

- [61] ANSHIKA Grover and A Amit. The big five personality traits and leadership: A comprehensive analysis. *International Journal For Multidisciplinary Research*, 6(1), 2024.
- [62] Wiktoria Mieszczenko-Kowszewicz, Dawid Płodowski, Filip Kołodziejczyk, Jakub Świstak, Julian Sienkiewicz, and Przemysław Biecek. The dark patterns of personalized persuasion in large language models: Exposing persuasive linguistic features for big five personality traits in llms responses. *arXiv preprint arXiv:2411.06008*, 2024.
- [63] Yang Yan, Lizhi Ma, Anqi Li, Jingsong Ma, and Zhenzhong Lan. Predicting the big five personality traits in chinese counselling dialogues using large language models. *arXiv preprint arXiv:2406.17287*, 2024.
- [64] Basile Garcia, Crystal Qian, and Stefano Palminteri. The moral turing test: Evaluating human-llm alignment in moral decision-making. *arXiv preprint arXiv:2410.07304*, 2024.
- [65] Jiseon Kim, Jea Kwon, Luiz Felipe Vecchietti, Alice Oh, and Meeyoung Cha. Exploring persona-dependent llm alignment for the moral machine experiment. *arXiv preprint arXiv:2504.10886*, 2025.
- [66] Jianchao Ji, Yutong Chen, Mingyu Jin, Wujiang Xu, Wenyue Hua, and Yongfeng Zhang. Moralbench: Moral evaluation of llms. *arXiv preprint arXiv:2406.04428*, 2024.
- [67] Yu Lei, Hao Liu, Chengxing Xie, Songjia Liu, Zhiyu Yin, Canyu Chen, Guohao Li, Philip Torr, and Zhen Wu. Fairmindsim: Alignment of behavior, emotion, and belief in humans and llm agents amid ethical dilemmas. *arXiv preprint arXiv:2410.10398*, 2024.
- [68] Elizaveta Tennant, Stephen Hailes, and Mirco Musolesi. Moral alignment for llm agents. *arXiv preprint arXiv:2410.01639*, 2024.
- [69] Kaiping Chen, Anqi Shao, Jirayu Burapachee, and Yixuan Li. A critical appraisal of equity in conversational ai: Evidence from auditing gpt-3’s dialogues with different publics on climate change and black lives matter. *ArXiv*, abs/2209.13627, 2022. URL <https://api.semanticscholar.org/CorpusID:252568261>.
- [70] Dominik Stambach, Philine Widmer, Eunjung Cho, Caglar Gulcehre, and Elliott Ash. Aligning large language models with diverse political viewpoints. *arXiv preprint arXiv:2406.14155*, 2024.
- [71] Falaah Arif Khan, Nivedha Sivakumar, Yinong Oliver Wang, Katherine Metcalf, Cezanne Camacho, Barry-John Theobald, Luca Zappella, and Nicholas Apostoloff. Uncovering intersectional stereotypes in humans and large language models. 2025.
- [72] J Shane Culpepper, Alistair Moffat, Sachin Pathiyan Cherumanal, Falk Scholer, and Johanne Trippas. The effects of demographic instructions on llm personas. 2025.
- [73] Helena A Haxvig. Concerns on bias in large language models when creating synthetic personae. *arXiv preprint arXiv:2405.05080*, 2024.
- [74] AI@Meta. Llama3 model card. <https://github.com/meta-llama/llama3>, 2024. Accessed: 2024-10-24.
- [75] IBM Granite Team. Granite: A new framework for language models. <https://github.com/ibm-granite/granite-3.0-language-models/blob/main/paper.pdf>, 2023. Accessed: 2024-10-24.
- [76] AQ Jiang, A Sablayrolles, A Mensch, C Bamford, DS Chaplot, D de las Casas, F Bressand, G Lengyel, G Lample, L Saulnier, et al. Mistral 7b (2023). *arXiv preprint arXiv:2310.06825*, 2023.
- [77] Daixuan Cheng, Yuxian Gu, Shaohan Huang, Junyu Bi, Minlie Huang, and Furu Wei. Instruction pre-training: Language models are supervised multitask learners. *arXiv preprint arXiv:2406.14491*, 2024.
- [78] Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. How to fine-tune bert for text classification? In *Chinese Computational Linguistics*, pages 194–206. Springer, 2019.
- [79] Yu-Hsiang Tseng, Pin-Er Chen, Da-Chen Lian, and Shu-Kai Hsieh. The semantic relations in llms: An information-theoretic compression approach. In *Workshop: Bridging Neurons and Symbols for NLP and KGR @ LREC-COLING*, pages 8–21, 2024.
- [80] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *JMLR*, 12:2825–2830, 2011.
- [81] Tadeusz Caliński and Jerzy Harabasz. A dendrite method for cluster analysis. *Communications Statistics–Theory and Methods*, 3(1):1–27, 1974.
- [82] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.
- [83] David L Davies and Donald W Bouldin. A cluster separation measure. *IEEE Transactions Pattern Analysis & Machine Intelligence*, (2):224–227, 1979.