References

1. Baldassini, F.B., Shukor, M., Cord, M., Soulier, L., Piwowarski, B.: What makes multimodal in-context learning work? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1539–1550 (2024)
2. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in Neural Information Processing Systems **33**, 1877–1901 (2020)
3. Cao, B., Lin, H., Han, X., Sun, L., Yan, L., Liao, M., Xue, T., Xu, J.: Knowledgeable or educated guess? revisiting language models as knowledge bases. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 1860–1874 (2021)
4. Ferber, D., Wölflein, G., Wiest, I.C., Ligero, M., Sainath, S., Ghaffari Laleh, N., El Nahhas, O.S., Müller-Franzes, G., Jäger, D., Truhn, D., et al.: In-context learning enables multimodal large language models to classify cancer pathology images. Nature Communications **15**(1), 10104 (2024)
5. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: International Conference on Machine Learning. pp. 1321–1330. PMLR (2017)
6. Han, Z., Hao, Y., Dong, L., Sun, Y., Wei, F.: Prototypical calibration for few-shot learning of language models. In: The Eleventh International Conference on Learning Representations (2023)
7. He, K., Long, Y., Roy, K.: Prompt-based bias calibration for better zero/few-shot learning of language models. In: Findings of the Association for Computational Linguistics: EMNLP 2024. pp. 12673–12691. Association for Computational Linguistics (2024)
8. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
9. Jin, R., Xu, Z., Zhong, Y., Yao, Q., Dou, Q., Zhou, S.K., Li, X.: FairmedFM: Fairness benchmarking for medical imaging foundation models. In: The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2024)
10. Johnson, A.E., Pollard, T.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Peng, Y., Lu, Z., Mark, R.G., Berkowitz, S.J., Horng, S.: Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. arXiv preprint arXiv:1901.07042 (2019)
11. Kovalyk, O., Morales-Sánchez, J., Verdú-Monedero, R., Sellés-Navarro, I., Palazón-Cabanes, A., Sancho-Gómez, J.L.: Papila: Dataset with fundus images and clinical data of both eyes of the same patient for glaucoma assessment. Scientific Data **9**(1), 291 (2022)
12. Kull, M., Perello Nieto, M., Kängsepp, M., Silva Filho, T., Song, H., Flach, P.: Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. Advances in Neural Information Processing Systems **32** (2019)
13. Lu, Y., Bartolo, M., Moore, A., Riedel, S., Stenetorp, P.: Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 8086–8098 (2022)

14. Luo, Y., Tian, Y., Shi, M., Pasquale, L.R., Shen, L.Q., Zebardast, N., Elze, T., Wang, M.: Harvard glaucoma fairness: a retinal nerve disease dataset for fairness learning and fair identity normalization. IEEE Transactions on Medical Imaging (2024)
15. Ma, H., Zhang, C., Bian, Y., Liu, L., Zhang, Z., Zhao, P., Zhang, S., Fu, H., Hu, Q., Wu, B.: Fairness-guided few-shot prompting for large language models. Advances in Neural Information Processing Systems **36** (2024)
16. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM Computing Surveys **54**(6), 1–35 (2021)
17. PhysioNet: Responsible use of mimic data with online services like GPT (2023), `https://physionet.org/news/post/gpt-responsible-use`
18. Shui, C., Szeto, J., Mehta, R., Arnold, D.L., Arbel, T.: Mitigating calibration bias without fixed attribute grouping for improved fairness in medical imaging analysis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 189–198. Springer (2023)
19. Team, G., Georgiev, P., Lei, V.I., Burnell, R., Bai, L., Gulati, A., Tanzer, G., Vincent, D., Pan, Z., Wang, S., et al.: Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv preprint arXiv:2403.05530 (2024)
20. Tian, Y., Shi, M., Luo, Y., Kouhana, A., Elze, T., Wang, M.: Fairseg: A large-scale medical image segmentation dataset for fairness learning using segment anything model with fair error-bound scaling. In: The Twelfth International Conference on Learning Representations (2024)
21. Tschandl, P., Rosendahl, C., Kittler, H.: The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific data **5**(1), 1–9 (2018)
22. Wang, X., Zhang, X., Cao, Y., Wang, W., Shen, C., Huang, T.: Seggpt: Towards segmenting everything in context. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1130–1140 (2023)
23. Wu, Z., Lin, X., Dai, Z., Hu, W., Shu, Y., Ng, S.K., Jaillet, P., Low, B.K.H.: Prompt optimization with EASE? efficient ordering-aware automated selection of exemplars. In: The Thirty-eighth Annual Conference on Neural Information Processing Systems (2024)
24. Xiong, M., Hu, Z., Lu, X., LI, Y., Fu, J., He, J., Hooi, B.: Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In: The Twelfth International Conference on Learning Representations (2024)
25. Xu, G., CHEN, Q., Ling, C., Wang, B., Shui, C.: Intersectional unfairness discovery. In: Forty-first International Conference on Machine Learning (2024)
26. Xu, Z., Peng, K., Ding, L., Tao, D., Lu, X.: Take care of your prompt bias! investigating and mitigating prompt bias in factual knowledge extraction. In: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). pp. 15552–15565 (2024)
27. Zhao, Z., Wallace, E., Feng, S., Klein, D., Singh, S.: Calibrate before use: Improving few-shot performance of language models. In: International Conference on Machine Learning. pp. 12697–12706. PMLR (2021)
28. Zong, Y., Yang, Y., Hospedales, T.: MEDFAIR: Benchmarking fairness for medical imaging. In: The Eleventh International Conference on Learning Representations (2023)