1. Ethical-Advice Taker: Do Language Models Understand Natural Language Interventions? (https://arxiv.org/pdf/2106.01465.pdf)
2. The Capacity for Moral Self-Correction in Large Language Models (https://arxiv.org/pdf/2302.07459.pdf)
3. Process for Adapting Language Models to Society (PALMS) with Values-Targeted Datasets (https://arxiv.org/pdf/2106.10328.pdf)
4. Self-Diagnosis and Self-Debiasing: A Proposal for Reducing Corpus-Based Bias in NLP (https://arxiv.org/pdf/2103.00453.pdf)
5. Forecasting potential misuses of language models for disinformation campaigns—and how to reduce risk, Josh A. Goldstein, Girish Sastry, Micah Musser, Renée DiResta, Matthew Gentzel, Katerina Sedova, Stanford Internet Observatory, Jan 2023
6. Exploring AI Ethics of ChatGPT: A Diagnostic Analysis https://arxiv.org/pdf/2301.12867.pdf
7. The Capacity for Moral Self-Correction in Large Language Models, Ganguli et al, Feb 2023: “We test the hypothesis that language models trained with reinforcement learning from human feedback (RLHF) have the capability to "morally self-correct" -- to avoid producing harmful outputs -- if instructed to do so.”
8. A Comprehensive Survey on Pretrained Foundation Models: A History from BERT to ChatGPT, Zhou et al., arXiv, Feb 2023

9. More than you've asked for: A Comprehensive Analysis of Novel Prompt Injection Threats to Application-Integrated Large Language Models, Greshake et al., arXiv, Feb 2023
10. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing, Liu et al, CMU/NUS, Jul 2021 Misc:
11. https://www.cnn.com/2023/02/17/tech/microsoft-bing-ai-changes, confrontational remarks and troubling fantasies
12. "Learning to Generate Human-Like Dialogue Response with Deep Reinforcement Learning" by Bing Liu, Ian Lane, and Frank Lin (2017).
13. "Preventing Discrimination in Language Models via Counterfactual Evaluation" by Daniel Khashabi, Tushar Khot, Ashish Sabharwal, and Oyvind Tafjord (2020).
14. **Investigating Gender Bias in Language Models Using Causal Mediation Analysis (**Part of Advances in Neural Information Processing Systems 33 (NeurIPS 2020)) **Authors:** *Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, Stuart Shieber*
15. D. Chong *et al*., "A real-time platform for contextualized conspiracy theory analysis," *2021 International Conference on Data Mining Workshops (ICDMW)*, Auckland, New Zealand, 2021, pp. 118-127, doi: 10.1109/ICDMW53433.2021.00021
16. Bing He, Mustaque Ahamad, Srijan Kumar, “Reinforcement Learning-based Counter-Misinformation Response Generation: A Case Study of COVID-19 Vaccine Misinformation”

17. J. Yao *et al*., "Edge-Cloud Polarization and Collaboration: A Comprehensive Survey for AI," in *IEEE Transactions on Knowledge and Data Engineering*, doi: 10.1109/TKDE.2022.3178211
18. Femi Emmanuel Ayo, Olusegun Folorunso, Friday Thomas Ibharalu, Idowu Ademola Osinuga, Machine learning techniques for hate speech classification of twitter data: State-of-the-art, future challenges and research directions, Computer Science Review, Volume 38, 2020, 100311, ISSN 1574-0137, https://doi.org/10.1016/j.cosrev.2020.100311
19. Velankar, A., Patil, H., & Joshi, R. (2022). A Review of Challenges in Machine Learning based Automated Hate Speech Detection. *arXiv*. https://doi.org/https://arxiv.org/abs/2209.05294v1
20. Michael Landon-Murray, Edin Mujkic & Brian Nussbaum (2019) Disinformation in Contemporary U.S. Foreign Policy: Impacts and Ethics in an Era of Fake News, Social Media, and Artificial Intelligence, Public Integrity, 21:5, 512-522, DOI: 10.1080/10999922.2019.1613832
21. https://www.cnn.com/2023/02/17/tech/microsoft-bing-ai-changes, confrontational remarks and troubling fantasies
22. Jail break ChatGPT
23. Nature news: What ChatGPT and generative AI mean for science
24. Detect AI-generated content:https://www.wired.com/story/how-to-spot-generative-ai-text-chatgpt/, https://gptzero.me/
25. CYBERCRIMINALS STARTING TO USE CHATGPT
26. ChatGPT generates paper ideas: https://twitter.com/emollick/status/1626021146406662146?s=20

27. ChatGPT for easier scientific research: https://consensus.app
28. ChatGPT Subs In as Security Analyst, Hallucinates Only Occasionally
29. GPTZero: The World's #1 AI Detector https://gptzero.me/

    "GPTZero is a classification model that predicts whether a document was written by a large language model, providing predictions on a sentence, paragraph, and document level. GPTZero was trained on a large, diverse corpus of human-written and AI-generated text, with a focus on English prose."
30. Learning to Prevent Neural Network Generation of Obscene Content" by Yen-Chun Chen, Yi-Pei Chen, Yu-Chen Kuo, Lun-Wei Ku, and Wen-Lian Hsu (2018).
31. "Controlling Offensive Language in Open-Ended User Simulations with Unlikelihood Training" by Adarsh Pyarelal, Kyunghyun Cho, and Jason Weston (2019).
32. "Adversarial Filtering for Neural Dialogue Generation" by Jiaao Chen, Kai Fan, Chengqi Zhao, and Duyu Tang (2018).
33. "Conan: Bridging Text and Knowledge with a Conversational Question Answering System" by Xiaodong Liu, Yelong Shen, Kevin Duh, and Jianfeng Gao (2017).
34. "Controlling GPT-2 Content with Semantic-Based Filtering" by Nikolaos Aletras, Daniil Sorokin, and Iryna Gurevych (2020).
35. "Training Deep Reinforcement Learning Models with Synthetic Data and Transfer Learning for Text-based Games" by Ondrej Skopek, Khalil Kidwai, and Michael M. Bronstein (2019).

36. "Learning Personalized Dialogue Generation with Neural Latent Variable Models" by Zhou Yu, Jun Yu, Yujie Qian, and Weinan Zhang (2017).
37. "Preventing undesirable behavior of intelligent agents through misbehaviour shaping" by Jelena Mirkovic, David Livingstone, and Tomasz M. Rutkowski (2019).
38. Online. In: Gregor, M., Mlejnková, P. (eds) Challenging Online Propaganda and Disinformation in the 21st Century. Political Campaigning and Communication. Palgrave Macmillan, Cham. https://doi.org/10.1007/978-3-030-58624-9_2