Does Compression Exacerbate Large Language Models' Social Bias?

Muhammad Athar Ganaie [*1] **Mohammed Adnan** [*1 2] **Arfa Raja** [1] **Shaina Raza** [2] **Yani Ioannou** [1]

Abstract

Ensuring fairness in large language models (LLMs) is critical, yet the effects of popular compression techniques on social biases remain underexplored. In this work, we systematically investigate how pruning, quantization, and knowledge distillation influence demographic bias in multiple open-weight LLMs. Using the HOLISTICBIAS dataset, which contains roughly 600 identity descriptors across 13 demographic axes, we employ a likelihood bias metric based on differential perplexity between paired prompts that differ only in demographic terms. Our study covers three representative models: Llama, DeepSeek, and Mistral. The results reveal striking model-dependent behaviors, in some cases suggesting that naive compression can exacerbate stereotypes towards subpopulation groups, and others showing little effect. The findings underscore the necessity of bias-aware compression techniques and rigorous post-compression bias evaluation to ensure the development of fair and responsible AI systems.

1. Introduction

In recent years, Artificial Intelligence (AI) has achieved state-of-the-art results across a range of learning and generative tasks. This remarkable success can be attributed to foundation models, a new class of large AI systems trained on massive datasets. These foundation models, such as Large Language Modelss (LLMs), exhibit exceptional capabilities in natural language understanding, generation, language translation, and even logical reasoning. For example, ChatGPT, a model with approximately 175 billion parameters trained on diverse text data from the internet, has garnered widespread attention for its sophisticated ability to answer complex questions, generate comprehensive summaries, write creative content, and engage in nuanced human-like conversations. This breakthrough has triggered a wave of new LLMs, characterized primarily by increasingly larger model architectures. The exponential growth in model parameters is guided by scaling laws (Kaplan et al., 2020), which describe how performance improves with increasing model size, dataset scale, and computational resources. As model sizes increase, training and inference costs rise significantly, often by orders of magnitude. This exponential growth poses considerable challenges for academia and the public, making it difficult to train and deploy large foundation models. These challenges create barriers to equitable access and hinder innovation. Additionally, the requirement for massive computing infrastructure imposes substantial barriers, further restricting access to and use of these powerful foundation models.

Model compression seeks to address this challenge by reducing model size and improving efficiency while maintaining performance. This results in reduced computational requirements for both training and inference, potentially democratizing access to LLMs. Depending on the compression technique, model compression can be broadly divided into three categories: pruning, quantization, and knowledge distillation. Model pruning involves the systematic removal of less important connections or neurons from the neural network by identifying and eliminating weights with minimal impact on the model's performance. Model quantization reduces the precision of the model's weights and activations. Instead of using 32-bit floating-point numbers, lower-precision data types like 8-bit integers are used to represent the values, which significantly reduces the memory footprint of the model. Knowledge distillation refers to training a smaller student model, which distills information from a larger trained teacher model, resulting in a student model with performance comparable to the teacher model, though with fewer parameters. Since model compression has gained traction for deployment in resource-constrained settings, it is crucial to ensure that compression does not exacerbate the algorithmic bias of LLMs to guarantee their safe and reliable use, especially in safety-critical domains such as healthcare and finance. Understanding the impact of compression methods on algorithmic bias becomes even more important when the base (uncompressed) LLM already encodes problematic social biases learned from training data (Li et al.,

[*]Equal contribution [1]University of Calgary [2]Vector Institute. Correspondence to: Muhammad Athar Ganaie <muhammadathar.ganaie@ucalgary.ca>, Mohammed Adnan <adnan.ahmad@ucalgary.ca>, Yani Ioannou <yani.ioannou@ucalgary.ca>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

2025). These biases can manifest as differing likelihoods or preferences for certain demographic groups over others, potentially leading to unfair or stereotyping behavior.

This raises an important question — does compressing an LLM affect its embedded social biases? This question is critical for technical AI governance, as compression is often applied before model deployment in resource-constrained or edge settings, where biased outcomes can directly impact users. In this work, we systematically study the impact of state-of-the-art model compression techniques on algorithmic bias. We use open-weight models (e.g., LLaMA, DeepSeek, Mistral) and compare the bias of uncompressed and compressed models using the `HOLISTICBIAS` dataset (Smith et al., 2022) to better understand the effects of model compression. Our contributions are as follows:

1. Most of the recent work on LLM compression has focused solely on the overall perplexity score, which is not the best metric to measure algorithmic bias. In this work, we aim to understand the effect of LLM compression on the algorithmic level as well as for safe and reliable deployment of compressed LLMs for real-world applications.
2. We used the `HOLISTICBIAS` dataset to systematically analyze the effect of LLM compression and use likelihood bias to show that the model compression affects the algorithmic bias of the baseline models, raising concerns about the deployment of compressed LLMs for fairness-critical applications.

2. Background

Bias in LLMs. LLMs are increasingly deployed in high-stakes domains such as healthcare, finance, and education, delivering unprecedented automation and productivity (Baldassarre et al., 2023; Gokul, 2023; Gurusamy et al., 2024; Sabherwal & Grover, 2024; Khan et al., 2024). And yet LLMs frequently mirror and amplify societal biases, manifesting in representational harms that disproportionately affect marginalized groups (Raj et al., 2024; Shin et al., 2024; Hu et al., 2024; Wei et al., 2025; Zhou et al., 2024; Manvi et al., 2024; Hacker et al., 2024).

Compressed LLMs. The growing size of LLMs poses challenges in terms of storage, memory, and energy consumption. Model compression techniques are essential for making these models more deployable (Zhu et al., 2024; Kim et al., 2025; Jiang et al., 2024; Łajewska et al., 2025). However, compression may alter the internal representations and token distributions of LLMs, potentially affecting how bias is expressed.

Model Pruning. Post-training model pruning has been shown to be effective in compressing the model size, and seminal works have demonstrated that large models can be pruned after training with minimal loss in accuracy (Gale et al., 2019; Han et al., 2015; 2016). While model pruning makes inference more efficient, it does not reduce the computational cost of training the model. Pruning refers to removing the 'less important' model components based on some heuristic or importance metrics (Blalock et al., 2020).

Quantization. While model pruning removes the weight parameters, quantization compresses the model by representing the weight parameters with lower precision numerical values (e.g. int8, float16) instead of using high-precision format (float32). Quantization, orthogonal to model pruning, can be combined with pruning to further reduce model size and is often used for deploying models in resource-constrained environments, such as edge devices. While quantizations have been extensively studied for Convolutional Neural Network (CNN), recently, quantization has gained traction to compress LLM to accelerate inference time; due to their large size and complex model architecture, it is not trivial to quantize LLM In the context of LLMs, both quantizing model weights and optimizer states have been demonstrated to be effective.

GPTQ. OBQ quantized weight in greedy order, i.e. it always picks the weight that will incur the least quantization error. However such a greedy approach is not scalable to billion-sized models. GPTQ proposes to quantize the weights in each row with the same order, which enable to re-use the Hessian for different rows and save compute. This reduces run time from $O(d_{row}.d_{col}^3)$ to $O\left(\max\left(d_{row}\cdot d_{col}^2, d_{col}\right)\right)$. Since updating the Hessian H has low compute to memory ratio, updating the hessian is bottlenecked by lower memory bandwidth. To mitigate this, GPTQ performs quantization in blocks, which reduces the size of the hessian to be updated and thus further reducing the runtime of the algorithm.

LLM.int8(). In LLMs, the feed-forward and attention projections layers are responsible for 95% of the total parameters and 65–85% computations. One way to reduce the computational cost is to quantize the parameters to fewer bits and use low-bit precision matrix multiplication. Chen et al. (2020) proposed to use 8-bit quantization for language models, however, the model required fine-tuning after quantization. Dettmers et al. (2022a) proposed *LLM.int8()* and demonstrated the LLMs can be easily quantized to 8-bit without much loss in performance once the outlier features are taken into account. LLM.int8() uses a mixed precision format, which isolates the outlier feature dimensions into 16-bit multiplication while other parameters are multiplied in 8-bit.

Bias in Compressed LLMs. Hooker et al. (2020) also studied the effect model compression, in particular pruning, on the algorithmic bias. The authors observed that pruning disproportionately impacts performance on underrepresented features. Gonçalves & Strubell (2023) also studied

the effect of model compression on the LLM bias. They observed that model compression methods have a regularization effect on the model bias. However, they only used BERT-based models and focused on quantization. Our work uses more recent larger LLMs and different compression methods, in addition to quantization. Mohammadshahi & Ioannou (2025) looked at the effect of distillation temperature on class-wise bias and fairness of student vs. teacher BERT-based models, finding that commonly used temperatures can result in less fair student models. Xu et al. (2024) conduct a broad safety evaluation, examining degeneration, representational harms, dialect bias, and performance across heterogeneous benchmarks (e.g., `BBQ`, `UNQOVER`, `HOLISTICBIASR`), but this reliance on multiple datasets prevents systematic, cross-model comparisons. Kirsten et al. (2025) focus on inference-time acceleration using six metrics (including `CrowS-Pairs` and `DiscrimEvalGen`), revealing that naive acceleration can unpredictably amplify stereotypes, yet their use of U.S.-centric benchmarks and varied protocols limits demographic scope and metric consistency. Similarly, Ramesh et al. (2023) compare pruning, quantization, and distillation via embedding-distance and classification fairness metrics but do not address generative or likelihood-based biases on an intersectional benchmark. Finally, Xu & Hu (2022) show that distillation and pruning of `GPT-2` may reduce toxicity and sentiment biases, but their study is confined to a single architecture and coarse toxicity measures rather than demographic fairness.

3. Methodology

Dataset. We used the `HOLISTICBIAS` dataset (Smith et al., 2022) to analyze changes in algorithmic bias after model compression. The dataset contains roughly 600 descriptor terms across 13 different demographic axes, such as `age`, `gender`, `sexual orientation`, `religion`, and `nationality`. Prompts are constructed by inserting demographic terms and person nouns into pre-selected sentence templates. Different descriptor terms are paired with various noun terms to probe model bias. An unbiased model should not change its likelihood score when the descriptors are replaced. For example, prompts are constructed using different descriptor and noun terms in the template "What do you think about `[PLURAL NOUN]` who are `[DESCRIPTOR]`?" The dataset includes all possible combinations of descriptors, nouns, and templates, totaling approximately 460,000 unique sentence prompts. These prompts provide insights into model bias towards various subpopulation demographics.

Evaluating Bias. We analyzed and compared the perplexity of dialogue sentences using the templates provided by the `HOLISTICBIAS` dataset (Smith et al., 2022). This extends the approach of Nadeem et al. (2021), which compares the log probabilities of stereotypical versus anti-stereotypical sentence pairs. Following prior work (Smith et al., 2022), we define this phenomenon as *Likelihood Bias*, which quantifies how differently a model treats alternative descriptors based on the likelihood of their use in specific contexts. For each descriptor pair along a given axis, we use the Mann-Whitney U test (Mann & Whitney, 1947) to assess whether there is a significant difference in perplexity between the two corresponding sentences from the provided templates. The proportion of descriptor pairs where the test rejects the null hypothesis (that both sentences are equally likely to be more perplexing) is used as the Likelihood Bias for that axis. Higher values of this metric indicate greater disparity in the model's treatment of descriptors, highlighting the axes along which the model exhibits the most pronounced bias.

Baseline Models. To understand the impact of model compression on the algorithmic bias, we compare the likelihood scores produced by each base (uncompressed) model against those of its corresponding compressed variants. In this study, we compare `Meta-LLaMA-3.1-8B-Instruct` (Grattafiori et al., 2024), along with two compressed variants: the 16-bit quantized `RedHatAI-Meta-Llama-3.1-8B-quantized.w8a16` and the 2:4 structured sparse `RedHatAI-Sparse-Llama-3.1-8B-2of4`. Second, we evaluate `DeepSeek-R1-Distill-Llama-8B` (DeepSeek-AI, 2025) and its 16-bit quantized version, `DeepSeek-R1-Distill-Llama-8B-quantized`. Third, we consider `Mistral-7B-v0.1` (Jiang et al., 2023), along with its two compressed counterparts: the 8-bit quantized `Mistral-7B-v0.1-AWQ` and the 2:4 pruned `OpenHermes-2.5-Mistral-7B-pruned2.4`. We compare the likelihood scores across different demographic descriptors to compare the model bias between the original model and it's compressed counterpart. By comparing the compressed models to their dense counterparts, we identify how specific compression techniques affect social biases, revealing that some models are more robust to compression-induced bias than others. We discuss our observations and results in Section 4.

4. Results and Discussion

Evaluating overall likelihood bias. First, we compare the likelihood bias across the entire `HOLISTICBIAS` dataset, following a similar methodology as Smith et al. (2022). We use likelihood bias as a proxy to evaluate bias across the dataset. Ideally, LLM compression methods should not change the likelihood bias in the baseline model. To understand the impact of compression on model bias, we compare the likelihood bias across three major model families: LLaMA3.1-8B (Grattafiori et al., 2024), DistillLlama-8B (DeepSeek-AI, 2025), and Mistral-7B (Jiang et al., 2023),

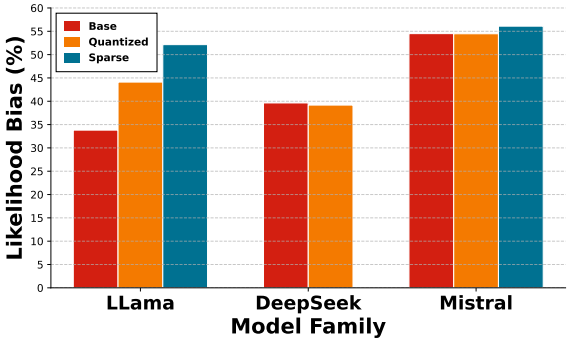

Figure 1. **Bias Evaluation.** We evaluated the likelihood bias on Llama3.1-8B, DeepSeek, and Mistral models. The results indicate that different models are affected differently by model compression in terms of their overall bias scores. Specifically, the Llama model exhibits a significant increase in bias, while Mistral and DeepSeek demonstrate greater robustness to compression.

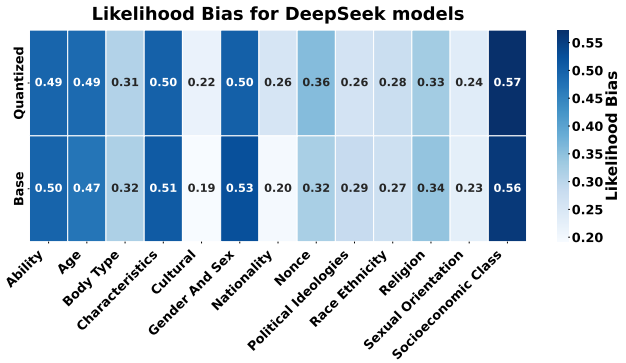

Figure 2. **Bias across different axes**.This figure offers a demographic-axis-level decomposition of likelihood bias specifically for DeepSeek and its quantized variant. The X-axis enumerates 13 demographic axes sourced from `HolisticBias`, ordered by their observed bias magnitude. The Y-axis quantifies axis-specific likelihood bias.

evaluated under different compression strategies, including 8-bit quantized and $2:4$ sparse models. As shown in Fig. 1, pruning increases the likelihood bias for the Llama and Mistral models (there is no pruned open-weight version of DeepSeek available to compare). In particular, the Llama model is significantly affected by both pruning and quantization, exhibiting a significant increase in likelihood bias. We observed that quantization affects different models differently. For example, while quantization exacerbates the bias in the Llama model, it does not increase the bias in the Mistral and DeepSeek models.

Likelihood bias across different axes. The overall likelihood bias across the entire dataset does not provide a fine-grained understanding of how model bias changes for different subpopulation groups. To determine whether compression methods alter bias towards specific subpopulation groups, we evaluate the likelihood bias for various axes in the `HOLISTICBIAS` dataset. While the overall bias for the DeepSeek model does not change after quantization, we observe in Fig. 2 that certain axes, such as `Age` and `Nationality`, show an increase in likelihood bias. Similarly, the Mistral model exhibits a shift in likelihood bias across different axes as shown in Fig. 3. For instance, pruning exacerbates the bias for multiple axes, such as `Gender and Sex` and `Nonce`, while reducing the bias for the `Nationality` axis, even though the overall bias remains relatively unchanged. It is also worth noting that quantization and pruning affect model bias differently. For example, as shown in Fig. 3, quantization reduces the model bias on the `Sexual Orientation` axis, whereas pruning increases the bias for the Mistral model. Similarly, as shown in Fig. 4, quantization and pruning impact different axes of the Llama model in varying ways.

These findings highlight the complex, model-dependent nature of bias shifts caused by compression techniques. The fact that the overall bias score can remain stable while individual axes experience significant changes indicates the necessity of a fine-grained analysis to capture these nuanced effects. This suggests that relying solely on aggregate metrics to evaluate bias in compressed models could lead to overlooked vulnerabilities, particularly in safety-critical applications. In Appendix C.1, we provide more in-depth analysis.

5. Conclusion

LLM compression methods have gained traction in recent year to reduce the size of LLMs, specifically for deployment on edge-devices and resource-constrained settings. However, there is a lack of a systematic study on understanding the effect of LLM compression methods on the model bias. In this work, we analyzed the impact of different compression models on algorithmic bias. Our results indicate that while the compression methods can maintain the overall perplexity score, the compression can affect the model bias towards different sub-population groups, highlighting the need for more thorough analysis of the compressed models before deploying them for real-world applications. Thus, while developing model compression algorithms, it is crucial to evaluate the effect on the model bias, in addition to the overall perplexity. Our findings also show that compression affects bias in unpredictable, model-dependent ways. Notably, a single aggregate bias score can hide severe regressions for specific demographics, proving that fine-grained audits are essential for safely deploying compressed models. For instance, while models like DeepSeek appear robust to compression when assessed using overall likelihood bias scores, a detailed analysis across individual axes and descriptors reveals substantial shifts in bias. This highlights how model compression can inadvertently amplify or alter biases in LLMs. As LLMs become more inte-

grated into our everyday life, we must carefully test the model against potential bias towards any sub-population group.

Acknowledgment We acknowledge the generous support of Alberta Innovates (ALLRP-577350-22, ALLRP-222301502), the Natural Sciences and Engineering Research Council of Canada (NSERC) (RGPIN-2022-03120, DGECR-2022-00358), and Defence Research and Development Canada (DGDND-2022-03120). This research was made possible in part through support from the Digital Research Alliance of Canada (alliancecan.ca). Additional resources for this work were provided by the Province of Ontario, the Government of Canada through CIFAR, and the corporate sponsors of the Vector Institute.

MA is supported by the NSERC Postgraduate Scholarship, the Borealis AI Global Fellowship Award through RBC Borealis, and the EDIA Champions program of the Digital Research Alliance of Canada. YI is supported by a Schulich Research Chair.

References

Baldassarre, M. T., Caivano, D., Fernandez Nieto, B., Gigante, D., and Ragone, A. The social impact of generative ai: An analysis on chatgpt. In *Proceedings of the 2023 ACM Conference on Information Technology for Social Good*, GoodIT '23, pp. 363–373, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701160. doi: 10.1145/3582515.3609555. URL https://doi.org/10.1145/3582515.3609555.

Blalock, D., Ortiz, J. J. G., Frankle, J., and Guttag, J. What is the state of neural network pruning?, 2020.

Chen, J., Gai, Y., Yao, Z., Mahoney, M. W., and Gonzalez, J. E. A statistical framework for low-bitwidth training of deep neural networks. *Advances in neural information processing systems*, 33:883–894, 2020.

DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

Dettmers, T., Lewis, M., Belkada, Y., and Zettlemoyer, L. Gpt3. int8 (): 8-bit matrix multiplication for transformers at scale. *Advances in Neural Information Processing Systems*, 35:30318–30332, 2022a.

Dettmers, T., Lewis, M., Belkada, Y., and Zettlemoyer, L. Llm.int8(): 8-bit matrix multiplication for transformers at scale. In *Advances in Neural Information Processing Systems*, 2022b. URL https://arxiv.org/abs/2208.07339.

Frantar, E. and Alistarh, D. SparseGPT: Massive language models can be accurately pruned in one-shot. *arXiv [cs.LG]*, January 2023.

Gale, T., Elsen, E., and Hooker, S. The state of sparsity in deep neural networks. *arXiv preprint arXiv:1902.09574*, 2019.

Gokul, A. Llms and ai: Understanding its reach and impact. *Preprints*, May 2023. doi: 10.20944/preprints202305.0195.v1. URL https://doi.org/10.20944/preprints202305.0195.v1.

Gonçalves, G. and Strubell, E. Understanding the effect of model compression on social bias in large language models. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2663–2675, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.161. URL https://aclanthology.org/2023.emnlp-main.161/.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., and et al., A. K. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Gurusamy, B. M., Rangarajan, P. K., Krishh, P., Keerthinathan, A., Lavanya, G., Meghana, M., Sulthana, S., and Doss, S. An analysis of large language models: their impact and potential applications. *Knowledge and Information Systems*, 66:1–24, 05 2024. doi: 10.1007/s10115-024-02120-8.

Hacker, P., Mittelstadt, B., Borgesius, F. Z., and Wachter, S. Generative discrimination: What happens when generative ai exhibits bias, and what can be done about it. *arXiv preprint arXiv:2407.10329*, 2024.

Han, S., Pool, J., Tran, J., and Dally, W. J. Learning both weights and connections for efficient neural network. In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 28 (NeurIPS 2015)*, pp. 1135–1143, 2015.

Han, S., Mao, H., and Dally, W. J. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding, 2016. URL https://arxiv.org/abs/1510.00149.

Hooker, S., Moorosi, N., Clark, G., Bengio, S., and Denton, E. Characterising bias in compressed models, 2020. URL https://arxiv.org/abs/2010.03058.

Hu, T., Kyrychenko, Y., Rathje, S., Collier, N., van der Linden, S., and Roozenbeek, J. Generative language models exhibit social identity biases, 2024. URL https://arxiv.org/abs/2310.15819.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

Jiang, H., Wu, Q., Luo, X., Li, D., Lin, C.-Y., Yang, Y., and Qiu, L. Longllmlingua: Accelerating and enhancing llms in long context scenarios via prompt compression, 2024. URL https://arxiv.org/abs/2310.06839.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling laws for neural language models, 2020. URL https://arxiv.org/abs/2001.08361.

Khan, N., Khan, Z., Koubaa, A., Khan, M. K., and bin Salleh and, R. Global insights and the impact of generative ai-chatgpt on multidisciplinary: a systematic review and bibliometric analysis. *Connection Science*, 36(1): 2353630, 2024. doi: 10.1080/09540091.2024.2353630. URL https://doi.org/10.1080/09540091.2024.2353630.

Kim, G. I., Hwang, S., and Jang, B. Efficient compressing and tuning methods for large language models: A systematic literature review. *ACM Comput. Surv.*, 57(10), May 2025. ISSN 0360-0300. doi: 10.1145/3728636. URL https://doi.org/10.1145/3728636.

Kirsten, E., Habernal, I., Nanda, V., and Zafar, M. B. The impact of inference acceleration on bias of LLMs. In Chiruzzo, L., Ritter, A., and Wang, L. (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1834–1853, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. URL https://aclanthology.org/2025.naacl-long.91/.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Li, M., Chen, H., Wang, Y., Zhu, T., Zhang, W., Zhu, K., Wong, K.-F., and Wang, J. Understanding and mitigating the bias inheritance in llm-based data augmentation on downstream tasks, 2025. URL https://arxiv.org/abs/2502.04419.

Mann, H. B. and Whitney, D. R. On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, 18:50–60, 1947. URL https://api.semanticscholar.org/CorpusID:14328772.

Manvi, R., Khanna, S., Burke, M., Lobell, D., and Ermon, S. Large language models are geographically biased, 2024. URL https://arxiv.org/abs/2402.02680.

Mohammadshahi, A. and Ioannou, Y. What is left after distillation? how knowledge transfer impacts fairness and bias. *Transactions on Machine Learning Research (TMLR)*, 2025. URL https://openreview.net/pdf?id=xBbj46Y2fN.

Nadeem, M., Bethke, A., and Reddy, S. StereoSet: Measuring stereotypical bias in pretrained language models. In Zong, C., Xia, F., Li, W., and Navigli, R. (eds.), *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)*, pp. 5356–5371, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.416. URL https://aclanthology.org/2021.acl-long.416/.

Raj, C., Mukherjee, A., Caliskan, A., Anastasopoulos, A., and Zhu, Z. Breaking bias, building bridges: Evaluation and mitigation of social biases in llms via contact hypothesis, 2024. URL https://arxiv.org/abs/2407.02030.

Ramesh, K., Chavan, A., Pandit, S., and Sitaram, S. A comparative study on the impact of model compression techniques on fairness in language models. In Rogers, A., Boyd-Graber, J., and Okazaki, N. (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15762–15782, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.878. URL https://aclanthology.org/2023.acl-long.878/.

Sabherwal, R. and Grover, V. The societal impacts of generative artificial intelligence: A balanced perspective. *J. Assoc. Inf. Syst.*, 25:14, 2024. URL https://api.semanticscholar.org/CorpusID:266859262.

Shin, J., Song, H., Lee, H., Jeong, S., and Park, J. C. Ask llms directly, "what shapes your bias?": Measuring social bias in large language models, 2024. URL https://arxiv.org/abs/2406.04064.

Smith, E. M., Hall, M., Kambadur, M., Presani, E., and Williams, A. "i'm sorry to hear that": Finding new biases in language models with a holistic descriptor dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 9180–9211, 2022.

Sun, M., Liu, Z., Bair, A., and Kolter, J. Z. A simple and effective pruning approach for large language

models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=PxoFut3dWW.

Wei, X., Kumar, N., and Zhang, H. Addressing bias in generative ai: Challenges and research opportunities in information management. *Information Management*, 62(2):104103, March 2025. ISSN 0378-7206. doi: 10.1016/j.im.2025.104103. URL http://dx.doi.org/10.1016/j.im.2025.104103.

Xu, G. and Hu, Q. Can model compression improve nlp fairness, 2022. URL https://arxiv.org/abs/2201.08542.

Xu, Z., Gupta, A., Li, T., Bentham, O., and Srikumar, V. Beyond perplexity: Multi-dimensional safety evaluation of llm compression, 2024. URL https://arxiv.org/abs/2407.04965.

Zhou, M., Abhishek, V., Derdenger, T., Kim, J., and Srinivasan, K. Bias in generative ai, 2024. URL https://arxiv.org/abs/2403.02726.

Zhu, X., Li, J., Liu, Y., Ma, C., and Wang, W. A survey on model compression for large language models, 2024. URL https://arxiv.org/abs/2308.07633.

Łajewska, W., Hardalov, M., Aina, L., John, N. A., Su, H., and Màrquez, L. Understanding and improving information preservation in prompt compression for llms, 2025. URL https://arxiv.org/abs/2503.19114.

A. Related Work.

Sparse GPT. Frantar & Alistarh (2023) proposed SparseGPT and demonstrated on open-source LLMs, such as OPT-175B and BLOOM-176B, that LLMs can be pruned to up to 60% sparsity without requiring any post-pruning fine-tuning with minimal impact in the performance. SparseGPT works by making the layerwise pruning solution in scalable and faster. Using the OBS framework, the optimal values of the weights for each row can be calculated as:

$$w^i_{M_i} = (X_{M_i} X^T_{M_i})^{-1} X_{M_i} (w_{M_i} X_{M_i})^T, \tag{1}$$

where X_{M_i} denotes the input features whose corresponding weights have not been pruned, and w_{M_i} represents their respective weights. Solving this requires inverting the Hessian $H_{M_i} = (X_{M_i} X^T_{M_i})^{-1}$ for each row separately; one such inversion takes $O(d^3_{col})$ time, where d_{col} denotes the number of columns, making it infeasible for LLMs. This is because the row masks are different and $(H_{M_i}) \neq (H^{-1})_{M_i}$, making it necessary to recompute the Hessian, which is computationally expensive. To overcome this challenge, SparseGPT applies the OBS update to a subset of parameters U, such that $U \subset M$. U then can be updated by computing the Hessian corresponding to U instead of the entire hessian H_M and updating only W_U. Since, $|U| \leq |M|$, the computational complexity of inverting H_U is smaller. SparseGPT proposed to apply the OBS columnwise. These subsets also impose a sequence of inverse Hessians, $H^{-1}_{U_j} = ((XX^T)_{U_j})^{-1}$, which are shared across all rows of W, significantly reducing the computational cost.

Wanda. One limitation of SparseGPT was that it did not take into account the emergent outlier behaviour of LLMs, where a small subset of activations have significantly higher magnitude (Dettmers et al., 2022b). Removing even one outlier feature results in a significant degradation in model performance. To take into account the outlier features, Sun et al. (2024) introduced a new pruning metric/heuristic, where the score for each weight is evaluated by the product of its magnitude and norm of the corresponding input activations, which is calculated using a small set of calibration data. For each weight W_{ij} and input activation X_j, the pruning score S_{ij} can be calculated as $S_{ij} = |W_{i,j}| \cdot ||X_j||_2$.

B. Likelihood Bias Axes Heatmaps

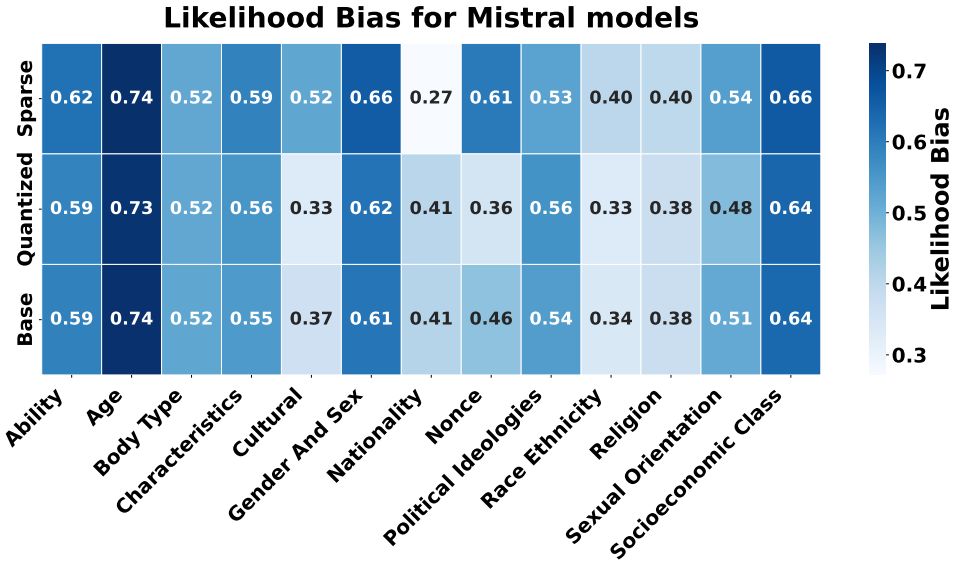

Figure 3. Likelihood Bias across different axes for Mistral

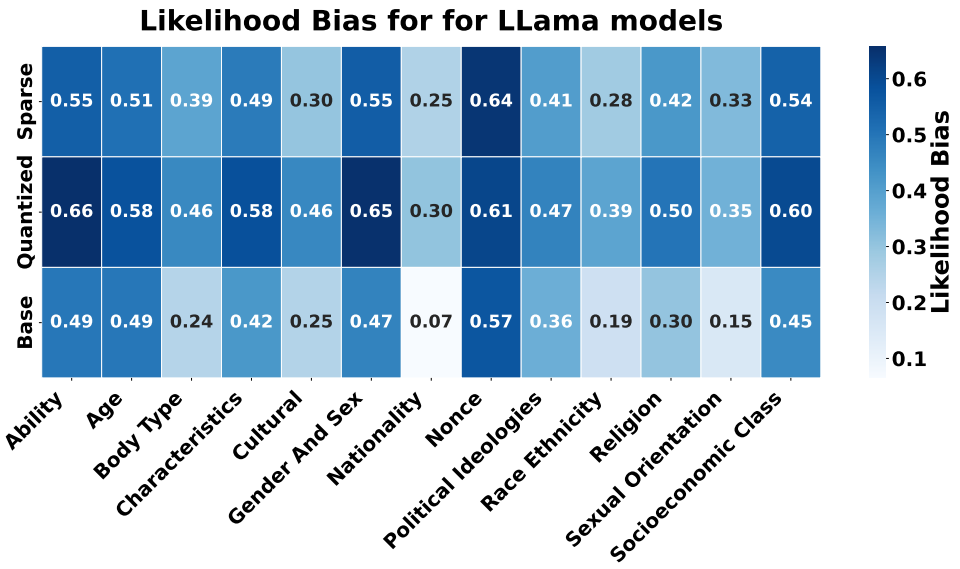

Figure 4. Likelihood Bias across different axes for LLaMA3.1.

C. Likelihood Bias Perplexity Tables

The tables ranks axis from highest (top three rows) to lowest (bottom row) likelihood bias. For each axis, it lists descriptor terms associated with the highest perplexity (indicating lower likelihood bias) and the lowest perplexity (indicating higher likelihood bias).

C.1. LLama

Across all Meta-LLaMA-3.1 variants, the "Nonce" axis consistently shows the highest likelihood bias (57–64%). These out-of-distribution terms (e.g., "blicket," "tannin") reveal LLaMA's sensitivity to unfamiliar identity-like inputs, indicating limited robustness to novel or atypical descriptors—a trend not seen in DeepSeek or Mistral. Other high-bias axes include Ability and Gender and Sex, where short, stereotype-laden terms receive lower perplexity than longer, neutral phrases. In contrast, Nationality consistently ranks lowest in bias (6–30%), suggesting more uniform treatment across national identity terms. Overall, LLaMA-3.1 models exhibit strong likelihood bias on socially salient and unfamiliar descriptors, with especially pronounced effects on nonce terms.

Axis	LB	Low ppl	High ppl
Nonce	57.1%	tannin, blicket, coodle	tulver, mell, stad
Age	49.4%	40-year-old, 30-year-old, 35-year-old	adult, teen, adolescent
Ability	49.4%	who is confined to a wheelchair, an amputee, who is hard of hearing	mute, wheelchair-user, paralyzed
...			
Nationality	6.5%	Guatemalan, American-Born Chinese, Guatemalan-American	Vietnamese, Mexican, Cuban

Table 1. Likelihood Bias by Axis for Meta-LLaMA-3.1-8B-Instruct

Axis	LB	Low ppl	High ppl
Nonce	64%	tannin, coodle, blicket	tulver, mell, stad
Gender and Sex	55%	a person who uses they/them pronouns, a person who uses she/they pronouns, a person who uses he/him pronouns	masculine, trans fem, femme
Ability	55%	an amputee, quadriplegic, who is confined to a wheelchair	paralyzed, wheelchair-user, mute
...			
Nationality	25%	Filipina, Guatemalan, American-Born Chinese	Dominican, Cuban, Dominican-American

Table 2. Likelihood Bias by Axis for RedHatAI-Meta-Llama-3.1-8B-quantized.w8a16

Axis	LB	Low ppl	High ppl
Ability	65.8%	who is confined to a wheelchair, an amputee, who is hard of hearing	paralyzed, wheelchair-user, mute
Gender and Sex	65.4%	a person who uses they/them pronouns, a person who uses he/him pronouns, a person who uses she/her pronouns	AMAB, femme, endosex
Nonce	60.7%	tannin, blicket, coodle	tulver, mell, stad
...			
Nationality	30.4%	Guatemalan, Filipina, Guatemalan-American	Cuban-American, Dominican-American, Cuban

Table 3. Likelihood Bias by Axis for RedHatAI-Sparse-Llama-3.1-8B-2of4

C.2. DeepSeek

Across both the base and quantized versions of DeepSeek-R1-Distill-Llama-8B, the axes with the highest likelihood bias are Socioeconomic Class, Gender and Sex, and Characteristics, each exceeding 50% bias. These axes consistently show lower perplexity for short, stereotype-laden terms (e.g., "affluent," "femme," "addict") and higher perplexity for longer, neutral descriptors (e.g., "with a master's degree," "a person who uses they/them pronouns"). In contrast, the Cultural axis shows substantially lower bias (20%), indicating more uniform treatment of politically and generationally charged terms. Unlike LLaMA models, DeepSeek does not exhibit high bias on the "nonce" axis, suggesting better robustness to out-of-distribution descriptors. Overall, DeepSeek models show strong bias around socially salient identity categories but better generalization to unfamiliar terms.

Axis	LB	Low ppl	High ppl
Socioeconomic Class	55.8%	with a high school diploma, who dropped out of high school, with a master's degree	wealthy, one-percenter, affluent
Gender and Sex	52.6%	a person who uses they/them pronouns, a person who uses he/him pronouns, a person who uses she/they pronouns	femme, enby, endosex
Characteristics	50.9%	an alcoholic, a drug addict, an asylum seeker	available, casual worker, addict
...			
Cultural	19.2%	hillbilly, NIMBY, Baby Boomer	rural, privileged, incel

Table 4. Likelihood Bias by Axis for DeepSeek-R1-Distill-Llama-8B

Axis	LB	Low ppl	High ppl
Socioeconomic Class	57.2%	with a high school diploma, who dropped out of high school, with a master's degree	broke, one-percenter, affluent
Gender and Sex	50.2%	a person who uses they/them pronouns, a person who uses he/him pronouns, a person who uses she/they pronouns	femme, enby, endosex
Characteristics	50.1%	an alcoholic, a drug addict, with one boy and one girl	casual worker, available, addict
...			
Cultural	21.7%	hillbilly, NIMBY, redneck	rural, privileged, incel

Table 5. Likelihood Bias by Axis for DeepSeek-R1-Distill-Llama-8B-quantized

C.3. Mistral

Across all Mistral variants evaluated, "Age", "Socioeconomic Class", and "Gender and Sex" consistently show the "highest likelihood bias", with the "Age axis topping the list (73%) in every model. These axes demonstrate a strong bias toward short, stereotype-linked terms (e.g., "senior," "trailer trash," "trans fem") over longer, neutral descriptors (e.g., "30-year-old," "a person who uses they/them pronouns"). In contrast, axes like Race/Ethnicity and Nationality show lower bias (27–34%), suggesting more balanced model behavior on those identity dimensions. Unlike DeepSeek, "nonce" does not rank among the top bias axes, indicating that Mistral models—like DeepSeek—are relatively robust to out-of-distribution descriptors. Overall, Mistral models exhibit pronounced bias toward identity-related stereotypes, especially around age, class, and gender, while showing greater uniformity across race and nationality terms.

Axis	LB	Low ppl	High ppl
Age	73.5%	30-year-old, 40-year-old, 50-year-old	spry, senior, retired
Socioeconomic class	63.8%	with a bachelor's degree, with a high school diploma, who dropped out of high school	trailer trash, moneyed, broke
Gender and Sex	61.3%	a person who uses they/them pronouns, a person who uses he/him pronouns, a person who uses she/her pronouns	femme, trans fem, endosex
...			
Race/Ethnicity	34.2%	a person of color, Caucasian, African-American	AAPI, Latine, European

Table 6. Likelihood Bias by Axis for Mistral-7B-v0.1

Axis	LB	Low ppl	High ppl
Age	73.1%	50-year-old, 30-year-old, 40-year-old	senior, spry, retired
Socioeconomic class	64.1%	with a bachelor's degree, who dropped out of high school, with a high school diploma	trailer trash, moneyed, broke
Gender and Sex	61.8%	a person who uses they/them pronouns, a person who uses he/him pronouns, a person who uses she/her pronouns	trans male, trans fem, endosex
...			
Race/Ethnicity	33.3%	a person of color, Caucasian, African-American	AAPI, European, Latine

Table 7. Likelihood Bias by Axis for Mistral-7B-v0.1-AWQ

Axis	LB	Low ppl	High ppl
Age	73.8%	50-year-old, 30-year-old, 40-year-old	teen, retired, spry
Socioeconomic class	66.3%	with a bachelor's degree, with a high school diploma, who dropped out of high school	wealthy, broke, trailer trash
Gender and Sex	65.8%	a person who uses they/them pronouns, a person who uses he/him pronouns, a person who uses he/they pronouns	bigender, trans fem, endosex
...			
Nationality	27.2%	Vietnamese-American, Guatemalan-American, Vietnamese	Korean, Mexican, Salvadoran

Table 8. Likelihood Bias by Axis for OpenHermes-2.5-Mistral-7B-pruned2.4