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

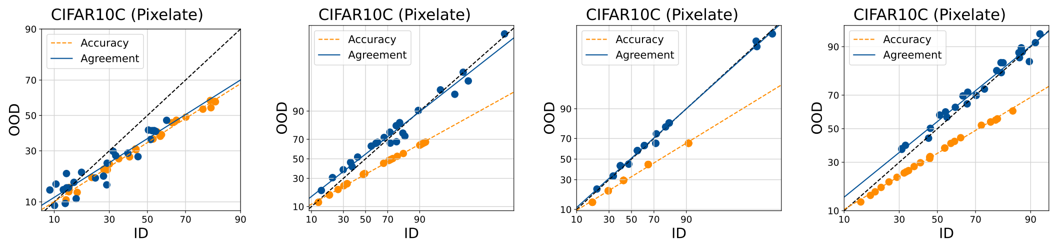
Estimating out-of-distribution (OOD) performance is critical to safely deploying machine learning models. Recently, Baek et al. [2] showed that the phenomenon "agreement-on-the-line" can be a reliable method for predicting OOD accuracy of models in an ensemble consisting largely of CNNs trained from 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 models 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, hyperparameter selection, data subsetting, and data shuffling) contribute to the agreement-on-the-line of the resulting model sets; we find, somewhat surprisingly, that it is typically possible to obtain strong agreement via random initialization of the linear head alone. Next, we study how *multiple* foundation models, pretrained on different data sets but fine-tuned on the same task, may or may not produce agreement; we show, again rather surprisingly, that the diversity of such models is already sufficient and not too disparate for them to 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 model (FM) approaches, where one first pretrains a large model on open world data then fine-tunes or prompts for a specific downstream task, have achieved state-of-the-art results on image classification [32, 25, 44], text classification [6], question answering [11], and others. They are particularly noted for their often strong performance on OOD data, that may vary substantially from the data used for fine-tuning (referred to as the in-distribution (ID) data) [5, 45]. Unfortunately, a significant practical problem arises precisely in this OOD setting: in many cases, one does not have access to labeled OOD data, but only has such data available in *unlabeled* form. Obtaining an explicitly labeled hold-out set for each potential OOD distribution shift is costly and impractical, and thus the field has explored other means for estimating OOD accuracy.

Recently, Baek et al. [2] proposed a method for estimating the accuracy of deep network classifiers on OOD data using unlabeled data alone, by analyzing the *agreement* between pairs of classifiers in some collection (i.e., measuring how often two classifiers make the same prediction, with slight variants for alternate metrics such as F1 score). They showed that empirically, the OOD and ID agreement rates often observe a strong linear correlation, reminiscent of a similar trend for OOD and ID accuracy [30], and that the slopes and biases for these agreement and accuracy lines were often

Submitted to R0-FoMo: Workshop on Robustness of Few-shot and Zero-shot Learning in Foundation Models at NeurIPS 2023. Do not distribute.



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

Figure 1: ACL/AGL for CIFAR10C "Pixelate" with CLIP linear probing fine-tuned using different sources of randomness

extremely similar. These effects are referred to as agreement-on-the-line (AGL) and accuracy-on-the-line (ACL) respectively, and together they provide a simple method for estimating OOD accuracy via unlabeled data alone. In particular, whenever the ID versus OOD accuracy is strongly linearly correlated, one may estimate the linear trend using agreement without labels. Unfortunately, the AGL approach requires a *diverse collection* of classifiers over which to compute agreement: classifiers must vary sufficiently in their incorrect predictions. As an extreme, consider an ensemble where ACL is observed and every pair of models achieves maximal ID and OOD agreement. Namely, say two models observe ID performances of 60% and 80% and OOD performances of 30% and 40%, respectively (linear fit of accuracy is $a_{\text{OOD}} = 0.5a_{\text{ID}}$). Then the maximum agreement rate achievable is 80% ID and 90% OOD. The agreement rate is higher OOD than ID and does not capture the linear trend of ID versus OOD accuracy, in particular the decay under distribution shift. Baek et al. [2] achieve this diversity through 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 foundation model for a downstream task. Such fine-tuning usually involves far fewer gradient steps than training from scratch and even after multiple runs would seemingly lead to highly correlated downstream models, making it unsuitable for AGL-based OOD performance estimation.

In this work, we develop methods for extending AGL performance estimation to foundation models, 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 foundation model. Key to our approach is a detailed empirical study of different types of randomness that we can inject into the fine-tuning process, so as to encourage the needed diversity amongst models. Specifically, we analyze four different potential sources of randomness: 1) random linear head initialization; 2) hyperparameter choice; 3) subsets of the ID data; and 4) permutations of the ID data. We find, somewhat surprisingly, that using different random linear heads is able to much more reliably induce AGL behavior for the resulting classifiers, despite all settings still resulting in the ACL phenomenon alone. We find that these results hold across multiple different foundation models and modalities, holding for CLIP-based image classification and LLM-based QA tasks. The end result is a simple and straightforward method for evaluating OOD performance for a fine-tuned foundation model, applicable to settings where we only one want to fine-tune a single such base model.

Second, we analyze the ability of AGL-based method to predict OOD performance when using *multiple* different pretrained foundation models. Here the likely problem seems to be opposite to what occurred previously: whereas before we expected to have too little diversity in models, 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, this work allows us to substantially expand 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 these settings where training models from scratch on tasks of interest is not feasible.

2 Preliminaries

We are interested in mapping an input $x \in \mathbb{X}$ to a discrete output $y \in \mathbb{Y}$. In particular, we consider fine-tuned foundation models. For a base model B, let $f(\mathsf{B})$ denote a fine-tuned version of B. In this work, we study a variety of base foundation models: GPT2 [32], GPT-Neo, OPT [48], Llama2 [42], and CLIP [33].

Fine-tuning We consider two types of fine-tuning techniques to adapt our foundation models for the downstream task: **linear probing (LP)** and **full fine-tuning (FFT)**. Given features B_θ from the base model B, a linear head v is attached on top to map features to confidence scores $f(\mathsf{B}) = v^\top \mathsf{B}_\phi(x)$. For classification tasks, $f(\mathsf{B}) \in \mathbb{R}^k$ where k refers to the total number of classes, while in extractive question answering tasks, $f(\mathsf{B}) \in \mathbb{R}^{2\times k}$ where k refers to the length of the context. [1] We refer to v as either a linear probe (classification) or span prediction head (question answering). For LP, the features are frozen and only the linear layer v is optimized by gradient updates. On the other hand, FFT updates *all parameters* including the backbone B_ϕ. When infeasible to update all parameters natively, we use parameter efficient *low-rank adaptation* (LoRA) [19] which still effectively updates the feature extractor B_ϕ. In this work, we do not distinguish between LoRA and FFT as they conceptually achieve the same effect, and seem to show similar empirical trends in our studies. Refer to Appendix 6.3 for specific fine-tuning parameters.

OOD performance estimation Given 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 for various tasks: Zero-one loss $\ell_{\text{0-1}}$ for classification and Macro-averaged F1 score ℓ_{F1} for question answering.

Accuracy and agreement on the line ACL is a striking phenomenon, however, it does not immediately provide a practical method to estimate OOD performance—computing the slope and bias of the linear correlation requires access to labeled samples from $\mathcal{D}_{\text{ood}}$. Baek et al. [2] propose AGL which uses *agreement between models* rather than accuracy to estimate OOD performance.

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

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

where ℓ is the appropriate performance metric of interest (e.g. 1 minus the zero-one loss for classification). Note that while accuracy requires access to the labels y, agreement only requires access to unlabeled data and a pair of models. The key observation in Baek et al. [2] is that ACL and AGL share the *same linear slope and bias*. More details on AGL can be found in Appendix 6.2 while a discussion on prior OOD performance estimation methods is in Appendix 6.9.

Since computing agreement does not require labels, one can compute the slope and bias using unlabeled data, then estimate the OOD performance when AGL and ACL hold by linearly transforming the ID validation performance. We refer the reader to [2] for formal ALine algorithms (ALine-S and ALine-D) to use AGL for OOD performance estimation (Appendix 6.7). Note that ACL is a prerequisite for good OOD performance estimation via ALine. However, as ACL only occurs coupled with AGL, we can only rely on ALine when agreements show strong linear correlation.

3 Predicting OOD performance: single base foundation model

Our first setting of interest concerns the case where we have a *single* foundation model that we would like to fine-tune for a given downstream task. Since AGL-methods cannot be applied to a single classifier (requiring a collection of classifiers over which to compute agreement between pairs), we need some method to introduce variability amongst multiple variants of this base model. Such variability can be introduced in many ways, but an overriding concern is that even with some randomness in the fine-tuning process, it may not be enough to overcome the underlying similarities in predictions due to the same base foundation model.

To address this problem, in this section we evaluate multiple different possible sources of diversity in the fine-tuning process, to see what approach (if any) can lead to AGL. Specifically, we analyze four

[1] The output of the foundation model for extractive QA is $2 \times k$ as the model predicts both the start and end of the context span that contains the ground truth answer.

Table 1: OOD accuracy prediction MAE (%) for image classification

OOD Dataset	ALine-D	ALine-S	Naive Agr	ATC	AC	DF
CIFAR10C (averaged across shifts)	**3.34**	3.40	15.46	8.00	23.37	10.85
CIFAR10.1 (averaged across v4, v6)	**0.63**	0.87	17.59	2.83	29.93	4.26
CIFAR100C (averaged across shifts)	3.11	**2.87**	11.94	4.04	21.86	10.48
ImageNetC (averaged across shifts)	**2.16**	2.87	11.94	4.04	21.86	10.48
ImageNet V2 (averaged across 3 format)	**1.30**	2.56	9.86	4.31	19.85	9.13
fMoW-WILDS (val OOD split)	0.99	**0.91**	20.39	2.66	9.59	1.26
Camelyon17-WILDS (val OOD split)	4.68	**4.50**	9.75	7.01	11.01	6.35
iWildCam-WILDS (val OOD split)	**4.91**	4.99	13.19	8.84	12.26	10.23

possible methods for introducing diversity into the fine-tuning process (which then lets us create a differentiated collection of classifiers by repeating the fine-tuning process multiple times):

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. **Different fine-tuning hyperparameters.** We use a variety of different learning rates and weight decays to encourage diversity of the resulting models.
3. **Data subsetting.** We present each fine-tuned model to be fine-tuned with an independent subset of the (ID) fine-tuning data.
4. **Data shuffling.** We present the same data to each model, but shuffle the order for the data differently within each fine-tuning optimization run.

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 the same (but random) initialization, but the data used for training is seen in different orders; and so on.

When models are trained from scratch, it is well established that independent data subsetting tends to lead to the greatest diversity of classifiers [31]. Nonetheless, in this setting we find rather surprisingly, that *just using different random linear heads* achieves the highest diversity. We show that this finding persists over multiple models, multiple tasks, and indeed multiple modalities entirely.

3.1 Experimental setup

Models Given its well-established 0-shot capabilities, we use linear probing atop CLIP [33], specifically the ViT-B/32 model trained on LAION-2B [40] for our image classification tasks. For our QA tasks, we evaluate a collection of 50 fully fine-tuned models, wherein each model is obtained by fine-tuning from the same checkpoint of GPT2-Medium (links to the base FMs are in Appendix 6.8).

Datasets We fine-tune and test our models on several different image classification datasets. We fine-tune models on CIFAR10 [23], and then evaluate on CIFAR10C and CIFAR10.1. We repeat the same for CIFAR100 [22], ImageNet-1k [38] and their respective shifted datasets CIFAR100C, ImageNetC [16], and ImageNetV2 [36]. We additionally validate our finding by testing on three real world shifts from the WILDS benchmark (FMoW, iWildCam, Camelyon17) [21]. For extractive QA, we fine-tune on the SQuAD v1.1 dataset [34]. We evaluate the fine-tuned LLMs on four distribution shifts present in SQuAD-Shifts (New Wiki, New York Times, Amazon, and Reddit) [29].

3.2 Results

In Figure 1, we observe the ID and OOD agreements and accuracies of linear probes trained on top of CIFAR10 CLIP representations. One may suspect that in this setting, the simple linear models would agree highly and AGL may break. For example, Baek et al. [2] has shown previously that AGL is a phenomenon that is specific to neural networks (e.g. linear models trained on top of the flattened CIFAR10 images do not observe AGL). Indeed, while ACL holds with strong correlation for each of the ensembles constructed with the four sources of diversity, AGL does not hold for all

Table 2: ALine-D MAE for fine-tuning with different sources of randomness for extractive QA

Source of Diversity	SQuAD-Shifts Amazon (%)	SQuAD-Shifts Reddit (%)
Random Linear Heads	**0.69**	**0.79**
Different fine-tuning hyperparameters	2.55	2.06
Data Shuffling	4.18	4.32
Data Subsetting	5.2	4.71

ensembles. However, AGL interestingly does hold strongly for the case of random head initialization. Thus, contrary to the findings of Baek et al. [2], even in linear models, when on top of neural network features (in this case CLIP) with the *right type of diversity*, one may observe AGL and use the related ALine algorithms to predict OOD estimation.

On the other hand, for the other sources of diversity, we observe a consistent trend where agreement is also strongly linearly correlated but the OOD agreement rate is too high, and the slope of the linear fit of agreement surpasses that of accuracy. In fact, all ensembles achieved through data subsetting, data shuffling, and hyperparameter changes, *strictly lie on the diagonal* $y = x$ *line*. In some sense, this is particularly very surprising for linear models. Intuition may suggest that independent data subsetting leads to the greatest diversity as the other sources of diversity optimize over the same convex landscape. Yet, even when we distribute the number of epochs trained to achieve a wide spread of ID accuracy models, AGL only holds for models that start at different random initialization. The averaged Mean Absolute Error (MAE) between the AGL-interpolated and actual OOD accuracies for the CIFAR10C shifts with these sources of diversity, can be found in Appendix 6.4, further quantifying these visually apparent results. We refer the reader to Appendix 6.10 which contains the ACL/AGL plots with the random-head initialized ensembles for other datasets. Furthermore, Table 1 shows the averaged MAE for the OOD accuracies as calculated using the ALine algorithms and other OOD performance estimation methods for the image classification dataset shifts. We find that when ACL holds, ALine estimates the OOD accuracy significantly better than baselines, thus lending support for utilizing AGL *induced by random initialization* to evaluate the performance of lightly fine-tuned models.

We similarly find that not all sources of diversity are equally likely to yield sufficient diversity in fully fine-tuned LLMs for extractive QA. As seen in CLIP linear probing, varying the random initialization of the span head consistently provides sufficient stochasticity during fine-tuning to obtain a suitably diverse ensemble that demonstrates AGL and enables accurate prediction of OOD accuracy. On the other hand, stochasticity arising from data shuffling, data subsetting, and from varying hyperparameters may not always yield an ensemble that is amenable to accurately estimating OOD accuracy (see Table 2). Specifically, these sources tend to yield ensembles with correlated errors which results in the agreement line often lying above the accuracy line and on the $y = x$ line, although the trend is less stark than the one observed in image classification by linear probing. We refer the reader to Appendix 6.5 to observe these trends on all shifts within the SQuaD-Shifts dataset.

4 Predicting OOD performance: multiple foundation models

Alternatively, with multiple base foundation models pretrained on different text corpora, agreement-on-the-line may potentially fail due to an opposite failure mode of different model pairs disagreeing too highly or in unstructured ways on OOD data. Moreover, models heavily pretrained on different corpora may lie on different accuracy lines to begin with. But to the contrary, we observe that foundation models fine-tuned from a wide range of base models *observe both ACL and AGL.*

4.1 Experimental Setup

Models We fine-tune 41 models on the extractive QA task with SQuAD v1.1 as the ID dataset and observe their OOD performance on SQuAD-Shifts; specifically OPT-125M, OPT-350M, OPT-1.3B, GPT2-XL, GPT2-Large, GPT2-Medium, GPT2, GPT-Neo-135M, Llama2-7B, Alpaca-7B, and Vicuna-7B. OPT was pretrained on a wide variety of data including BookCorpus [49], Stories [43], a subset of PILE [13], CCNews v2 corpus, and PushShift.io Reddit [3]. GPT2 was pretrained on BookCorpus while GPT-Neo was trained on PILE. Llama2 was trained on an undisclosed set of publicly available data. Finally, Alpaca and Vicuna are additionally trained from Llama2 on instruction-following demonstrations and user-shared conversations from ShareGPT, respectively.

4.2 Results

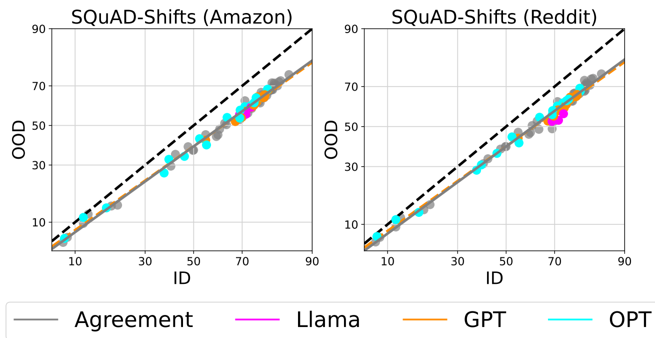


Figure 2: ACL/AGL when using different base models for SQuAD-Shifts

In Figure 2, we see that base LLM models pretrained on different sources of text corpora lead to fine-tuned models that lie on the *same linear trend in accuracy* on SQuAD. This is in contradiction to previous works that indicate (benchmarking the performance of foundation models on image classification tasks [33, 41]) that models heavily pretrained on differerent image corpora may lie on different lines. We suspect that the pretraining datasets for the models in our study exhibit much more homogeneity. Second, the ID versus OOD agreement for pairs of fine-tuned models (even with different bases) retain a strong linear correlation, and the slope and bias closely match that of accuracy; i.e., different pretraining does not break AGL (also see Appendix 6.6). As reported in Table 3, using ALine-S and ALine-D with AGL yields better OOD estimation performance than other baselines over SQuAD-Shifts overall.

Table 3: OOD accuracy prediction MAE (%) for extractive QA

OOD Dataset	ALine-D	ALine-S	Naive Agr	ATC	AC	DF
SQuAD-Shifts Reddit	**0.76**	1.19	9.18	6.21	24.35	2.99
SQuAD-Shifts Amazon	**0.97**	1.44	9.22	7.15	24.86	3.69
SQuAD-Shifts New York Times	**0.52**	0.68	9.56	1.32	19.94	1.54
SQuAD-Shifts New Wiki	1.97	1.98	10.01	2.42	21.03	**0.71**

5 Conclusion

We develop methods for extending AGL to foundation models to enable OOD performance prediction in this emerging paradigm. We found that applying AGL directly may sometimes fail and properly utilizing this phenomenon for performance estimation requires careful tuning of the distribution of models in the 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 head initialization, is crucial for lightly fine-tuned foundation models. Second, though Baek et al. [2] posed AGL as a model centric phenomenon that is specifically only observed in neural network ensembles, we find that linear models could 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, a phenomenon that is of independent interest. Recent works that study the effect of pretraining on ACL [33, 41] indicate that models pretrained on different datasets 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 (OPT, GPT2, GPT2-Neo, and Llama2) lie on the *same* accuracy and agreement line. This is particularly intriguing because it goes against the common wisdom that the amount of pretraining data determines the effective robustness. Additionally, though our findings help us utilize AGL for predicting the performance of foundation models, they also raise potential concerns about the robustness of fine-tuned foundation models – even light linear probing over these base models could lead to models disagreeing highly on OOD data. 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] Jiefeng Chen, Frederick Liu, Besim Avci, Xi Wu, Yingyu Liang, and Somesh Jha. Detecting errors and estimating accuracy on unlabeled data with self-training ensembles. *arXiv preprint arXiv:2106.15728*, 2021.

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

[9] Weijian Deng and Liang Zheng. Are labels always necessary for classifier accuracy evaluation? In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15064–15073. IEEE Computer Society, 2021. doi: 10.1109/CVPR46437.2021.01482.

[10] Weijian Deng, Stephen Gould, and Liang Zheng. What does rotation prediction tell us about classifier accuracy under varying testing environments? *arXiv preprint arXiv:2106.05961*, 2021.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

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

[17] 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.

[18] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *International Conference on Learning Representations*, 2017.

[19] 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.

[20] Yiding Jiang, Vaishnavh Nagarajan, Christina Baek, and J Zico Kolter. Assessing generalization of sgd via disagreement. *International Conference on Learning Representations*, 2022.

[21] 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.

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

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

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

[25] 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.

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

[27] Omid Madani, David Pennock, and Gary Flake. Co-validation: Using model disagreement on unlabeled data to validate classification algorithms. *Advances in neural information processing systems*, 17, 2004.

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

[29] 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.

[30] 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.

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

[32] 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.

[33] 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.

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

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

[36] 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.

[37] 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.

[38] 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`.

[39] Sebastian Schelter, Tammo Rukat, and Felix Biessmann. Learning to validate the predictions of black box classifiers on unseen data. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, page 1289–1299, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450367356.

[40] 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.

[41] 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.

[42] 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.

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

[44] 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.

[45] 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.

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

[47] Yaodong Yu, Zitong Yang, Alexander Wei, Yi Ma, and Jacob Steinhardt. Predicting out-of-distribution error with the projection norm, 2022.

[48] 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.

[49] 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.

6 Appendix

6.1 Background on OOD accuracy estimation

There is a rich literature on OOD performance estimation, with a variety of proposed approaches. One family of approaches attempts to quantify the degree of distribution shift through data and/or model dependent metrics e.g. uniform convergence bounds using metrics such as $\mathcal{H}$-divergence [4, 28, 8, 24]. However, these approaches only provide upper bounds on the OOD error, and these bounds tend to be loose when evaluated on deep networks used in practice [30].

Another line of work looks at leveraging the model's own softmax predictions a.k.a the model's confidence to predict the OOD performance [17, 16, 14, 12, 15]. Since models are typically over-confident, it is common practice to first calibrate these models using ID validation data to further improve the reliability of such approaches. 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 [14]. There are other heuristic OOD estimation strategies that are reported to work in some datasets such as using performance on auxiliary self-supervised tasks [39, 9, 10, 47] or leveraging characteristics of self-trained models on the OOD data [47, 7].

6.2 Accuracy on the Line

In recent work, Baek et al. [2] propose a different approach for estimating OOD performance, that is empirically reliable across a variety of shifts and outperforms prior approaches. This approach is based on an earlier intriguing observation from [30, 35, 36, 37, 46, 41, 29]—there is a strong linear correlation between the ID and OOD performance of models for several distribution shifts. We call this phenomenon "accuracy-on-the-line" (ACL). ACL has been observed for image classification shifts such as some common corruptions on CIFAR10, ImageNetV2, FMoW-WILDS, and question answering shifts such as SQuAD-Shifts. However, ACL does not always hold e.g. Camelyon-WILDS [30] and SearchQA [1] do not show ACL.

6.3 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 [26]. 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 4: 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
Finetuning hyperparameters	LR: $2 \times 10^{-6} - 2 \times 10^{-4}$ WD: $1 \times 10^{-5} - 1 \times 10^{-2}$	DS: fixed LS: fixed DP: 90% EP: 0.2 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 5: Finetuning specifics for image classification (LR: learning rate, WD: weight decay, LS: linear head initialization seed, DS: data shuffling seed, DP: data subsetting proportion, EP: epochs, B: batch size)

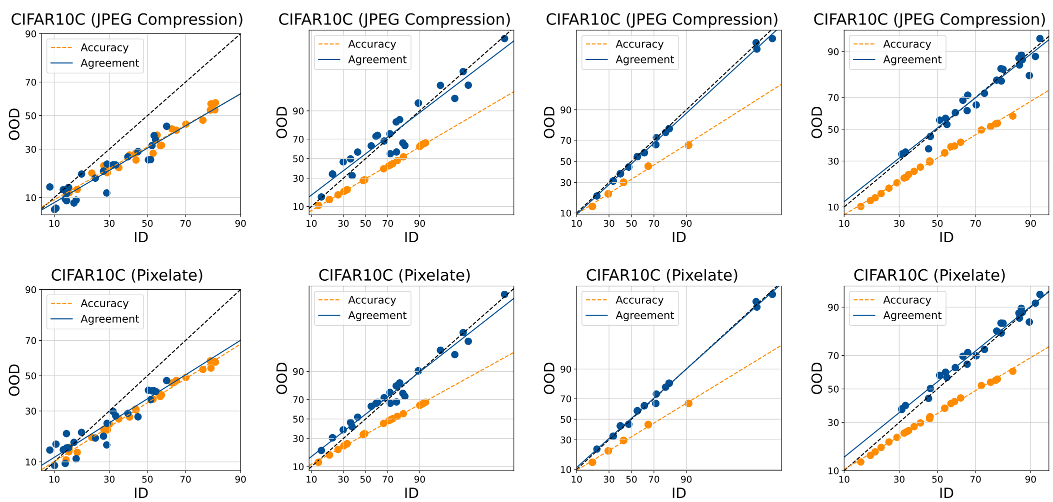
Source of Diversity	**CLIP + ViT-B/32 (LAION-2B)**	
	Varied	**Fixed**
Random linear heads	LS: varied	LR: different per dataset WD: 0 DS: fixed DP: 100% EP: 1–100 B: 1024
Finetuning hyperparameters	LR: $1 \times 10^{-4} - 1 \times 10^{-3}$ WD: $0 - 0.5$	DS: fixed LS: fixed DP: 100% EP: 1–100 B: 1024
Data 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

6.4 Sources of Diversity (Image Classification)

Figure 3 shows the four 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 for CLIP linear fine-tuned for CIFAR10 image classification with different sources of diversity. Note that the reported MAE is averaged across all 19 CIFAR10C shifts.

Source of Diversity	CIFAR10C (%)
Random linear heads	3.96
Different fine-tuning hyperparameters	12.47
Data shuffling	11.09
Data subsetting	10.91

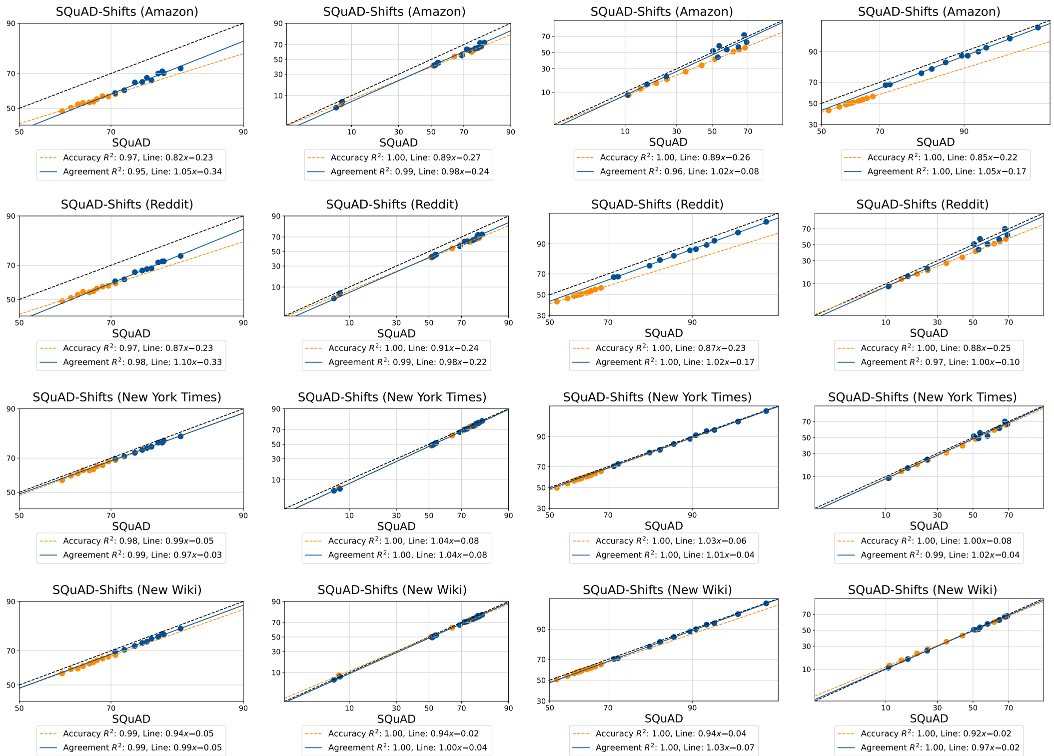


(a) Random Head (b) Hyperparameters (c) Data Shuffling (d) Data Subsetting

Figure 3: The ACL and AGL plots for the "JPEG Compression" (top row) and "Pixelate" (bottom row) fine-tuned using different sources of randomness

6.5 Sources of Diversity (Question Answering)

Figure 4 shows the four sources of diversity for all SQuAD-Shifts OOD datasets.



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

Figure 4: 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.

6.6 Multiple Foundation Models

Figure 5 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 6.8.

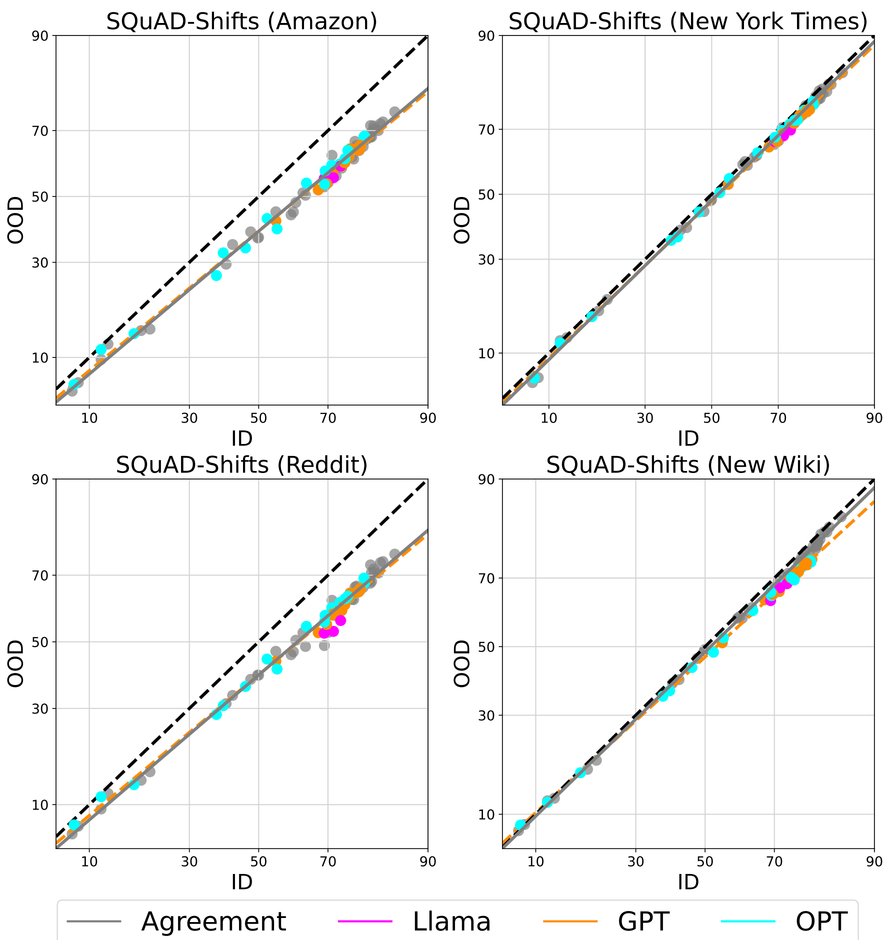


Figure 5: AGL when using different base models for SQuAD-Shifts

6.7 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.

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

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

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

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.

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

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

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

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

6.8 Model Links

Here are the links to the pretrained base foundation models we finetuned: CLIP (`https://github.com/mlfoundations/open_clip`), 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`)

6.9 OOD Accuracy Estimation Methods (Baselines)

With sufficient diversity residing in the ensemble, we observe that ALine succeeds over other OOD estimation baselines in terms of predicting the performance of the models in the ensemble. We compare the algorithms ALine-S and ALine-D [2] on this sufficiently diverse ensemble of models to other existing methods that estimate the accuracy of OOD performance: ATC [14], AC [18] and DOC-Feat [15] that utilize model confidence to estimate OOD accuracy in addition to directly using agreement to predict accuracy, dubbed naive agreement [20] [27]. We observe that with sufficient diversity in the ensembles, variants of the ALine algorithm surpass confidence/probability based methods by achieving the lowest error of predicting the OOD performance of fine-tuned foundation models on all tasks as seen in Table 7. For this comparison, the lowest error rate picked from the errors found prior and post the application of temperature scaling is reported for confidence based methods. Though temperature scaling can be applied to calibrate models in terms of their accuracy, calibrating models for the F1 score by temperature scaling is not directly obvious. As a result, we observe that for extractive QA datasets, confidence based methods particularly suffer.

Table 7: OOD accuracy prediction MAE (%) of various methods

OOD Dataset	ALine-D	ALine-S	Naive Agr	ATC	AC	DF
SQuAD-Shifts Reddit	**0.76**	1.19	9.18	6.21	24.35	2.99
SQuAD-Shifts Amazon	**0.97**	1.44	9.22	7.15	24.86	3.69
SQuAD-Shifts Nyt	**0.52**	0.68	9.56	1.32	19.94	1.54
SQuAD-Shifts New Wiki	1.97	1.98	10.01	2.42	21.03	**0.71**
CIFAR10C (averaged across shifts)	**3.34**	3.40	15.46	8.00	23.37	10.85
CIFAR10.1 (averaged across v4, v6)	**0.63**	0.87	17.59	2.83	29.93	4.26
CIFAR100C (averaged across shifts)	3.11	**2.87**	11.94	4.04	21.86	10.48
ImageNetC (averaged across shifts)	**2.16**	2.87	11.94	4.04	21.86	10.48
ImageNet V2 (averaged across 3 format)	**1.30**	2.56	9.86	4.31	19.85	9.13
fMoW-WILDS (val OOD split)	0.99	**0.91**	20.39	2.66	9.59	1.26
Camelyon17-WILDS (val OOD split)	4.68	**4.50**	9.75	7.01	11.01	6.35
iWildCam-WILDS (val OOD split)	**4.91**	4.99	13.19	8.84	12.26	10.23

6.10 CLIP

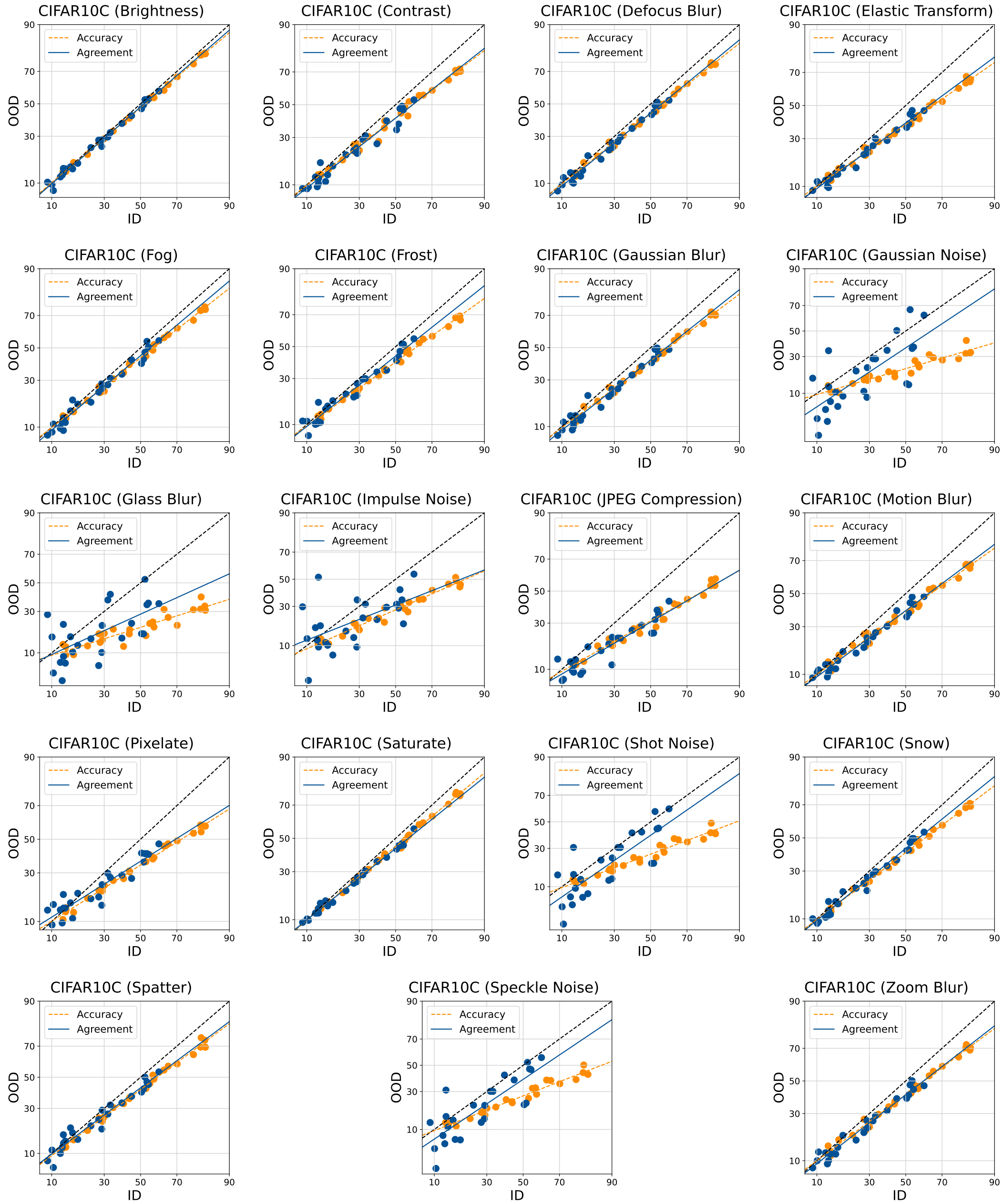


Figure 6: AGL and ACL for all CIFAR10C shifts with random head initialization fine-tuning.

Figure 7: AGL and ACL for the CIFAR10.1 shifts with random head initialization fine-tuning.

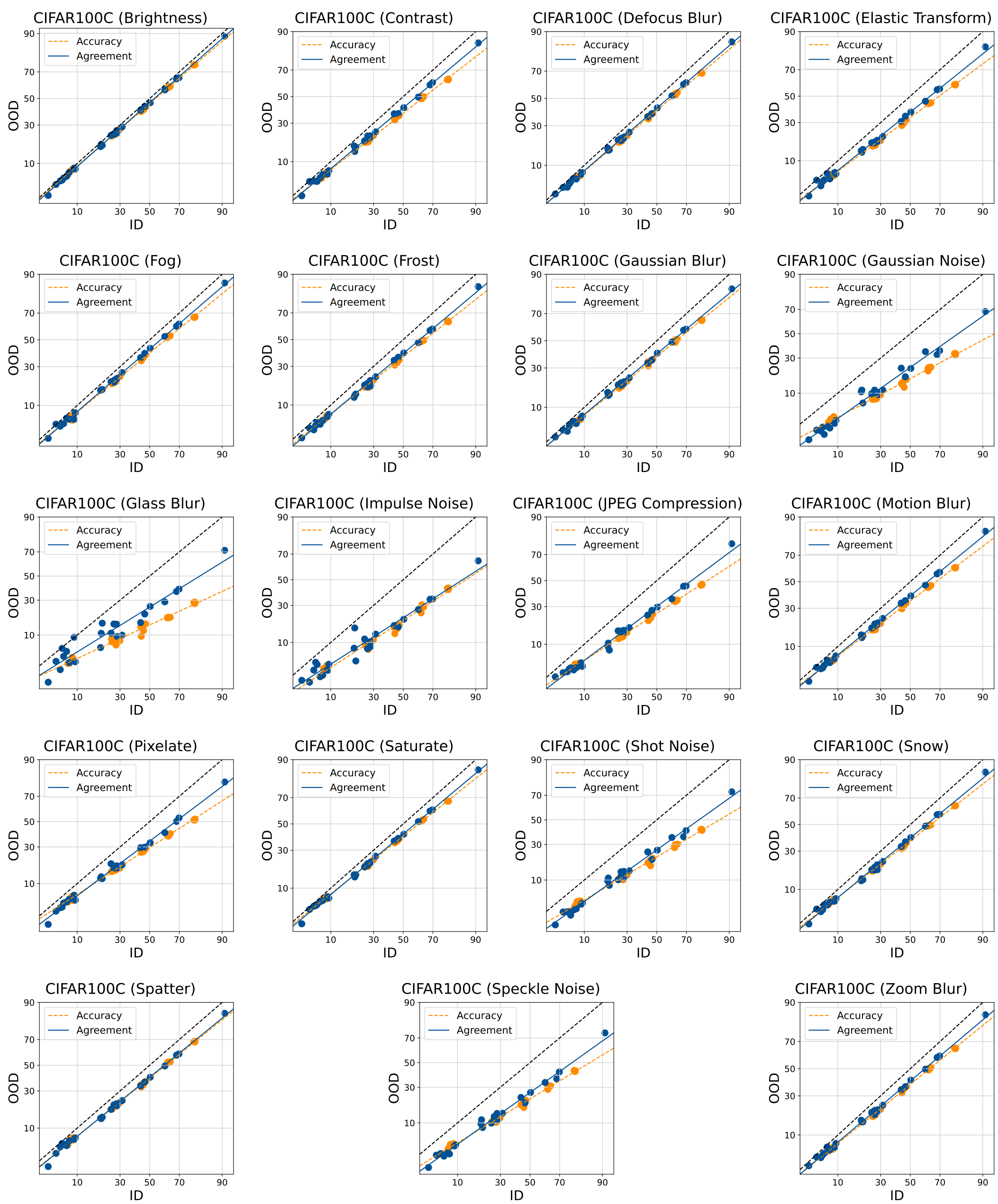


Figure 8: AGL and ACL for the CIFAR100C shifts with random head initialization fine-tuning.

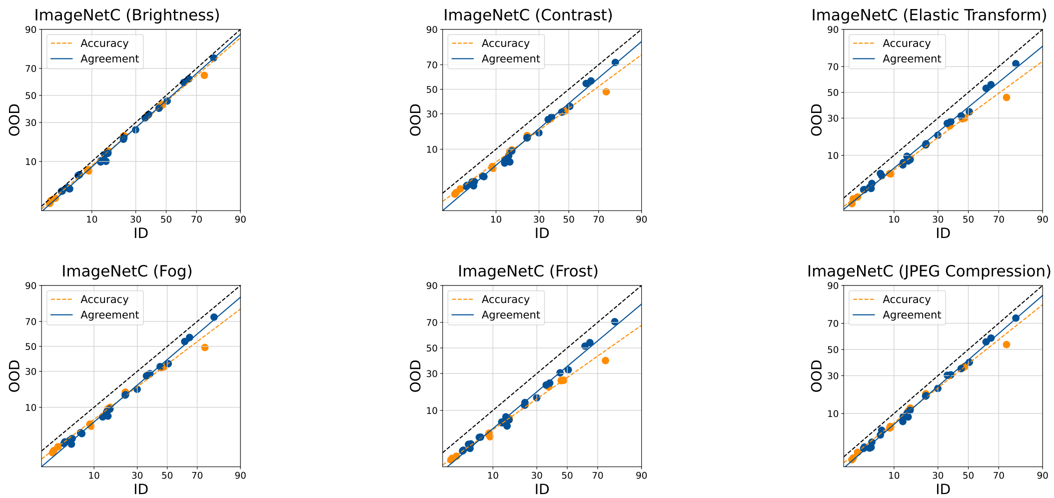


Figure 9: AGL and ACL for the ImageNetC shifts with random head initialization fine-tuning.

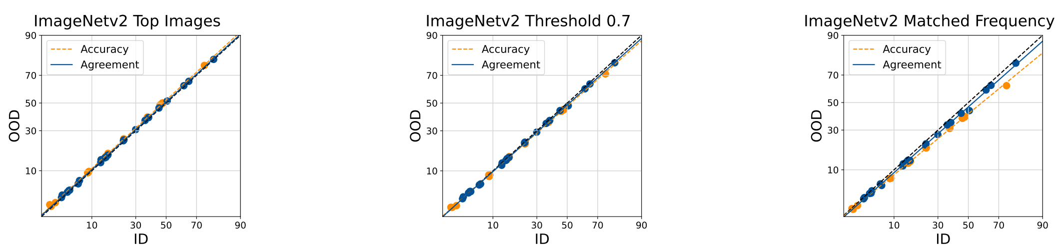


Figure 10: AGL and ACL for the ImageNet V2 shifts with random head initialization fine-tuning.

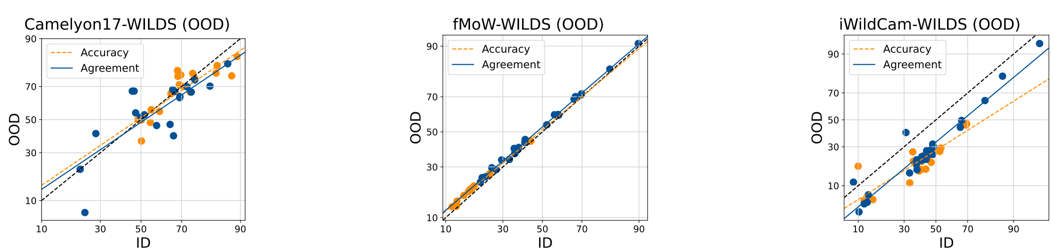


Figure 11: AGL and ACL for 3 benchmarks from the WILDS dataset with random head initialization fine-tuning.