, we asked: how does reducing weight precision to variable bit-widths change the bias exhibited by models?

Our 8-bit experiment shows that standard dynamic quantization can reliably reduce CrowS-Pairs bias relative to full-precision baselines. In particular, the quantized model showed a statistically significant drop of roughly five percentage points, with non-overlapping bootstrap confidence intervals between the FP32 and INT8 models. Extending beyond our initial single-point comparison, our N-bit sweep reveals that moderate quantization (approximately 6-12 bits) tends to lower both CrowS-Pairs and StereoSet stereotype scores while preserving language modeling quality (i.e., keeping a high LM-OK). Simultaneously, extremely aggressive quantization (2-4 bits) destabilizes behavior, leading to a sharp drop in utility and stereotyping significantly above full-precision levels. Taken together, these results suggest that quantization can act as a bias regularizer within a restricted precision window. However, at a sufficiently low bit width, quantization can become harmful.

We emphasize that the findings are not prescriptive; the study merely highlights what we observed. Bit-width should not be used as a standalone debiasing knob. Before deployment, quantized models should be thoroughly tested to ensure safety. Our conclusions are limited to the models and datasets we tested.

6 Limitations

Our study features several limitations that are important to consider when analyzing our results.

Models and Quantization Scope. We focused on a small set of encoder-only masked language models (BERT, DistilBERT, RoBERTa, and DistilRoBERTa) in English. Our findings may not directly transfer to non-encoder-only models, instruction-tuned models, models of significantly different size (larger or smaller), multilingual models, or models trained with a different end objective. Furthermore, the specific behavior we observe is also shaped by the training data used for these models; different pretraining corpora may interact with compression in different ways. Our study also focused on post-training, weight-only quantization using uniform per-tensor schemes (plus PyTorch's dynamic 8-bit baseline). Other compression methods (knowledge distillation, pruning, etc.) and quantization of activations or other components may yield different patterns of bias.

Bias Metrics. Our notion of "bias" is derived from the metrics in the CrowS-Pairs and StereoSet datasets. These datasets cover a limited set of demographics and stereotypes, are exclusively in English, and focus on sentence-level preferences under controlled templates rather than more complex behaviors (e.g., downstream decision-making or interactive conversation). The CrowS Overall Stereotype Percentage, CrowS Group-Specific Stereotype Percentages, Stereoset Stereotype Percentage, and StereoSet LM-OK Percentage are merely means to quantify phenomena that we observe in the outputs of the LLM. These metrics are proxies that necessarily compress a wide range of phenomena into a few aggregate scores, potentially obscuring important per-group or per-example effects. Scores, high or low, should not be interpreted as evidence that a given model is unbiased or extremely biased.

Compression Metrics. For non-8-bit settings, we report the computed model size and theoretical compression ratio based on the bits-per-weight ratio, assuming all parameters are stored at the target precision. As our custom N-bit pipeline dequantizes weights back to floating-point for evaluation, these values do not reflect the actual on-disk sizes. The PyTorch 8-bit baseline also uses a particular implementation of dynamic quantization; alternative libraries may yield different trade-offs between accuracy, bias, and efficiency. Our results should therefore be interpreted as characterizing how bias metrics respond to changes in numerical precision, not as definitive engineering guidance for deployment.

Experiment and Results. Our experiments evaluate a discrete set of bit-widths with a finite set of models. We do not exhaustively explore all possible model sizes or all possible bit-widths. The statistical analyses for 8-bit quantization (OLS regression, Welch's t-test, and bootstrap confidence intervals) are performed on a relatively small sample of models and benchmarks, and we do not correct for multiple comparisons across all metrics and bit-widths. Consequently, some marginal effects may be sensitive to sampling noise or design choices.

Societal Response. Our work should be considered a diagnostic tool rather than a prescriptive one. We do not propose quantization as a debiasing method; we are merely showing what our study has observed with the models, metrics, and values that we observed with quantization.

7 Ethical Considerations

This work studies how post-training quantization affects demographic bias in LLMs. Our core motivation is to better understand how compression, quantization in particular, affects the fairness, specifically the bias, in LLMs. As we did not deploy and further test the models, these findings should not be taken as a certification that the quantized models studied are fair or safe.

Datasets. We rely heavily on CrowS-Pairs (Nangia et al., 2020), and StereoSet (Nadeem et al., 2021), both of which contain harmful stereotypes about a wide range of demographics. These datasets were used to measure the level of bias in LLMs. The use of these in testing does not endorse or promote the use of these stereotypes.

Fairness and Compression. Our experiment shows how quantization can change measured bias under a fixed evaluation protocol. This should not be interpreted as evidence that compression is inherently fair or unfair, nor that adjusting bit-width is an adequate debiasing strategy. Additional rigorous fairness and quality evaluations should accompany any attempt to use quantization. Our study was merely to inform such an evaluation of the effects of bit-width on bias.

Contributions

Owen was responsible for everything related to reading existing literature, writing code to perform and score results from the quantization, and running models with 16 or fewer bits. Jonah was responsible for helping with the implementation design of variable-bit quantization, statistical analysis, running all models with more than 16 bits, and developing scripts to generate plots.

Acknowledgments

We want to thank our CS-1684 professor at the University of Pittsburgh (Dr. Xiang Lorraine Li), our CS-1684 course grader (Rn. Nefrianna), and our peers in the University of Pittsburgh Fall 2025 section of CS-1684 for their feedback on this project.

References

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery.

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. In *Advances in Neural Information Processing Systems 29*.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. LLM.int8(): 8-bit matrix multiplication for transformers at scale. *arXiv preprint arXiv:2208.07339*.

Sunipa Dev, Tao Li, Jeff M. Phillips, and Vivek Srikumar. 2020. On measuring and mitigating biased inferences of word embeddings. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

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

Elias Frantar and Dan Alistarh. 2023. SparseGPT: Massive language models can be accurately pruned in one-shot. *arXiv preprint arXiv:2301.00774*.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. 2022. GPTQ: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323*.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024a. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Tong Yu, Hanieh Deilamsalehy, Ruiyi Zhang, Sungchul Kim, and Franck Dernoncourt. 2024b. Self-debiasing large language models: Zero-shot recognition and reduction of stereotypes. *arXiv preprint arXiv:2402.01981*.

Prakhar Ganesh, Yao Chen, Xin Lou, Mohammad Ali Khan, Yin Yang, Hassan Sajjad, Preslav Nakov, Deming Chen, and Marianne Winslett. 2020. Compressing large-scale transformer-based models: A case study on BERT. *arXiv preprint arXiv:2002.11985*.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*.

Gustavo Goncalves and Emma Strubell. 2023. Understanding the effect of model compression on social bias in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2468–2488, Singapore. Association for Computational Linguistics.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *arXiv preprint arXiv:1903.03862*.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. 2023. MiniLLM: Knowledge distillation of large language models. *arXiv preprint arXiv:2306.08543*.

Sara Hooker, Nyalleng Moorosi, Gregory Clark, Samy Bengio, and Emily Denton. 2020. Characterising bias in compressed models. *arXiv preprint arXiv:2010.03058*.

Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Social biases in NLP models as barriers for persons with disabilities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5491–5501.

Renren Jin, Jiangcun Du, Wuwei Huang, Wei Liu, Jian Luan, Bin Wang, and Deyi Xiong. 2024. A comprehensive evaluation of quantization strategies for large language models. *arXiv preprint arXiv:2402.16775*.

Zhen Li, Yupeng Su, Runming Yang, Congkai Xie, Zheng Wang, Zhongwei Xie, Ngai Wong, and Hongxia Yang. 2025. Quantization meets reasoning: Exploring LLM low-bit quantization degradation for mathematical reasoning. *arXiv preprint arXiv:2501.03035*.

Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2021. Towards understanding and mitigating social biases in language models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. 2023. AWQ: Activation-aware weight quantization for LLM compression and acceleration. *arXiv preprint arXiv:2306.00978*.

Peiyu Liu, Zikang Liu, Ze-Feng Gao, Dawei Gao, Wayne Xin Zhao, Yaliang Li, Bolin Ding, and Ji-Rong Wen. 2023. Do emergent abilities exist in quantized large language models: An empirical study. *arXiv preprint arXiv:2307.08072*.

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. *arXiv preprint*, arXiv:1907.11692.

Federico Marcuzzi, Xuefei Ning, Roy Schwartz, and Iryna Gurevych. 2025. How quantization shapes bias in large language models. *arXiv preprint arXiv:2508.18088*.

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

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

Roberto Navigli, Simone Conia, and Björn Ross. 2023. Biases in large language models: Origins, inventory, and discussion. *ACM Journal of Data and Information Quality*, 15(2):1–21.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. *arXiv preprint arXiv:1808.07231*.

Vinay Uday Prabhu and Abeba Birhane. 2020. Large image datasets: A pyrrhic win for computer vision? *arXiv preprint arXiv:2006.16923*.

Irina Proskurina, Guillaume Metzler, and Julien Velcin. 2025. Fair-gptq: Bias-aware quantization for large language models. *arXiv preprint arXiv:2509.15206*.

Nithya Sambasivan, Shivani Kapania, Hannah Highfill, and 1 others. 2021. "everyone wants to do the model work, not the data work": Data cascades in high-stakes AI. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv preprint*, arXiv:1910.01108.

Timo Schick, Sahana Udupa, and Hinrich Schütze. 2021. Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in NLP. *arXiv preprint arXiv:2103.00453*.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2020. Towards controllable biases in language generation. *arXiv preprint arXiv:2005.00268*.

Qianli Wang, Mingyang Wang, Nils Feldhus, Simon Ostermann, Yuan Cao, Hinrich Schütze, Sebastian Möller, and Vera Schmitt. 2025. Through a compressed lens: Investigating the impact of quantization on LLM explainability and interpretability. *arXiv preprint arXiv:2505.13963*.

Johannes Welbl, Amelia Glaese, Jonathan Uesato, and 1 others. 2021. Challenges in detoxifying language models. *arXiv preprint arXiv:2109.07445*.

Canwen Xu and Julian McAuley. 2022. A survey on model compression and acceleration for pretrained language models. *arXiv preprint arXiv:2202.07105*.

Peng Xu, Wenqi Shao, Mengzhao Chen, Shitao Tang, Kaipeng Zhang, Peng Gao, Fengwei An, Yu Qiao, and Ping Luo. 2024a. BESA: Pruning large language models with blockwise parameter-efficient sparsity allocation. *arXiv preprint arXiv:2402.16880*.

Xinyang Xu and 1 others. 2024b. A survey on knowledge distillation of large language models. *arXiv preprint arXiv:2402.13116*.

Yuzhuang Xu, Xu Han, Zonghan Yang, Shuo Wang, Qingfu Zhu, Zhiyuan Liu, Weidong Liu, and Wanxiang Che. 2024c. Onebit: Towards extremely low-bit large language models. *arXiv preprint arXiv:2402.11295*.

Navya Yarrabelly, Vinay Damodaran, and Feng-Guang Su. 2024. Mitigating gender bias in contextual word embeddings. *arXiv preprint arXiv:2411.12074*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordóñez, and Kai-Wei Chang. 2018a. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018b. Learning gender-neutral word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4847–4853.

Xunyu Zhu, Jian Li, Yong Liu, Can Ma, and Weiping Wang. 2023. A survey on model compression for large language models. *arXiv preprint arXiv:2308.07633*.

A Full Bias Tables by Bit-Width

In the following tables, each number represents a different stereotype score from the Crows-Pairs dataset. They are defined as follows: 0.) Race-Color, 1.) Socioeconomic, 2.) Gender, 3.) Disability, 4.) Nationality, 5.) Sexual Orientation, 6.) Physical Appearance, 7.) Religion, and 8.) Age. Similar to the overall score, 0 indicates full favoring of the anti-stereotype, 50 indicates neutrality, and 100 indicates full favoring of the stereotype.

A.1 No Quantization

Model	Size (MB)	CR	CrowS Overall	0	1	2	3	4	5	6	7	8	SS Stereo	SS LM-OK
BERT-Base	417.827	N/A	60.942	59.690	58.721	57.252	78.333	46.541	77.381	69.841	73.333	60.920	60.594	95.916
DistilBERT-Base	255.563	N/A	60.809	57.364	61.628	58.015	80.0	51.572	76.190	65.079	72.381	59.770	62.777	96.439
RoBERTa-Base	475.747	N/A	61.273	58.140	66.279	57.634	71.667	57.233	66.667	63.492	69.524	64.368	63.922	96.866
DistilRoBERTa-Base	313.488	N/A	61.472	60.271	65.116	57.252	71.667	55.975	69.048	69.841	68.571	55.172	63.366	97.341
BERT-Large	1278.713	N/A	62.798	59.690	61.628	59.542	81.667	51.572	72.619	76.190	76.190	65.517	62.598	96.486
RoBERTa-Large	1355.914	N/A	67.507	68.605	72.674	62.214	70.0	55.975	67.857	74.603	75.238	71.264	64.404	97.246

Table 1: Bias evaluation results with no quantization. CR = Compression Ratio.

A.2 30-Bit Quantization

Model	Size (MB)	CR	CrowS Overall	0	1	2	3	4	5	6	7	8	SS Stereo	SS LM-OK
BERT-Base	391.713	1.067	60.942	59.690	58.721	57.252	78.333	46.541	77.381	69.841	73.333	60.920	60.594	95.916
DistilBERT-Base	239.590	1.067	60.809	57.364	61.628	58.015	80.0	51.572	76.190	65.079	72.381	59.770	62.777	96.439
RoBERTa-Base	446.0127	1.067	61.273	58.140	66.279	57.634	71.667	57.233	66.667	63.492	69.524	64.368	63.922	96.866
DistilRoBERTa-Base	293.895	1.067	61.472	60.271	65.116	57.252	71.667	55.975	69.048	69.841	68.571	55.172	63.366	97.341

Table 2: Bias evaluation results with 30-bit quantization. CR = Theoretical Compression Ratio.

A.3 28-Bit Quantization

Model	Size (MB)	CR	CrowS Overall	0	1	2	3	4	5	6	7	8	SS Stereo	SS LM-OK
BERT-Base	365.599	1.143	60.942	59.690	58.721	57.252	78.333	46.541	77.381	69.841	73.333	60.920	60.594	95.916
DistilBERT-Base	223.618	1.143	60.809	57.364	61.628	58.015	80.0	51.572	76.190	65.079	72.381	59.770	62.777	96.439
RoBERTa-Base	416.279	1.143	61.273	58.140	66.279	57.634	71.667	57.233	66.667	63.492	69.524	64.368	63.922	96.866
DistilRoBERTa-Base	274.302	1.143	61.472	60.271	65.116	57.252	71.667	55.975	69.048	69.841	68.571	55.172	63.366	97.341

Table 3: Bias evaluation results with 28-bit quantization. CR = Theoretical Compression Ratio.

A.4 24-Bit Quantization

Model	Size (MB)	CR	CrowS Overall	0	1	2	3	4	5	6	7	8	SS Stereo	SS LM-OK
BERT-Base	313.371	1.333	60.942	59.690	58.721	57.252	78.333	46.541	77.381	69.841	73.333	60.920	60.594	95.916
DistilBERT-Base	191.672	1.333	60.809	57.364	61.628	58.015	80.0	51.572	76.190	65.079	72.381	59.770	62.777	96.439
RoBERTa-Base	235.116	1.333	61.472	60.271	65.116	57.252	71.667	55.975	69.048	69.841	68.571	55.172	63.366	97.341
DistilRoBERTa-Base	356.810	1.333	61.273	58.140	66.279	57.634	71.667	57.233	66.667	63.492	69.524	64.368	63.922	96.866

Table 4: Bias evaluation results with 24-bit quantization. CR = Theoretical Compression Ratio.

A.5 20-Bit Quantization

Model	Size (MB)	CR	CrowS Overall	0	1	2	3	4	5	6	7	8	SS Stereo	SS LM-OK
Bert-Base	261.142	1.6	60.942	59.690	58.721	57.252	78.333	46.541	77.381	69.841	73.333	60.920	60.594	95.916
DistilBERT-Base	127.781	2.0	60.809	57.364	61.628	58.015	80.0	51.572	76.190	65.079	72.381	59.770	62.777	96.439
RoBERTa-Base	297.342	1.6	61.273	58.140	66.279	57.634	71.667	57.233	66.667	63.492	69.524	64.368	63.922	96.866
DistilRoBERTa-Base	195.930	1.6	61.472	60.271	65.116	57.252	71.667	55.975	69.048	69.841	68.571	55.172	63.366	97.341

Table 5: Bias evaluation results with 20-bit quantization. CR = Theoretical Compression Ratio.

A.6 16-Bit Quantization

Model	Size (MB)	CR	CrowS Overall	0	1	2	3	4	5	6	7	8	SS Stereo	SS LM-OK
BERT-Base	208.91	2.0	60.942	59.690	58.721	57.252	78.333	46.541	77.381	69.841	73.333	60.920	60.545	95.916
DistilBERT-Base	127.781	2.0	60.809	57.364	61.628	58.015	80.0	51.572	76.190	65.079	72.381	59.770	62.777	96.439
RoBERTa-Base	237.873	2.0	61.273	58.140	66.279	57.634	71.667	57.233	66.667	63.492	69.524	64.368	63.922	96.866
DistilRoBERTa-Base	156.744	2.0	61.472	60.271	65.116	57.252	71.667	55.975	69.0476	69.841	68.571	55.172	63.366	97.341

Table 6: Bias evaluation results with 16-bit quantization. CR = Theoretical Compression Ratio.

A.7 12-Bit Quantization

Model	Size (MB)	CR	CrowS Overall	0	1	2	3	4	5	6	7	8	SS Stereo	SS LM-OK
BERT-Base	156.685	2.667	60.875	59.496	58.721	57.634	78.333	46.541	77.381	69.841	72.381	60.920	60.525	95.869
DistilBERT-Base	95.836	2.667	60.676	57.364	61.628	58.015	80.0	50.943	76.190	65.079	72.381	58.621	62.740	96.344
RoBERTa-Base	178.405	2.667	61.141	57.946	66.279	56.870	73.333	57.233	66.667	63.492	70.476	63.218	63.971	96.866
DistilRoBERTa-Base	117.558	2.667	61.141	59.884	65.116	56.870	71.667	55.975	67.857	68.254	68.571	55.172	63.41	97.341

Table 7: Bias evaluation results with 12-bit quantization. CR = Theoretical Compression Ratio.

A.8 8-Bit Quantization

Model	Size (MB)	CR	CrowS Overall	0	1	2	3	4	5	6	7	8	SS Stereo	SS LM-OK
BERT-Base	195.545	2.137	57.228	55.814	56.977	53.435	73.333	53.459	77.381	57.143	61.905	48.276	58.903	95.204
DistilBERT-Base	154.757	1.651	55.438	51.357	62.791	52.672	65.0	45.283	73.810	55.556	62.857	58.621	59.653	95.916
RoBERTa-Base	267.925	1.776	56.432	55.233	60.465	54.198	60.0	50.314	61.905	66.667	53.333	62.0689	61.163	96.344
DistilRoBERTa-Base	227.142	1.380	57.029	54.264	60.465	55.725	66.667	53.459	63.095	73.016	59.048	50.575	60.835	96.629
BERT-Large	441.619	2.896	57.427	55.814	59.302	50.0	78.333	48.428	75.0	68.254	64.762	54.023	61.436	95.916
RoBERTa-Large	538.101	2.520	61.340	60.078	65.698	62.218	60.0	52.201	67.857	68.254	64.762	59.770	63.351	96.914

Table 8: Bias evaluation results with 8-bit quantization. CR = Compression Ratio.

A.9 6-Bit Quantization

Model	Size (MB)	CR	CrowS Overall	0	1	2	3	4	5	6	7	8	SS Stereo	SS LM-OK
BERT-Base	78.34	5.333	50.0	51.550	41.860	59.160	38.333	34.591	72.619	42.857	44.762	55.172	55.496	80.342
DistilBERT-Base	47.918	5.333	51.790	46.124	47.674	59.160	46.667	45.283	71.429	57.143	58.095	56.322	56.234	91.785
RoBERTa-Base	89.203	5.333	55.968	52.326	63.953	55.344	61.667	49.686	65.476	68.254	56.190	52.874	63.255	97.436
DistilRoBERTa-Base	58.779	5.333	52.321	49.031	56.977	58.779	68.333	40.252	57.143	60.317	44.762	52.874	60.693	97.246

Table 9: Bias evaluation results with 6-bit quantization. CR = Theoretical Compression Ratio.

A.10 4-Bit Quantization

Model	Size (MB)	CR	CrowS Overall	0	1	2	3	4	5	6	7	8	SS Stereo	SS LM-OK
BERT-Base	52.228	8.0	37.666	36.240	26.744	48.092	33.333	36.478	23.810	36.508	31.429	63.218	50.621	61.159
DistilBERT-Base	31.945	8.0	42.905	47.674	28.488	49.237	35.0	32.704	23.810	42.857	60.952	44.828	52.326	65.337
RoBERTa-Base	59.468	8.0	46.751	42.829	46.512	48.473	36.667	42.138	45.238	55.556	60.0	59.770	50.615	65.575
DistilRoBERTa-Base	39.186	8.0	51.260	51.357	58.140	45.420	40.0	46.541	60.714	47.619	69.524	42.529	52.718	63.770

Table 10: Bias evaluation results with 4-bit quantization. CR = Theoretical Compression Ratio.

A.11 2-Bit Quantization

Model	Size (MB)	CR	CrowS Overall	0	1	2	3	4	5	6	7	8	SS Stereo	SS LM-OK
BERT-Base	26.114	16.0	11.008	8.527	16.860	8.779	21.667	13.208	11.905	9.524	5.714	16.092	84.895	95.252
DistilBERT-Base	15.973	16.0	10.875	8.527	15.116	8.779	20.0	13.208	11.905	9.524	7.619	16.092	84.895	95.252
RoBERTa-Base	29.734	16.0	16.512	11.434	16.279	18.321	21.667	18.239	26.190	22.222	18.095	19.540	84.940	94.587
DistilRoBERTa-Base	19.593	16.0	15.186	10.271	14.535	17.939	21.667	15.723	25.0	22.222	16.190	16.092	85.028	94.824

Table 11: Bias evaluation results with 2-bit quantization. CR = Theoretical Compression Ratio.

B Best and Worst Models

B.1 Best Models

Model	Bits	LM OK	Bias Score	$\Delta 50$
DistilBERT-Base-Uncased	6	91.785	51.790	1.790
DistilRoBERTa-Base	6	97.246	52.321	2.321
DistilBERT-Base-Uncased	8	95.916	55.438	5.438
RoBERTa-Base	6	97.436	55.968	5.968
RoBERTa-Base	8	96.344	56.432	6.432
DistilRoBERTa-Base	8	96.629	57.029	7.029
BERT-Base-Uncased	8	95.204	57.228	7.228
BERT-Large-Uncased	8	95.916	57.427	7.427
DistilBERT-Base-Uncased	12	96.344	60.676	10.676
DistilBERT-Base-Uncased	16	96.439	60.809	10.809

Table 12: Best Models where Bits is the Quantization Bit-Width, Bias Score is the CrowS Pairs Overall Score, and $\Delta 50$ is the $|50 - \text{Bias Score}|$

B.2 Worst Models

Model	Bits	LM OK	Bias Score	$\Delta 50$
DistilBERT-Base-Uncased	2	95.252	10.875	39.125
BERT-Base-Uncased	2	95.252	11.008	38.992
DistilRoBERTa-Base	2	94.824	15.186	34.814
RoBERTa-Base	2	94.587	16.512	33.488
DistilRoBERTa-Base	28	97.341	61.472	11.472
DistilRoBERTa-Base	30	97.341	61.472	11.472
DistilRoBERTa-Base	24	97.341	61.472	11.472
DistilRoBERTa-Base	20	97.341	61.472	11.472
DistilRoBERTa-Base	16	97.341	61.472	11.472
RoBERTa-large	8	96.914	61.340	11.340

Table 13: Worst Models where Bits is the Quantization Bit-Width, Bias Score is the CrowS Pairs Overall Score, and $\Delta 50$ is the $|50 - \text{Bias Score}|$

C Figures

C.1 Crows Pairs Overall vs Bit-Width

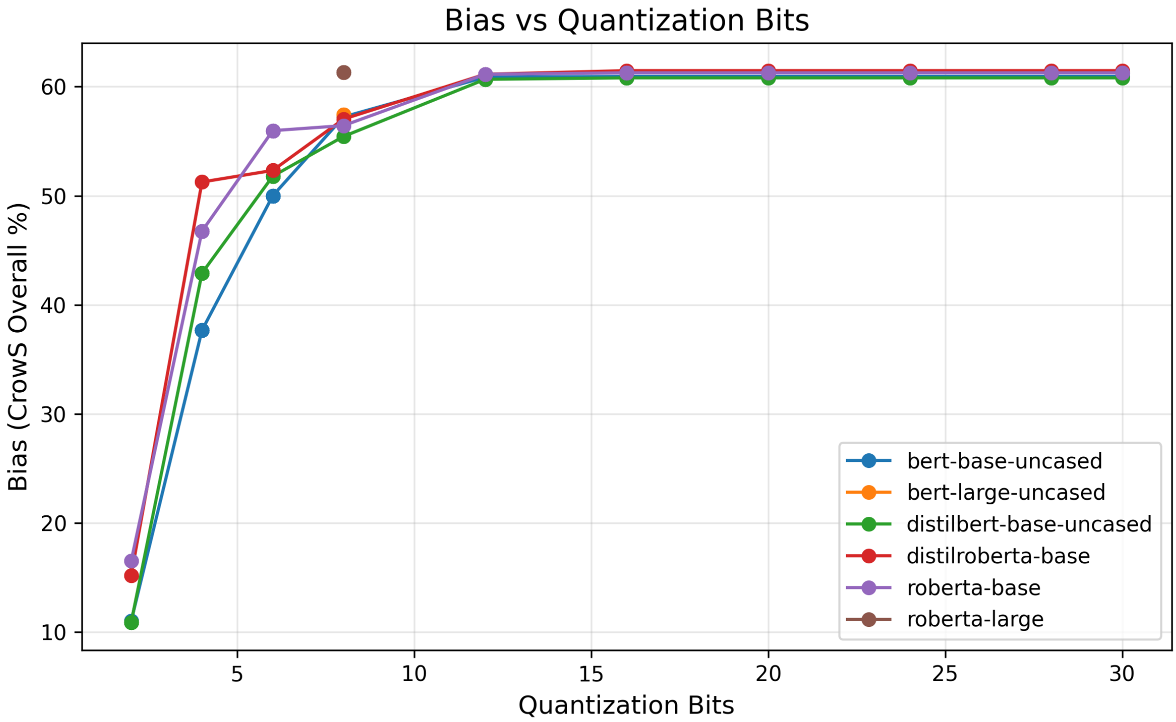

Figure 4: CrowS-Pairs Overall Stereotype Percentage vs. Quantization Bit-Widths.

C.2 StereoSet Stereotype Score vs Bit-Width

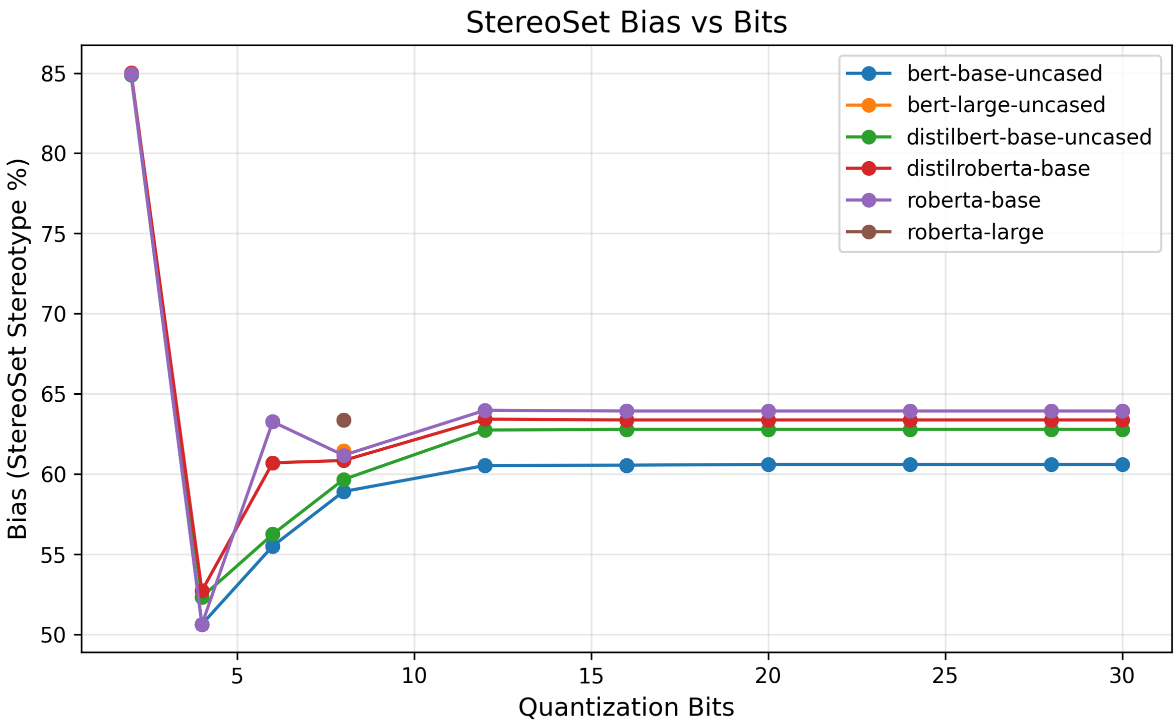

Figure 5: StereoSet Stereotype Score vs. Quantization Bit-Width.

C.3 StereoSet LM-OK (Utility Score) vs Bit-Width

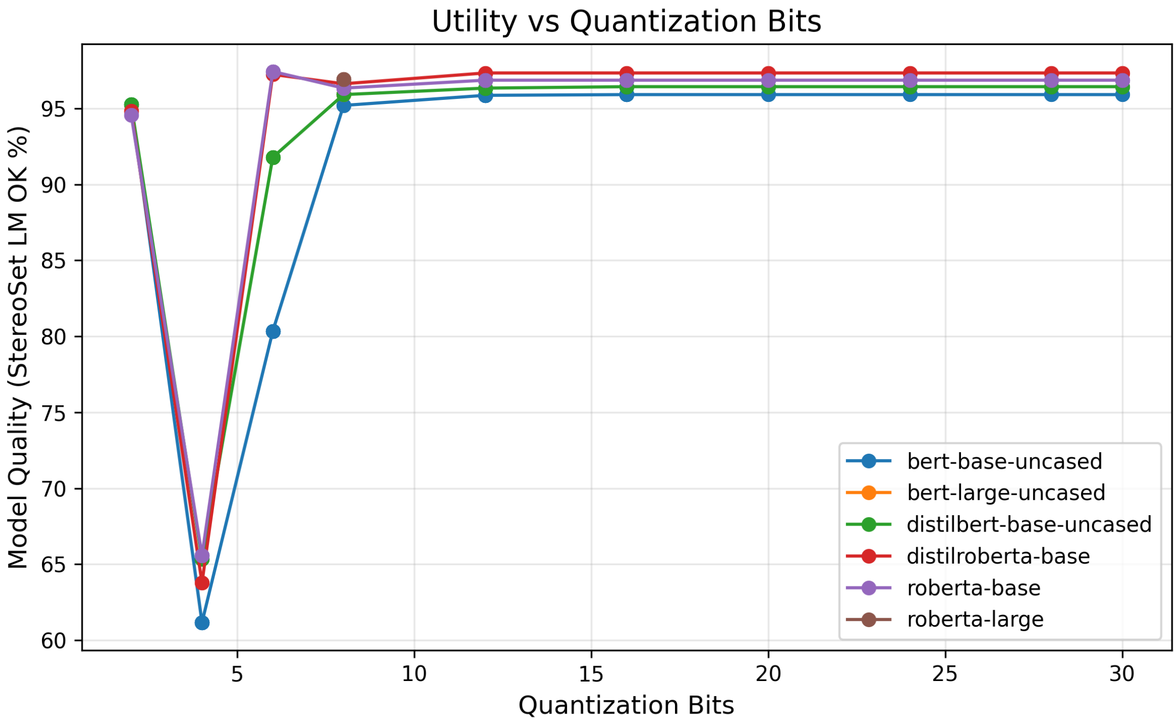

Figure 6: StereoSet LM-OK (Utility Score) vs. Quantization Bit-Width.

C.4 CrowS Pairs By Stereotype vs Bit-Width

Below is the per-stereotype breakdown of the scores with the N-bit quantization. As with the tables in A, we classify the scores by demographic as follows: 0.) Race-Color, 1.) Socioeconomic, 2.) Gender, 3.) Disability, 4.) Nationality, 5.) Sexual Orientation, 6.) Physical Appearance, 7.) Religion, and 8.) Age. Similar to the overall score, 0 indicates full favoring of the anti-stereotype, 50 indicates neutrality, and 100 indicates full favoring of the stereotype. The models are colored as follows:

- Blue: **BERT-Base-Uncased**
- Orange: **BERT-Large-Uncased**
- Green: **DistilBERT-Base-Uncased**
- Red: **DistilRoBERTa-Base**
- Purple: **RoBERTa-Base**
- Brown: **RoBERTa-Large**

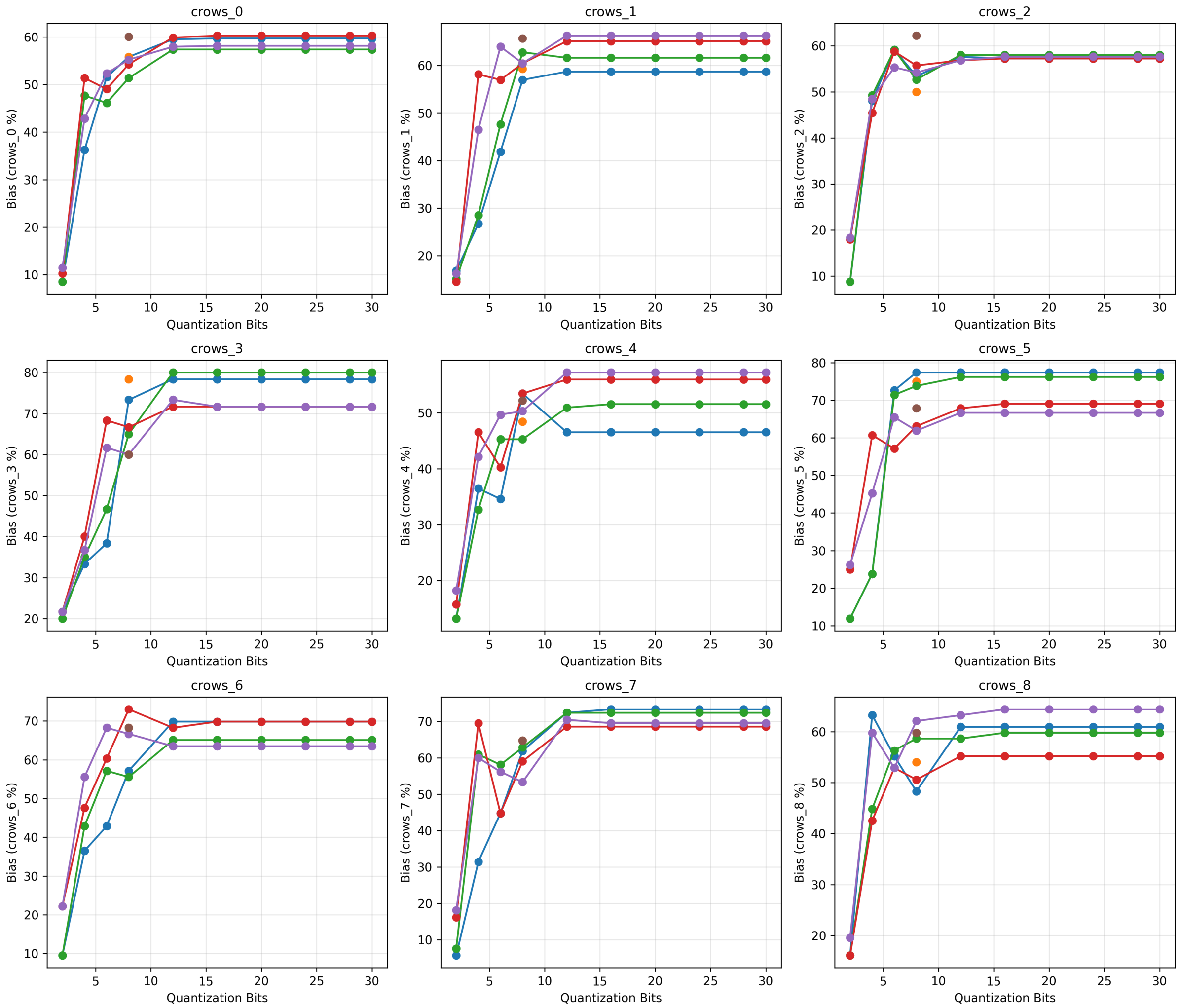

Figure 7: Stereotype Score vs. Quantization Bit-Width.