5 Discussion

Our work underlines the task-dependent nature of evaluating and mitigating output homogenization. We find that our task-anchored sampling technique outperforms more general sampling approaches in terms of increasing response diversity only when desired. Our results show that without task-dependence, previous methods to reduce output homogenization often (1) misconceptualize output diversity (2) reduce homogenization in tasks where homogenization should be preserved and (3) maintain homogenization in tasks where more pluralism is desired. Further, our results show that task-anchored sampling does not result in a significant diversity-quality trade-off under task-based metrics. These results challenge the common assumption in the literature of a diversity-quality tradeoff. In this section, we discuss the implications of our work and avenues for future research.

5.1 Our framework improves homogenization evaluation

We have developed a taxonomy of task categories that clarifies how a model can conceptualize diversity based on the categorization. For example, evaluating homogenization in math problem-solving should measure variety in solution strategies, whereas evaluating homogenization in advice or opinions should measure variety in viewpoints or perspectives. We improve upon previous studies that rely on generic measures of diversity (vocabulary or embedding differences), which is particularly meaningful when evaluating diversity loss in alignment and diversity-promoting methods. Our findings suggest that using general metrics without accounting for task dependence does not capture meaningful functional diversity and may falsely show a diversity-quality tradeoff. Future research may further explore how this applies to alignment. While previous studies show that token entropy collapses during alignment (Lanchantin et al., 2025b), our preliminary experiments show that functional diversity does not necessarily collapse (Appendix Figures 15-16).

We highlight the importance of evaluating prompts across our taxonomy when analyzing output homogenization. When studies limit their evaluation to tasks where diversity is desired, there may be unintended effects (e.g. confabulations) when those methods are applied to tasks which rely on homogenization being preserved. Hence, not adopting a task-dependent approach could result in less robust evaluation and present safety or ethical concerns downstream. Our taxonomy is one example of a categorization that anchors task dependence. An important limitation is that our taxonomy is English-centric; we only define functional diversity concepts in English, and we only evaluate English prompts. Future work may adapt or expand our taxonomy and evaluation approach. To modify the taxonomy and run new evaluations, one simply needs to edit or add new task-anchored prompts for classification, sampling, and evaluation.

Our evaluation relies on LLM-judges to measure task-based diversity and quality, which has known limitations (Shi et al., 2024; Li et al., 2025a). Future work could include a user study to confirm that task-based functional diversity aligns with human judgments of what constitutes a meaningful difference between responses. Similarly, future work may further explore how task-dependent quality metrics align with human preferences (Wei et al., 2025; Lin et al., 2025).

5.2 Our framework improves homogenization mitigation

There are many ways to apply task-dependence in mitigating homogenization and our approach could be applied at inference-time automatically. For instance, when given a prompt, the model could be instructed to determine its task categorization and output responses according to the task-based conceptualization of output homogenization. Our main improvement is in clarifying model instructions for output homogenization behavior in terms of the task category. Instead of assuming the model does this inherently, it may be important to clarify and steer expected behavior.

Although we focus on prompt-based strategies, our task-anchored approach may be applied to other diversity-promoting methods that modify the alignment process. For example, Lanchantin et al. (2025a) propose a method for improving diversity through preference pair construction (x, y^+, y^-) in DPO. This approach could be modified to construct pairs in a task-informed way that avoids learning undesired semantic preferences that might reduce functional diversity. Slocum et al. (2025) also propose modifying the RLHF or DPO optimization objective to include a penalty for lower token-level entropy. This penalty could be selectively

applied to certain task categories where vocabulary diversity is desired, such as random generation and creative writing. Furthermore, Li et al. (2025b) propose Diversity Aware Reinforcement Learning (DARLING) to jointly optimize for response quality and semantic diversity. While they use a general semantic diversity classifier, this approach could be modified to use task-dependent functional diversity. We evaluate and explore modifying DARLING using our task-dependent framework in Appendix A.7.

Future work may further explore how to embed task-anchored homogenization considerations directly into a model's learning or reasoning process. Our task-anchored sampling strategies could be incorporated into a chain-of-thought instruction, with models first reasoning about task-appropriate functional diversity. A reasoning model could also be trained to directly reason about the functional diversity requirements for a given task before generating a response. Future work in this direction could be quite impactful in terms of preventing problematic occurrences. With task-dependent reasoning about functional diversity, the model may avoid undesirable behavior such as confabulations or increasing diversity when it is culturally or socially inappropriate to do so. Ultimately, we offer a simple but important improvement to the field's conceptualization of output homogenization by grounding it in task dependence.

References

Dhruv Agarwal, Mor Naaman, and Aditya Vashistha. Ai suggestions homogenize writing toward western styles and diminish cultural nuances. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–21, 2025.

Barrett R Anderson, Jash Hemant Shah, and Max Kreminski. Homogenization Effects of Large Language Models on Human Creative Ideation. In *Proceedings of the 16th conference on creativity & cognition*, pages 413–425, 2024.

Rishi Bommasani, Kathleen A. Creel, Ananya Kumar, Dan Jurafsky, and Percy S Liang. Picking on the same person: Does algorithmic monoculture lead to outcome homogenization? In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 3663–3678. Curran Associates, Inc., 2022. https://proceedings.neurips.cc/paper_files/paper/2022/file/17a234c91f746d9625a75cf8a8731ee2-Paper-Conference.pdf.

Aaron Chatterji, Thomas Cunningham, David J Deming, Zoe Hitzig, Christopher Ong, Carl Yan Shan, and Kevin Wadman. How people use chatgpt. Technical report, National Bureau of Economic Research, 2025.

Daiwei Chen, Yi Chen, Aniket Rege, and Ramya Korlakai Vinayak. Pal: Pluralistic alignment framework for learning from heterogeneous preferences. *arXiv preprint arXiv:2406.08469*, 2024.

John Joon Young Chung, Vishakh Padmakumar, Melissa Roemmele, Yuqian Sun, and Max Kreminski. Modifying Large Language Model Post-Training for Diverse Creative Writing. *arXiv preprint arXiv:2503.17126*, 2025.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, et al. Ultrafeedback: Boosting language models with scaled ai feedback. In *Forty-first International Conference on Machine Learning*, 2023.

Harry Dong, David Brandfonbrener, Eryk Helenowski, Yun He, Mrinal Kumar, Han Fang, Yuejie Chi, and Karthik Abinav Sankararaman. Generalized Parallel Scaling with Interdependent Generations. *arXiv preprint arXiv:2510.01143*, 2025.

Anil R Doshi and Oliver P Hauser. Generative AI Enhances Individual Creativity but Reduces the Collective Diversity of Novel Content. *Science advances*, 10(28):eadn5290, 2024.

Esin Durmus, Karina Nguyen, Thomas I Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, et al. Towards Measuring the Representation of Subjective Global Opinions in Language Models. *arXiv preprint arXiv:2306.16388*, 2023.

Sina Fazelpour and Will Fleisher. The Value of Disagreement in AI Design, Evaluation, and Alignment. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, pages 2138–2150, 2025.

Evan Frick, Tianle Li, Connor Chen, Wei-Lin Chiang, Anastasios Nikolas Angelopoulos, Jiantao Jiao, Banghua Zhu, Joseph E Gonzalez, and Ion Stoica. How to evaluate reward models for rlhf. In *The Thirteenth International Conference on Learning Representations*, 2025.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring Mathematical Problem Solving with the MATH Dataset. *arXiv preprint arXiv:2103.03874*, 2021.

Aspen K Hopkins, Alex Renda, and Michael Carbin. Can LLMs generate random numbers? evaluating LLM sampling in controlled domains. In *ICML 2023 workshop: sampling and optimization in discrete space*, 2023.

Shomik Jain, Kathleen Creel, and Ashia Camage Wilson. Position: Scarce Resource Allocations That Rely On Machine Learning Should Be Randomized. In *Forty-first International Conference on Machine Learning*, 2024a. https://openreview.net/forum?id=44qxX6Ty6F.

Shomik Jain, Vinith Suriyakumar, Kathleen Creel, and Ashia Wilson. Algorithmic Pluralism: A Structural Approach to Equal Opportunity. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 197–206, 2024b.

Shomik Jain, Margaret Wang, Kathleen Creel, and Ashia Wilson. Allocation Multiplicity: Evaluating the Promises of the Rashomon Set. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, pages 2040–2055, 2025.

Elliot Kim, Avi Garg, Kenny Peng, and Nikhil Garg. Correlated Errors in Large Language Models. *arXiv preprint arXiv:2506.07962*, 2025.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. Understanding the effects of RLHF on LLM generalisation and diversity. In *The Twelfth International Conference on Learning Representations*, 2024. https://openreview.net/forum?id=PXD3FAVHJT.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, et al. Tulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*, 2024.

Jack Lanchantin, Angelica Chen, Shehzaad Dhuliawala, Ping Yu, Jason Weston, Sainbayar Sukhbaatar, and Ilia Kulikov. Diverse Preference Optimization. *arXiv preprint arXiv:2501.18101*, 2025a.

Jack Lanchantin, Angelica Chen, Janice Lan, Xian Li, Swarnadeep Saha, Tianlu Wang, Jing Xu, Ping Yu, Weizhe Yuan, Jason E Weston, et al. Bridging Offline and Online Reinforcement Learning for LLMs. *arXiv preprint arXiv:2506.21495*, 2025b.

Chang-Yu Lee and I-Wei Lai. Enhancing solution diversity in arithmetic problems using fine-tuned ai language model. In *2024 International Conference on Consumer Electronics-Taiwan (ICCE-Taiwan)*, pages 515–516. IEEE, 2024.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, et al. From generation to judgment: Opportunities and challenges of llm-as-a-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2757–2791, 2025a.

Tianjian Li, Yiming Zhang, Ping Yu, Swarnadeep Saha, Daniel Khashabi, Jason Weston, Jack Lanchantin, and Tianlu Wang. Jointly Reinforcing Diversity and Quality in Language Model Generations. *arXiv preprint arXiv:2509.02534*, 2025b.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's Verify Step By Step. In *The Twelfth International Conference on Learning Representations*, 2023.

Bill Yuchen Lin, Yuntian Deng, Khyathi Chandu, Abhilasha Ravichander, Valentina Pyatkin, Nouha Dziri, Ronan Le Bras, and Yejin Choi. WildBench: Benchmarking LLMs with Challenging Tasks from Real Users in the Wild. In *The Thirteenth International Conference on Learning Representations*, 2025.

Kevin Ma, Daniele Grandi, Christopher McComb, and Kosa Goucher-Lambert. Exploring the capabilities of large language models for generating diverse design solutions. *arXiv preprint arXiv:2405.02345*, 2024.

Kibum Moon. Homogenizing Effect of Large Language Model on Creativity: An Empirical Comparison of Human and ChatGPT Writing, 2024.

Kibum Moon, Kostadin Kushlev, Andrew Bank, and Adam Green. Impersonal statements: Llm-era college admissions essays exhibit deep homogenization despite lexical diversity, 2025.

Vishakh Padmakumar and He He. Does Writing with Language Models Reduce Content Diversity? In *12th International Conference on Learning Representations, ICLR 2024*, 2024.