BiasICL: In-Context Learning and Demographic Biases of Vision Language Models

Sonnet Xu[*1⋆], Joseph D. Janizek[*1,2], Yixing Jiang[1], and Roxana Daneshjou[1]

[1] Stanford University, Stanford, CA
[2] Virginia Mason Internal Medicine Residency, Seattle, WA
[*]Equal Contribution

Abstract. Vision language models (VLMs) show promise in medical diagnosis, but their performance across demographic subgroups when using in-context learning (ICL) remains poorly understood. We examine how the demographic composition of demonstration examples affects VLM performance in two medical imaging tasks: skin lesion malignancy prediction and pneumothorax detection from chest radiographs. Our analysis reveals that ICL influences model predictions through multiple mechanisms: (1) ICL allows VLMs to learn subgroup-specific disease base rates from prompts and (2) ICL leads VLMs to make predictions that perform differently across demographic groups, even after controlling for subgroup-specific disease base rates. Our empirical results inform best-practices for prompting current VLMs (specifically examining demographic subgroup performance, and matching base rates of labels to target distribution at a bulk level and within subgroups), while also suggesting next steps for improving our theoretical understanding of these models. `https://github.com/DaneshjouLab/BiasICL`

Keywords: In-context learning · Fairness · Vision Language Models.

1 Introduction and background

In-context learning (ICL), or the capacity of large language models (LLMs) to adapt to new tasks from a handful of demonstrations in the prompt, has emerged as an exciting development in AI [4]. For medical AI tasks, this technique circumvents the large datasets required for supervised learning, which are costly due to privacy regulations and the clinical expertise required for data annotation. While traditional deep learning often demands tens or hundreds of thousands of labeled samples, ICL enables models to be rapidly customized to a new task using only a few examples. Although first demonstrated predominantly in text-based medical tasks [17], recent work has shown that ICL also improves the performance of vision-language models (VLMs) for tasks such as cancer pathology image classification [8], chest radiograph classification, and dermatology image classification [13].

⋆ Corresponding author: sonnet@stanford.edu

Despite its promise, ICL remains imperfectly understood and can lead to high variance in predictive accuracy based on seemingly minor modifications in prompts. In text-only settings, prior research demonstrates that LLMs may over-predict labels that appear more frequently in the prompt, appear last in the prompt, or are simply more common in their pre-training data [26].

Prior to deployment in medical settings, VLM safety and bias need to be better understood. Substantial work has been devoted to understanding whether machine learning models have problems with demographic biases or fairness issues. This includes work on previous generations of supervised deep learning vision models [20], text-based large language models [19], and now even vision-based large multi-modal VLMs [22,21].

Our work differs from prior bias and fairness work in the following way – we specifically aim to understand how prompting VLMs using ICL impacts demographic fairness, as opposed to prior work that investigates text models or investigates multi-modal models in the 0-shot setting. Consequently, we focus on large, commercial API-based VLMs, with the capacity to handle interleaved images and text with sufficiently large contexts to handle many demonstrations, in line with prior work from Jiang et al. [13].

This paper investigates how ICL with VLMs may inadvertently shift predictive distributions in ways that influence demographic fairness. First, we demonstrate that VLMs show a "majority label bias," similar to what has been observed in LLMs. Second, we find that VLMs exhibit a *demographic group* majority label bias, indicating sensitivity to base rates not just overall but also within specific demographic subgroups. This sensitivity seems to depend on how accurately the models can identify the demographic subgroups. Finally, we show that even after ensuring subgroup balance in prompts, ICL can increase the level of bias in predictions. Surprisingly, in this setting, ICL can lead to improvement in accuracy of prediction of one subgroup directly to the detriment of accuracy in another.

2 Methods

2.1 Datasets

The Diverse Dermatology Images (DDI) dataset contains clinical images of skin lesions with biopsy-proven benign or malignant labels and Fitzpatrick skin type (FST) labels [5]. Following Daneshjou et al. [5], we focus on FST I–II (light skin tones) and FST V–VI (dark skin tones), comprising 208 (159 benign, 49 malignant) and 207 (159 benign, 48 malignant) images, respectively (see Fig. 1a). We exclude FST III–IV images to highlight performance across more distinct skin tones. We split these into a 311-image demo set and a 104-image test set, ensuring equal representation of FST I–II and V–VI and maintaining a balanced $\tilde{2}5\%$ malignancy rate in each group.

CheXpert is a large dataset of 224,316 chest radiographs from 65,240 patients, with 14 labels automatically extracted from radiology reports [11]. Because we focus on demographic differences in prompt demonstrations, we limit our analysis

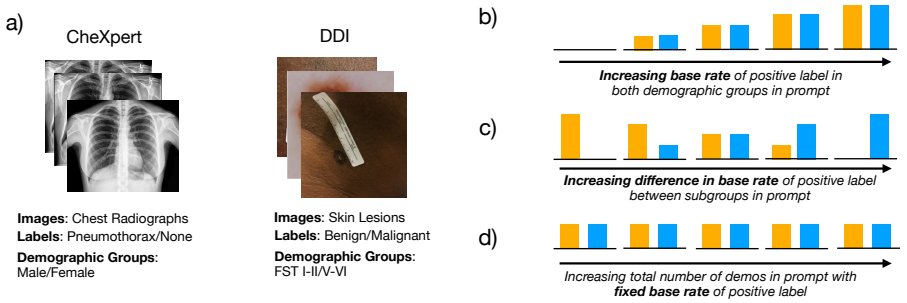

Fig. 1: **Overview**. CheXpert and DDI **(a)** were used to investigate a variety of different biases, including: **(b)** Majority label bias, or the tendency of models to predict more prevalent labels in the prompt more frequently; **(c)** a new bias introduced in our paper called group majority label bias, or the tendency of models to be swayed by the majority label seen using ICL *within a particular demographic subgroup* when encountering test examples from that same subgroup; and **(d)** ICL bias, or the extent to which models learn disparities between groups as the number of demos in a prompt increases. In (b-d), orange and blue bars represent different demographic groups, and the height of each bar represents the fraction of positive labels in the prompt within that subgroup.

to a small subset: 400 patients in a demo set and 100 in a test set, each split evenly by sex. In each demographic subgroup (in both demo and test sets), half the radiographs are labeled "pneumothorax," and half are not (see Fig. 1a).

2.2 Models

Our study focuses on API-based, commercial VLMs. We investigate three models from three different providers: GPT-4o (with the specific endpoint "gpt-4o-2024-05-13"), Gemini 1.5 Pro (with the specific endpoint "gemini-1.5-pro-preview-0409"), and Claude3.5-Sonnet (with the specific endpoint "claude-3-5-sonnet-20241022"). We use the API service provided by OpenAI for GPT-4o, the API service provided by Google Cloud on Vertex AI for Gemini 1.5 Pro, and the API service provided by Anthropic for Claude3.5-Sonnet. We set the temperature to zero for all models and a random seed for GPT-4o to obtain more deterministic responses. These models were selected because they all have large contexts, high accuracy across many multi-modal benchmark tasks, and most importantly, have the capacity in their context to handle many interleaved images and texts.

2.3 Prompting and evaluation

LLMs and VLMs can be sensitive to prompting and evaluation strategies [3], so we provide the exact prompts in our GitHub. Following Jiang et al. [13], our prompts include: (1) a preamble specifying response format; (2) demonstration

examples (image, question, possible answers, correct answer); and (3) test images/questions. We also use Batch Querying [13], a method explored in a variety of prior works that leads to more efficient and cheaper inference by batching multiple test questions together in a single prompt.

To evaluate models, we parse answers from text completions because many API-based models do not provide logprob access [15,14]. When formatting issues arose, we simply re-sent queries, as in Jiang et al. [13], rather than using an LLM fallback parser [15].

2.4 Bias measures

We considered three different types of bias when assessing in-context learning (ICL). The first, **majority label bias**, first observed by Zhao et al. [26], measures how sensitive models are to the frequency of positive labels in the demonstration set (see Fig. 1b). Specifically, we measure how the average binary classification prediction made by the model ($f(x)$) over all images in the test set, $\mathbb{E}_{(x)\sim\mathcal{D}_{\text{test}}}[f(x)]$, is impacted by varying the proportion of positive labels (y) in the demonstrations in the prompt, $\mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{demo}}}[y]$. We assess this bias since recent work has suggested that newer models may be more robust to this bias [10].

The second, which we termed **group majority label bias**, measures how sensitive models are to the frequency of positive labels within *each demographic group* in the demonstration set (see Fig. 1c). Specifically, we measure how the difference in the average prediction made by the model over all images in the test set conditional on subgroup membership ($\mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{test}}}[f(x) \mid g(x) = 1] - \mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{test}}}[f(x) \mid g(x) = 0]$) is impacted by varying the difference in the proportion of positive labels between demographic groups in the prompt ($\mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{demo}}}[y \mid g(x) = 1] - \mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{demo}}}[y \mid g(x) = 0]$). Here, $g(x)$ is an indicator function evaluating as 1 when sample x belongs to a particular demographic group (0 otherwise).

Finally, to isolate the effects of number of demonstrations in the prompt, we fix the base rate of positive labels in the prompt equal to the base rate in the test set, and fix the base rates equal across both demographic subgroups. Then we measure how the difference in the average prediction made by the model over all images in the test set conditional on subgroup membership ($\mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{test}}}[f(x) \mid g(x) = 1] - \mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{test}}}[f(x) \mid g(x) = 0]$) is impacted by increasing the number of demonstrations in the prompt. We refer to this as **ICL bias** (see Fig. 1d). We also examined the difference in the predictive performance (measured by F1 score) between demographic groups as the number of prompt demonstrations are increased.

3 Results

3.1 VLMs learn a majority label bias

When we prompted models with a constant number of demonstrations, but increased the frequency of demonstrating examples with positive labels, we see

that the models' outputs become biased towards that prediction (see Fig. 2). While this relationship appears to mostly be nearly linear, we observe two outlier points in the CheXpert dataset in the two cases when all demonstrating examples are positive. We also omitted Claude 3.5 Sonnet results on the CheXpert dataset in this experiment, as the model tended to abstain too frequently to obtain reliable results.

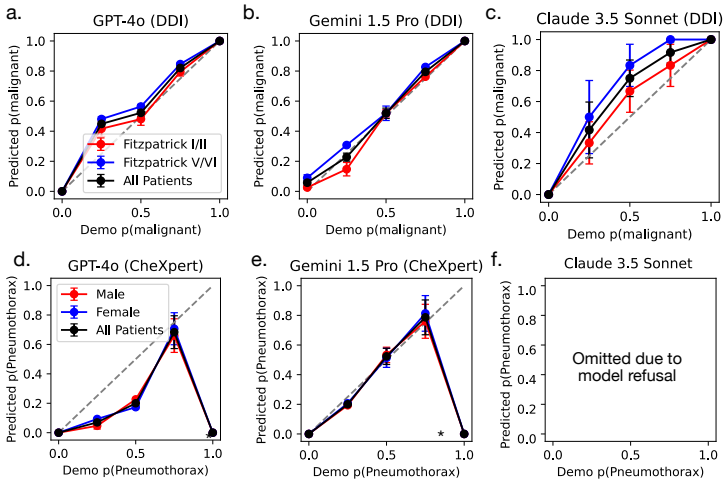

Fig. 2: Majority label bias – models more frequently predict labels that are more frequent in the prompt. (a-c) Prediction of malignancy on the DDI dataset. (d-f) Prediction of pneumothorax on the CheXpert dataset. Error bars = standard error over three independent runs with different random seeds for selection of demonstration examples from the dataset and ordering of demonstrations in the prompts.

3.2 VLMs learn a demographic group majority label bias

After demonstrating that VLMs are sensitive to the overall base rate of label frequency in their prompts, we wanted to investigate whether it is important to pay attention to the base rate of different labels *within* different subgroups. We refer to this property as a demographic group majority label bias. In Fig. 3, we hold the total number of samples in the prompt constant and increase the *difference* between the base rate of positive labels in the two demographic subgroups. Across most model-dataset pairs, we see that models do learn this bias from the prompt. We notice that this effect is more pronounced in the DDI dataset, where the maximum difference in subgroup mean prediction is 30% per model tested (Fig. 3a-c), compared to the CheXpert dataset, where the maximum difference is closer to 10% (Fig. 3e-g).

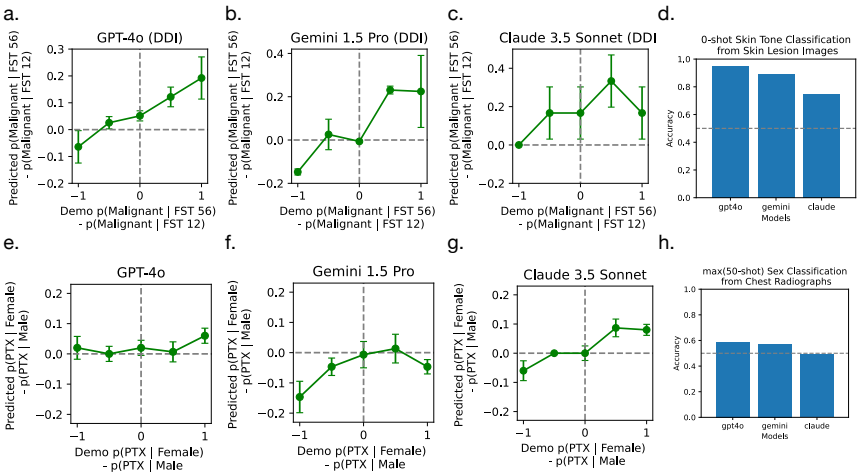

Fig. 3: Demographic group majority label bias. (a-c) Malignancy prediction on DDI dataset; (e-g) Pneumothorax prediction on CheXpert. Error bars = standard error over three independent runs with different random seeds for demonstration selection and prompt ordering. (d) 0-shot accuracy for patient Fitzpatrick skin type prediction from dermatology images; (h) maximum 0-to-50-shot accuracy for patient sex prediction from chest radiographs.

We hypothesized that this might be due to differences in the ability of models to detect and predict those demographic subgroups in the first place, particularly as there has been extensive study of demographic biases in *supervised* deep learning models for medical imaging, showing that these models can learn patient attributes such as race, sex, and age [23,9,16]. Across models tested, in the 0-shot setting, models can highly accurately classify patients' Fitzpatrick skin type (see Fig 3d). We also investigate the ability of models to identify demographic subgroups from chest radiographs, namely, patients' sex. Because predictive performance was low in the 0-shot setting, we increased the number of demonstrating examples in this experiment up to 50, and plotted the max accuracy per model over that range. We found that while models could predict sex with greater-than-random accuracy (see Fig 3h), the accuracy was generally much lower than for skin tone prediction.

Finally, given that previous work had demonstrated that supervised, deep convolutional neural network models could predict patients' self-reported race from chest radiographs with very high accuracy [9], we also probed the ability of VLMs to make this prediction. Despite substantial efforts to re-engineer prompts, all three VLMs invariably refused to attempt to make this classification.

These findings suggest that VLMs are less capable of identifying demographic subgroup identities than previous generations of supervised deep learning models, and also that as their capability in this area grows, their susceptibility to this

particular group majority label bias will increase as well. This is in keeping with prior work suggesting that supervised learning models may have a tendency to use demographic groups as "shortcuts" or proxies in their predictions [2,24,6,12].

3.3 ICL alone can increase demographic subgroup bias in VLMs

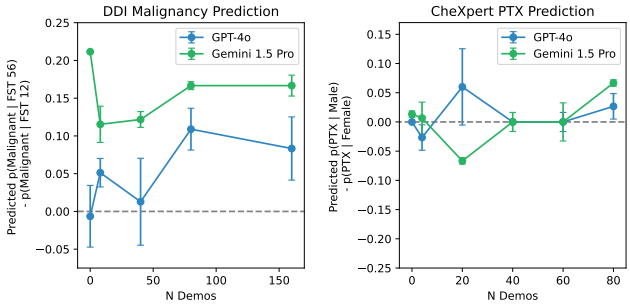

Fig. 4: The impact of ICL on the difference between models' average predictions across subgroups when the base rate of positive labels is set equal between subgroups in the prompt.

We initially hypothesized that adding demos from patients across different demographic subgroups to the prompt with ICL might be able to decrease models' inherent bias. This was because prior work on supervised learning models had shown that fine-tuning using more diverse data could improve subgroup performance [5]. However, our analysis showed that this was not necessarily the case.

For the task of malignant skin lesion prediction on the DDI dataset, GPT-4o demonstrated minimal bias in its predictions across demographic subgroups in the 0-shot setting. However, when provided with additional in-context examples, even those with balanced malignancy rates, the model developed a systematic bias toward predicting higher malignancy rates in patients with Fitzpatrick Skin Types V/VI (roughly 10%, see Fig. 4, left). In contrast, Gemini 1.5 Pro exhibited strong 0-shot bias toward predicting malignancy in FST V/VI patients (around 20%), and while this bias persisted with the addition of balanced in-context examples, it showed modest attenuation. The models' behavior differed substantially in the CheXpert pneumothorax prediction task (see Fig. 4, right). Both GPT-4o and Gemini 1.5 Pro maintained relatively unbiased predictions across gender, independent of number of demonstrating examples. Claude 3.5 Sonnet results were again excluded from this experiment as refusals were too numerous to have a reliable sample size.

In addition to considering bias in terms of the difference in average predictions between groups, we also looked at the *predictive performance* of models across

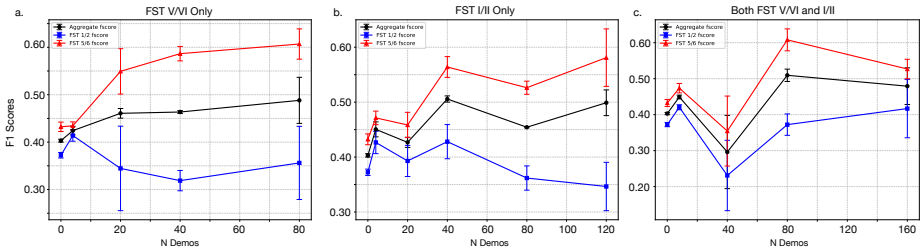

Fig. 5: The impact of ICL on GPT-4o's predictive performance on the DDI dataset when (a) adding only FST V/VI demos, (b) adding only FST I/II demos, and (c) adding equal numbers of both.

subgroups. Our most interesting results were on the DDI dataset with the GPT-4o model (see Fig. 5). In this experiment, the base rate of malignancy in the prompt was fixed equal to the base rate of malignancy in the test set, and the number of demonstrating examples were increased. Independently of whether the demonstrating examples in the prompt were (a) all of skin type FST V/VI, (b) all of skin type FST I/II, or (c) even numbers of both both, ICL significantly increased the predictive performance (as measured by F1 score) for FSV V/VI patients at the expense of predictive performance for FST I/II patients.

4 Discussion and limitations

This empirical work has several immediately-relevant implications for prompting. For medical vision systems that are task-adapted using ICL, developers must consider not only the overall base rate of the labels in the prompt, but also the demographic subgroup-specific base rate of the labels in the prompt. Additionally, even after carefully controlling the base rates of labels per demographic subgroups in prompts, ICL can still lead to exacerbation of differences in predictions across demographic subgroups. This highlights the ongoing importance of lessons learned from the supervised learning era – hidden stratification, or the tendency of models to have poor performance on important subsets of a population [18], is still a relevant property of VLMs "trained" with ICL. Consequently, developers should evaluate their models' performance stratified on different relevant subgroups.

Limitations of our current work also suggest future directions for research. While our work shows that ICL with large state-of-the-art API-based models can lead to exacerbated bias between subgroups, it does not uncover the mechanism of *why* this occurs. Future work with open source models [1], for which the weights and pre-training data can be directly accessed, will allow us to run ablation experiments to determine what aspects of a models training corpus, training process, and architecture primarily contribute to the observed phenomena. We also acknowledge the limitations of our work in evaluating the full social/societal impacts of the *fairness* of these models. We measure bias between subgroups as a

quantifiable and mathematical property of these models and datasets. However, a full analysis of the impact of these models would require further investigation of the real sociotechnical contexts in which they might be used [7]. Finally, various improvements to the basic ICL scheme, in particular regarding output calibration, have also been proposed [27,25,28]; further adaptations of these works as possible solutions to the biases observed here will likely represent useful future work.

Acknowledgments. This study was funded by UCB and Apple.

Disclosure of Interests. SX and JJ have no competing interests to declare that are relevant to the content of this article. RD has served as an advisor to MDAlgorithms, Pair, and Revea and received consulting fees from Pfizer, L'Oreal, Frazier Healthcare Partners, and DWA, and research funding from UCB and declares no non-financial competing interests.

References

1. Awadalla, A., Gao, I., Gardner, J., Hessel, J., Hanafy, Y., Zhu, W., Marathe, K., Bitton, Y., Gadre, S., Sagawa, S., et al.: Openflamingo: An open-source framework for training large autoregressive vision-language models. arXiv preprint arXiv:2308.01390 (2023)
2. Banerjee, I., Bhattacharjee, K., Burns, J.L., Trivedi, H., Purkayastha, S., Seyyed-Kalantari, L., Patel, B.N., Shiradkar, R., Gichoya, J.: "shortcuts" causing bias in radiology artificial intelligence: causes, evaluation and mitigation. Journal of the American College of Radiology (2023)
3. Biderman, S., Schoelkopf, H., Sutawika, L., Gao, L., Tow, J., Abbasi, B., Aji, A.F., Ammanamanchi, P.S., Black, S., Clive, J., et al.: Lessons from the trenches on reproducible evaluation of language models. arXiv preprint arXiv:2405.14782 (2024)
4. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners. CoRR **abs/2005.14165** (2020), `https://arxiv.org/abs/2005.14165`
5. Daneshjou, R., Vodrahalli, K., Novoa, R.A., Jenkins, M., Liang, W., Rotemberg, V., Ko, J., Swetter, S.M., Bailey, E.E., Gevaert, O., et al.: Disparities in dermatology ai performance on a diverse, curated clinical image set. Science advances **8**(31), eabq6147 (2022)
6. DeGrave, A.J., Janizek, J.D., Lee, S.I.: Ai for radiographic covid-19 detection selects shortcuts over signal. Nature Machine Intelligence **3**(7), 610–619 (2021)
7. Diao, J.A., He, Y., Khazanchi, R., Tiako, M.N., Witonsky, J.I., Pierson, E., Rajpurkar, P., Elhawary, J.R., Melas-Kyriazi, L., Yen, A., et al.: Implications of race adjustment in lung-function equations. The New England journal of medicine **390**(22), 2083 (2024)

8. 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)
9. Gichoya, J.W., Banerjee, I., Bhimireddy, A.R., Burns, J.L., Celi, L.A., Chen, L.C., Correa, R., Dullerud, N., Ghassemi, M., Huang, S.C., et al.: Ai recognition of patient race in medical imaging: a modelling study. The Lancet Digital Health **4**(6), e406–e414 (2022)
10. Gupta, K., Roychowdhury, S., Kasa, S.R., Kasa, S.K., Bhanushali, A., Pattisapu, N., Murthy, P.S.: How robust are llms to in-context majority label bias? arXiv preprint arXiv:2312.16549 (2023)
11. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., et al.: Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 590–597 (2019). https://doi.org/10.71718/y7pj-4v93
12. Janizek, J.D., Erion, G., DeGrave, A.J., Lee, S.I.: An adversarial approach for the robust classification of pneumonia from chest radiographs. In: Proceedings of the ACM conference on health, inference, and learning. pp. 69–79 (2020)
13. Jiang, Y., Irvin, J., Wang, J.H., Chaudhry, M.A., Chen, J.H., Ng, A.Y.: Many-shot in-context learning in multimodal foundation models. arXiv preprint arXiv:2405.09798 (2024)
14. Laurent, J.M., Janizek, J.D., Ruzo, M., Hinks, M.M., Hammerling, M.J., Narayanan, S., Ponnapati, M., White, A.D., Rodriques, S.G.: Lab-bench: Measuring capabilities of language models for biology research. arXiv preprint arXiv:2407.10362 (2024)
15. Mirza, A., Alampara, N., Kunchapu, S., Ríos-García, M., Emoekabu, B., Krishnan, A., Gupta, T., Schilling-Wilhelmi, M., Okereke, M., Aneesh, A., et al.: Are large language models superhuman chemists? arXiv preprint arXiv:2404.01475 (2024)
16. Munk, M.R., Kurmann, T., Marquez-Neila, P., Zinkernagel, M.S., Wolf, S., Sznitman, R.: Assessment of patient specific information in the wild on fundus photography and optical coherence tomography. Scientific reports **11**(1), 8621 (2021)
17. Nori, H., Lee, Y.T., Zhang, S., Carignan, D., Edgar, R., Fusi, N., King, N., Larson, J., Li, Y., Liu, W., et al.: Can generalist foundation models outcompete special-purpose tuning? case study in medicine. arXiv preprint arXiv:2311.16452 (2023)
18. Oakden-Rayner, L., Dunnmon, J., Carneiro, G., Ré, C.: Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In: Proceedings of the ACM conference on health, inference, and learning. pp. 151–159 (2020)
19. Omiye, J.A., Lester, J.C., Spichak, S., Rotemberg, V., Daneshjou, R.: Large language models propagate race-based medicine. NPJ Digital Medicine **6**(1), 195 (2023)
20. Seyyed-Kalantari, L., Liu, G., McDermott, M., Chen, I.Y., Ghassemi, M.: Chexclusion: Fairness gaps in deep chest x-ray classifiers. In: BIOCOMPUTING 2021: proceedings of the Pacific symposium. pp. 232–243. World Scientific (2020)
21. Wu, P., Liu, C., Chen, C., Li, J., Bercea, C.I., Arcucci, R.: Fmbench: Benchmarking fairness in multimodal large language models on medical tasks. arXiv preprint arXiv:2410.01089 (2024)
22. Yang, Y., Liu, Y., Liu, X., Gulhane, A., Mastrodicasa, D., Wu, W., Wang, E.J., Sahani, D.W., Patel, S.: Demographic bias of expert-level vision-language foundation models in medical imaging. arXiv preprint arXiv:2402.14815 (2024)

23. Yi, P.H., Wei, J., Kim, T.K., Shin, J., Sair, H.I., Hui, F.K., Hager, G.D., Lin, C.T.: Radiology "forensics": determination of age and sex from chest radiographs using deep learning. Emergency Radiology **28**, 949–954 (2021)
24. Zech, J.R., Badgeley, M.A., Liu, M., Costa, A.B., Titano, J.J., Oermann, E.K.: Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. PLoS medicine **15**(11), e1002683 (2018)
25. Zhang, H., Zhang, Y.F., Yu, Y., Madeka, D., Foster, D., Xing, E., Lakkaraju, H., Kakade, S.: A study on the calibration of in-context learning (2024), https://arxiv.org/abs/2312.04021
26. 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)
27. Zhou, H., Wan, X., Proleev, L., Mincu, D., Chen, J., Heller, K.A., Roy, S.: Batch calibration: Rethinking calibration for in-context learning and prompt engineering. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=L3FHMoKZcS
28. Zu, W., Xie, S., Zhao, Q., Li, G., Ma, L.: Embedded prompt tuning: Towards enhanced calibration of pretrained models for medical images. Medical Image Analysis **97**, 103258 (2024). https://doi.org/https://doi.org/10.1016/j.media.2024.103258, https://www.sciencedirect.com/science/article/pii/S136184152400183X