

- Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, Sang T. Truong, Simran Arora, Mantas Mazeika, Dan Hendrycks, Zinan Lin, Yu Cheng, Sanmi Koyejo, Dawn Song, and Bo Li. 2023. [DecodingTrust: A comprehensive assessment of trustworthiness in GPT models](#). *CoRR*, abs/2306.11698.
- Haoran Wang and Kai Shu. 2023. [Backdoor activation attack: Attack large language models using activation steering for safety-alignment](#). *CoRR*, abs/2311.09433.
- Xinwei Wu, Weilong Dong, Shaoyang Xu, and Deyi Xiong. 2024. [Mitigating privacy seesaw in large language models: Augmented privacy neuron editing via activation patching](#). In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 5319–5332.
- Shaoyang Xu, Junzhuo Li, and Deyi Xiong. 2023. [Language representation projection: Can we transfer factual knowledge across languages in multilingual language models?](#) In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 3692–3702. Association for Computational Linguistics.
- Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. [mT5: A massively multilingual pre-trained text-to-text transformer](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 483–498. Association for Computational Linguistics.
- Zheng-Xin Yong, Cristina Menghini, and Stephen H Bach. 2023. [Low-resource languages jailbreak GPT-4](#). *CoRR*, abs/2310.02446.
- Linhao Yu, Yongqi Leng, Yufei Huang, Shang Wu, Haixin Liu, Xinmeng Ji, Jiahui Zhao, Jinwang Song, Tingting Cui, Xiaoqing Cheng, Liutao Liutao, and Deyi Xiong. 2024. [CMoralEval: A moral evaluation benchmark for chinese large language models](#). In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 11817–11837.
- Xiang Zhang, Senyu Li, Bradley Hauer, Ning Shi, and Grzegorz Kondrak. 2023. [Don’t trust ChatGPT when your question is not in english: A study of multilingual abilities and types of LLMs](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 7915–7927.
- Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xu Wang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. 2023a. [Representation engineering: A top-down approach to AI transparency](#). *CoRR*, abs/2310.01405.
- Andy Zou, Zifan Wang, J. Zico Kolter, and Matt Fredrikson. 2023b. [Universal and transferable adversarial attacks on aligned language models](#). *CoRR*, abs/2307.15043.