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*, 36:53728–53741, 2023.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose Opinions Do Language Models Reflect? In *International Conference on Machine Learning*, pages 29971–30004. PMLR, 2023.

Farhana Shahid, Stella Zhang, and Aditya Vashistha. Llms homogenize values in constructive arguments on value-laden topics. *arXiv preprint arXiv:2509.10637*, 2025.

Chantal Shaib, Joe Barrow, Jiuding Sun, Alexa F Siu, Byron C Wallace, and Ani Nenkova. Standardizing the measurement of text diversity: A tool and a comparative analysis of scores. *arXiv preprint arXiv:2403.00553*, 2024.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Nikhil Sharma, Q Vera Liao, and Ziang Xiao. Generative echo chamber? effect of llm-powered search systems on diverse information seeking. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–17, 2024.

Judy Hanwen Shen, Archit Sharma, and Jun Qin. Towards Data-Centric RLHF: Simple Metrics for Preference Dataset Comparison. *arXiv preprint arXiv:2409.09603*, 2024.

Lin Shi, Chiyu Ma, Wenhua Liang, Xingjian Diao, Weicheng Ma, and Soroush Vosoughi. Judging the judges: A systematic study of position bias in LLM-as-a-judge. *arXiv preprint arXiv:2406.07791*, 2024.

Alexander Shypula, Shuo Li, Botong Zhang, Vishakh Padmakumar, Kayo Yin, and Osbert Bastani. Evaluating the Diversity and Quality of LLM Generated Content. In *ICLR Workshop on Deep Learning for Code*, 2025.

Stewart Slocum, Asher Parker-Sartori, and Dylan Hadfield-Menell. Diverse Preference Learning for Capabilities and Alignment. In *The Thirteenth International Conference on Learning Representations*, 2025.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, et al. A Roadmap to Pluralistic Alignment. In *Proceedings of the 41st International Conference on Machine Learning*, pages 46280–46302, 2024.

Peiqi Sui, Eamon Duede, Hoyt Long, and Richard Jean So. Critical confabulation: Can llms hallucinate for social good? *arXiv preprint arXiv:2511.07722*, 2025.

Alex Tamkin, Miles McCain, Kunal Handa, Esin Durmus, Liane Lovitt, Ankur Rathi, Saffron Huang, Alfred Mountfield, Jerry Hong, Stuart Ritchie, et al. Clio: Privacy-preserving insights into real-world ai use. *arXiv preprint arXiv:2412.13678*, 2024.

Yufei Tian, Abhilasha Ravichander, Lianhui Qin, Ronan Le Bras, Raja Marjieh, Nanyun Peng, Yejin Choi, Thomas L Griffiths, and Faeze Brahman. MacGyver: Are Large Language Models Creative Problem Solvers? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5303–5324, 2024.

Angelina Wang, Jamie Morgenstern, and John P Dickerson. Large Language Models That Replace Human Participants Can Harmfully Misportray and Flatten Identity Groups. *Nature Machine Intelligence*, pages 1–12, 2025a.

Qihan Wang, Shidong Pan, Tal Linzen, and Emily Black. Multilingual Prompting for Improving LLM Generation Diversity. *arXiv preprint arXiv:2505.15229*, 2025b.

Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. Measuring short-form factuality in large language models. *arXiv preprint arXiv:2411.04368*, 2024.

Tianjun Wei, Wei Wen, Ruizhi Qiao, Xing Sun, and Jianghong Ma. RocketEval: Efficient Automated LLM Evaluation via Grading Checklist. In *The Thirteenth International Conference on Learning Representations*, 2025.

Emily Wenger and Yoed Kenett. We're Different, We're the Same: Creative Homogeneity Across LLMs. *arXiv preprint arXiv:2501.19361*, 2025.

Dustin Wright, Sarah Masud, Jared Moore, Srishti Yadav, Maria Antoniak, Chan Young Park, and Isabelle Augenstein. Epistemic diversity and knowledge collapse in large language models. *arXiv preprint arXiv:2510.04226*, 2025.

Fan Wu, Emily Black, and Varun Chandrasekaran. Generative Monoculture in Large Language Models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Chenyang Yang, Yike Shi, Qianou Ma, Michael Xieyang Liu, Christian Kästner, and Tongshuang Wu. What prompts don't say: Understanding and managing underspecification in llm prompts. *arXiv preprint arXiv:2505.13360*, 2025.

Binwei Yao, Zefan Cai, Yun-Shiuan Chuang, Shanglin Yang, Ming Jiang, Diyi Yang, and Junjie Hu. No Preference Left Behind: Group Distributional Preference Optimization. In *The Thirteenth International Conference on Learning Representations*, 2025.

Lily Hong Zhang, Smitha Milli, Karen Jusko, Jonathan Smith, Brandon Amos, Wassim, Bouaziz, Manon Revel, Jack Kussman, Lisa Titus, Bhaktipriya Radharapu, Jane Yu, Vidya Sarma, Kris Rose, and Maximilian Nickel. Cultivating Pluralism In Algorithmic Monoculture: The Community Alignment Dataset. *arXiv preprint arXiv: 2507.09650*, 2025a.

Yiming Zhang, Avi Schwarzschild, Nicholas Carlini, J Zico Kolter, and Daphne Ippolito. Forcing diffuse distributions out of language models. In *First Conference on Language Modeling*, 2024. https://openreview.net/forum?id=9JY1QLVFPZ.

Yiming Zhang, Harshita Diddee, Susan Holm, Hanchen Liu, Xinyue Liu, Vinay Samuel, Barry Wang, and Daphne Ippolito. NoveltyBench: Evaluating Language Models for Humanlike Diversity. In *The Conference on Language Modeling (COLM)*, 2025b.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. WildChat: 1M ChatGPT Interaction Logs in the Wild. In *The Twelfth International Conference on Learning Representations*, 2024.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-Tuning Language Models from Human Preferences. *arXiv preprint arXiv:1909.08593*, 2019.

Appendix

Appendix A includes the following supplementary material about our experiment details.

Appendix B includes the following supplementary tables and figures about our experiment results.

- Figure 5, 6, 7, 8: Functional Diversity for Claude-4-Sonnet, Gemini-2.5-Flash, Llama-3.1-8B-Instruct, and Mistral-7B-Instruct-v0.3 (c.f. Figure 2 for GPT-4o)
- Figure 9, 10, 11, 12: Claude-4-Sonnet, Gemini-2.5-Flash, Llama-3.1-8B-Instruct, and Mistral-7B-Instruct-v0.3 (c.f. Figure 3 for GPT-4o)
- Figure 13: Diversity-Quality Tradeoff Using Embeddings
- Figure 14: Diversity-Quality Tradeoff for Varying Number of Generated Responses (5-10)
- Figures 15-16: Functional Diversity for Llama-3.1-8B-Instruct with DPO and GRPO
- Figure 17: Diversity-Quality Tradeoff for Llama-3.1-8B-Instruct with DPO and GRPO
- Figures 18-19: Functional Diversity for Llama-3.1-8B-Instruct with DARLING alignment
- Figure 20: Diversity-Quality Tradeoff for Llama-3.1-8B-Instruct w/DARLING alignment
- Table 14-15: Functional Diversity by Model, Sampling Strategy & Task Category
- Table 16-17: Vocabulary Diversity by Model, Sampling Strategy & Task Category
- Table 18-19: Embedding Diversity by Model, Sampling Strategy & Task Category
- Table 22-23: Checklist-Based Quality by Model, Sampling Strategy & Task Category
- Table 24-25: Athene-RM-8B Reward by Model, Sampling Strategy & Task Category
- Table 26-27: Accuracy by Model, Sampling Strategy & Task Category (for verifiable tasks)
- Table 28, 29, 30: Functional Diversity As Calculated by Single LLM-Judges
- Table 31, 32, 33: Checklist-Based Quality As Calculated by Single LLM-Judges
- Table 34-35: Functional Diversity & Checklist-Based Quality for 10 Generated Responses
- Table 36: Functional Diversity for Llama-3.1-8B-Instruct with DPO and GRPO
- Table 37 Functional Diversity for Llama-3.1-8B-Instruct with DARLING alignment