[38] Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. Personallm: Investigating the ability of large language models to express personality traits. *arXiv preprint arXiv:2305.02547*, 2023.

[39] Myra Cheng, Esin Durmus, and Dan Jurafsky. Marked personas: Using natural language prompts to measure stereotypes in language models. *arXiv preprint arXiv:2305.18189*, 2023.

[40] Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. Toxicity in chatgpt: Analyzing persona-assigned language models. In *EMNLP*, pages 1236–1270, 2023.

[41] Emily Sheng, Josh Arnold, Zhou Yu, Kai-Wei Chang, and Nanyun Peng. Revealing persona biases in dialogue systems. *arXiv preprint arXiv:2104.08728*, 2021.

[42] Preetam Prabhu Srikar Dammu, Hayoung Jung, Anjali Singh, Monojit Choudhury, and Tanushree Mitra. "they are uncultured": Unveiling covert harms and social threats in llm generated conversations. *arXiv preprint arXiv:2405.05378*, 2024.

[43] Ashutosh Kumar, Shiv Vignesh Murthy, Sagarika Singh, and Swathy Ragupathy. The ethics of interaction: Mitigating security threats in llms. *arXiv preprint arXiv:2401.12273*, 2024.

[44] Huandong Wang, Wenjie Fu, Yingzhou Tang, Zhilong Chen, Yuxi Huang, Jinghua Piao, Chen Gao, Fengli Xu, Tao Jiang, and Yong Li. A survey on responsible llms: Inherent risk, malicious use, and mitigation strategy. *arXiv preprint arXiv:2501.09431*, 2025.

[45] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *naacL-HLT*, volume 1, 2019.

[46] Patrick Schramowski, Cigdem Turan, Sophie Jentzsch, Constantin Rothkopf, and Kristian Kersting. Bert has a moral compass: Improvements of ethical and moral values of machines. *arXiv preprint arXiv:1912.05238*, 2019.

[47] Asma Ghandeharioun, Avi Caciularu, Adam Pearce, Lucas Dixon, and Mor Geva. Patchscopes: A unifying framework for inspecting hidden representations of language models. In *ICML*, 2024. URL `https://openreview.net/forum?id=5uwBzcn885`.

[48] Chandan Singh, Jeevana Priya Inala, Michel Galley, Rich Caruana, and Jianfeng Gao. Rethinking interpretability in the era of large language models. *arXiv preprint arXiv:2402.01761*, 2024.

[49] Adam Scherlis, Kshitij Sachan, Adam S Jermyn, Joe Benton, and Buck Shlegeris. Polysemanticity and capacity in neural networks. *arXiv preprint arXiv:2210.01892*, 2022.

[50] Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. Linear algebraic structure of word senses, with applications to polysemy. *Transactions of the Association for Computational Linguistics*, 6:483–495, 2018.

[51] Miriam Rateike, Celia Cintas, John Wamburu, Tanya Akumu, and Skyler Speakman. Weakly supervised detection of hallucinations in llm activations. *arXiv preprint arXiv:2312.02798*, 2023.

[52] Victor Akinwande, Celia Cintas, Skyler Speakman, and Srihari Sridharan. Identifying audio adversarial examples via anomalous pattern detection. *Workshop on Adversarial Learning Methods for ML and DM, KDD*, 2020.

[53] Celia Cintas, Skyler Speakman, Girmaw Abebe Tadesse, Victor Akinwande, Edward McFowland III, and Komminist Weldemariam. Pattern detection in the activation space for identifying synthesized content. *Pattern Recognition Letters*, 153:207–213, 2022.

[54] Lewis R Goldberg. An alternative "description of personality": The big-five factor structure. In *Personality and personality disorders*, pages 34–47. Routledge, 2013.

[55] Oliver P John, Laura P Naumann, and Christopher J Soto. Paradigm shift to the integrative big five trait taxonomy. *Handbook of personality: Theory and research*, 3(2):114–158, 2008.

[56] Robert R McCrae and Paul T Costa. The stability of personality: Observations and evaluations. *Current Directions in Psychological Science*, 3(6):173–175, 1994.

[57] Gönül Kaya Özbağ. The role of personality in leadership: Five factor personality traits and ethical leadership. *Procedia-Social and Behavioral Sciences*, 235:235–242, 2016.

[58] Timothy A Judge and Cindy P Zapata. The person–situation debate revisited: Effect of situation strength and trait activation on the validity of the big five personality traits in predicting job performance. *Academy of Management Journal*, 58(4):1149–1179, 2015.

[59] Bronti Baptiste. The relationship between the big five personality traits and authentic leadership. 2018.

[60] Zana Hasan Babakr and Nabi Fatahi. Big five personality traits and risky decision-making: A study of behavioural tasks among college students. *Passer Journal of Basic and Applied Sciences*, 5(2):298–303, 2023.