Wenxiang Jiao, Wenxuan Wang, Jen tse Huang, Xing Wang, Shuming Shi, and Zhaopeng Tu. Is chatgpt a good translator? yes with gpt-4 as the engine, 2023. URL https://arxiv.org/abs/2301.08745.

Zdeněk Kasner and Ondřej Dušek. Beyond traditional benchmarks: Analyzing behaviors of open llms on data-to-text generation, 2024. URL https://arxiv.org/abs/2401.10186.

Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang. Benchmarking cognitive biases in large language models as evaluators, 2023. URL https://arxiv.org/abs/2309.17012.

Khiem Le, Zhichun Guo, Kaiwen Dong, Xiaobao Huang, Bozhao Nan, Roshni Iyer, Xiangliang Zhang, Olaf Wiest, Wei Wang, and Nitesh V. Chawla. Molx: Enhancing large language models for molecular learning with a multi-modal extension, 2024. URL https://arxiv.org/abs/2406.06777.

Y. Leo. Emerton-dpo-pairs-judge. https://huggingface.co/datasets/yleo/emerton_dpo_pairs_judge/viewer, 2024. Accessed: 2024-07-15.

Alice Li and Luanne Sinnamon. Examining query sentiment bias effects on search results in large language models. In *The Symposium on Future Directions in Information Access (FDIA) co-located with the 2023 European Summer School on Information Retrieval (ESSIR)*, 2023.

Ruosen Li, Teerth Patel, and Xinya Du. Prd: Peer rank and discussion improve large language model based evaluations. *arXiv preprint arXiv:2307.02762*, 2023a.

Yuan Li, Yue Huang, Yuli Lin, Siyuan Wu, Yao Wan, and Lichao Sun. I think, therefore i am: Benchmarking awareness of large language models using awarebench, 2024a. URL https://arxiv.org/abs/2401.17882.

Yuan Li, Yue Huang, Hongyi Wang, Xiangliang Zhang, James Zou, and Lichao Sun. Quantifying ai psychology: A psychometrics benchmark for large language models. *arXiv preprint arXiv:2406.17675*, 2024b.

Zongjie Li, Chaozheng Wang, Pingchuan Ma, Daoyuan Wu, Shuai Wang, Cuiyun Gao, and Yang Liu. Split and merge: Aligning position biases in large language model based evaluators, 2023b. URL https://arxiv.org/abs/2310.01432.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods, 2022. URL https://arxiv.org/abs/2109.07958.

Xiao Liu, Xuanyu Lei, Shengyuan Wang, Yue Huang, Zhuoer Feng, Bosi Wen, Jiale Cheng, Pei Ke, Yifan Xu, Weng Lam Tam, et al. Alignbench: Benchmarking chinese alignment of large language models. *arXiv preprint arXiv:2311.18743*, 2023a.

Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo, Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. Trustworthy llms: a survey and guideline for evaluating large language models' alignment, 2024. URL https://arxiv.org/abs/2308.05374.

Zhengliang Liu, Yue Huang, Xiaowei Yu, Lu Zhang, Zihao Wu, Chao Cao, Haixing Dai, Lin Zhao, Yiwei Li, Peng Shu, et al. Deid-gpt: Zero-shot medical text de-identification by gpt-4. *arXiv preprint arXiv:2303.11032*, 2023b.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering, 2022. URL https://arxiv.org/abs/2209.09513.

John Macnicol. *Age Discrimination: An Historical and Contemporary Analysis*. 01 2006. ISBN 9780521847773. doi: 10.1017/CBO9780511550560.

Yu Meng, Mengzhou Xia, and Danqi Chen. Simpo: Simple preference optimization with a reference-free reward. *arXiv preprint arXiv:2405.14734*, 2024.

OpenAI. Gpt-4 technical report, 2024a. URL https://arxiv.org/abs/2303.08774.

OpenAI. Gpt-3.5-turbo model documentation, 2024b. URL https://platform.openai.com/docs/models.

OpenAI. Hello gpt-4o, 2024c. URL https://openai.com/index/hello-gpt-4o/.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.

Jiawen Shi, Zenghui Yuan, Yinuo Liu, Yue Huang, Pan Zhou, Lichao Sun, and Neil Zhenqiang Gong. Optimization-based prompt injection attack to llm-as-a-judge. *arXiv preprint arXiv:2403.17710*, 2024a.

Lin Shi, Weicheng Ma, and Soroush Vosoughi. Judging the judges: A systematic investigation of position bias in pairwise comparative assessments by llms, 2024b. URL https://arxiv.org/abs/2406.07791.

Paul Slovic, Melissa Finucane, Ellen Peters, and Donald G. MacGregor. *The Affect Heuristic*, pp. 397–420. Cambridge University Press, 2002.

Rickard Stureborg, Dimitris Alikaniotis, and Yoshi Suhara. Large language models are inconsistent and biased evaluators, 2024. URL https://arxiv.org/abs/2405.01724.

Lichao Sun, Yue Huang, Haoran Wang, Siyuan Wu, Qihui Zhang, Yuan Li, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, Xiner Li, Zhengliang Liu, Yixin Liu, Yijue Wang, Zhikun Zhang, Bertie Vidgen, Bhavya Kailkhura, Caiming Xiong, Chaowei Xiao, Chunyuan Li, Eric Xing, Furong Huang, Hao Liu, Heng Ji, Hongyi Wang, Huan Zhang, Huaxiu Yao, Manolis Kellis, Marinka Zitnik, Meng Jiang, Mohit Bansal, James Zou, Jian Pei, Jian Liu, Jianfeng Gao, Jiawei Han, Jieyu Zhao, Jiliang Tang, Jindong Wang, Joaquin Vanschoren, John Mitchell, Kai Shu, Kaidi Xu, Kai-Wei Chang, Lifang He, Lifu Huang, Michael Backes, Neil Zhenqiang Gong, Philip S. Yu, Pin-Yu Chen, Quanquan Gu, Ran Xu, Rex Ying, Shuiwang Ji, Suman Jana, Tianlong Chen, Tianming Liu, Tianyi Zhou, William Wang, Xiang Li, Xiangliang Zhang, Xiao Wang, Xing Xie, Xun Chen, Xuyu Wang, Yan Liu, Yanfang Ye, Yinzhi Cao, Yong Chen, and Yue Zhao. Trustllm: Trustworthiness in large language models, 2024. URL https://arxiv.org/abs/2401.05561.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. Commonsenseqa: A question answering challenge targeting commonsense knowledge, 2019. URL https://arxiv.org/abs/1811.00937.

Toughdata. Quora question answer dataset. https://huggingface.co/datasets/toughdata/quora-question-answer-dataset, 2023.

Yixin Wan, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang, and Nanyun Peng. "kelly is a warm person, joseph is a role model": Gender biases in llm-generated reference letters, 2023. URL https://arxiv.org/abs/2310.09219.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. Is chatgpt a good nlg evaluator? a preliminary study, 2023a. URL https://arxiv.org/abs/2303.04048.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators, 2023b. URL https://arxiv.org/abs/2305.17926.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023. URL https://arxiv.org/abs/2201.11903.

Minghao Wu and Alham Fikri Aji. Style over substance: Evaluation biases for large language models, 2023. URL https://arxiv.org/abs/2307.03025.

Siyuan Wu, Yue Huang, Chujie Gao, Dongping Chen, Qihui Zhang, Yao Wan, Tianyi Zhou, Xiangliang Zhang, Jianfeng Gao, Chaowei Xiao, et al. Unigen: A unified framework for textual dataset generation using large language models. *arXiv preprint arXiv:2406.18966*, 2024a.

Yuanwei Wu, Yue Huang, Yixin Liu, Xiang Li, Pan Zhou, and Lichao Sun. Can large language models automatically jailbreak gpt-4v? *arXiv preprint arXiv:2407.16686*, 2024b.

xDAN. xdan-sft-dpo-roleplay-nsfw-with-lf. https://huggingface.co/datasets/xDAN2099/xDAN-SFT-DPO-Roleplay-NSFW-with-lf, 2024. Accessed: 2024-07-15.

Wenda Xu, Guanglei Zhu, Xuandong Zhao, Liangming Pan, Lei Li, and William Wang. Pride and prejudice: LLM amplifies self-bias in self-refinement. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15474–15492, Bangkok, Thailand, August 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.acl-long.826.

Xinghua Zhang, Bowen Yu, Haiyang Yu, Yangyu Lv, Tingwen Liu, Fei Huang, Hongbo Xu, and Yongbin Li. Wider and deeper llm networks are fairer llm evaluators, 2023. URL https://arxiv.org/abs/2308.01862.

Chenyang Zhao, Xueying Jia, Vijay Viswanathan, Tongshuang Wu, and Graham Neubig. Self-guide: Better task-specific instruction following via self-synthetic finetuning. *arXiv preprint arXiv:2407.12874*, 2024.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Lmsys chat platform. https://chat.lmsys.org/, 2023.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36, 2024.

Terry Yue Zhuo, Yujin Huang, Chunyang Chen, and Zhenchang Xing. Exploring ai ethics of chatgpt: A diagnostic analysis. *ArXiv*, abs/2301.12867, 2023. URL https://api.semanticscholar.org/CorpusID:256390238.

A Related Works

A.1 LLM-as-a-Judge

Recent studies have demonstrated that LLMs can serve as high-quality evaluators for various NLP tasks (Li et al., 2023a; Kasner & Dušek, 2024; Huang et al., 2024a; Wang et al., 2023a), and Zheng et al. (2024) proposed the concept of LLM-as-a-Judge. As an evaluation method that does not require reference texts, it has demonstrated performance on open-ended questions that highly match human preference. Recent research has focused on exploring its fairness, for instance, Shi et al. (2024a) introduced JudgeDeceiver, emphasizing the vulnerabilities in the evaluation process. Zhang et al. (2023) conducted research indicates that wider and deeper LLM networks often provide more fair evaluations. Liu et al. (2023a) proposed ALIGNBENCH for the multi-dimensional evaluation of LLMs' fairness.

A.2 Fairness in Trustworthy LLMs

Ensuring the trustworthiness of LLMs is of great significance Liu et al. (2024); Shi et al. (2024a); Huang et al. (2024b); Gao et al. (2024); Wu et al. (2024b). In recent research, it has been discovered that LLMs may exhibit stereotypes against certain groups or make erroneous judgments based on specific statistical patterns (Zhuo et al., 2023; Ferrara, 2023; Liu et al., 2024), which highlights the