Are Sample-Efficient NLP Models More Robust?

Nelson F. Liu♠ Ananya Kumar♠ Percy Liang♠ Robin Jia♥
♠Computer Science Department, Stanford University, Stanford, CA
♥Department of Computer Science, University of Southern California, Los Angeles, CA
{nflu,ananya,pliang}@cs.stanford.edu
robinjia@usc.edu

Abstract

Recent results in image classification and extractive question answering have observed that pre-trained models trained on less in-distribution data have better out-of-distribution performance. However, it is unclear how broadly these trends hold. We conduct a large empirical study across three tasks, three broadly-applicable modeling interventions (increasing model size, using a different adaptation method, and pre-training on more data), and 14 diverse datasets to investigate the relationship between sample efficiency (amount of data needed to reach a given ID accuracy) and robustness (how models fare on OOD evaluation). We find that higher sample efficiency is only correlated with better average OOD robustness on some modeling interventions and tasks, but not others. On individual datasets, models with lower sample efficiency can even be more robust. These results suggest that general-purpose methods for improving sample efficiency are unlikely to yield universal OOD robustness improvements, since such improvements are highly dataset- and task-dependent. Even in an era of large, multi-purpose pre-trained models, task-specific decisions may often be necessary for OOD generalization.

1 Introduction

NLP models perform well when evaluated on data drawn from their training distribution (in-distribution / ID), but they typically suffer large drops in performance when evaluated on data distributions unseen during training (out-of-distribution / OOD; Blitzer, 2008).

How does exposure to ID training examples affect the ID-OOD gap? If two models have the same ID performance, will models trained on fewer ID examples (higher sample efficiency) also have higher OOD performance (higher robustness)? At one extreme, zero-shot models will not learn ID-specific patterns because they are not exposed to any labeled ID examples. Similarly, few-shot models trained on very few ID examples may also rely less on ID-specific patterns; if a model never sees the token “cat” while training on SNLI, then it will not learn that its presence is spuriously predictive of the contradiction label (Gururangan et al., 2018; Utama et al., 2021). Supporting this intuition, recent work in image classification (Radford et al., 2021) and extractive question answering (Awadalla et al., 2022) show that zero-shot inference and few-shot fine-tuning improve average robustness across a range of OOD test sets. However, it is unclear how universal these trends are across various tasks and methods for reducing exposure to ID examples, or how predictive they are for any individual test set of interest. Figure 1 illustrates this central question.

We conduct a broad empirical study over 14 datasets across three tasks to investigate the relationship between exposure to ID training examples (sample efficiency) and robustness. We experiment with three modeling interventions that improve sample efficiency: (1) using natural language prompts for zero-shot prediction and during fine-tuning (Brown et al., 2020; Schick and Schütze, 2021; Gao et al., 2021); (2) fine-tuning models of increasing size; (3) fine-tuning models pre-trained on increasing amounts of data.

We find that higher sample efficiency is only sometimes correlated with better robustness, and the effect of specific modeling interventions varies by task. For example, increasing pre-trained model size substantially improves sample efficiency and results in higher average robustness in sentiment experiments, but these sample efficiency gains do not translate to higher average robustness in NLI and extractive QA experiments. On individual datasets, models with better sample efficiency can even be less robust (e.g., increasing model size when training on SST-2 and evaluating OOD on IMDb).

Overall, these results indicate that general-
Figure 1: In this example, model B has higher sample efficiency than model A, since model B requires less ID training data to reach a given ID performance threshold (top). In this particular example, model B is also more robust than model A (bottom), since it has higher OOD performance for a given ID performance threshold.

Figure 1 provides a schematized example, with model B having better sample efficiency and robustness than model A.

3 Experimental Setup

We study three modeling interventions—using natural language prompts, increasing pre-trained model size, and pre-training on more data—on 14 total datasets spanning natural language inference (NLI), sentiment analysis, and extractive question answering (QA). See Appendix A for further details about experimental settings.

Tasks and Datasets. In our natural language inference (NLI) experiments, we use MultiNLI (Williams et al., 2018), SNLI (Bowman et al., 2015), and MedNLI (Romanov and Shivade, 2018). For sentiment analysis, we use IMDb reviews Maas et al. (2011), SST-2 (Socher et al., 2013), and reviews from the “Movies and TV” subsection of the Amazon Reviews corpus (Ni et al., 2019). Lastly, for extractive question answering, we use SQuAD (Rajpurkar et al., 2016), NaturalQuestions (Kwiatkowski et al., 2019), TriviaQA, BioASQ (Tsatsaronis et al., 2015), and the four SQuAD-Shifts test sets (Miller et al., 2020).

Modeling Interventions. To understand the effect of a particular modeling intervention on sample efficiency and robustness, we evaluate pre-trained models that differ only along the axis of interest (e.g., model size or fine-tuning method). Since the optimal fine-tuning hyperparameters depend on the ID training dataset size, we separately tune hyperparameters for each model on each training dataset subsample size, taking the models that achieve the best held-out ID performance for each setting. See

2 Measuring Sample Efficiency and Robustness.

Consider two data distributions $D_{id}$ and $D_{ood}$. Let $M$ be a model trained on examples drawn from $D_{id}$ (i.e., the ID training data). We study the relationship between three properties of $M$: (1) the number of ID examples it was trained on; (2) $M$’s performance on held-out examples from $D_{id}$ (i.e., the ID performance); (3) $M$’s performance on examples from $D_{ood}$ (i.e., the OOD performance).

Let $M_1$ and $M_2$ be two models with equivalent performance on held-out ID data. If $M_1$ was trained on fewer ID examples than $M_2$, then it has higher sample efficiency. If $M_1$ has higher OOD performance than $M_2$, it has higher effective robustness (henceforth “robustness”; Taori et al., 2020). Comparing models with equivalent ID performance controls for its effect on OOD performance, since improving ID performance usually yields commensurate improvements on OOD performance—in this study, we focus on OOD performance improvements beyond what is expected from ID gains.

Satisfying this equivalent-ID constraint is often difficult in practice: given an arbitrary model $M_1$ and its corresponding ID performance, it is difficult to produce a different model $M_2$ with identical ID performance. Rather than explicitly training models to identical ID performance, we train models on varying-size subsamples of a given ID dataset and interpolate between the results to estimate (1) the number of labeled ID training examples necessary to achieve a particular ID performance (sample efficiency) and (2) OOD performance, given ID performance (robustness). These interpolated curves approximate the ideal setting of training a model for every possible ID value. Figure 1 provides a schematized example, with model B having better sample efficiency and robustness than model A.
Appendix B for details about hyperparameter optimization.

4 Results and Discussion

Our results show that models with higher sample efficiency may not necessarily have higher average OOD robustness—different tasks and modeling interventions affect robustness in different ways (Figures 2-4). For example, prompt-based fine-tuning consistently improves both sample efficiency and average robustness, but only in low-data settings (Figure 2). In contrast, increasing model size improves sample efficiency across the range of training dataset sizes and tasks, but only improves average robustness on sentiment analysis (Figure 3). On individual datasets, we even observe cases where models with lower sample efficiency have higher robustness (Figure 3d). See Appendix C for full results on every ID-OOD setting.

Natural Language Prompting. We compare BERT\textsubscript{BASE} models using (1) standard fine-tuning, (2) prompt-based fine-tuning, and (3) zero-shot prompting. We also compare these results with zero-shot prompting of text-davinci-001, a much larger model trained on substantially more data. We run experiments on NLI and sentiment analysis, since extractive QA is not amenable to prompt-based fine-tuning with masked language models.

Figures 2a and 2b plot the average performance on all OOD datasets as a function of ID performance and the ID performance as a function of the number of labeled training examples. Sample efficiency improvements from prompt-based fine-tuning also translate to higher average robustness. However, these improvements only apply in the few-shot setting. As the size of the training dataset increases, the improvements in sample efficiency and average robustness steadily diminish. When using sufficiently large training datasets, models trained with prompt-based fine-tuning yield essentially the same sample efficiency and robustness results as standard fine-tuning (∼1K examples for NLI, ∼130 examples for sentiment).

However, results on individual OOD test sets can significantly differ from averaged-OOD trends. For example, Figure 2c shows that prompt-based fine-tuning on MNLI and evaluating on SNLI improves sample efficiency in the few-shot setting but without any robustness improvements.

Surprisingly, we also find that zero-shot inference does not necessarily improve average robustness over prompt-based fine-tuning—zero-shot performance lies on or below the trend line formed by prompt-based fine-tuning, despite not using any ID-specific data at all. See Appendix C.1 for full results of increasing pre-trained model size for every ID-OOD setting.

Increasing Pre-Trained Model Size. We run experiments with the checkpoints of Turc et al. (2019), who pre-train BERT models with various numbers of transformer layers (L) and hidden embedding sizes (H). We run experiments on NLI, sentiment analysis, and extractive QA to compare pre-trained models of five sizes: (1) Large (L=24, H=1024), (2) Base (L=12, H=768), (3) Medium
(L=8, H=512), (4) Small (L=4, H=512), and (5) Tiny (L=2, H=128). Although increasing the pre-trained model size improves sample efficiency on every task, it does not always improve average robustness (Figure 3). In particular, increasing model size minimally affects average robustness in NLI and extractive QA (Figure 3a,3c), but substantially improves average robustness on sentiment analysis (Figure 3b). However, results on individual ID-OOD pairs can again significantly differ from average OOD performance trends. For example, when training on SST-2 and evaluating on IMDb, larger models actually have lower OOD performance. This occurs because SST-2 examples (single sentences) are significantly shorter than IMDb examples (paragraphs). As a result, models trained on the shorter SST-2 examples struggle when evaluated on IMDb because this particular ID-OOD pair requires length extrapolation, and increasing pre-trained model size does not help models generalize to longer input sequences. As a result, effective robustness decreases because larger models have higher ID (SST-2) performance but unchanged OOD (IMDb) performance. See Appendix C.2 for full results of natural language prompting for every ID-OOD setting.

**Pre-Training on More Data.** We conduct NLI, sentiment, and QA experiments with RoBERTa Large models pre-trained on 10M, 100M, and 1B tokens of web text (Zhang et al., 2021).

Pre-training on more data consistently improves sample efficiency, but only yields average robustness improvements in NLI and sentiment analysis (Figure 4a,b). In extractive QA experiments, varying the amount of pre-training data does not significantly change average robustness (Figure 4c). Again, we find that results on average OOD performance are not predictive of results on individual test sets—despite unchanged average OOD robustness when pre-training on more data, OOD performance can be higher on individual extractive QA test sets (e.g., SQuAD → BioASQ; Figure 4d). See Appendix C.3 for full results of pre-training on RoBERTa Large.
Figure 4: Pre-training on more data is an effective method for improving sample efficiency, but these sample efficiency improvements are not always accompanied by robustness improvements. In NLI and sentiment analysis experiments, these sample efficiency gains correlate with improved average robustness (a,b). However, there are no average robustness gains in extractive QA (c). Despite no average robustness improvement in extractive QA, pre-training on more data can still improve robustness on particular test sets (e.g., BioASQ; d).

5 Conclusion

We study the relationship between sample efficiency and robustness across three tasks and three modeling interventions, finding that sample efficiency improvements often fail to translate to improved robustness. As larger models quickly become more sample efficient, our results caution that sample efficiency and robustness are different axes of improvement and that optimizing for sample efficiency will not necessarily always yield robustness gains.

Acknowledgments

We thank the anonymous reviewers for their feedback and comments that helped improve this work. We also thank Kevin Lin and Eric Wallace for their feedback and useful discussions. NL was supported by an NSF Graduate Research Fellowship under grant number DGE-1656518. Other funding was provided by a PECASE Award and the Open Philanthropy Project.

Limitations

Our study focuses on natural language understanding tasks, though it may also be interesting to study whether these trends apply in natural language generation tasks (e.g., summarization). In particular, it’s possible that zero- or few-shot pre-trained models may do better on generation tasks because these tasks are more similar to the models’ original pre-training objective (e.g., language modeling).

Furthermore, we compared few-shot prompt-based fine-tuning, zero-shot inference, and standard fine-tuning. However, other methods of adapting models to labeled ID data can have very different sample efficiency properties (e.g., in-context learning). Future work could explore whether these results hold with few-shot in-context learning or parameter-efficient fine-tuning tuning (e.g., adapaters; Houlsby et al., 2019).

References

Anas Awadalla, Mitchell Wortsman, Gabriel Ilharco, Sewon Min, Hannaneh Hajishirzi, and Ludwig Schmidt. 2022. Exploring the landscape of distri-
butional robustness for question answering models. In *Findings of EMNLP*.


Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proc. of ACL*.


Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proc. of EMNLP*.


Yian Zhang, Alex Warstadt, Xiaocheng Li, and Samuel R. Bowman. 2021. When do you need billions of words of pretraining data? In *Proc. of ACL*. 
A Experimental Setup Details

Natural Language Inference. We use MultiNLI (Williams et al., 2018) and SNLI (Bowman et al., 2015) as ID datasets. We use MultiNLI, SNLI and MedNLI (Romanov and Shivade, 2018) as OOD test sets. All of our ID datasets have three labels (entailment, contradiction, neutral).

We also evaluate OOD on HANS (McCoy et al., 2019), a diagnostic dataset targeting lexical overlap, an ID-specific pattern in SNLI and MultiNLI. In MultiNLI and SNLI, the majority of examples with high lexical overlap between the NLI premise and hypothesis have the “entailment” label. In HANS, 50% of examples support this heuristic, and 50% contradict it, so a model that exclusively relies on the word overlap heuristic would have an accuracy of 50%. but HANS has two labels (entailment, non-entailment). To evaluate our 3-class models on 2-class HANS, we follow McCoy et al. (2019) and translate contradiction or neutral model predictions to non-entailment.

We train on the MultiNLI and SNLI training sets. We evaluate on the MultiNLI matched development set, the SNLI test set, and the HANS evaluation split. When evaluating OOD on MedNLI, we evaluate on the training set (~11K examples) because the development and test sets are quite small (~1.5K examples each).

Sentiment Analysis. We use the IMDb reviews dataset of (Maas et al., 2011), SST-2 (Socher et al., 2013) as ID datasets. We use IMDb, SST-2, and reviews from the “Movies and TV” subsection of the Amazon Reviews corpus (Ni et al., 2019) as OOD datasets.

These datasets are all binary classification, where reviews are labeled as positive or negative sentiment. To construct the “Movies and TV” Amazon review sentiment dataset, we randomly select one- or two-star (negative) reviews and four- or five-star (positive) reviews from the full Amazon Reviews corpus, using 25,000 examples for training, 10,000 examples for development, and 10,000 examples for testing. Each of these splits is balanced.

We train on the IMDb, SST, and Amazon Reviews training splits, and use the corresponding evaluation splits to measure ID performance. When evaluating OOD on SST, we use the concatenation of the train and test sets (4977 examples in total), since the original test set is quite small (1518 examples). Beyond this exception, we use each dataset’s evaluation split for OOD evaluation.

Extractive Question Answering. We use SQuAD (Rajpurkar et al., 2016) and NaturalQuestions (Kwiatkowski et al., 2019) as ID datasets. We use SQuAD, NaturalQuestions, TriviaQA, BioASQ (Tsatsaronis et al., 2015), and the SQuADShifts test sets of Miller et al. (2020) as OOD datasets.

The SQuADShifts test sets were constructed following the original SQuAD crowdsourcing procedure, but with passages drawn from both the original Wikipedia domain, as well as the New York Times (NYT), Amazon reviews, and Reddit. For NaturalQuestions, we only consider questions over paragraphs (as opposed to those over tables and lists). We use the MRQA 2019 Shared Task versions of TriviaQA and BioASQ (Fisch et al., 2019). We also use the MRQA 2019 Shared Task version of NaturalQuestions, but only include examples questions over paragraphs (removing those with questions over tables or lists). In all of these extractive QA datasets, models are given a passage and a question and tasked with identifying a substring of the passage that answers the question.

We train on the SQuAD and NaturalQuestions training splits, and use the corresponding evaluation splits to measure ID performance. When evaluating OOD on BioASQ, we use the concatenation of the train, development, and test sets (3977 examples in total), since the original test set is quite small (1518 examples). Beyond this exception, we use each dataset’s evaluation split for OOD evaluation.

B Hyperparameter Optimization Details

We conduct extensive hyperparameter optimization when training models on a particular ID dataset (or a subsample thereof). We re-tune hyperparameters for each subsample size, since the optimal value of certain hyperparameters may depend on number of available training examples (e.g., batch size and learning rate). For each experimental setting, we use a combination of (1) previously-reported hyperparameters (taken from prior work) and (2) random search (10 samples) over a pre-defined grid of reasonable hyperparameter values. For each experiment, we take the checkpoint with the best ID performance.
Natural Language Inference. For every NLI ID-OOD setting, we run experiments with the cross-product of learning rates in \{1e-5, 2e-5, 3e-5\} with batch sizes of \{16, 32\}. We also sample additional runs from the following grid:

- Random seed: [0, 100000]
- Learning rate: \{1e-5, 2e-5, 3e-5\}
- Batch size: \{16, 32\}
- Number of training epochs: \{10\}

Sentiment Analysis. For every sentiment analysis ID-OOD setting, we run experiments with the cross-product of learning rates in \{1e-5, 2e-5, 3e-5, 5e-5\} with batch sizes of \{16, 32\} and training for \{20, 50\} epochs. We also sample additional runs from the following grid:

- Random seed: [0, 100000]
- Learning rate: \{1e-5, 2e-5, 3e-5, 5e-5\}
- Batch size: \{16, 32\}
- Number of training epochs: \{20, 50\}

Extractive Question Answering. For every extractive question answering ID-OOD setting, we run experiments with the cross-product of learning rates in \{2e-5, 3e-5, 5e-5\} with batch sizes of \{16, 32\}. We also sample additional runs from the following grid:

- Random seed: [0, 100000]
- Learning rate: \{2e-5, 3e-5, 5e-5\}
- Batch size: \{16, 32\}
- Number of training epochs: \{4\}
C Results of All Methods on All ID-OOD Settings

C.1 Natural Language Prompting

Figure 5: Results on all NLI ID-OOD settings when comparing zero-shot prompting, prompt-based fine-tuning, and standard fine-tuning.
Figure 6: Results on all sentiment analysis ID-OOD settings when comparing zero-shot prompting, prompt-based fine-tuning, and standard fine-tuning.
C.2 Increasing Pre-Trained Model Size

Figure 7: Results on all NLI ID-OOD settings when increasing pre-trained model size.
Figure 8: Results on all sentiment analysis ID-OOD settings when increasing pre-trained model size.
Figure 9: Results on all extractive QA OOD settings when training on SQuAD with pre-trained models of increasing size.
Figure 10: Results on all extractive QA OOD settings when training on NaturalQuestions with pre-trained models of increasing size.
C.3 Pre-Training on More Data

Figure 11: Results on all NLI ID-OOD settings when increasing the amount of pre-training data.
Figure 12: Results on all sentiment analysis ID-OOD settings when increasing the amount of pre-training data.
Figure 13: Results on all extractive QA OOD settings when training on SQuAD with models pre-trained on varying amounts of data.
Figure 14: Results on all extractive QA OOD settings when training on NaturalQuestions with models pre-trained on varying amounts of data.