Predicting the Performance of Foundation Models via Agreement-on-the-Line

Aman Mehra*[1] **Rahul Saxena***[1] **Taeyoun Kim***[1] **Christina Baek**[1]
Zico Kolter[1,2] **Aditi Raghunathan**[1]
Carnegie Mellon University[1], Bosch Center for AI[2]
{amanmehr, rsaxena2, taeyoun3, kbaek, zkolter, raditi}@andrew.cmu.edu

Abstract

Estimating out-of-distribution performance is critical to safely deploy machine learning models. Recently, Baek et al. showed that the phenomenon "agreement-on-the-line" can be a reliable method for predicting the OOD accuracy of models in an ensemble consisting largely of CNNs trained from scratch. However, it is now increasingly common to lightly fine-tune foundation models, and it is unclear whether such fine-tuning is sufficient to produce enough diversity in model predictions for such agreement-based methods to work properly. In this paper, we develop methods for reliably applying agreement-on-the-line-based performance estimation to fine-tuned foundation models. In particular, we first study the case of fine-tuning a single foundation model, where we extensively study how different types of randomness (linear head initialization, data shuffling, and data subsetting) contribute to agreement-on-the-line of the resulting model sets. Somewhat surprisingly, we find that it is possible to obtain strong agreement via random initialization of the linear head alone. Next, we find how *multiple* foundation models, pretrained on different data sets but fine-tuned on the same task, also observe agreement-on-the-line. Again rather surprisingly, the diversity of such models is not too disparate, and they all lie on the same agreement line. In total, these methods enable reliable and efficient estimation of OOD accuracy for fine-tuned foundation models, without leveraging any labeled OOD data.

1 Introduction

Foundation models (FM) approaches, where one first pretrains a large model on open world data then fine-tunes for a specific downstream task, have achieved state-of-the-art results on image classification [27, 21, 38], text classification [6], question answering [8], and others. They are particularly noted for their often strong performance on out-of-distribution (OOD) data, that may vary substantially from the data used for fine-tuning (referred to as the in-distribution (ID) data) [5, 39]. Unfortunately, a substantial practical problem arises in this OOD setting: in many cases, one does not have access to labeled OOD data, and thus the field has explored other means for estimating OOD accuracy.

Interestingly, across a variety of distribution shift benchmarks, models often observe strong linear correlation between the ID and OOD accuracies of models, a phenomenon dubbed "Accuracy-on-the-line" (ACL) [25, 31, 32]. Recently, Baek et al. [2] empirically demonstrated that for ensembles of deep network classifiers trained from scratch, the rates of ID and OOD agreement also show a strong linear correlation with the same slope and bias. Baek et al. [2] used this to estimate the accuracies of models in such ensembles, thus providing a simple method for estimating OOD accuracy via unlabeled data alone. Thus, whenever the ID versus OOD accuracy is strongly linearly correlated, one may estimate the linear OOD performance trend using agreement without ground truth labels.

Submitted to the Workshop on Distribution Shifts (DistShift@NeurIPS 2023). Do not distribute.

Unfortunately, the AGL approach requires a *diverse collection* of classifiers over which to compute agreement: classifiers must vary in their predictions. Baek et al. [2] achieve this by training various models of different architectures from scratch. However, in the case of fine-tuned FMs, this diversity is seemingly lacking: we often want to *lightly* fine-tune just a single base FM for a downstream task, which even after multiple runs would seemingly lead to highly correlated downstream models, making them unsuitable for AGL-based OOD performance estimation.

In this work, we develop methods for extending AGL performance estimation to FMs, thus enabling practitioners to estimate the OOD performance of fine-tuned models without any labeled data. We first investigate the ability to estimate performance using a *single* base FM. We present a detailed empirical study of three potential sources of randomness during fine-tuning: 1) random linear head initialization; 2) random orderings of the fine-tuning data; and 3) random i.i.d subsets of the fine-tuning data. We find, somewhat surprisingly, that using random linear heads is able to reliably induce AGL behavior for the resulting classifiers, with the result holding across multiple different FMs and modalities (image classification and question answering a.k.a QA tasks). The result is a simple and straightforward method for evaluating OOD performance for a fine-tuned FM, applicable to settings where we only one want to fine-tune a single such base FM.

Second, we analyze the ability of the AGL-based method to predict OOD performance when using *multiple* different pretrained FMs. Here we encounter a setting where the different base models are pretrained on potentially entirely different data sets, using different architectures, and different training regiments. We show, however, that this degree of diversity is *also* sufficient for producing AGL behavior. Thus, for settings where multiple pretrained models exist, they can all be fine-tuned for a given downstream task, and AGL can allow us to estimate their accuracies.

In total, our contributions are as follows:

1. We propose a state-of-the-art method for unsupervised accuracy estimation under distribution shift when using large pretrained foundation models that are lightly fine-tuned for specific tasks. Prior works have primarily dealt with models trained from scratch, and hence are not directly applicable in this setting.
2. Our work leverages Agreement-on-the-line (AGL) [2] for OOD estimation, but extends it in important ways to apply to finetuned foundation models. The key to making AGL work is obtaining the right ensemble. In Baek et al. [2], multiple models were trained independently from scratch, an unfeasible step for FMs. We show that creating an ensemble with randomly initialized linear heads and then fine-tuning, also allows for AGL behavior, while other similar forms of ensembling (such as data ordering or data subsetting) do not.
3. We also identify several interesting phenomena underlying AGL that go beyond previous knowledge. Prior work Baek et al. [2] claimed that AGL does not hold for linear models. However, we find the contrary when using pretrained CLIP features. Furthermore, other prior work Miller et al. [25] suggests that the effective robustness (i.e. the linear fit between ID and OOD accuracy) would change depending on the pretraining data. We find that this is not the case for question answering with different pretrained FMs.

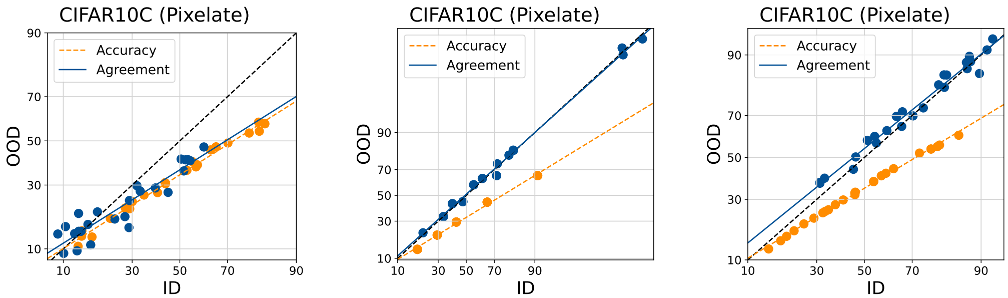
In total, this work substantially expands the set of problems and models for which AGL-based OOD performance estimation is practical, and allows us to leverage much more powerful models for settings where training models from scratch on tasks of interest is not feasible.

2 Background and related work

OOD performance estimation of FMs. Numerous tasks of interest boil down to mapping an input $x \in \mathbb{X}$ to a discrete output $y \in \mathbb{Y}$. In particular, consider a base FM $\mathsf{B} : \mathbb{X} \mapsto \mathbb{R}^d$ that we fine-tune to get $f(\mathsf{B}) : \mathbb{X} \mapsto \mathbb{Y}$. In this work, we consider a variety of foundation models: BERT [9], GPT2 [27], GPT-Neo, OPT [41], Llama2 [36], and CLIP [28].

Given access to a labeled validation set from $\mathcal{D}_{\text{ID}}$ and *unlabeled* samples from a different distribution $\mathcal{D}_{\text{ood}}$, our goal is to estimate performance on $\mathcal{D}_{\text{ood}}$. We consider the standard performance metrics: Accuracy $\ell_{\text{0-1}} : \mathbb{Y} \mapsto \mathbb{Y}$ for classification, and Macro-averaged F1 score $\ell_{\text{F1}} : \mathbb{Y} \mapsto \mathbb{Y}$ for QA.

There are a variety of proposed approaches for OOD performance estimation. One family of approaches attempts to quantify the degree of distribution shift through data and/or model dependent



(a) Random Head (b) Data Shuffling (c) Data Subsetting

Figure 1: The ID vs OOD trends for accuracy and agreement on the CIFAR10C "Pixelate" shift for linear-probe CLIP models, fine-tuned on CIFAR10 using the respective source of randomness: random linear heads, data shuffling, and independent data subsets. Clearly, the use of random linear heads is the only method producing AGL behaviour (i.e. matching bias and slope of the two lines)

metrics [4, 23, 7, 20]. However, these approaches only provide upper bounds on the OOD error, and the bounds tend to be loose when evaluated on deep networks [25]. Another line of work looks at leveraging the model's softmax predictions to predict the OOD performance [15, 14, 12, 10, 13]. While these approaches show empirical promise in some settings, they are not expected to work in general and often fail in the presence of large shifts [12].

ACL and AGL Baek et al. [2] propose AGL, a recent approach for estimating OOD performance, that outperforms prior approaches across a variety of shifts. It is based on an earlier intriguing observation from [25, 30, 31, 32, 40, 35, 24]—there is a strong linear correlation between the probit-scaled ID and OOD performances of models across many distribution shift benchmarks (ACL).

Interestingly, Baek et al. [2] observes that when where ACL holds, the probit-scaled *agreement between models* is also strongly correlated and observe the *same* slope and bias. Furthermore, when accuracies do not show a linear correlation, agreements also do not. This phenomenon was called "agreement-on-the-line" (AGL).

Formally, given a pair of models f_1 and f_2 that map inputs to labels, accuracy and agreement can be defined as

$$\mathsf{Acc}(f_1) = \mathbb{E}_{x,y\sim\mathcal{D}}[\ell(f_1(x), y)], \;\; \mathsf{Agr}(f_1, f_2) = \mathbb{E}_{x,y\sim\mathcal{D}}[\ell(f_1(x), f_2(x))], \tag{1}$$

where ℓ is the appropriate performance metric of interest. Note that while accuracy requires access to the labels y, agreement only requires access to unlabeled data and a pair of models. Thus, one can compute this line using OOD unlabeled data, and then estimate the OOD performance by linearly transforming the ID performance measured on ID validation data. See Appendix 5.3 for formal ALine methods to use AGL for OOD estimation.

Training from scratch vs fine-tuning A crucial component for AGL is the *diversity* of the ensemble predictions over which agreements are evaluated. If the models are not diverse enough, AGL is bound to fail. As an extreme, consider an ensemble of effectively identical models. Their ID and OOD agreement will always be 1, and there is no linear fit to estimate. Prior work on AGL has exclusively focused on training from scratch for several epochs, a very different regime from light fine-tuning. In this work, we focus on how to introduce sufficient diversity during *just* the fine-tuning process which can start from the *same* base FM and usually involves far fewer gradient steps.

3 Experiments and Results

Fine-tuning. In this work, we consider **linear probing (LP)** and **full fine-tuning (FFT)**. For LP, given features B_θ from the base model B, we train just the linear head v on top of frozen features such that the final classifier maps the score $v^\top \mathsf{B}_\phi(x)$ to a predicted class. We refer to v as either a linear probe (classification) or span prediction head (QA). For FFT, we attach a linear head v and

Table 1: The MAPE (%) of predicting OOD performance using ALine and other baseline methods. Evaluations on QA tasks (SQuAD-Shifts) are performed over a set of models finetuned from multiple base FMs (LlaMa, GPT, OPT). Evaluations on the image classification datasets are conducted with CLIP models fine-tuned with linear probing.

OOD Dataset	ALine-D	ALine-S	Naive Agr	ATC	AC	DF
SQuAD-Shifts (averaged across 4 shifts)	**1.68**	2.55	19.48	9.16	45.04	4.54
CIFAR10C (averaged across shifts)	6.99	**6.92**	44.33	31.28	48.66	32.79
CIFAR10.1 (averaged across v4, v6)	**2.42**	3.03	41.52	6.48	54.57	8.51
CIFAR100C (averaged across shifts)	**11.94**	12.67	46.13	18.69	80.81	37.36
ImageNet V2 (averaged across 3 format)	**4.96**	5.03	47.65	8.96	77.34	7.86
WILDS (averaged across 3 benchmarks)	**11.52**	12.91	50.12	21.73	42.18	27.54

optimize the suitable loss function, but we *update all parameters* of the backbone such that the feature extract B_ϕ is updated. When infeasible to update all parameters natively, we perform *low-rank adaptation* (LoRA) [16] which uses trainable rank decomposition matrices to reduce the number of trainable parameters while still effectively updating the feature extractor B_ϕ. In this work, we do not distinguish between LoRA and FFT as they conceptually achieve the same effect, and show similar empirical trends in our studies. Refer to Appendix 5.1 for details on fine-tuned models and Appendix 5.2 for specific fine-tuning parameters.

Datasets We study AGL for the tasks of QA and Image Classification. For QA, we fine-tune on the SQuAD v1.1 dataset [29] and evaluate on four distribution shifts present in SQuAD-Shifts (New Wiki, New York Times, Amazon, and Reddit) [24]. For image classification, we fine-tune on CIFAR10 [19], and then test on CIFAR10C [14], a dataset with 19 corruptions, some natural (Snow), and some synthetic (JPEG compression). We also test on the CIFAR10.1 dataset [30], which contains newer images for the same labels. We repeat the same for CIFAR100 [18], ImageNet-1k [33]. We additionally validate our finding by testing on three natural shifts from the WILDS benchmark (FMoW, iWildCam, Camelyon17) [17].

3.1 Predicting OOD performance: single base foundation model

Consider the case where we have a *single* base FM to fine-tune. An overriding concern when calculating agreement is that even some randomness in the fine-tuning process may not be enough to overcome the underlying similarities in predictions due to the same base FM. To address this problem, we evaluate three possible methods for introducing diversity in the fine-tuning, to see what approach (if any) can lead to AGL behavior:

1. **Random linear heads.** Before fine-tuning, we initialize the last layer of the network (i.e., the linear head) randomly, instead of via some zero-shot or pre-specified manner.
2. **Data shuffling.** We present the same data to each model, but shuffle the order for the data differently within each fine-tuning optimization run.
3. **Data subsetting.** We fine-tune each model with an independently sampled subset of the ID data. All models are trained on subsets of the same size.

Note that we perturb only one source of diversity at a time. For example, in the random linear head setting, all models start with a different initialization, but the data used for training is the same and seen in the same order. In the data shuffling setting, all models start with an identical arbitrary initialization, but the data used for training is seen in different orders; and so on.

For our study of image classification, we train a linear probe atop of CLIP, specifically the ViT-B/32 model trained on LAION-2B [34]. For QA, we evaluate a collection of 50 fine-tuned models, all obtained by fine-tuning from the same checkpoint of a GPT2-Medium. We repeat the same procedure for OPT and BERT architectures, the details of which can be found in the Appendix (Sections 5.1 and 5.5).

For the case of training models from scratch, it is well established that independent data subsetting tends to lead to the greatest diversity of classifiers [26]. Nonetheless, in this setting we find rather surprisingly, that model pairs trained with different *randomly initialized linear heads* achieve the lowest OOD agreement for the same ID agreement. In fact, the ID versus OOD agreement matching the slope of ID versus OOD accuracy. On the other hand, data ordering and data shuffling observe ID versus OOD agreement that lies closer to the diagonal $y = x$ and away from the accuracy linear fit. We show that this finding persists over numerous models and tasks.

3.2 Predicting OOD performance: multiple base foundation models

When multiple base foundation models (pretrained on different data) are accessible, it is unclear if models with different bases would lie on different or similar accuracy lines, even if fine-tuned on the same ID data. We observe that for certain extractive QA shifts, foundation models fine-tuned from a wide range of base models *exhibit both ACL and AGL* (See Appendix 5.6 for details)

3.3 Results

Figure 1 shows the ACL/AGL trends for linear probes trained on top of CIFAR10 CLIP representations. One may suspect that such linear models would agree highly and AGL may break. However, we see that contrary to the findings of Baek et al. [2], even linear models, when on top of neural network features with the *right type of diversity*, may exhibit AGL. Interestingly enough, for the other sources of diversity, we observe ACL and strongly linearly correlated agreement, but the latter at a much higher rate OOD. We refer the reader to Appendix 5.9 for a more exhaustive evaluation. The same observations, however not as stark, can be made for the fine-tuned LLMs. We refer the reader to Appendix 5.5 to observe these trends on all four shifts within the SQuaD-Shifts dataset.

When considering multiple base foundation models, we first observe that base LLMs pretrained on different corpora also lead to fine-tuned models that exhibit ACL. This is in contrast to the findings of previous works [28, 35]. Second, the ID versus OOD agreement for pairs of models in this ensemble, including pairs of different base foundation models, retains a strong linear correlation and the slope and bias closely matches that of accuracy. As a result, different pretraining does not break AGL.

Table 1 shows the averaged MAPE (Mean Absolute Percentage Error) as calculated using the ALine algorithm and other baseline methods for some dataset shifts (the full version for all datasets can be found in Appendix 5.8). The QA ensembles are generated by fine-tuning multiple foundation models, and the image classification ones are all CLIP linear-probes. Since AGL is demonstrated to hold well in all these ensembles, the ALine MAE is able to surpass other methods; thus lending support to our method to get AGL to hold for lightly fine-tuned models, and using it to estimate OOD performance.

4 Conclusion

We develop methods for extending AGL to lightly fine-tuned FMs to enable OOD performance prediction in this emerging paradigm. We found that applying AGL directly may sometimes fail, and proper utilization of this phenomena requires a careful tuning of the distribution of models in an ensemble for their errors to be uncorrelated. Unlike the original paradigm of AGL, where models observed tens or hundreds of epochs of training on the in-distribution dataset, we find that stochasticity in specific optimization choices, specifically random initialization, is crucial for observing AGL in lightly fine-tuned FMs. Second, though Baek et al. [2] posed AGL as a model centric phenomena that is specifically only observed in neural network ensembles, we find that linear models can also observe AGL when the data and the distribution shift contain certain structures (as is possible in the CLIP representation space).

Our conclusion on AGL also sheds light on ACL (i.e. accuracy-on-the-line) in the presence of foundation models, a phenomenon that is of independent interest. Some recent works have studied the effect of different forms of fine-tuning on ACL [28, 1]. The main finding reported is that different forms of fine-tuning lead to different slopes in the linear correlations, a term that is often called “effective robustness”. In our results, we find that when fine-tuned the same way, models obtained from *different base foundation models* all lie on the *same* line. This is particularly intriguing because it goes against the common wisdom that the amount of pretraining data determines the effective robustness. We leave these questions for future analysis.

References

[1] Anas Awadalla, Mitchell Wortsman, Gabriel Ilharco, Sewon Min, Ian Magnusson, Hannaneh Hajishirzi, and Ludwig Schmidt. Exploring the landscape of distributional robustness for question answering models. *arXiv preprint arXiv:2210.12517*, 2022.

[2] Christina Baek, Yiding Jiang, Aditi Raghunathan, and J Zico Kolter. Agreement-on-the-line: Predicting the performance of neural networks under distribution shift. *Advances in Neural Information Processing Systems*, 35:19274–19289, 2022.

[3] Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. The pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 830–839, 2020.

[4] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19, 2006.

[5] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[7] Corinna Cortes, Yishay Mansour, and Mehryar Mohri. Learning bounds for importance weighting. *Advances in neural information processing systems*, 23, 2010.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[10] Hady Elsahar and Matthias Gallé. To annotate or not? predicting performance drop under domain shift. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2163–2173, 2019.

[11] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

[12] Saurabh Garg, Sivaraman Balakrishnan, Zachary C Lipton, Behnam Neyshabur, and Hanie Sedghi. Leveraging unlabeled data to predict out-of-distribution performance. *International Conference on Learning Representations*, 2022.

[13] Devin Guillory, Vaishaal Shankar, Sayna Ebrahimi, Trevor Darrell, and Ludwig Schmidt. Predicting with confidence on unseen distributions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1134–1144, 2021.

[14] Dan Hendrycks and Thomas G. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *7th International Conference on Learning Representations, ICLR*, 2019.

[15] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017.

[16] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[17] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664. PMLR, 2021.

[18] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-100 (canadian institute for advanced research). URL `http://www.cs.toronto.edu/~kriz/cifar.html`.

[19] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[20] Ilja Kuzborskij and Francesco Orabona. Stability and hypothesis transfer learning. In *International Conference on Machine Learning*, pages 942–950. PMLR, 2013.

[21] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

[22] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[23] Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. *arXiv preprint arXiv:0902.3430*, 2009.

[24] John Miller, Karl Krauth, Benjamin Recht, and Ludwig Schmidt. The effect of natural distribution shift on question answering models. In *International conference on machine learning*, pages 6905–6916. PMLR, 2020.

[25] John P Miller, Rohan Taori, Aditi Raghunathan, Shiori Sagawa, Pang Wei Koh, Vaishaal Shankar, Percy Liang, Yair Carmon, and Ludwig Schmidt. Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization. In *International Conference on Machine Learning*, pages 7721–7735. PMLR, 2021.

[26] Preetum Nakkiran and Yamini Bansal. Distributional generalization: A new kind of generalization. *arXiv preprint arXiv:2009.08092*, 2020.

[27] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[29] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.

[30] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do cifar-10 classifiers generalize to cifar-10? *arXiv preprint arXiv:1806.00451*, 2018.

[31] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pages 5389–5400. PMLR, 2019.

[32] Rebecca Roelofs, Vaishaal Shankar, Benjamin Recht, Sara Fridovich-Keil, Moritz Hardt, John Miller, and Ludwig Schmidt. A meta-analysis of overfitting in machine learning. *Advances in Neural Information Processing Systems*, 32, 2019.

[33] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *CoRR*, abs/1409.0575, 2014. URL `http://arxiv.org/abs/1409.0575`.

[34] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

[35] Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems*, 33:18583–18599, 2020.

[36] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023.

[37] Trieu H Trinh and Quoc V Le. A simple method for commonsense reasoning. *arXiv preprint arXiv:1806.02847*, 2018.

[38] Dequan Wang, Xiaosong Wang, Lilong Wang, Mengzhang Li, Qian Da, Xiaoqiang Liu, Xiangyu Gao, Jun Shen, Junjun He, Tian Shen, et al. Medfmc: A real-world dataset and benchmark for foundation model adaptation in medical image classification. *arXiv preprint arXiv:2306.09579*, 2023.

[39] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7959–7971, 2022.

[40] Chhavi Yadav and Léon Bottou. Cold case: The lost mnist digits. *Advances in neural information processing systems*, 32, 2019.

[41] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. Opt: Open pre-trained transformer language models, 2022.

[42] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27, 2015.

5 Appendix

5.1 Models

Extractive QA We evaluate a collection of 125 fine-tuned models for our experiments in this section. Each model is obtained by fine-tuning from the same checkpoint of a GPT2-Medium, OPT-125M, and BERT. We individually present findings on both these families of models in the following sections. Huggingface links to the base models we trained are in Appendix 5.7.

Image Classification We use CLIP [28], specifically the ViT-B/32 model trained on LAION-2B [34] for our image classification tasks. Given its well-established 0-shot capabilities, a popular method of fine-tuning CLIP for downstream tasks is to simply employ linear probing on top of the CLIP representation. Thus, we are interested in evaluating the OOD performance of an ensemble of models where the only difference is the linear head.

Multiple Models We train 41 models on the extractive QA benchmark SQuAD as in the previous section, and observe their OOD performance to SQuAD-Shifts. We fine-tune OPT-125M [41], OPT-350M, OPT-1.3B, GPT2-XL, GPT2-Large, GPT2-Medium, GPT2 [27], GPT-Neo-135M, Llama2-7B [36], Alpaca-7B, and Vicuna-7B to extractive QA. OPT was pretrained on a wide variety of data including BookCorpus [42], Stories [37], a subset of PILE [11], CCNews v2 corpus, and PushShift.io Reddit [3]. Similarly, GPT2 was pretrained on BookCorpus while GPT-Neo was trained on PILE. Llama2 was trained on an undisclosed set of publicly available data. Sprouting from Llama2, Alpaca is additionally trained from Llama2 on instruction-following demonstrations while Vicuna is additionally trained from Llama2 on user-shared conversations from ShareGPT.

5.2 Finetuning Specifics

We state here the specific parameters used in finetuning GPT2-Medium for extractive QA and CLIP for image classification. Across the four different sources of diversity, the epochs are varied regardless of the experiment. We train with AdamW as the optimizer [22]. For randomly initializing linear heads we vary the seed for the head and keep all other values fixed. For changing the finetuning hyperparameters, we vary the learning rate and weight decay. To shuffle the data, we change the data seed that control the data ordering during training. And finally for data subsetting, we get different proportions of the dataset which are independently sampled.

For the GPT2-Medium models, we train a total of 50 models for studying the sources of diversity. For the CLIP models, we fine-tune upwards of 200 models (i.e. linear heads on top of the CLIP representation) for the different vision datasets.

Table 2: Finetuning specifics for extractive QA (LR: learning rate, WD: weight decay, LS: linear head initialization seed, DS: data shuffling seed, DP: data subsetting proportion, EP: epochs, B: batch size)

Source of Diversity	**GPT2-Medium**	
	Varied	**Fixed**
Random linear heads	LS: varied	LR: 3×10^{-6} WD: 2×10^{-4} DS: fixed DP: 20% EP: 0–3 B: 4
Data shuffling	DS: varied	LR: 4×10^{-6} WD: 1×10^{-4} LS: fixed DP: 10% EP: 0–3 B: 4
Data subsetting	DP: $4.5\% - 50\%$	LR: 2×10^{-6} WD: 1×10^{-4} DS: varied LS: fixed EP: 1 B: 4

Table 3: Finetuning specifics for OPT-125M (LR: learning rate, WD: weight decay, LS: linear head initialization seed, DS: data shuffling seed, SS: random subsetting seed, EP: epochs)

Source of Diversity	**OPT-125M**	
	Varied	**Fixed**
Random linear heads	LS: varied	LR: 4×10^{-7} WD: 1×10^{-5} DS: fixed SS: fixed EP: 10
Data shuffling	DS: varied	LR: 4×10^{-7} WD: 1×10^{-5} LS: fixed SS: fixed EP: 10
Data subsetting	SS: varied	LR: 4×10^{-7} WD: 1×10^{-5} DS: varied LS: fixed EP: 10

Table 4: Finetuning specifics for BERT (LR: learning rate, WD: weight decay, LS: linear head initialization seed, DS: data shuffling seed, SS: random subsetting seed, EP: epochs)

Source of Diversity	**BERT** **Varied**	 **Fixed**
Random linear heads	LS: varied	LR: 2×10^{-7} WD: 1×10^{-5} DS: fixed SS: fixed EP: 10
Data shuffling	DS: varied	LR: 2×10^{-7} WD: 1×10^{-5} LS: fixed SS: fixed EP: 10
Data subsetting	SS: varied	LR: 2×10^{-7} WD: 1×10^{-5} DS: varied LS: fixed EP: 10

Table 5: Finetuning specifics for CLIP (LR: learning rate, WD: weight decay, LS: linear head initialization seed, DS: data shuffling seed, DP: data subsetting proportion, EP: epochs, B: batch size)

Source of Diversity	**CLIP + ViT-B/32 (LAION-2B)** **Varied**	 **Fixed**
Random linear heads	LS: varied	LR: different per dataset WD: 0 DS: fixed DP: 100% EP: 1–100 B: 1024
Data shuffling	DS: varied	LR: different per dataset WD: 0 LS: fixed DP: 100% EP: 1–100 B: 1024
Data subsetting	DP: $10\% - 50\%$	LR: different per dataset WD: 0 DS: varied LS: fixed EP: 1–100 B: 1024

5.3 ALine-S/D

ALine is the OOD accuracy estimating metric that utilizes AGL [2]. There are two methods within ALine: ALine-S and ALine-D

Given $Acc_{ID}(f_1)$ and $Agr_{OOD}(f_1, f_2)$, when agreement holds, the relationship between the agreement line and accuracy line is as follows.

$$\Phi^{-1}(\mathrm{Acc}_{\mathrm{OOD}}(f_1)) = a\cdot\Phi^{-1}(\mathrm{Acc}_{\mathrm{ID}}(f_1))+b \Leftrightarrow \Phi^{-1}(\mathrm{Agr}_{\mathrm{OOD}}(f_1, f_2)) = a\cdot\Phi^{-1}(\mathrm{Agr}_{\mathrm{ID}}(f_1, f_2))+b \tag{2}$$

To find $Acc_{OOD}(f_2)$, we can estimate the slope a and bias b as follows and

$$\hat{a}, \hat{b} = \arg\min_{a,b\in\mathbb{R}} \sum_{i\neq j} \left(\Phi^{-1}(\hat{\mathrm{Agr}}_{\mathrm{OOD}}(h_i, h_j)) - a\cdot\Phi^{-1}(\hat{\mathrm{Agr}}_{\mathrm{ID}}(h_i, h_j)) - b\right)^2 \tag{3}$$

With $\hat{a}$ and $\hat{b}$, we can find $Acc_{OOD}(f_2)$ with the estimator for the ID accuracy $\hat{Acc}_{ID}(f_1)$. This method is called Aline-S.

A similar method, ALine-D, uses pointwise accuracies and agreement of the model of interest instead of estimating the entire agreement line. If the models of interest are h and h', then the following holds.

$$\frac{1}{2}\left(\Phi^{-1}(\mathrm{Acc}_{\mathrm{OOD}}(h)) + \Phi^{-1}(\mathrm{Acc}_{\mathrm{OOD}}(h'))\right) = \frac{a}{2}\left(\Phi^{-1}(\mathrm{Acc}_{\mathrm{ID}}(h)) + \Phi^{-1}(\mathrm{Acc}_{\mathrm{ID}}(h'))\right) + \frac{b}{2} \tag{4}$$

With the fact that $b = \Phi^{-1}(\mathrm{Agr}_{\mathrm{OOD}}(h, h')) - a\cdot\Phi^{-1}(\mathrm{Agr}_{\mathrm{ID}}(h, h'))$, we have

$$\begin{aligned} &\frac{1}{2}\left(\Phi^{-1}(\mathrm{Acc}_{\mathrm{OOD}}(h)) + \Phi^{-1}(\mathrm{Acc}_{\mathrm{OOD}}(h'))\right) \\ &= \Phi^{-1}(\mathrm{Agr}_{\mathrm{OOD}}(h, h')) + a\cdot\left(\frac{\Phi^{-1}(\mathrm{Acc}_{\mathrm{ID}}(h)) + \Phi^{-1}(\mathrm{Acc}_{\mathrm{ID}}(h'))}{2} - \Phi^{-1}(\mathrm{Agr}_{\mathrm{ID}}(h, h'))\right) \end{aligned} \tag{5}$$

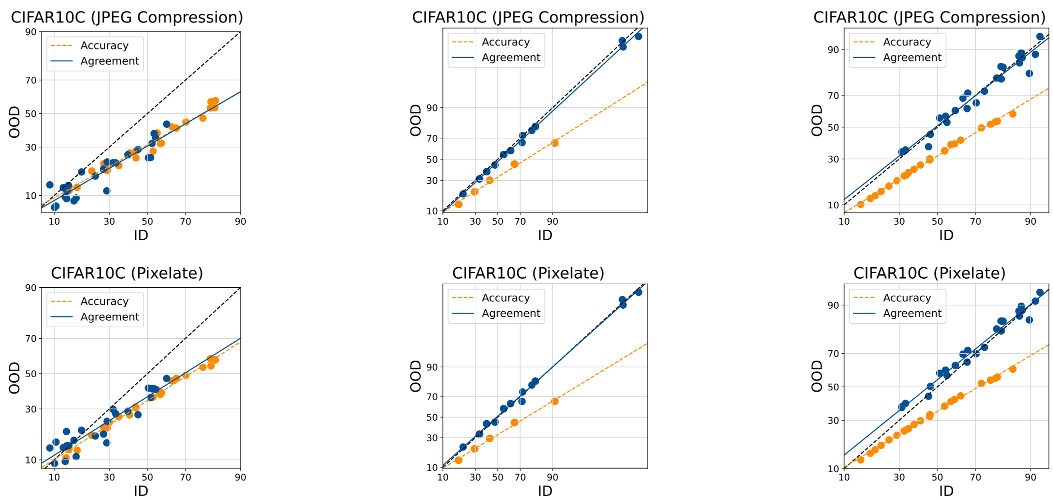
With the two unknowns, $\mathrm{Acc}_{\mathrm{OOD}}(h)$ and $\mathrm{Acc}_{\mathrm{OOD}}(h')$, and one equations we cannot find the unknowns. However, with more overlapping pairs, we can get the same number equations as variables and find the OOD accuracy of a model of interest.

5.4 Sources of Diversity (Image Classification)

Figure 2 shows the three sources of diversity for the "Pixelate" and "JPEG-Compression" shifts in the CIFAR 10C OOD dataset. Table 6 shows the ALine-D MAE (%) for image classification on CIFAR10C (average across all 19 shifts).

Table 6: ALine-D MAE and MAPE for CLIP linear probing on CIFAR10 image classification. Note that the reported MAE and MAPE is averaged across all 19 CIFAR10C evaluated shifts.

Source of Diversity	CIFAR10C MAPE (%)	CIFAR10C MAE (%)
Random linear heads	**15.88**	**5.74**
Data shuffling	74.16	22.61
Data subsetting	25.94	7.39



(a) Random Head (b) Data Shuffling (c) Data Subsetting

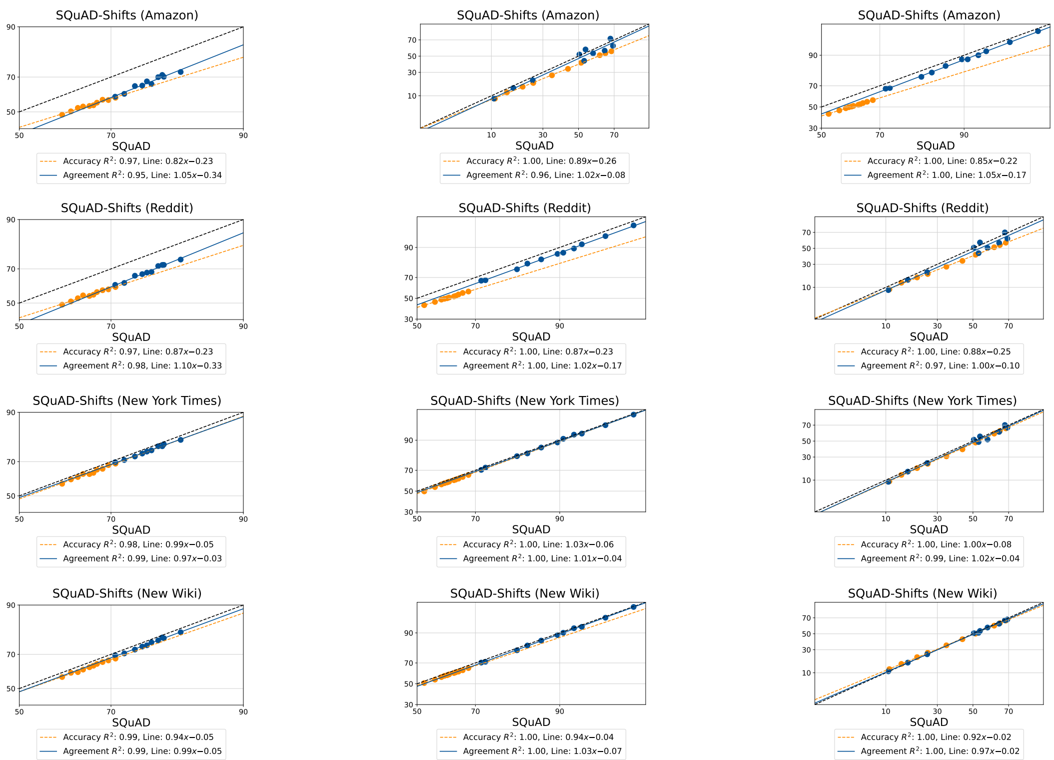
Figure 2: The ACL and AGL plots for the "JPEG Compression" (top row) and "Pixelate" (bottom row) fine-tuned using different sources of randomness

5.5 Sources of Diversity (Question Answering)

Figure 3 shows the three sources of diversity for all SQuAD-Shifts OOD datasets. Table 7 shows the ALine-D MAE for SQuAD-Shifts Amazon and Reddit.

Table 7: ALine-D MAPE(%) and MAE (%) on the SQuAD-Shifts Amazon and Reddit datasets when applied to sets of fully-finetuned models, trained using different sources of randomness

Source of Diversity	SQuAD-Shifts Amazon MAPE (%)	SQuAD-Shifts Amazon MAE (%)	SQuAD-Shifts Reddit MAPE (%)	SQuAD-Shifts Reddit MAE (%)
Random Linear Heads	**6.34**	**0.69**	**3.48**	**0.79**
Data Shuffling	10.30	4.18	9.59	4.32
Data Subsetting	16.21	5.2	13.94	4.71

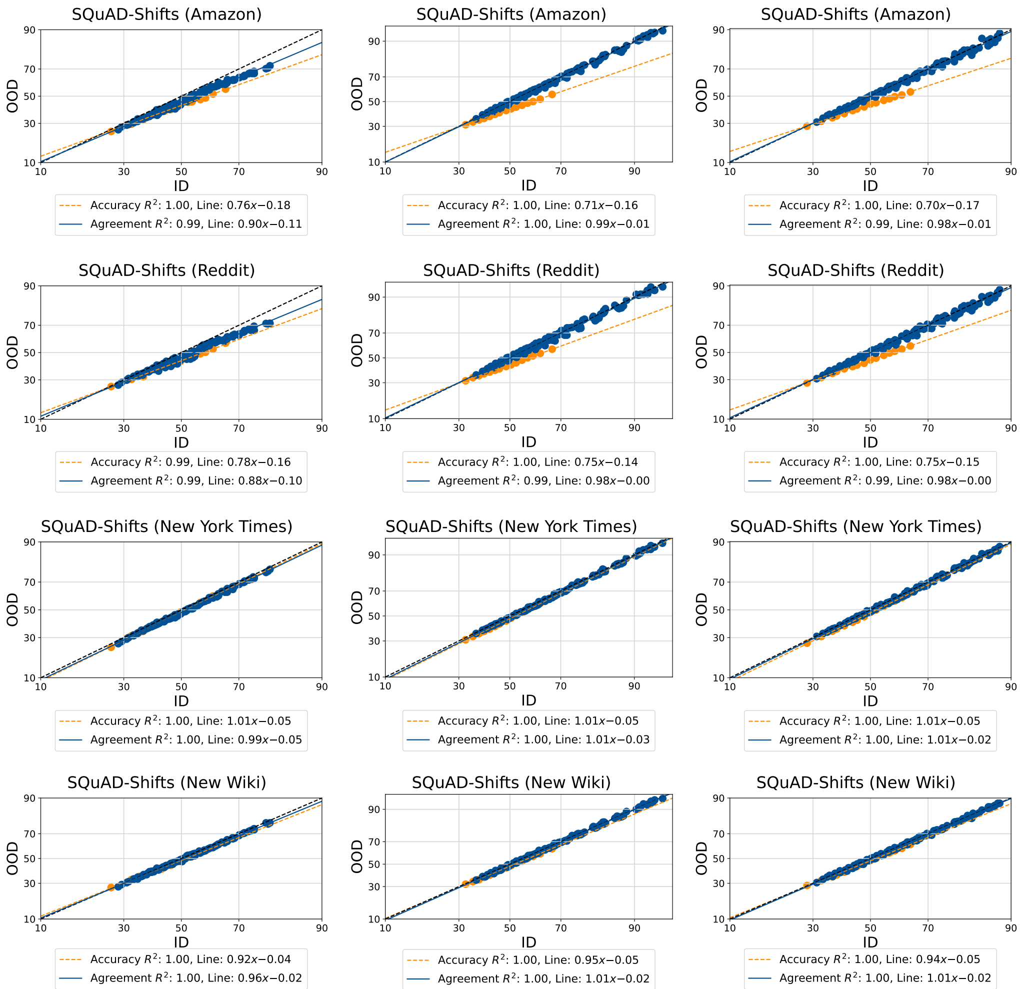


(a) Random Head (b) Data Shuffling (c) Data Subsetting

Figure 3: ID vs OOD trends of accuracy and agreement of LLMs finetuned for Question Answering from a single pretrained base model. Each column presents trends for different sources of stochasticity employed to obtain a diverse ensemble of finetuned models.

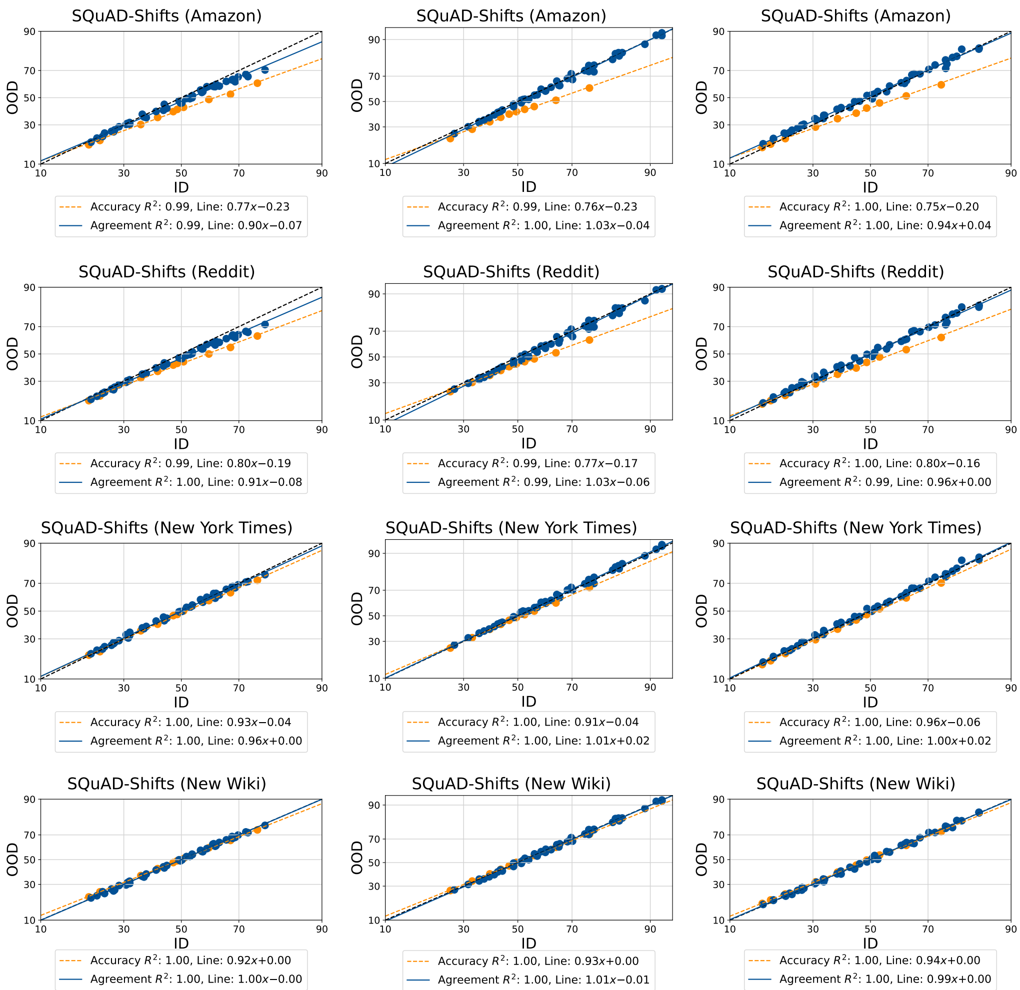
In this section, we also expand our evaluations to finetuned OPT-125M and BERT models for the extractive question answering task discussed in Section **??**. For both of these base foundation models, we consider the three sources of diversity for finetuning i.e. using random linear heads, random ordering, and independent data subsetting, and plot the respective ID vs OOD accuracy of models and agreement between pairs of models in the resultant model set.

These experiments also afford us the chance to analyse the similarites and differences between the ACL/AGL trends exhibited by the model sets with GPT2-Medium, OPT-125M, and BERT as the base FM respectively. In particular, AGL is slightly worse for OPT-125M and BERT, and thus ALine has a higher error on OPT-125M and BERT than GPT2-Medium. However, we still see a consistent trend where AGL holds the best for random head initialization compared to data shuffling and data subsetting; thus implying that the ALine error for random head initialization is the smallest out of all diversity sources. Thus, the importance of random head initialization applies to all models regardless of architecture in AGL.



(a) Random Head (b) Data Shuffling (c) Data Subsetting

Figure 4: ID vs OOD trends of accuracy and agreement of LLMs finetuned for Question Answering from a single pretrained base model (OPT-125M). Similar to the GPT2-Medium results, these show that random linear head initialization is the best method to obtain model sets exhibiting AGL



(a) Random Head (b) Data Shuffling (c) Data Subsetting

Figure 5: ID vs OOD trends of accuracy and agreement of LLMs finetuned for Question Answering from a single pretrained base model (BERT). Similar to the GPT2-Medium results, these show that random linear head initialization is the best method to obtain model sets exhibiting AGL

5.6 Multiple Foundation Models

Figure 6 shows AGL and ACL for different base models for all SQuAD-Shifts OOD datasets. We have fine-tuned OPT-125M, OPT-350M, OPT-1.3B, GPT2-XL, GPT2-Large, GPT2-Medium, GPT2, GPT-Neo-135M, Llama2-7B, Alpaca-7B, and Vicuna-7B. The links to the models are in Appendix 5.7.

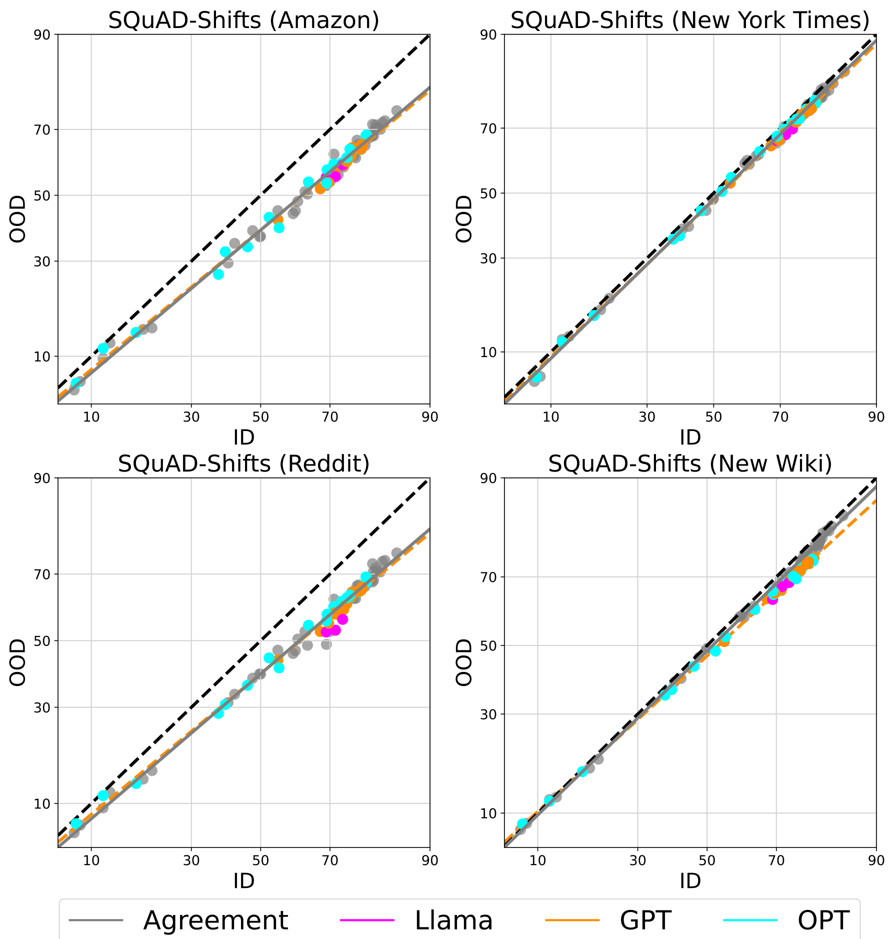


Figure 6: AGL when using different base models for SQuAD-Shifts

5.7 Huggingface Links

Here are the Huggingface links to the pretrained base foundation models we finetuned: GPT2 (https://huggingface.co/gpt2), GPT2-Medium (https://huggingface.co/gpt2-medium), GPT2-Large (https://huggingface.co/gpt2-large), GPT2-XL (https://huggingface.co/gpt2-xl), GPT-Neo-125M (https://huggingface.co/EleutherAI/gpt-neo-125m), GPT-Neo-1.3B (https://huggingface.co/EleutherAI/gpt-neo-1.3B), OPT-125M (https://huggingface.co/facebook/opt-125m), OPT-1.3B (https://huggingface.co/facebook/opt-1.3b), Llama2-7B (https://huggingface.co/meta-llama/Llama-2-7b-hf), Alpaca-7B (https://huggingface.co/WeOpenML/Alpaca-7B-v1), Vicuna-7B (https://huggingface.co/lmsys/vicuna-7b-v1.3), BERT (https://huggingface.co/bert-base-uncased)

5.8 OOD Accuracy Estimation Methods

Table 8: The MAPE (%) of predicting OOD performance using ALine and other baseline methods. Evaluations on QA tasks (SQuAD-Shifts) are performed over a set of models finetuned from multiple base FMs (LlaMa, GPT, OPT). Evaluations on the image classification datasets are conducted with CLIP models fine-tuned with linear probing.

OOD Dataset	ALine-D	ALine-S	Naive Agr	ATC	AC	DF
SQuAD-Shifts Reddit	**1.20**	2.60	20.21	12.74	49.25	6.09
SQuAD-Shifts Amazon	**1.64**	3.10	20.40	15.35	51.06	7.39
SQuAD-Shifts Nyt	**0.82**	1.33	18.46	3.11	38.61	3.18
SQuAD-Shifts New Wiki	3.08	3.18	18.87	5.46	41.26	**1.50**
Average	**1.68**	2.55	19.48	9.16	45.04	4.54
CIFAR10C (averaged across shifts)	6.99	**6.92**	44.33	31.28	48.66	32.79
CIFAR10.1 (averaged across v4, v6)	**2.42**	3.03	41.52	6.48	54.57	8.51
CIFAR100C (averaged across shifts)	**11.94**	12.67	46.13	18.69	80.81	37.36
ImageNetC (averaged across shifts)	**10.91**	11.04	56.76	27.25	79.00	37.86
ImageNet V2 (averaged across 3 format)	**4.96**	5.03	47.65	8.96	77.34	7.86
fMoW-WILDS (val OOD split)	**2.59**	2.74	83.94	9.03	44.59	5.86
iWildCam-WILDS (val OOD split)	**22.05**	25.29	46.42	37.25	57.31	69.58
Camelyon17-WILDS (val OOD split)*	9.93	10.71	19.99	18.92	24.64	**7.18**

5.9 Using Random-Head initialized fine-tuned CLIP models for other datasets

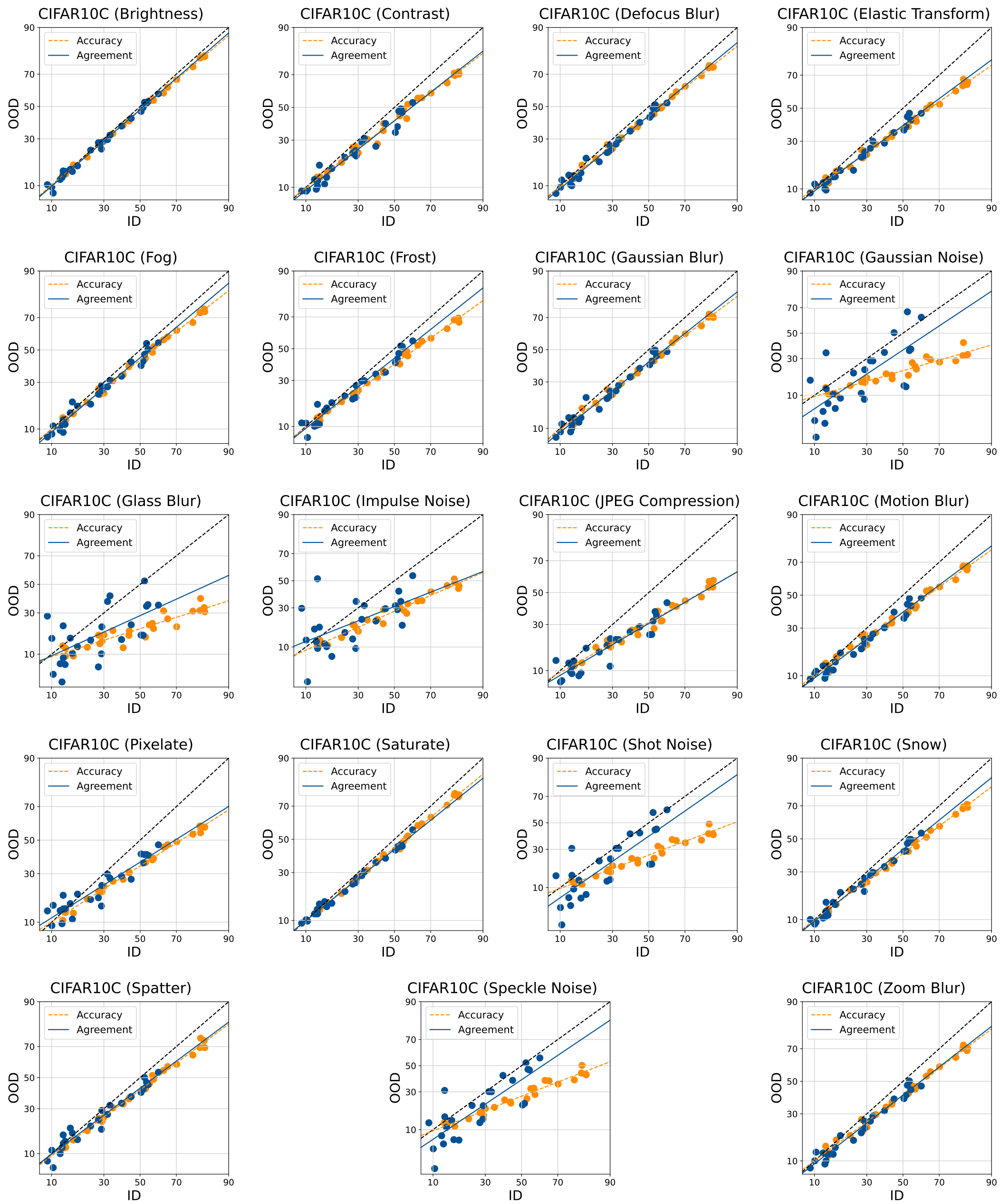


Figure 7: AGL and ACL for all C10C shifts with random head initialization fine-tuning.

Figure 8: AGL and ACL for the C10.1 shifts with random head initialization fine-tuning.

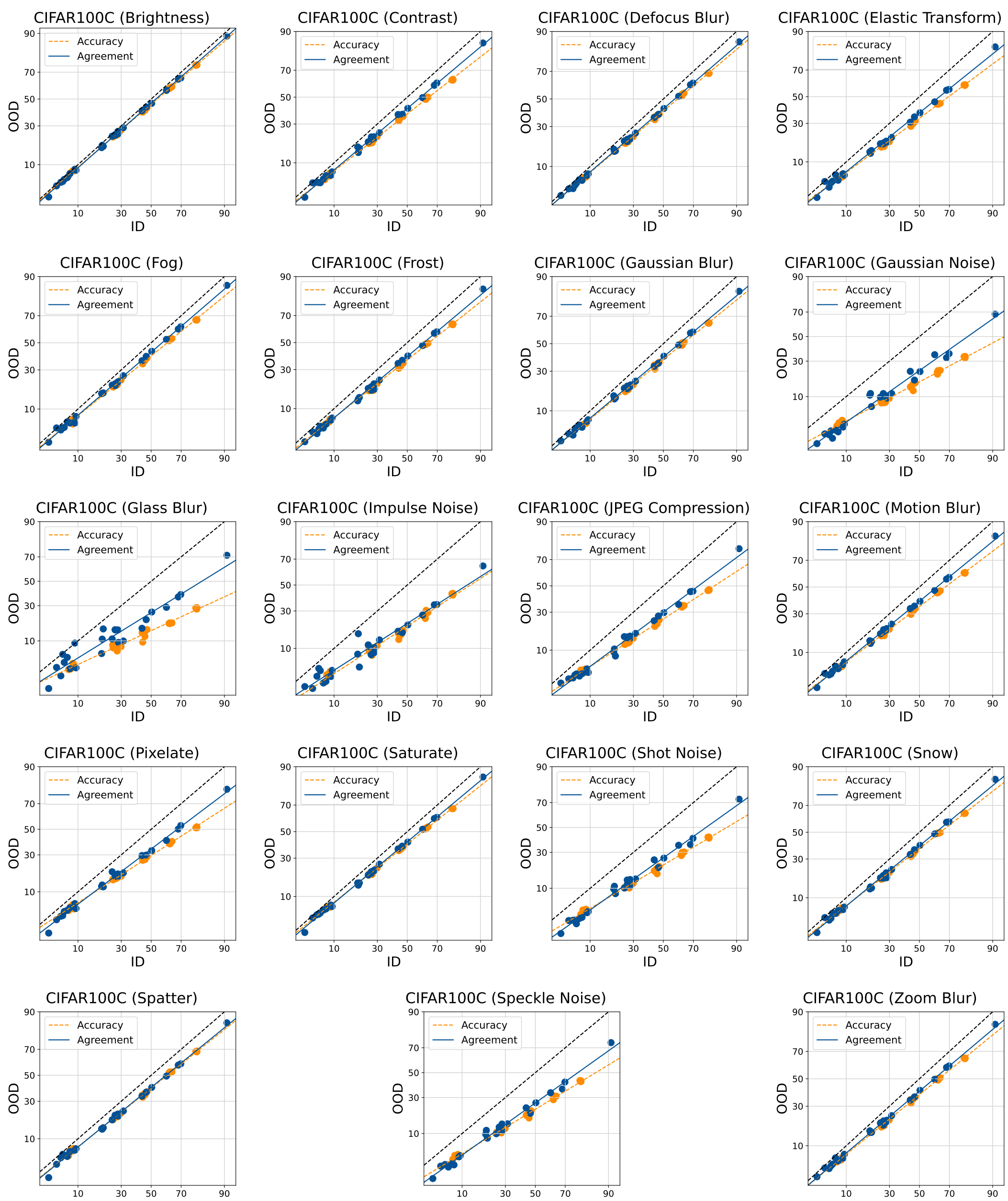


Figure 9: AGL and ACL for the C100C shifts with random head initialization fine-tuning.

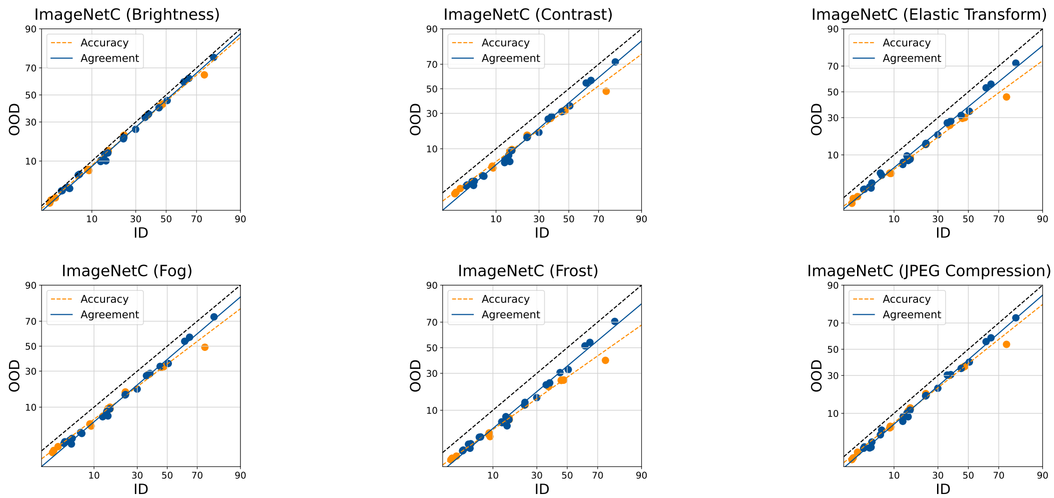


Figure 10: AGL and ACL for the ImageNetC shifts with random head initialization fine-tuning.

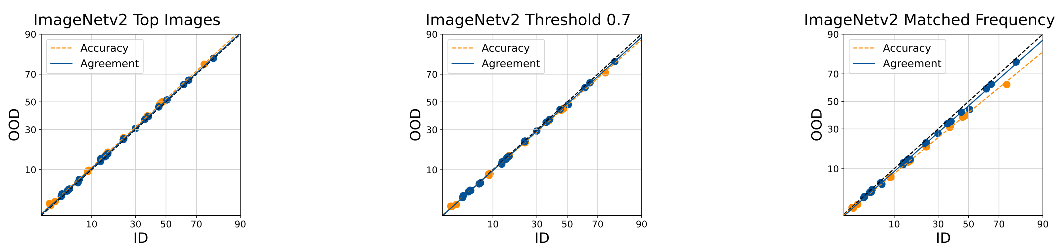


Figure 11: AGL and ACL for the ImageNet V2 shifts with random head initialization fine-tuning.

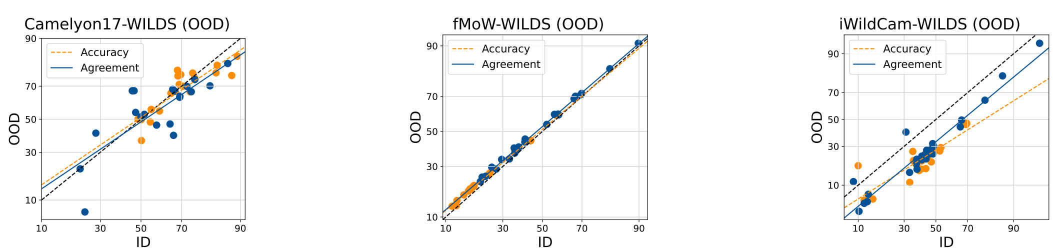


Figure 12: AGL and ACL for 3 benchmarks from the WILDS dataset with random head initialization fine-tuning.