Abid, A., Farooqi, M., & Zou, J. (2021). Persistent Anti-Muslim Bias in Large Language Models. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 298–306. https://doi.org/10.1145/3461702.3462624

Bevara, R. V. K., Mannuru, N. R., Karedla, S. P., & Xiao, T. (2024). Scaling Implicit Bias Analysis across Transformer-Based Language Models through Embedding Association Test and Prompt Engineering. *Applied Sciences*, *14*(8). https://doi.org/10.3390/app14083483

Deshmukh, A., & Raut, A. (2024). Applying BERT-Based NLP for Automated Resume Screening and Candidate Ranking. *Annals of Data Science.* https://doi.org/10.1007/s40745-024-00524-5

Dhamala, J., Sun, T., Kumar, V., Krishna, S., Pruksachatkun, Y., Chang, K.-W., & Gupta, R. (2021). BOLD: Dataset and Metrics for Measuring Biases in Open-Ended Language Generation. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 862–872. https://doi.org/10.1145/3442188.3445924

Esiobu, D., Tan, X., Hosseini, S., Ung, M., Zhang, Y., Fernandes, J., Dwivedi-Yu, J., Presani, E., Williams, A., & Smith, E. (2023). ROBBIE: Robust Bias Evaluation of Large Generative Language Models. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 3764–3814. https://doi.org/10.18653/v1/2023.emnlp-main.230

Giordano, C., Brennan, M., Mohamed, B., Rashidi, P., Modave, F., & Tighe, P. (2021). Accessing Artificial Intelligence for Clinical Decision-Making. *Frontiers in Digital Health, 3*. https://doi.org/10.3389/fdgth.2021.645232

Jacob, & Turner, A. (2024). *I found >800 orthogonal "write code" steering vectors*. LessWrong. https://www.lesswrong.com/posts/CbSEZSpjdpnvBcEvc/i-found-greater-than-800-orthogonal-write-code-steering

Khan, S., Zakir, M., Bashir, S., & Ali, R. (2024). Artificial Intelligence and Machine Learning in Legal Research: A Comprehensive Analysis. *Qlantic Journal of Social Sciences, 5*(1), 307-317. https://doi.org/10.55737/qjss.203679344

Kotek, H., Dockum, R., & Sun, D. (2023). Gender bias and stereotypes in Large Language Models. *Proceedings of The ACM Collective Intelligence Conference*, 12–24. https://doi.org/10.1145/3582269.3615599

Lu, D., & Rimsky, N. (2024). *Investigating Bias Representations in Llama 2 Chat via Activation Steering*. ArXiv. https://doi.org/10.48550/arXiv.2402.00402

Olah, C., Mordvintsev, A., & Schubert, L. (2017). Feature Visualization. *Distill*, *2*(11). https://doi.org/10.23915/distill.00007

Panickssery, N., Gabrieli, N., Schulz, J., Tong, M., Hubinger, E., & Turner, A. M. (2023). *Steering Llama 2 via Contrastive Activation Addition.* ArXiv. https://doi.org/10.48550/ARXIV.2312.06681

Papakyriakopoulos, O., Hegelich, S., Serrano, J. C. M., & Marco, F. (2020). Bias in word embeddings. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 446–457. https://doi.org/10.1145/3351095.3372843

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J. T., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Luca Antiga, Alban Desmaison, Kopf, A., Yang, E. S., DeVito, Z., Raison, M., Tejani, A., Sasank Chilamkurthy, Steiner, B., Fang, L., & Bai, J. (2019). P*yTorch: An Imperative Style, High-Performance Deep Learning Library.* ArXiv. https://doi.org/10.48550/arxiv.1912.01703

Raiaan, M., Mukta, S., Fatema, K., Fahad, N., Sakib, S., Mim, M., Marufatul, J., Ahmad, J., Ali, M. E., & Azam, S. (2023). A Review on Large Language Models: Architectures, Applications, Taxonomies, Open Issues and Challenges. *IEEE Access, 12*, 26839-26874. https://doi.org/10.1109/ACCESS.2024.3365742

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., & Fuller, B. (2023). *Llama 2: Open Foundation and Fine-Tuned Chat Models.* ArXiv. https://doi.org/10.48550/arXiv.2307.09288

Turner, A. M., Thiergart, L., Leech, G., Udell, D., Vazquez, J. J., Mini, U., & MacDiarmid, M. (2023). *Steering Language Models With Activation Engineering* (Version 5). ArXiv. https://doi.org/10.48550/ARXIV.2308.10248

Xue, M., Liu, D., Yang, K., Dong, G., Lei, W., Yuan, Z., Zhou, C., & Zhou, J. (2023). *OccuQuest: Mitigating Occupational Bias for Inclusive Large Language Models*. ArXiv. https://doi.org/10.48550/arXiv.2310.16517

Yang, Y., Liu, X., Jin, Q., Huang, F., & Lu, Z. (2024). Unmasking and quantifying racial bias of large language models in medical report generation. *Communications Medicine*, *4*(1), 176. https://doi.org/10.1038/s43856-024-00601-z

Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., Pan, A., Yin, X., Mazeika, M., Dombrowski, A., Goel, S., Li, N., Byun, M.J., Wang, Z., Mallen, A.T., Basart, S., Koyejo, S., Song, D., Fredrikson, M., Kolter, Z., & Hendrycks, D. (2023). *Representation Engineering: A Top-Down Approach to AI Transparency*. ArXiv. https://doi.org/10.48550/arXiv.2310.01405