Exploring Visual Prompt Tuning for Demographic Adaptation in Foundation Models for Medical Imaging

Artur Parkhimchyk
EECS
York University
North York, ON M3J 1P3
arturp@yorku.ca

Amirreza Naziri
EECS
York University
North York, ON M3J 1P3
naziriam@yorku.ca

Laleh Seyyed-Kalantari
EECS
York University
North York, ON M3J 1P3
lsk@yorku.ca

Abstract

Pre-trained medical foundation models are large, and they require significant computational resources for training. Visual Prompt Tuning (VPT) allows foundation models to efficiently adapt to new tasks with minimal changes to the model's architecture, reducing the need for extensive fine-tuning. Here, we explore demographic (race) adaptation of foundation models (MAE and MoCoV3) for disease classification in medical imaging using naturally imbalanced data. We compare three adaptation strategies: linear probing, full fine-tuning, and VPT. We find that VPT obtains a clear boost in performance, starting with prompt length 5 over linear probing. In the case of race demographics (e.g. Asian with 5.7% of the full dataset), a VPT model trained on a demographic (Asian) performed similarly to a fully fine-tuned model trained on same dateset. A fully fine-tuned foundation model on a diverse and large dataset performs better than a model adapted only for a specific subset of data. However, it needs large data and computing resources, which may not always be available. These findings show that VPT can efficiently adapt foundation models for small datasets, achieving performance comparable to full fine-tuning.

1 Introduction

Foundation models are large pre-trained models using self-supervised learning that can be fine-tuned for various tasks. They are increasingly used in medical imaging Sellergren et al. [2022], but their re-training or fine-tuning demands significant computation time. The most common approach to adapting foundation models includes full fine-tuning and linear probing, with a recent addition of Visual Prompt Tuning (VPT) Jia et al. [2022]. VPT offers a more efficient approach by enabling quick adaptation to new tasks with minimal changes to the model's architecture, reducing the need for extensive fine-tuning and making the process computationally efficient.

Concurrently, there are numerous challenges facing AI in medical imaging, with one of the most novel being fairness. Fairness is the disparate outcome of AI models across different sub-populations, which often leads to the discrimination of AI models against vulnerable sub-populations Gichoya et al. [2023], Banerjee et al. [2023], Seyyed-Kalantari et al. [2021a]. In particular, in medical imaging, it has been shown that AI models can predict the demographic features of the patients Gichoya et. al. [2022], Abbasi Bavil et al. [2024] from medical images and under-diagnose historically undeserved patients (e.g female or Black) patients Seyyed-Kalantari et al. [2021b].

While the accuracy of models upon training on all datasets regardless of patient demographics is constantly improving, it has been shown that unfairness toward subgroups persists. This brings our attention to the point that we need a model trained on all patients regardless of demographics. Can we do a demographic adaption of foundation models, which are inherently trained on large amounts

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

of data, to perform better for individual demographics rather than all? Such adaptation may be challenging for minority demographics (e.g. Asian), and we may not have a large amount of data from them, which may lead to more unfairness. As a result, efficient adaption techniques such as VPT may resolve the need for extensive fine-tuning or retraining, making it computationally efficient, even for minority demographics. At the same time, using VPT can preserve the pre-trained weights and mitigate the risk of catastrophic forgetting of the foundation models if fine-tuned per subgroups.

In this work, we examine the three different foundation model adaptability strategies in the medical imaging domain. In particular, we focus on models trained on Chest X-rays, CheXpert Irvin et al. [2019] and MIMIC-CXR Johnson et al. [2019]. The first adaptation strategy is the widely used method of linear probing, which uses a single linear layer classification head. The second method is full fine-tuning, which updates the whole backbone along with a single-layer classification head. The last method is Visual Prompt Tuning (VPT), which appends a tuning prompt to the input of the model, which is trained alongside a single linear layer classification head. We utilize the mentioned methods to design a personalized demographics-tuned model for disease multi-label classification of different demographics instead of using a single model for all. Therefore, the models are expected to do better on personalized demographic models for individuals (biased model) rather than general models. However, we observed that a large, diverse dataset performed better than intentionally biased models.

2 Related Works

In Saeed et al. [2023], the adaptation of a foundation model trained on an original dataset, and validated on a dataset that is gathered in another medical center, has been evaluated for Head and Neck Cancer segmentation using VPT. They demonstrated that prompt tuning could adapt models trained on original data to data from new clinics in medical image segmentation tasks. The paper compares various tuning approaches, including no tuning, partial fine-tuning, full fine-tuning, and prompt tuning. Results show that while the no-tuning model performs well on the original dataset, it struggled with the data from a new clinic. A fully fine-tuned model on the new clinic dataset performed well on new data but poorly on the original dataset, meaning full fine-tuning suffers from catastrophic forgetting. However, prompt tuning did not exhibit this weakness as the backbone model trained on original data kept frozen, and only prompts are fine-tuned. VPT models retain good accuracy for the old dataset and achieve similar results to fully fine-tuned models on the new datasets.

There are some studies that have not directly used prompt tuning, but their approach has similar impact. Fairness triggers append a border or patch within the input image and optimizes their values in training to reach more overall fairness Zhang et al. [2022]. They have done this using ConvNets (ResNet-18) on CelebA dataset Liu et al. [2015]. FairVPT has been also introduced Park and Byun [2024] where they have proposed a variant of VPT with the goal of increasing fairness, and it was evaluated on natural images. FairVPT demonstrates higher generalization performance in terms of balanced accuracy and equality of odds. However, the overall accuracy of the model does decrease.

For foundation models comparison, there have been a number of papers comparing MAE He et al. [2021] and MoCoV3 Chen et al. [2021] on CheXpert and MIMIC datasets; the most notable works are Khan et al. [2023], Xiao et al. [2023], Gupta et al. [2024], Sowrirajan et al. [2021]. These works evaluate the foundation models using linear probing and full fine-tuning. Our work explores VPT on merged CheXpert and MIMIC-CXR datasets and examines the adaptation of demographics as downstream tasks.

3 Method

Datasets Our foundation model adaptation evaluation (training and inference) is performed on the CheXpert and MIMIC-CXR Chest X-ray datasets. CheXpert was generated at Stanford University Medical Center between October 2002 and July 2017, consisting of 223'648 Chest X-rays associated with 65,240 patients Irvin et al. [2019]. The MIMIC-CXR dataset contains 227,835 Chest C-ray imaging studies with 377,110 corresponding images obtained from Beth Israel Deaconess Medical Center between 2011 and 2016 Johnson et al. [2019]. The CheXpert and MIMIC-CXR datasets were combined to create a single Chest X-Ray dataset. The dataset was further analyzed based on race, three most sizable races remained (Asian, Black, and White), while others were grouped under other. The dataset split followed an 80% training, 10% validation, and 10% testing distribution, and stratified on Race. Table 2 shows the race breakdown, per dataset and combined.

SSL-Pretraining For foundation models, we chose the momentum contrastive learning model MoCoV3 Chen et al. [2021] and a masked autoencoder MAE He et al. [2021] to cover a variety of self-supervised learning methods. Both models were pre-trained without labels. MAE, was pre-trained using the approach described in Xiao et al. [2023], taking 4 RTX 6000 GPUs and 15 days. For MoCoV3, we performed a hyper-parameter search through different optimizers, learning rates, weight decays, and batch sizes. We used adamw optimizer, learning rate of 0.00005, weight decay of 0.1, and batch size 4096. As MoCoV3 does not retain a dictionary look up as proposed in original MoCo model Chen et al. [2021], we chose to go for a large batch size. The pretraining was done using 64 RTX 6000 GPUs in approximately 3 days. The table in the appendix Table 3 shows the results of the hyper-parameters search for pretraining and adaptation methods.

Downstream Tasks To adapt the pre-trained foundation models, we utilize three methods: linear probing, Deep VPT Jia et al. [2022], and full fine-tuning. Deep VPT refers to adding prompts at each layer. Onwards, we refer to Deep VPT as VPT. Linear probing and full fine-tuning were adapted using the entire dataset. Linear probing, VPT, and full fine-tuning were additionally used to adapt to demographic-specific data subsets. For each foundation model and adaptation method, we performed a wide hyper-parameter search on the Asian data subset and a more narrow search on the rest of the data subsets. Then, we chose the optimal learning rate and weight decay. The models with the same hyper-parameters across all demographics were used for evaluation.

For downstream task performance evaluation, we used AUC instead of accuracy due to disease label imbalance. All models performed multi-label classification on 14 available labels as part of both CheXpert and MIMIC-CXR. AUC was computed for each class and then averaged. The combined Asian, Black and White output was concatenated, the AUC was computed for each class and then averaged across all labels.

4 Results

4.1 Prompt Size analysis

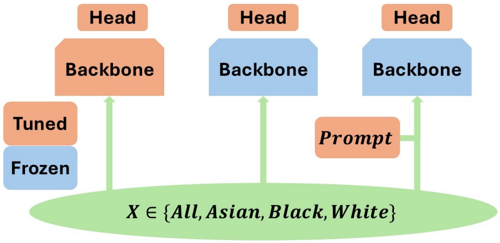

Figure 1: Visualization of adaptation strategies

(a) Full Fine-Tuning (Left)

(b) Linear Probing (Middle)

(c) VPT (Right)

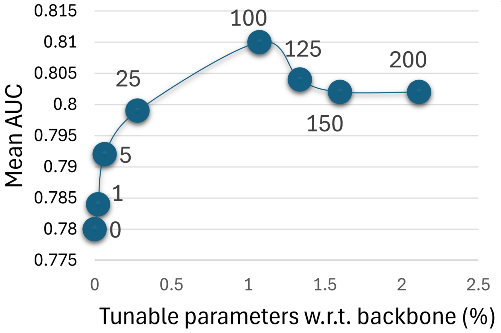

Figure 2: The mean AUC across all labels on Asian dataset subset by changing the prompt length

The main hyper-parameter for VPT is prompt length, therefore in this section we investigate the impact of prompt lengths in VPT on Asian subset of the Chest X-ray datasets. This was done to look at the prompt length impact on the most disadvantageous demographic as the optimal prompt length depends on the task Jia et al. [2022]. We also note that we start to observe an increase in performance with even a single prompt over a simple linear probing. In plot Figure 2 we see that the optimal length for this task is prompt length 100. This optimal prompt length is used for all VPT tasks in the next section.

Table 1: Summary of evaluation on foundation models using various adaptability methods. FT refers to a fine-tuned backbone. Combined column refers to Asian, Black and White subsets evaluated together

#	Backbone	Head	FT	VPT	Strategy	Test set, Mean AUC			
						Asian	Black	White	Combined
1	MAE	All	x	x	Figure 1b	0.783	0.797	0.783	0.792
2		Asian				0.78	x	x	0.79
3		Black				x	0.793	x	
4		White				x	x	0.783	
5		Asian		Asian	Figure 1c	0.81	x	x	**0.816**
6		Black		Black	Figure 1c	x	0.812	x	
7		White		White	Figure 1c	x	x	0.818	
8	MoCoV3	All		x	Figure 1b	0.766	0.775	0.759	0.765
9		Asian				0.769	x	x	0.789
10		Black				x	0.786	x	
11		White				x	x	0.788	
12		Asian		Asian	Figure 1c	0.788	x	x	**0.808**
13		Black		Black	Figure 1c	x	0.811	x	
14		White		White	Figure 1c	x	x	0.809	
15	MAE	All	All	x	Figure 1a	0.834	0.843	0.828	0.833
16		Asian	Asian			0.809	x	x	0.824
17		Black	Black			x	0.82	x	
18		White	White			x	x	0.823	
19		Asian	All	Asian	Figure 1c	0.834	x	x	**0.835**
20		Black		Black	Figure 1c	x	0.845	x	
21		White		White	Figure 1c	x	x	0.831	
22	MoCoV3	All	All	x	Figure 1a	0.826	0.842	0.829	0.833
23		Asian	Asian			0.797	x	x	0.823
24		Black	Black			x	0.813	x	
25		White	White			x	x	0.824	
26		Asian	All	Asian	Figure 1c	0.823	x	x	**0.836**
27		Black		Black	Figure 1c	x	0.836	x	
28		White		White	Figure 1c	x	x	0.828	

4.2 Evaluation of Adaptation Strategies

After pre-training a MoCoV3 and an MAE backbone, we first utilized linear probing on the entire dataset, then linear probing on individual (Asian, Black and White) demographics, and VPT on individual demographics to create intentionally biased models per demographics. The first half of the table Table 1 summarizes the mean AUC results on the same pre-trained backbone. Here, we observe that using a biased linear probing does not guarantee a performance increase; comparing biased linear model (rows 2-4) to linear mode trained on all data (row 1). Similarly for MoCoV3 the demographic biased model is only 0.03%-0.29% higher, which is negligible. In the case of MoCoV3, only the White demographic sees an improvement (0.788 vs 0.759), while other demographics observe a similar pattern to MAE. VPT has the highest performance (rows 5-7, 12-14) of both backbones by utilizing them to a greater extent.

Next, we fully fine-tune the model on the entire dataset, and additionally on the individual demographics to train biased models. Similarly, as with linear probing, the biased models fully fine-tuned on their respective demographics (rows 16-18 compared to corresponding each in row 15 for MAE and rows 23-25 vs 22 in MoCoV3) do not see any increase in performance, and instead, we see degradation in some scenarios specially where the dataset subset (Asian, Black) size is much smaller. The combined results for biased models, rows 16-18 is larger than the average between the AUC of each demographic because the AUC for combined was computed by concatenating all results, then computing AUC of each class, and taking the average of all labels.

Lastly, to see if the model can be improved any further after fine-tuning on all dataset, we applied additional VPT training on the individual demographics (rows 19-21 vs 15 for MAE and 26-28 vs 22 for MoCoV3). The fully fine-tuned backbone with VPT sees no further improvement instead, the results are very similar to the fully fine-tuned model.

5 Limitations

The limitation of this work is the lack of multiple runs through each method using the same hyper-parameters. The training time on a complete dataset takes a couple of weeks, even using a large number of GPUs, and for that reason we were not able to do that. However, since our datasets are very large (around half a million images), we expect calculating the confidence interval does not change the results much as it has been the case in former studies in these large-scale datasets Seyyed-Kalantari et al. [2021a,b], Gichoya et. al. [2022]. We also have not done the study on the Chest X-ray downstream tasks for different data scales, which we saved for future work.

6 Discussion & Conclusion

The results demonstrate a performance overview of the foundation models MAE and MoCoV3 pre-trained on Chest X-ray datasets, using linear, full fine-tuning and VPT. First, we can conclude that VPT at the cost of an increase of 1 percent in parameter count and longer training time Jia et al. [2022] shows a great improvement over linear probing.

We additionally observe that a VPT model on a demographic with a small amount of data performs similarly to a fully fine-tuned backbone on the same demographic. Specifically, MAE VPT on Asians compared to fine-tuned MAE on the same demographic shows similar performance. On the other hand, in demographics with larger representation, the performance is not far off. This demonstrates to us that VPT is an efficient adaptation method that allows foundation models to be adapted quickly, with minimal changes to the model's architecture. Unlike full fine-tuning, VPT retains the core model intact, using fewer additional parameters, resulting in a more lightweight adaptation.

On the other hand, we saw that a fully fine-tuned model on a large and diverse dataset leads to better performance, as was seen with the fully fine-tuned model on all Chest X-ray data. Finally, we see that a fully fine-tuned model sees no further improvement with the addition of prompts. This is likely due to model parameters and architecture being utilized to their biggest extent, where squeezing even a small amount of performance is not feasible.

In summary, VPT provides an efficient, scalable, and lightweight method for adapting foundation models for specific tasks while preserving general knowledge stored in the pre-trained foundation model. In our case, it's ideal for scenarios requiring task-specific tuning with limited data.

Acknowledgement

We sincerely thank the Natural Sciences and Engineering Research Council of Canada (NSERC) for the Discovery Grant (to Dr. Laleh Seyyed-Kalantari) and the Connected Minds CFREF (to Dr. Laleh Seyyed-Kalantari). The authors also express their gratitude to the Vector Institute for providing high-performance computing platforms.

References

E. Abbasi Bavil, M. Ahluwalia, L. Seyyed-Kalantari, B. Fine, and M. Abdalla. Body mass index prediction from chest radiographs and associated performance gaps in chest radiograph abnormality prediction. In *CAR 2024 Annual Scientific Meeting (ASM)*, Montréal, Canada, 2024.

I. Banerjee, K. Bhattacharjee, J. L. Burns, H. Trivedi, S. Purkayastha, L. S. Kalantari, B. N. Patel, R. Shiradkar, and J. Gichoya. "shortcuts" causing bias in radiology artificial intelligence: Causes, evaluation, and mitigation. *Journal of the American College of Radiology*, 20(9), September 2023.

X. Chen, S. Xie, and K. He. An empirical study of training self-supervised vision transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9620–9629, Los Alamitos, CA, USA, oct 2021. IEEE Computer Society. doi: 10.1109/ICCV48922.2021.00950. URL `https://doi.ieeecomputersociety.org/10.1109/ICCV48922.2021.00950`.

Judy Wawira Gichoya, Kaesha Thomas, Leo Anthony Celi, Nabile Safdar, Imon Banerjee, John D Banja, Laleh Seyyed-Kalantari, Hari Trivedi, and Saptarshi Purkayastha. Ai pitfalls and what not to do: mitigating bias in ai. *British Journal of Radiology*, 96(1150), October 2023.

Judy Wawira Gichoya et. al. Ai recognition of patient race in medical imaging: a modelling study. *The Lancet. Digital health*, 4:e406–e414, 2022. doi: 10.1016/S2589-7500(22)00076-2.

Anubhav Gupta, Islam Osman, Mohamed S Shehata, and John W Braun. Medmae: A self-supervised backbone for medical imaging tasks. *arXiv preprint arXiv:2407.14784*, 2024.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners, 2021. URL `https://arxiv.org/abs/2111.06377`.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David Mong, Safwan Halabi, Jesse Sandberg, Ricky Jones, David Larson, Curtis Langlotz, Bhavik Patel, Matthew Lungren, and Andrew Ng. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:590–597, 07 2019. doi: 10.1609/aaai.v33i01.3301590.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, pages 709–727, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-19827-4. URL `https://link.springer.com/chapter/10.1007/978-3-031-19827-4_41`.

Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1):317, Dec 2019. ISSN 2052-4463. doi: 10.1038/s41597-019-0322-0. URL `https://doi.org/10.1038/s41597-019-0322-0`.

Muhammad Osama Khan, Muhammad Muneeb Afzal, Shujaat Mirza, and Yi Fang. How fair are medical imaging foundation models? In Stefan Hegselmann, Antonio Parziale, Divya Shanmugam, Shengpu Tang, Mercy Nyamewaa Asiedu, Serina Chang, Tom Hartvigsen, and Harvineet Singh, editors, *Proceedings of the 3rd Machine Learning for Health Symposium*, volume 225 of *Proceedings of Machine Learning Research*, pages 217–231. PMLR, 10 Dec 2023. URL `https://proceedings.mlr.press/v225/khan23a.html`.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Sungho Park and Hyeran Byun. Fair-vpt: Fair visual prompt tuning for image classification. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12268–12278, 2024. URL `https://api.semanticscholar.org/CorpusID:271668477`.

Numan Saeed, Muhammad Ridzuan, Roba Al Majzoub, and Mohammad Yaqub. Prompt-based tuning of transformer models for multi-center medical image segmentation of head and neck cancer. *Bioengineering*, 10(7), 2023. ISSN 2306-5354. doi: 10.3390/bioengineering10070879. URL `https://www.mdpi.com/2306-5354/10/7/879`.

A. B. Sellergren, C. Chen, Z. Nabulsi, and et al. Li. Simplified transfer learning for chest radiography models using less data. *Radiology*, 305(2):454–465, 2022.

Laleh Seyyed-Kalantari, Guanxiong Liu, Matthew B. A. McDermott, and Marzyeh Ghassemi. Chexclusion: Fairness gaps in deep chest x-ray classifiers. *Pacific Symposium on Biocomputing*, 26:232–243, 2021a.

Laleh Seyyed-Kalantari, Haoran Zhang, and et al. McDermott. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature Medicine*, 27(12):2176–2182, 2021b.

Hari Sowrirajan, Jingbo Yang, Andrew Y. Ng, and Pranav Rajpurkar. Moco-cxr: Moco pretraining improves representation and transferability of chest x-ray models, 2021. URL `https://arxiv.org/abs/2010.05352`.

J. Xiao, Y. Bai, A. Yuille, and Z. Zhou. Delving into masked autoencoders for multi-label thorax disease classification. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3577–3589, Los Alamitos, CA, USA, jan 2023. IEEE Computer Society. doi: 10.1109/WACV56688.2023.00358. URL `https://doi.ieeecomputersociety.org/10.1109/WACV56688.2023.00358`.

Guanhua Zhang, Yihua Zhang, Yang Zhang, Wenqi Fan, Qing Li, Sijia Liu, and Shiyu Chang. Fairness reprogramming. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 34347–34362. Curran Associates, Inc., 2022. URL `https://proceedings.neurips.cc/paper_files/paper/2022/file/de08b3ee7c0043a76ee4a44fe68e90bc-Paper-Conference.pdf`.

A Dataset Demographic Count

Deeper look at the exact values of each demographic, and there total counts in the dataset.

Table 2: CheXpert and MIMIC-CXR Race counts

		Count				
		Asian	Black	White	Other	All
Dataset	CheXpert	23’298	11’970	125’624	62’756	223’648
	MIMIC-CXR	10’930	55’837	218’203	92’140	377’110
	All	34’228	67’807	343’827	154’896	600’758

B Parameter Search

Table 3: Hyper-parameters used for all stages of training

Backbone	Training Stage	Optimizer	Learning Rate	Weight Decay	Batch Size	Epochs
MAE	Pretraining	adamw	0.00015	0.05	128	800
	Linear probing	adamw	0.00025	0.05	128	75
	Full Fine-tuning	adamw	0.00025	0.05	128	75
	VPT	sgd	0.1	0.0001	128	75
	VPT on Full Fine-tuned	sgd	0.1	0.01	128	75
MoCoV3	Pretraining	adamw	0.00005	0.1	4096	300
	Linear probing	sgd	0.1	0.0001	128	75
	Full Fine-tuning	sgd	0.1	0.0	128	75
	VPT	sgd	0.3	0.0001	128	75
	VPT on Full Fine-tuned	sgd	0.1	0.001	128	75

For the hyper-parameter search, if the parameter is not in the Table 3, then a default value was used. The only exception is for VPT methods, we use prompt length 100. The pretraining of the MAE was done in similar manner to Xiao et al. [2023], but with our data split. The MAE was pretrained on 4 RTX 6000 24GB GPUs over 15 days. The MoCoV3 pretraining was done using 64 RTX 6000 24GB GPUs over 3 days. The tuning of the models was done using 8 RTX 6000 24GB GPUs. The duration for tuning ranged between 10 hours for a smaller subset, to 2 weeks for the entire dataset.