References list

[1] Z. Tang, K. Zhou, P. Wang, Y. Ding, J. Li, and Minzhang, “Detoxify language model step-by-step,” *arXiv [cs.CL]*, 2023.

[2] S. Hallinan, A. Liu, Y. Choi, and M. Sap, “Detoxifying text with MaRCo: Controllable revision with experts and anti-experts,” in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2023.

[3] D. Dale *et al.*, “Text detoxification using large pre-trained neural models,” in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021.

[4] G. Floto *et al.*, “DiffuDetox: A mixed diffusion model for text detoxification,” in *Findings of the Association for Computational Linguistics: ACL 2023*, 2023.

[5] Dataset: https://github.com/hexinz/SI630_final_project/tree/main/Data

[6] Fill-mask tutorial: https://huggingface.co/learn/nlp-course/chapter7/3?fw=pt

[7] Zero-shot tutorial: https://joeddav.github.io/blog/2020/05/29/ZSL.html