Better Sample Efficiency Does Not Imply Out-of-Distribution Robustness

Anonymous ACL submission

Abstract

We study the relationship between sample efficiency and out-of-distribution performance—if two models have the same in-distribution performance, does the model trained on fewer labeled training examples (higher sample efficiency) perform better out-of-distribution? First, we find that models with higher sample efficiency can have *worse* out-of-distribution robustness than models that are less sample-efficient. We then empirically study the correlation between sample efficiency and out-of-distribution robustness across three tasks, 23 total ID-OOD settings, and four broadly-applicable methods that change sample efficiency: (1) changing the pre-training data source; (2) using natural language prompts; (3) increasing model size; and (4) increasing the amount of pre-training data. Given that better sample efficiency does not necessarily give rise to robust models, our results underscore the importance of developing and evaluating whether interventions jointly improve both.

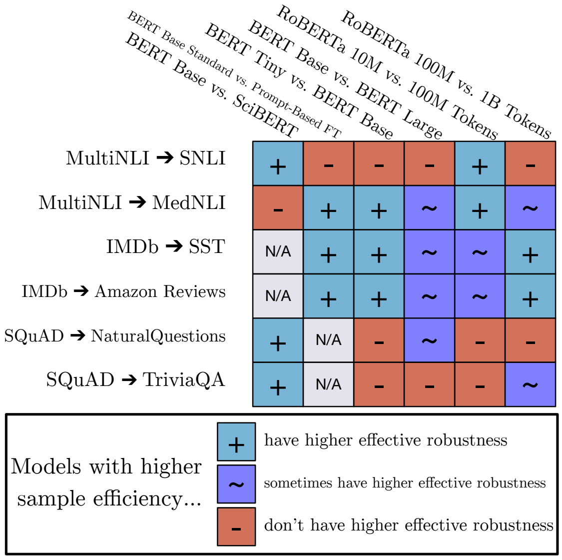


Figure 1: A summary of representative results from our empirical survey. Higher sample efficiency does not imply higher effective robustness.

1 Introduction

State-of-the-art 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; Jia and Liang, 2017). One potential cause of this ID-OOD performance gap is that models may learn to use ID-specific patterns that are predictive in-distribution but do not hold out-of-distribution. For example, the presence of the token *"sleeping"* is a strong indicator of the `contradiction` label in the SNLI dataset, but this feature is unlikely to hold in OOD test data (Gururangan et al., 2018). Models that rely on such ID-specific patterns may attain high ID performance, but at the cost of considerably lower OOD performance.

Does improving (in-distribution) sample efficiency, thereby reducing exposure to ID examples, also improve effective robustness on NLP tasks? As an extreme example, zero-shot models are much less likely to learn and use ID-specific patterns that do not hold in OOD settings because they are not exposed to *any* labeled ID examples. In a similar vein, one might expect that few-shot models trained on very small datasets may also rely less on ID-specific patterns. For example, if a model never sees the token "sleeping" while training on SNLI, then it is unlikely to learn that its presence is spuriously predictive of the contradiction label (Utama et al., 2021). Supporting this intuition, recent computer vision results show that zero-shot prediction with large pre-trained models can yield much better OOD performance than fine-tuning the same models on ID examples—fine-tuning on ID examples can actually *decrease* OOD performance (Radford et al., 2021).

In this paper, we study this relationship between sample efficiency and OOD robustness. Given models with the same ID performance, will the models trained on fewer ID examples (higher *sample efficiency*) also have better OOD performance (higher *effective robustness*; Taori et al., 2020)? For example, $\text{BERT}_{\text{BASE}}$ trained on 50,000 MultiNLI examples achieves 79% MultiNLI accuracy, but $\text{BERT}_{\text{LARGE}}$ requires only 10,000 examples to obtain the same accuracy. Which model will have higher OOD performance on SNLI? Despite the difference in sample efficiency, we find that these two models have roughly the same OOD performance on SNLI.

Although higher sample efficiency itself does not always imply higher effective robustness, the two may be empirically *correlated* for a wide range of ID and OOD datasets. We experimentally survey the extent of this correlation across three NLP tasks (23 total ID-OOD settings) and four methods that affect sample efficiency:

1. Changing the pre-training data source (§4.1).
2. Using natural language prompts for zero-shot prediction and during fine-tuning (Brown et al., 2020; Schick and Schütze, 2021; Gao et al., 2021; §4.2).
3. Fine-tuning models of increasing size (§4.3).
4. Fine-tuning models pre-trained on increasing amounts of data (§4.4).

First, we show that models pre-trained on data similar to the ID dataset can have higher sample efficiency, but worse effective robustness, than models pre-trained on data similar to the OOD dataset. This demonstrates that higher sample efficiency by itself does not always yield better effective robustness, since ID-specific inductive biases may improve sample efficiency, but not improve effective robustness because they do not apply OOD.

Next, we find that models trained with prompt-based fine-tuning often have better sample efficiency and effective robustness than models trained with standard fine-tuning. When evaluating OOD on *diagnostic datasets* (e.g., HANS; McCoy et al., 2019), *zero-shot* prompting yields even better effective robustness—in fact, we find that prompt-based fine-tuning on ID examples *reduces* effective robustness, corroborating the intuition that zero-shot models may be less reliant on ID-specific patterns. In contrast, when evaluating OOD on *standard benchmarks* (e.g., MultiNLI and SNLI), zero-shot prompting yields lower effective robustness than prompt-based fine-tuning.

Fianally, increasing the pre-trained model size or amount of pre-training data improves sample efficiency, but may not increase effective robustness. For example, while larger models consistently improve sample efficiency, they improve effective robustness when training on SNLI and testing on MultiNLI but not when training on MultiNLI and testing on SNLI. Similarly, while pre-training on more data yields higher sample efficiency, it slightly improves effective robustness in natural language inference experiments, but leads to no improvement on some extractive question answering datasets.

In general, the existence and magnitude of effective robustness gains depends on the particular sample efficiency intervention in question, the choice of ID and OOD dataset, and the amount of ID training data used (Hendrycks et al., 2021). Since it is empirically difficult to predict whether a particular intervention will reduce the ID-OOD performance gap, our results also emphasize the importance of collecting evaluation data from particular OOD distributions of interest. In order to better predict when interventions reduce the ID-OOD gap, future work should strive to better characterize ID-OOD shifts and better understand how interventions affect models.

Taken together, our results show that improving sample efficiency will not necessarily improve effective robustness, underscoring the importance of assessing whether proposed interventions jointly improve both.[1]

2 Measuring and Comparing Sample Efficiency and Robustness

Consider two models A and B with equivalent performance on held-out ID data. We say that a model A has higher *sample efficiency* than a model B if obtaining A requires fewer labeled ID examples than obtaining B, and we say that a model A has higher *effective robustness* than a model B if A outperforms B on held-out OOD data.

Given these definitions, we can only compare the sample efficiency and effective robustness of two models A and B if they have equivalent ID performance. This equivalent-ID constraint controls for the effect of ID performance on OOD performance, since ID gains usually yield commensurate OOD

[1]We plan to release all datasets, code, and models at omitted.link.

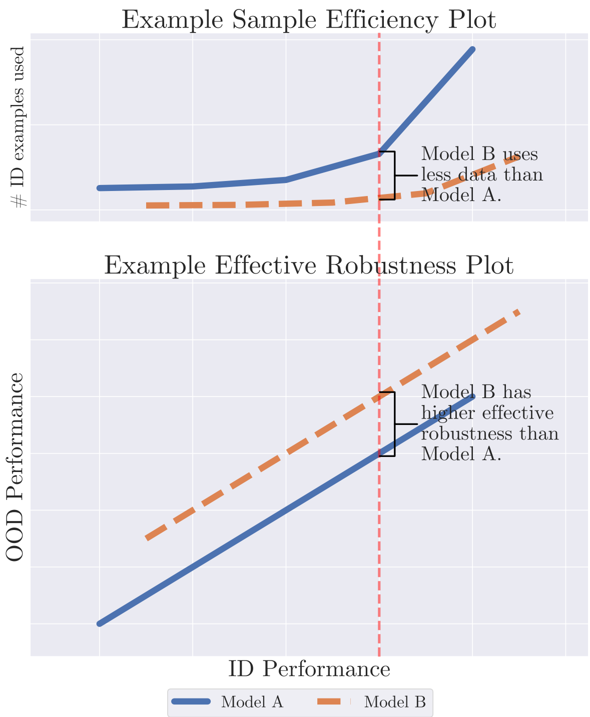


Figure 2: In this schematic example, model B has higher effective robustness and sample efficiency than model A. By plotting OOD performance (effective robustness) and the number of ID training examples used (sample efficiency) against ID performance, we can control for ID performance (vertical slice of plot) to relate sample efficiency to effective robustness.

gains (Taori et al., 2020; Miller et al., 2021).

We train models on varying-size subsamples of a given ID dataset and record the ID and OOD accuracy. We plot our results on effective robustness scatter plots, where each point is a model trained on some amount of data—the model's ID performance is its x-axis value, and its OOD performance is its y-axis value. To relate sample efficiency to ID and OOD performance, we also plot the number of training examples used against each models' ID performance. By placing this sample efficiency plot above the ID-OOD scatter plot, we can examine vertical slices to see (1) which equivalent-ID-performance model(s) have higher OOD performance, and (2) whether these models that do better OOD also use less ID training data.

Figure 2 provides an schematized example. In this example, model B has higher sample efficiency than model A. This is reflected in the top subfigure by the dashed orange series being *below* the solid blue series (uses less ID training data, given equivalent ID performance). Model B also has higher effective robustness than model A. In the bottom figure, the orange series is accordingly above the blue series (better absolute OOD performance, given equivalent ID performance).

3 Experimental Setup

3.1 Tasks and Datasets

To investigate the correlation between sample efficiency and effective robustness improvements for various interventions, we experiment with natural language inference (NLI; Dagan et al., 2005; Bowman et al., 2015), sentiment analysis, and extractive question answering (QA). We use "`[ID dataset]` $\rightarrow$ `[OOD dataset]`" to denote training and evaluating on a particular ID-OOD setting. See Appendix A for further details.

Natural Language Inference. We use MultiNLI (Williams et al., 2018) and SNLI (Bowman et al., 2015) as ID datasets. We use MultiNLI, SNLI, MedNLI (Romanov and Shivade, 2018), and HANS (McCoy et al., 2019) as OOD test sets.

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.

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.

3.2 Models

We experiment with various pre-trained masked language models. To understand the effect of a particular pre-training or fine-tuning intervention on sample efficiency and effective robustness, we evaluate models that differ along only the axis of interest (e.g., model size or pre-training corpus).

Since the optimal fine-tuning model 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 subsample size.

4 Is Sample Efficiency Empirically Correlated with Effective Robustness?

We empirically survey four methods for modulating sample efficiency (changing the pre-training data source, using natural language prompts, increasing pre-trained model size, and pre-training on more data) across 23 ID-OOD settings, showing that increasing sample efficiency can sometimes help but sometimes even *hurt* effective robustness. For the sake of brevity, we report on a representative subset of our results here—see Appendix B for results on all ID-OOD settings.

4.1 Changing the Pre-Training Data Source

Setup. To investigate how changing the pre-training data source affects sample efficiency and OOD robustness, we experiment with models pre-trained, fine-tuned, and evaluated on different data sources. We compare three different models: (1) $\text{BERT}_{\text{BASE}}$, which is pre-trained on the BookCorpus and English Wikipedia (Devlin et al., 2019); (2) SciBERT, which is pre-trained on scientific papers (Beltagy et al., 2019); and (3) Legal-BERT, which is pretrained on a variety of English legal texts (Chalkidis et al., 2020). We run experiments on NLI and extractive QA, since there are no suitable binary sentiment classification datasets for biomedical or legal text (to our knowledge).

Results and Discussion. When training on MultiNLI and testing on SNLI, we find that $\text{BERT}_{\text{BASE}}$ has higher sample efficiency and higher effective robustness than SciBERT or Legal-BERT (Figure 3a). Intuitively, pre-training on data similar to the ID dataset will improve sample efficiency, and pre-training on data similar to the OOD dataset will improve effective robustness. Indeed, when training on MultiNLI and testing on SNLI, we find that $\text{BERT}_{\text{BASE}}$ has higher sample efficiency and higher effective robustness than SciBERT or Legal-BERT (Figure 3a), possibly because the pre-training corpus for $\text{BERT}_{\text{BASE}}$ is most similar to the data in MultiNLI and SNLI (which contain premises from varying genres and internet captions, respectively). On the other hand, on MultiNLI $\rightarrow$ MedNLI, $\text{BERT}_{\text{BASE}}$ has higher sample efficiency but lower effective robustness than SciBERT (Figure 3b), since the $\text{BERT}_{\text{BASE}}$ pre-training corpus is similar to MultiNLI (improving sample efficiency), but dissimilar to the MedNLI OOD dataset, leading to lower effective robustness than SciBERT.

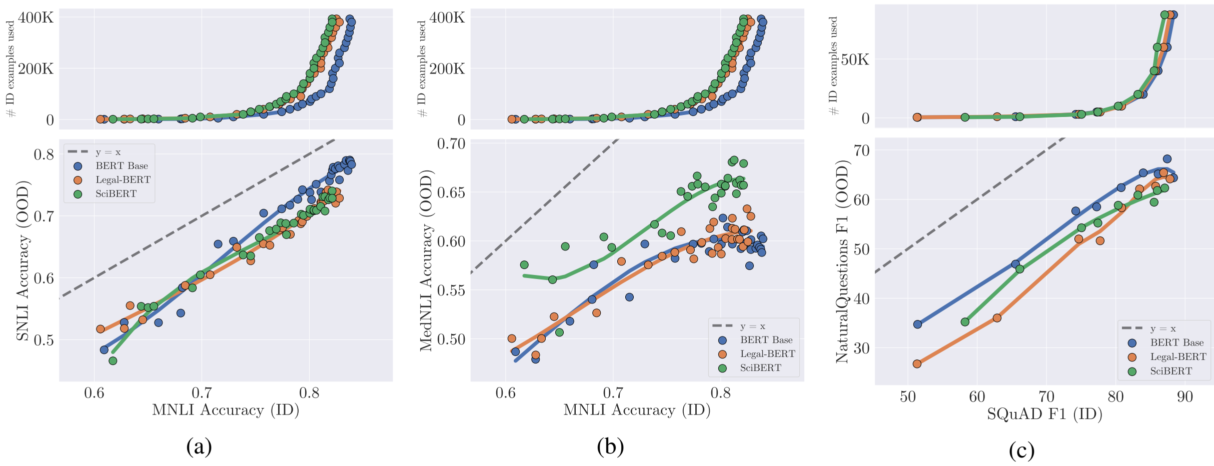
We see similar trends in extractive QA experiments. On SQuAD $\rightarrow$ NaturalQuestions, we see that $\text{BERT}_{\text{BASE}}$ has higher sample efficiency and effective robustness than SciBERT or Legal-BERT because the passages in both datasets are from English Wikipedia. (Figure 3c). However, on SQuAD $\rightarrow$ BioASQ (biomedical passages), SciBERT models have much higher effective robustness than $\text{BERT}_{\text{BASE}}$ models, despite being less sample-efficient (Figure 3d).

4.2 Natural Language Prompting

Setup. Models that use natural language prompts may have higher sample efficiency than models trained with standard fine-tuning, but do such models also have higher effective robustness? We investigate this question by comparing $\text{BERT}_{\text{BASE}}$ models using (1) standard fine-tuning, (2) zero-shot prompting, and (3) prompt-based fine-tuning. We refer readers to Gao et al. (2021) for additional background on these methods. We run experiments on NLI and sentiment analysis, since prompt-based fine-tuning with masked language models has not yet been applied to extractive QA.

Results and Discussion. To better understand how prompting affects the extent to which models learn ID-specific patterns, we first evaluate MultiNLI- and SNLI-trained models on the HANS diagnostic dataset. We first find that zero-shot prompting yields the highest effective robustness—prompt-based fine-tuning on MultiNLI or SNLI examples rapidly *reduces* HANS performance (while improving ID performance). Next, we see that models trained with prompt-based fine-tuning can have higher sample efficiency than models trained with standard fine-tuning models, and such models also have higher effective robustness. (Figure 4a-b).

In contrast to our results on diagnostic datasets, experiments on standard sentiment analysis and NLI benchmark datasets show that zero-shot prompting does not always yield higher effective robustness than prompt-based fine-tuning, despite its higher sample efficiency—prompt-based fine-tuning frequently improves both absolute ID and OOD performance over zero-shot prompting (Figure 4c-f). Even zero-shot prompting of GPT-3 (175B), a dramatically larger model trained on substantially more data, yields lower effective robustness than models trained with either prompt-based fine-tuning or standard fine-tuning, underscoring that zero-shot prediction does not always yield the best effective robustness.



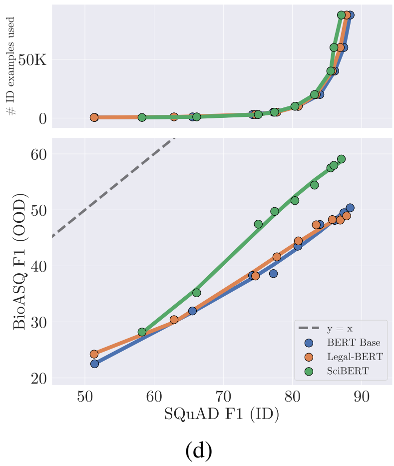


Figure 3: Although models pre-trained on data similar to the ID dataset have higher sample efficiency than models pre-trained on data similar to the OOD dataset, the latter yields better effective robustness. Better sample efficiency does not necessarily imply better effective robustness, since models with ID-specific inductive biases may have higher sample efficiency, but worse OOD performance, than models with OOD-specific inductive biases.

On standard benchmarks, we continue to see that prompt-based models with higher sample efficiency also often have higher effective robustness than their counterparts trained with standard fine-tuning (Figure 4d-f). However, as prompt-based models are trained with more examples, they lose their sample efficiency advantage and produce similar results to standard fine-tuning.

However, there exist ID-OOD settings where few-shot prompt-based fine-tuning improves sample efficiency, but not effective robustness, over standard fine-tuning. For example, on MultiNLI $\rightarrow$ SNLI, few-shot prompt-based fine-tuning models can have higher sample efficiency than models trained with standard fine-tuning, but the models achieve approximately the same absolute ID and OOD performance (Figure 4c).

4.3 Increasing Pretrained Model Size

Setup. To study how increasing pre-trained model size affects sample efficiency and effective robustness, 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) on a fixed pre-training dataset with a fixed optimization procedure. We run experiments on NLI, sentiment analysis, and extractive QA over five different pre-trained model sizes: (1) Large (L=24, H=1024), (2) Base (L=12, 768), (3) Medium (L=8, H=512), (4) Mini (L=4, H=256), and (5) Tiny (L=2, H=128).

Results and Discussion. In experiments on NLI datasets, we find that using larger models does not consistently improve effective robustness, despite improving sample efficiency. For example, larger models have higher sample efficiency and higher effective robustness on SNLI $\rightarrow$ MultiNLI (Figure 5b), but similar effective robustness as smaller models on MultiNLI $\rightarrow$ SNLI (Figure 5a).

In sentiment analysis experiments, larger models improve both sample efficiency and effective robustness on IMDb $\rightarrow$ SST (Figure 5c). However, on IMDb $\rightarrow$ Amazon reviews, increasing model size yields diminishing effective robustness gains as the ID-OOD gap shrinks (i.e., the models approach $y = x$; Figure 5d). Moving from $\text{BERT}_{\text{TINY}}$ to $\text{BERT}_{\text{MINI}}$ to $\text{BERT}_{\text{MEDIUM}}$ improves both sample efficiency and effective ro-

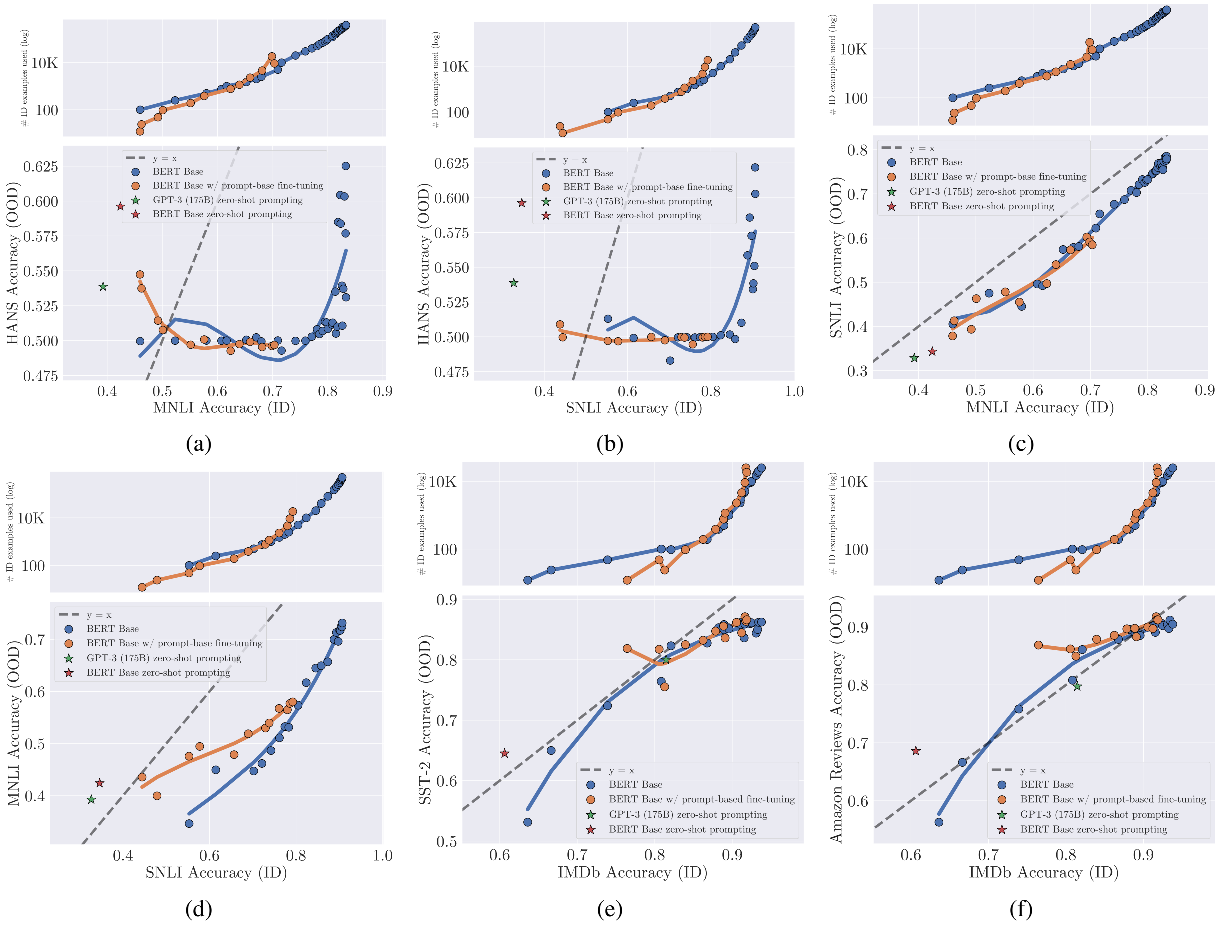


Figure 4: (a-b): When evaluating OOD on diagnostic datasets, zero-shot prompting yields the highest absolute OOD performance (and effective robustness)—prompt-based fine-tuning *decreases* OOD performance (while increasing ID performance). However, when prompt-based fine-tuning models are more sample-efficient than standard fine-tuning, they also have higher effective robustness. (c-f): In contrast, when evaluating OOD on standard NLI and sentiment analysis datasets, zero-shot prompting does not have better effective robustness than prompt-based fine-tuning.

bustness, but further increasing model size to $\text{BERT}_{\text{BASE}}$ and $\text{BERT}_{\text{LARGE}}$ yields substantially smaller gains. In fact, the effective robustness of $\text{BERT}_{\text{LARGE}}$ can *decrease* when using the full ID training set, since absolute OOD performance saturates before ID performance.

Finally, in extractive QA experiments, we find that larger models often do not yield effective robustness improvements, despite their higher sample efficiency. For example, on SQuAD → NaturalQuestions and SQuAD → TriviaQA, larger models have the same effective robustness as smaller models (Figure 5e-f).

4.4 Pre-Training on More Data

Setup. To study how pre-training on more data affects sample efficiency and effective robustness, we experiment with the RoBERTa models pre-trained on 10M, 100M, and 1B tokens of data drawn from Wikipedia and SmashWords (Zhang et al., 2021).

Results and Discussion. In our NLI experiments, we find that increasing the amount of pre-training data slightly improves sample efficiency and effective robustness. For example, using more pre-training data improves both sample efficiency and effective robustness on SNLI → MultiNLI (Figure 6b). However, there are diminishing returns on effective robustness from adding more pre-training data—pre-training on 10M vs. 100M tokens has a much larger impact than pre-training on 100M or 1B tokens. We see these same relative trends on MultiNLI → SNLI, though the absolute OOD performance improvements are smaller (Figure 6a).

Additional pre-training data also slightly improves sample efficiency and effective robustness on sentiment analysis datasets. On IMDb → SST and IMDb → Amazon reviews, increasing the pre-

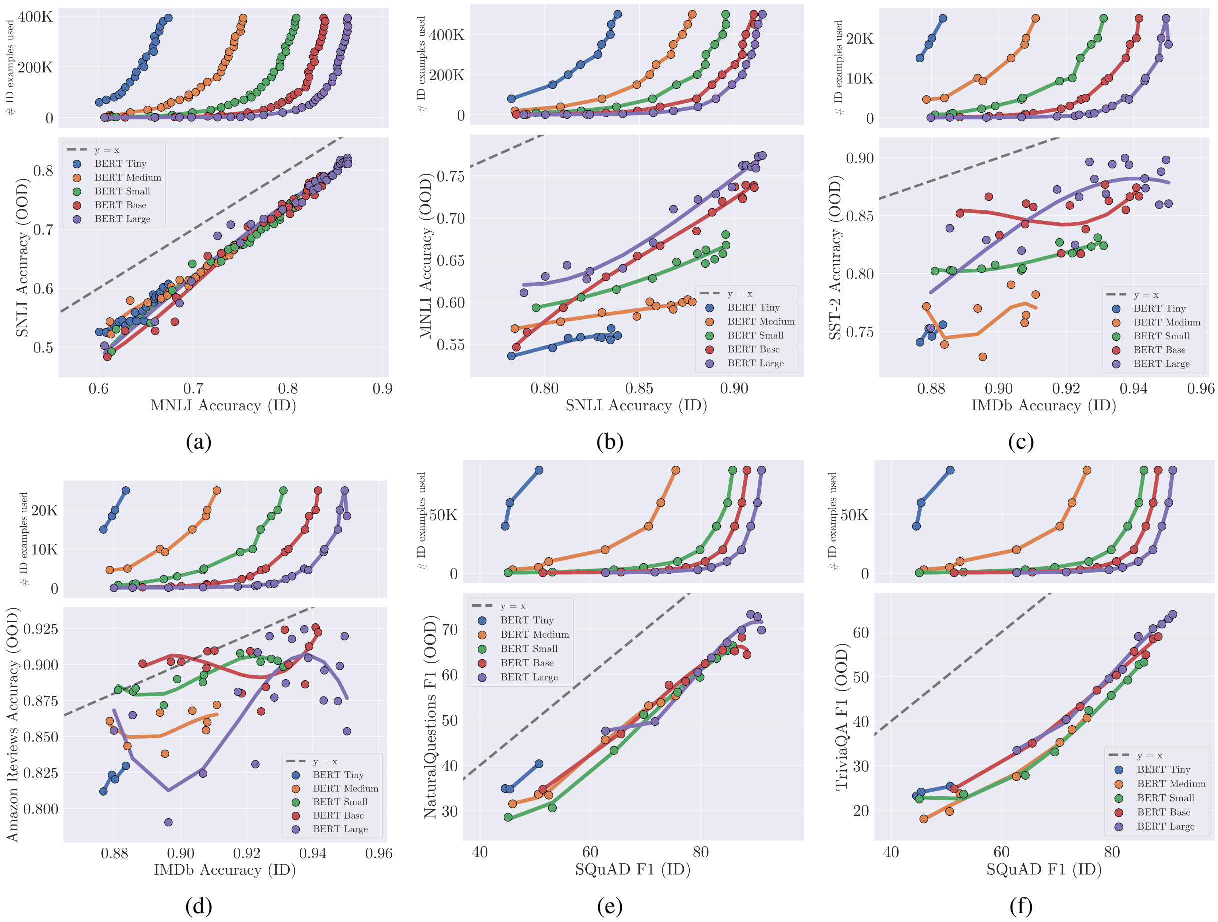


Figure 5: Increasing model size improves sample efficiency, but larger models may not have higher effective robustness. For example, larger models have higher sample efficiency and effective robustness on SNLI → MultiNLI, but they do not improve effective robustness on MultiNLI → SNLI. Similarly, increasing model size when training on IMDb and evaluating on Amazon reviews does not improve effective robustness, perhaps because smaller models already have a small ID-OOD gap in this setting.

training dataset size from 10M to 100M has little effect, but moving to 1B tokens yields proportionally larger effective robustness improvements (Figure 6c-d).

On extractive QA datasets, we find that pre-training on larger datasets improves sample efficiency but only marginally improves effective robustness, if at all. On SQuAD → NaturalQuestions, models pre-trained on 10M tokens have higher effective robustness than those pre-trained on 100M or 1B tokens; the latter two models have largely the same effective robustness (Figure 6e). In a similar vein, on SQuAD → TriviaQA, models pre-trained on 10M, 100M, and 1B tokens have largely the same effective robustness (Figure 6f).

5 Discussion

Predicting Intervention Efficacy Requires Better Characterizing ID-OOD Shifts. Our results are dependent on the particular ID-OOD pair, because choosing different ID or OOD datasets can dramatically change the challenges involved in overcoming the distribution shift. For example, while sample efficiency and effective robustness are positively correlated when training on IMDb and evaluating OOD on SST, having higher sample efficiency actually *reduces* effective robustness when training on SST and evaluating OOD on IMDb.

Since examples in SST are sentences, whereas examples in IMDb are multi-paragraph reviews, generalizing from SST to IMDb requires extrapolating from shorter sequences to much longer ones. Interventions that improve sample efficiency but do not help with length extrapolation—a seemingly orthogonal skill—therefore would not also improve effective robustness.

To better predict whether interventions will increase effective robustness and sample efficiency,

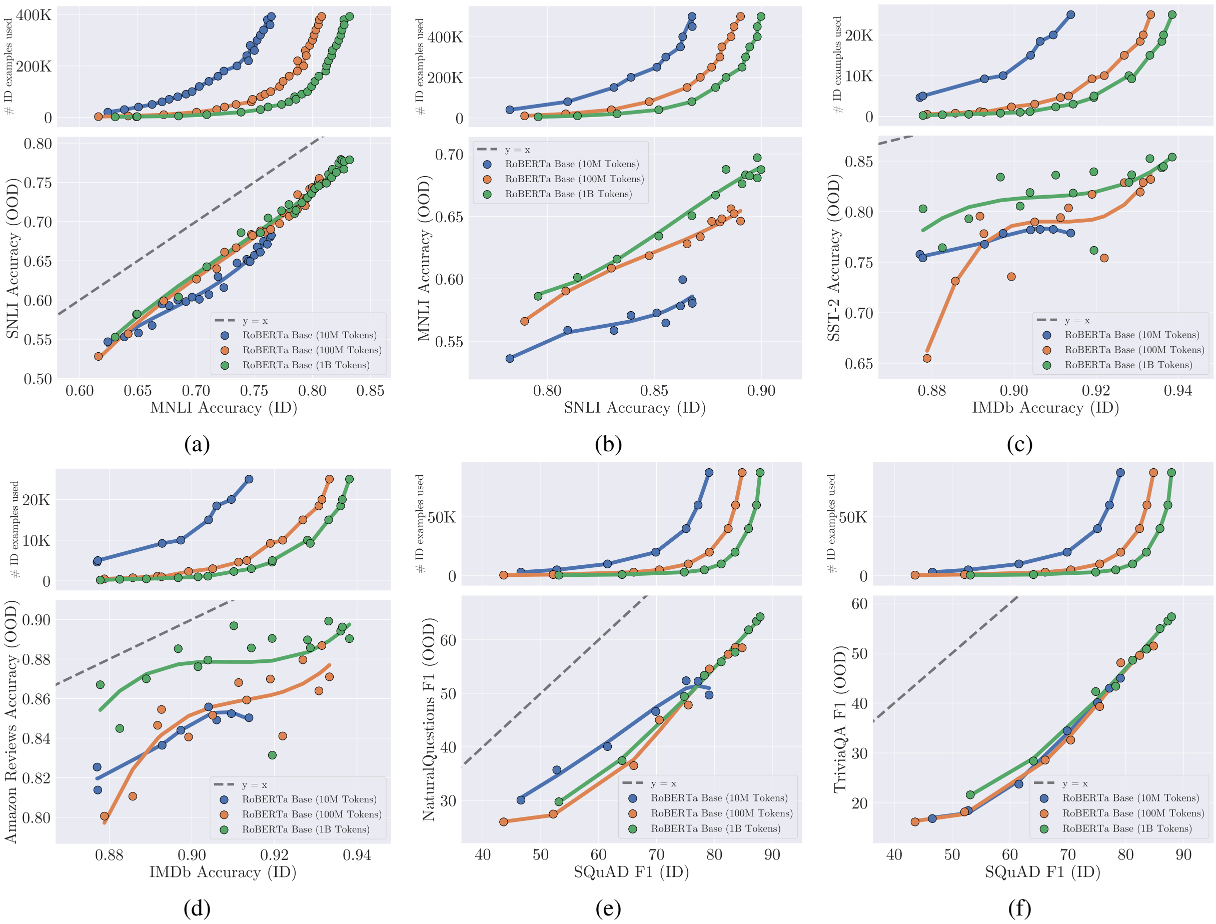


Figure 6: Pre-training on more data improves sample efficiency, but does not always improve effective robustness. The two are correlated in NLI experiments, but the effective robustness improvements are only apparent when moving to 1B tokens for sentiment analysis experiments and barely noticeable in extractive QA experiments.

future work should strive to better characterize ID-OOD shifts and better understand how interventions improve models, paving a path for reasoning about whether particular interventions are appropriate or useful for particular shifts. In the absence of such predictive powers, these results underscore the importance of collecting evaluation data from the OOD distribution(s) of interest.

Why Study Effective Robustness? Since ID performance is often strongly correlated with OOD performance, training the strongest model with the most data will generally yield the best absolute OOD performance (Fisch et al., 2019; Taori et al., 2020; Miller et al., 2021). However, training models with strong OOD performance in the face of practical resource constraints (e.g., the desire to minimize data annotation cost, engineering person-hours, and computation time) requires better understanding how different methods for improving ID performance might also affect OOD improvements; effective robustness is a useful tool for understanding this relationship.

6 Conclusion

In this work, we empirically study the relationship between sample efficiency and effective robustness. We find that better sample efficiency unto itself does not imply improved effective robustness, and survey the extent of their correlation for four interventions. Even on natural distribution shifts, we find that better sample efficiency is often not correlated with better effective robustness, underscoring the importance of developing and evaluating whether interventions jointly improve both sample efficiency and robustness.

References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proc. of EMNLP*.

John Blitzer. 2008. *Domain adaptation of natural lan-*

guage processing systems. Ph.D. thesis, University of Pennsylvania.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proc. of EMNLP*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proc. of NeurIPS*.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of law school. In *Findings of EMNLP*.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The PASCAL recognising textual entailment challenge. In *Machine Learning Challenges Workshop*, pages 177–190. Springer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of NAACL*.

Adam Fisch, Alon Talmor, Robin Jia, Minjoon Seo, Eunsol Choi, and Danqi Chen. 2019. MRQA 2019 shared task: Evaluating generalization in reading comprehension. In *Proc. of MRQA*.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Proc. of ACL*.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proc. of NAACL*.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. 2021. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proc. of ICCV*.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proc. of EMNLP*.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proc. of ACL*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

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*.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proc. of ACL*.

John Miller, Karl Krauth, Benjamin Recht, and Ludwig Schmidt. 2020. The effect of natural distribution shift on question answering models. In *Proc. of ICML*.

John Miller, Rohan Taori, Aditi Raghunathan, Shiori Sagawa, Pang Wei Koh, Vaishaal Shankar, Percy Liang, Yair Carmon, and Ludwig Schmidt. 2021. Accuracy on the line: On the strong correlation between out-of-distribution and in-distribution generalization. In *Proc. of ICML*.

Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proc. of EMNLP*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. ArXiv:2103.00020.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proc. of EMNLP*.

Alexey Romanov and Chaitanya Shivade. 2018. Lessons from natural language inference in the clinical domain. In *Proc. of EMNLP*.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proc. of EACL*.

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*.

Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. 2020. Measuring robustness to natural distribution shifts in image classification. In *Proc. of NeurIPS*.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R. Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, Yannis Almirantis, John Pavlopoulos, Nicolas Baskiotis, Patrick Gallinari, Thierry Artiéres, Axel-Cyrille Ngonga Ngomo, Norman Heino, Eric Gaussier, Liliana Barrio-Alvers, Michael Schroeder, Ion Androutsopoulos, and Georgios Paliouras. 2015. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC bioinformatics*, 16(1):1–28.

Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Well-read students learn better: On the importance of pre-training compact models. ArXiv:1908.08962.

Prasetya Ajie Utama, Nafise Sadat Moosavi, Victor Sanh, and Iryna Gurevych. 2021. Avoiding inference heuristics in few-shot prompt-based finetuning. In *Proc. of EMNLP*.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proc. of NAACL*.

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 exclusivly 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* ($\sim$11K examples) because the development and test sets are quite small ($\sim$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 (8471 examples in total), since the original test set is quite small (1821 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 NaturalQuetsions, 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 Results of All Methods on All ID-OOD Settings

B.1 Changing the Pre-Training Data Source

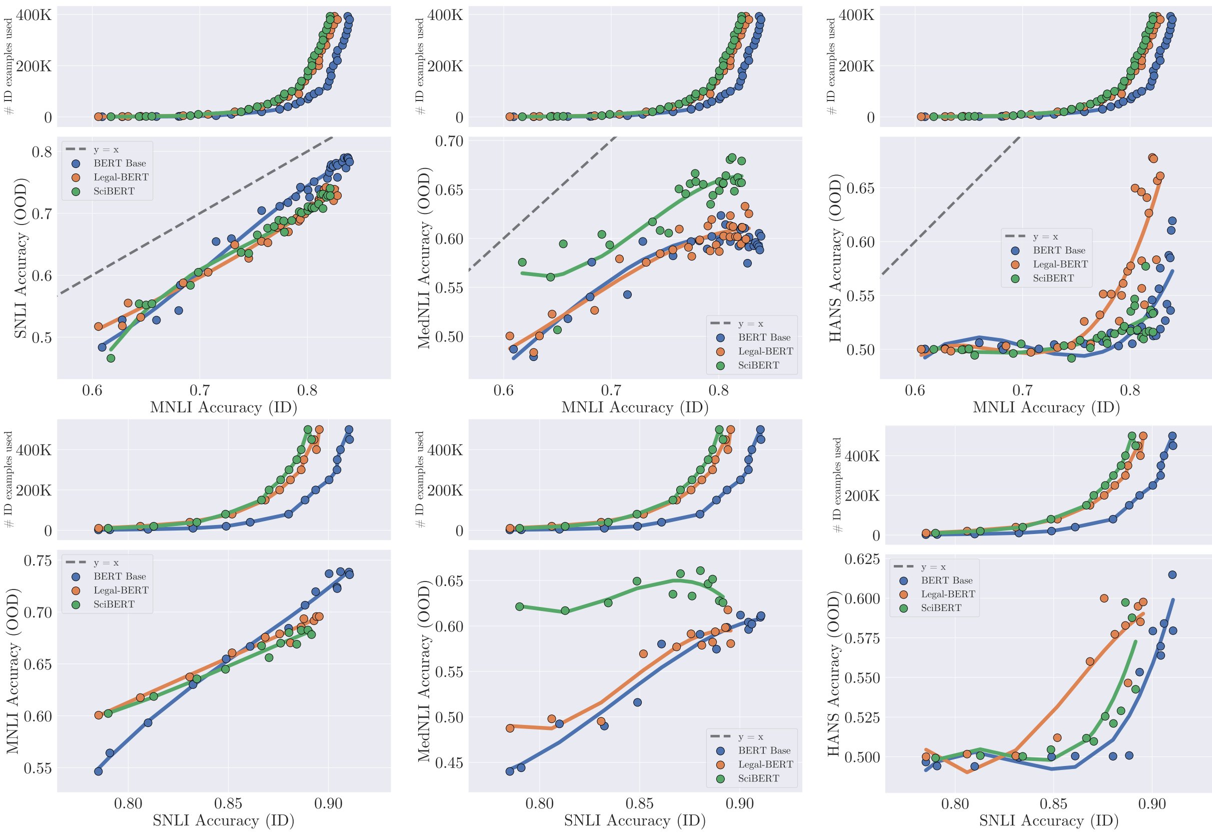


Figure 7: Results on all NLI ID-OOD settings when changing the pre-training data source.

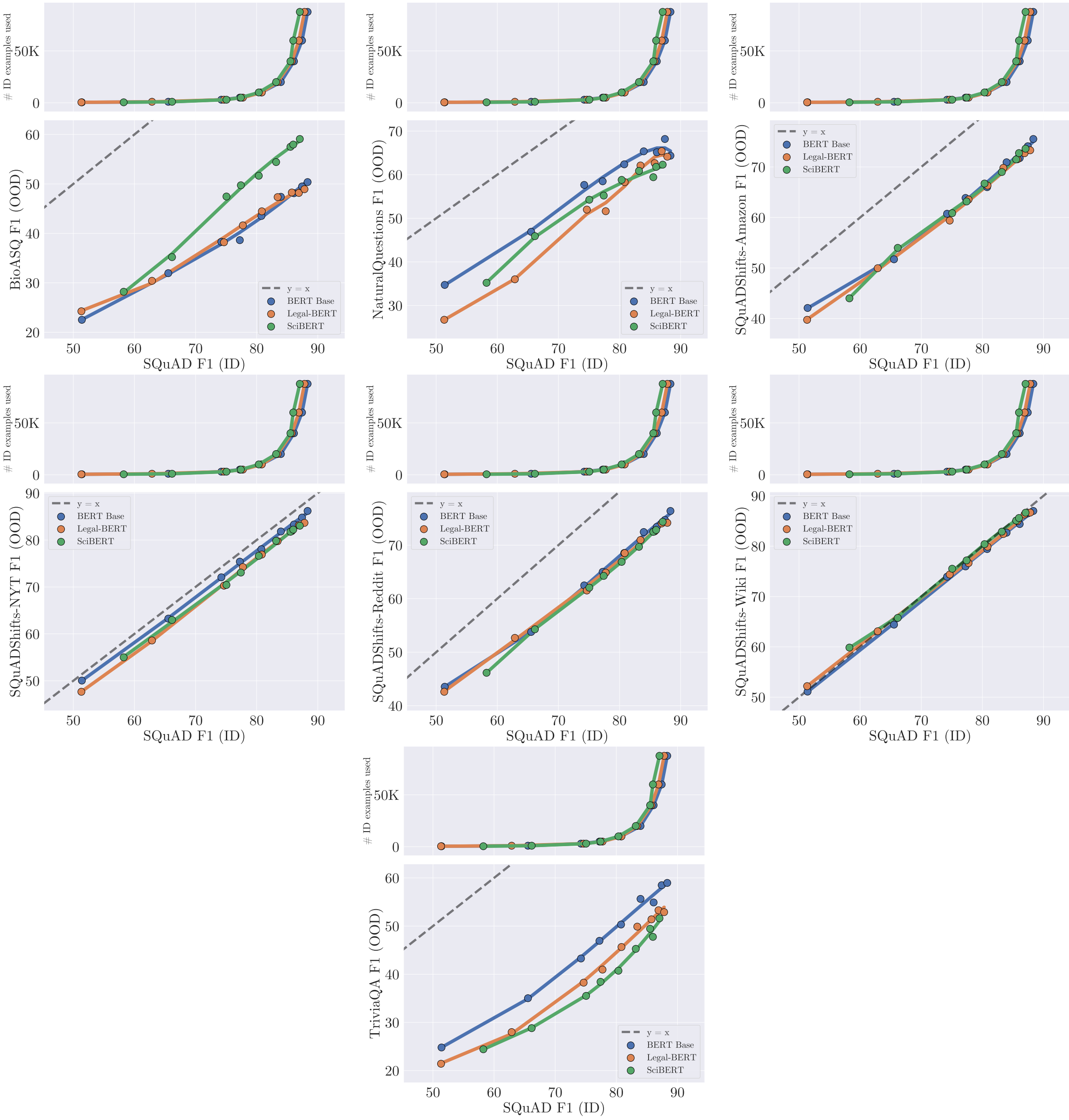


Figure 8: Results on all extractive QA OOD settings when training on SQuAD and changing the pre-training data source.

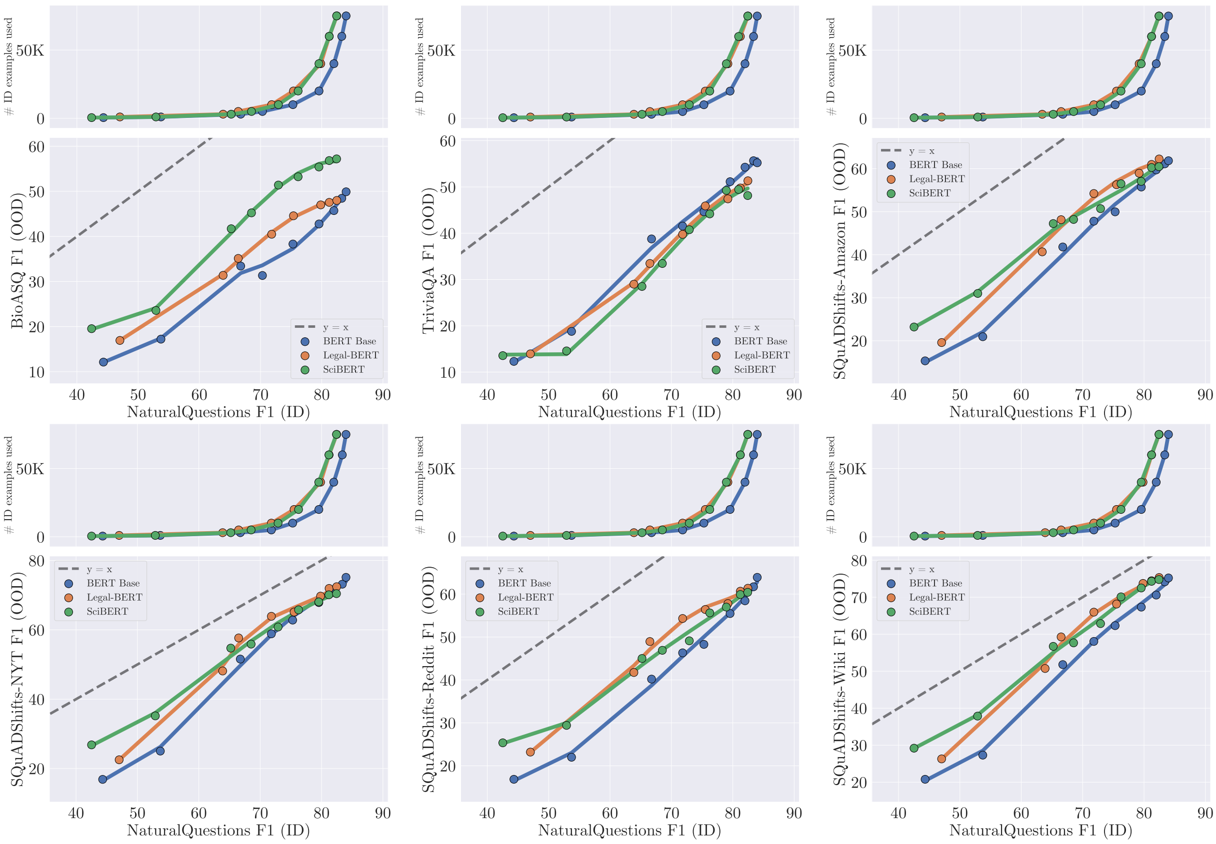


Figure 9: Results on all extractive QA OOD settings when training on NaturalQuestions and changing the pre-training data source.

B.2 Natural Language Prompting

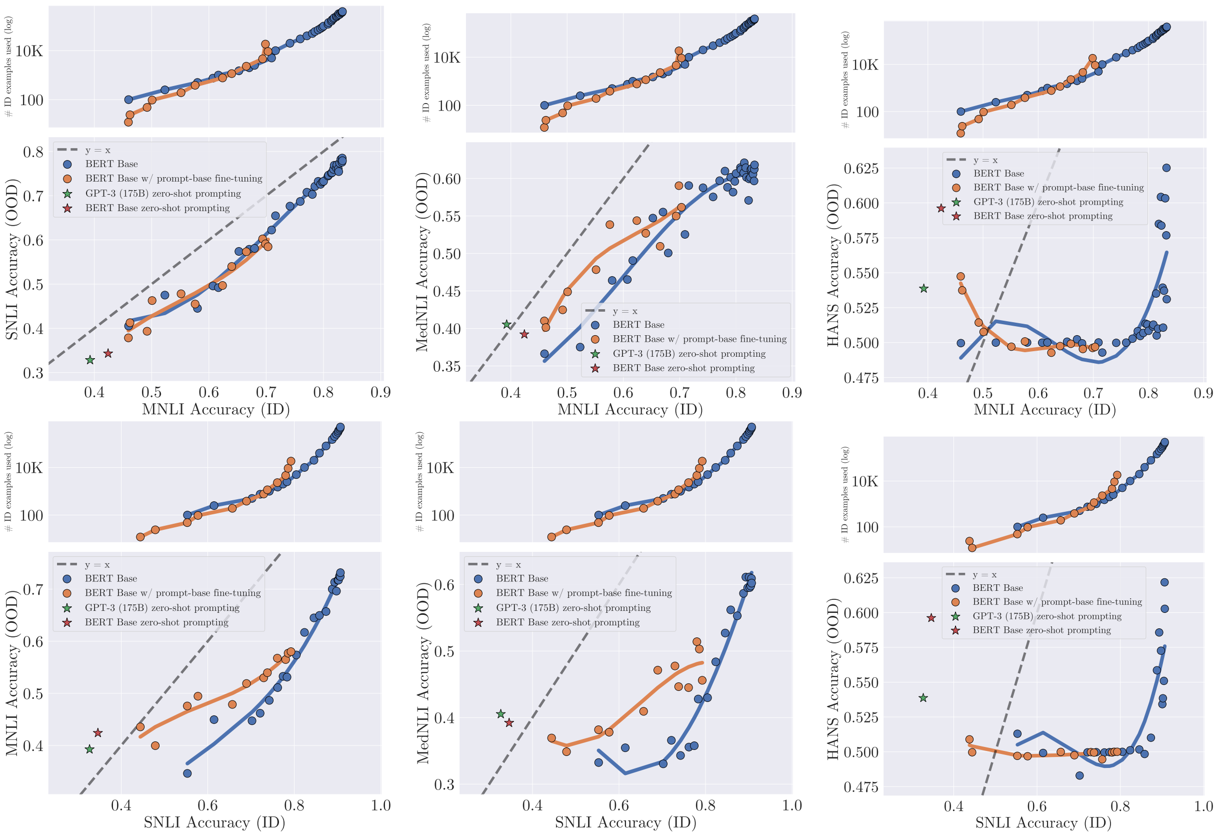


Figure 10: Results on all NLI ID-OOD settings when comparing zero-shot prompting, prompt-based fine-tuning, and standard fine-tuning.

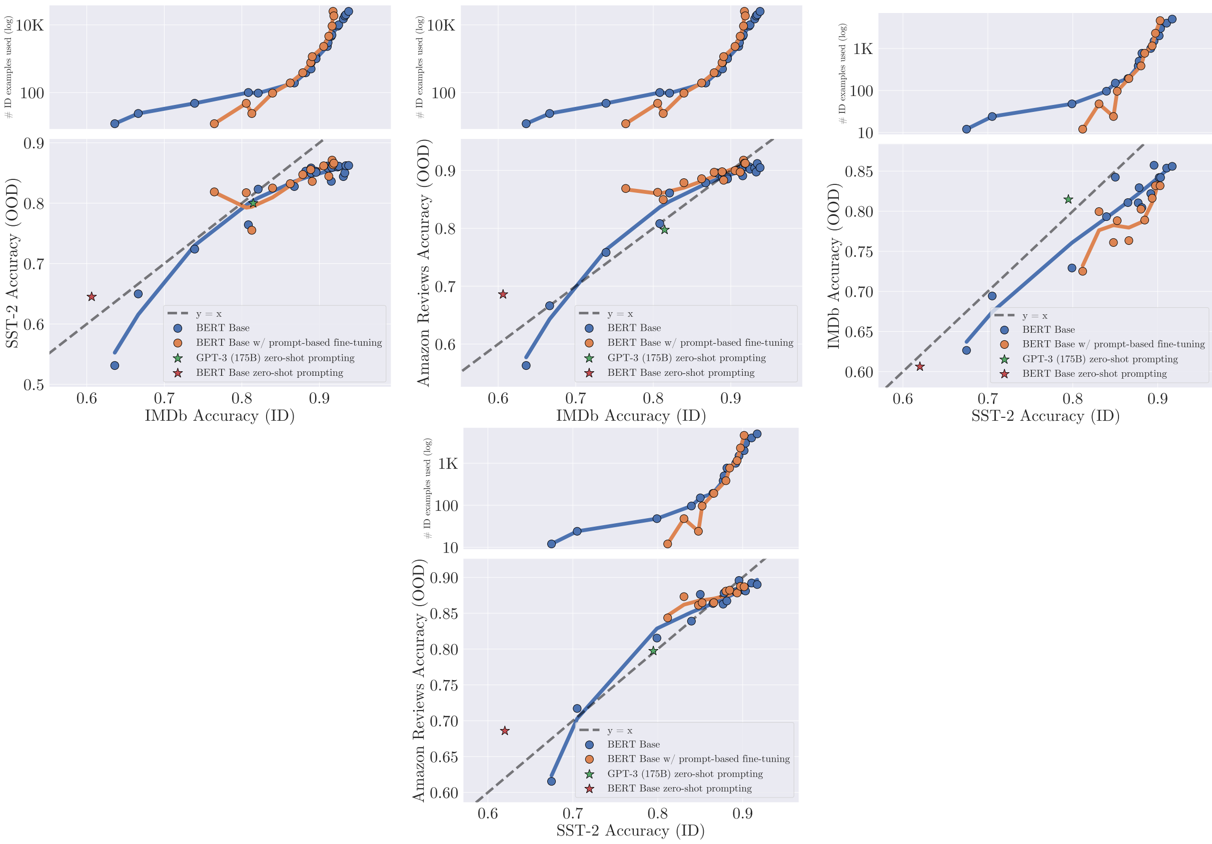


Figure 11: Results on all sentiment ID-OOD settings when comparing zero-shot prompting, prompt-based fine-tuning, and standard fine-tuning.

B.3 Increasing Pre-Trained Model Size

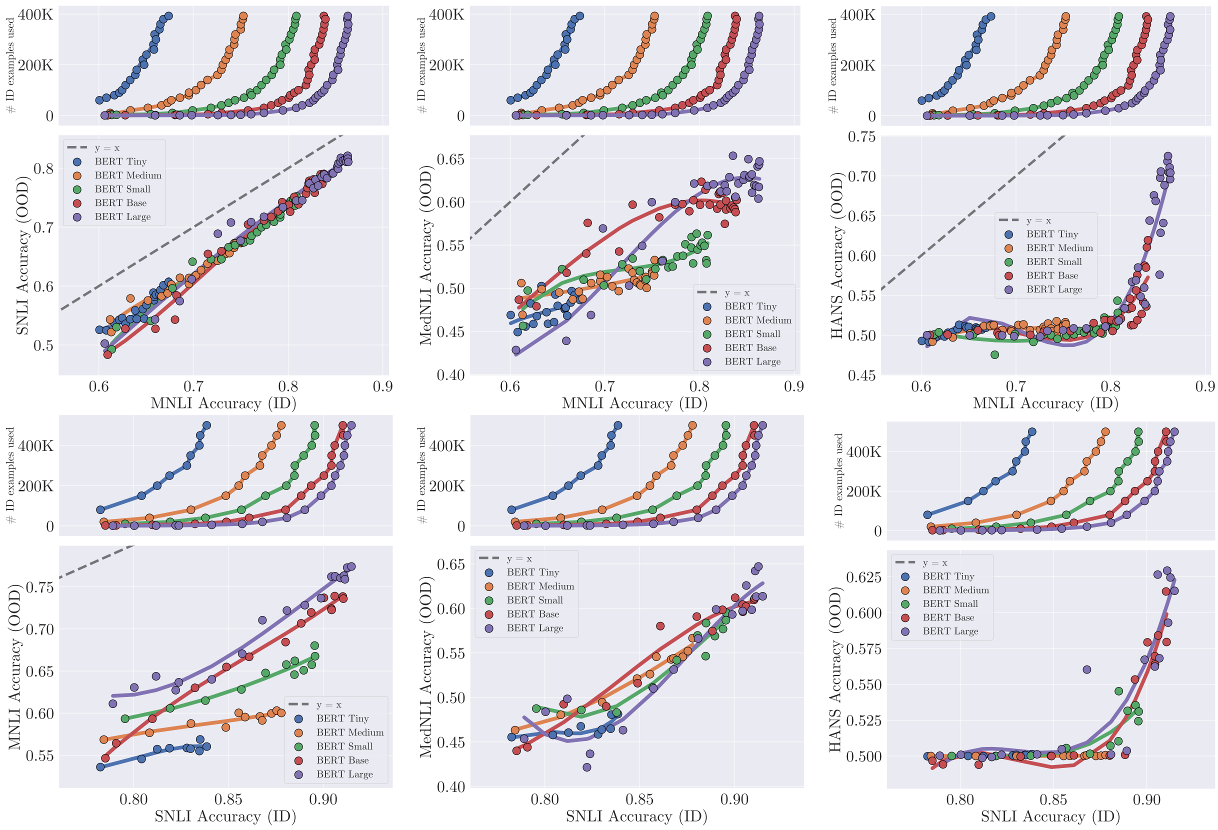


Figure 12: Results on all NLI ID-OOD settings when increasing pre-trained model size.

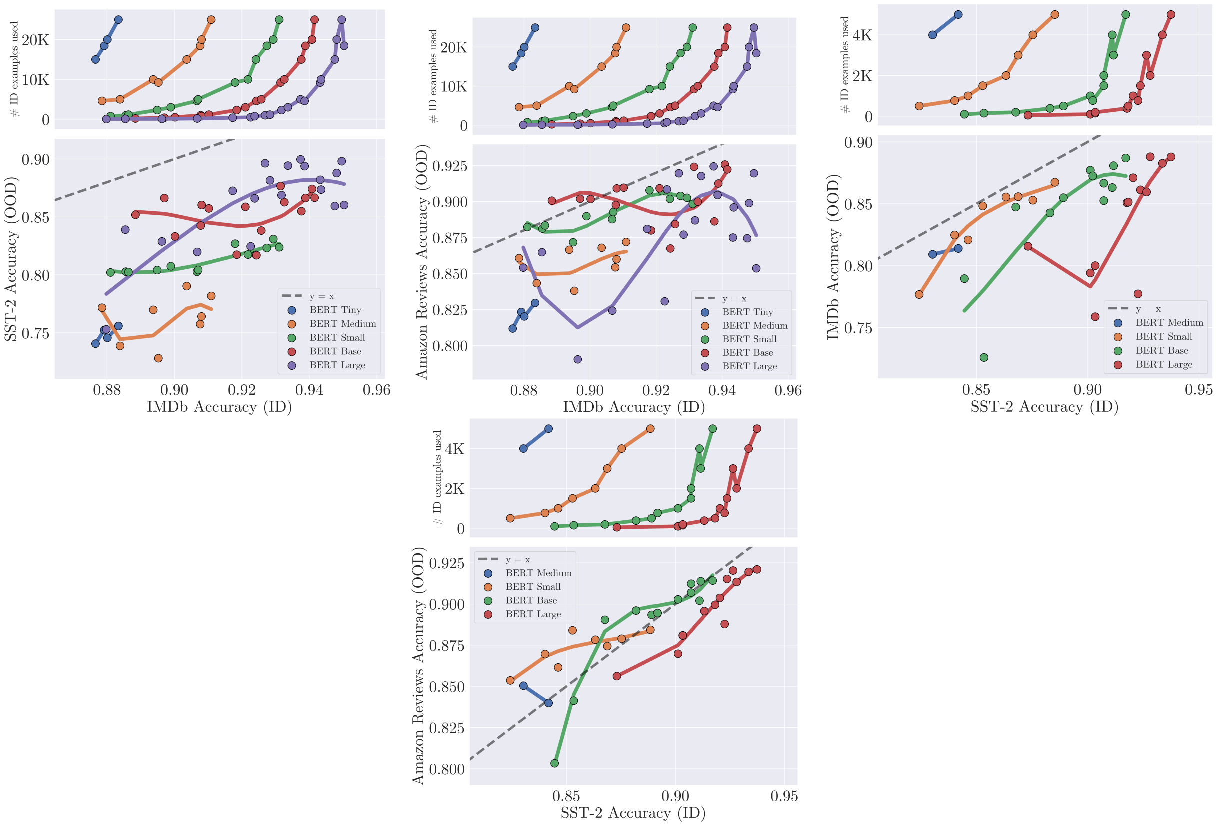


Figure 13: Results on all sentiment ID-OOD settings when increasing pre-trained model size.

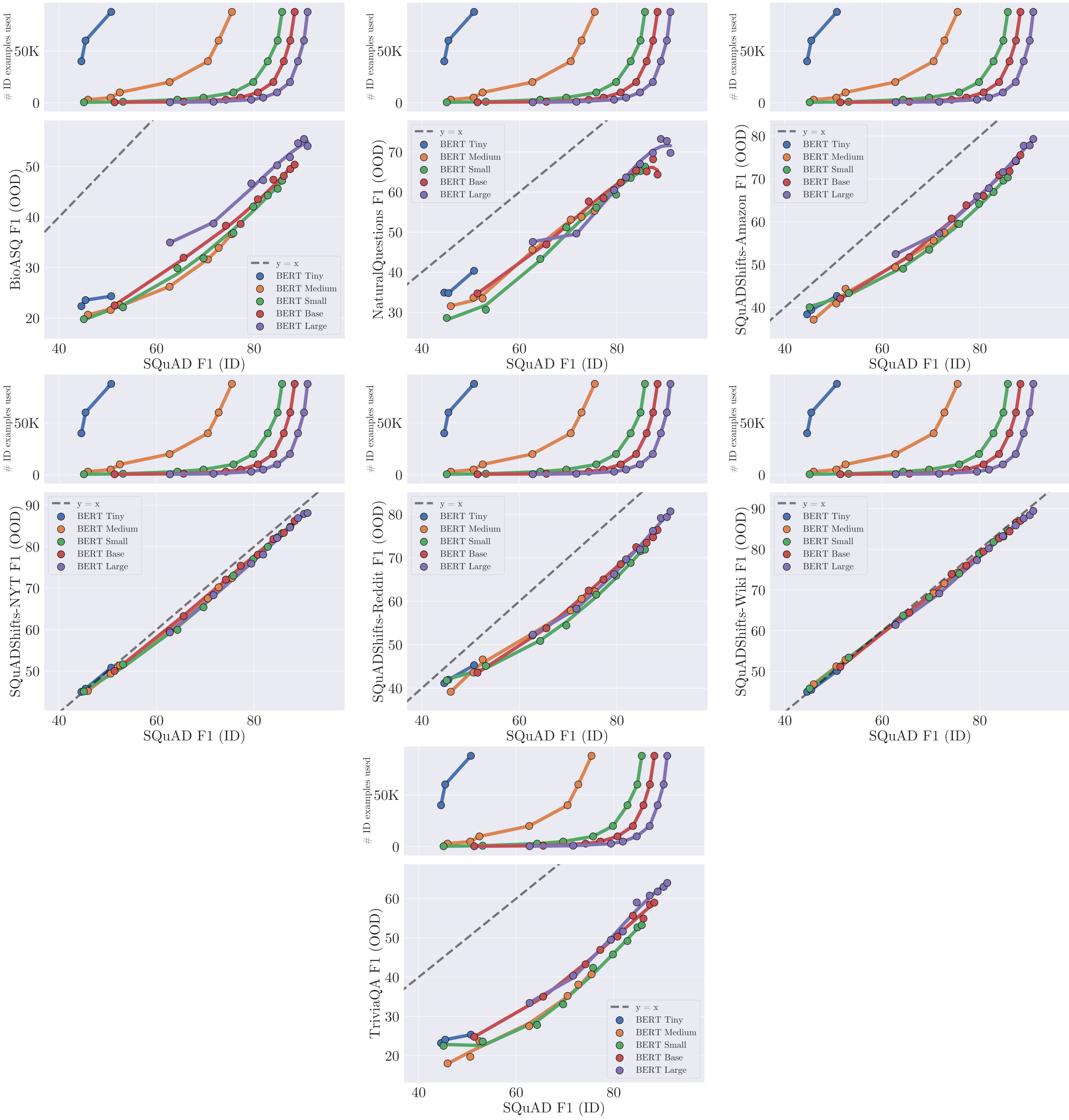


Figure 14: Results on all extractive QA OOD settings when training on SQuAD with pre-trained models of increasing size.

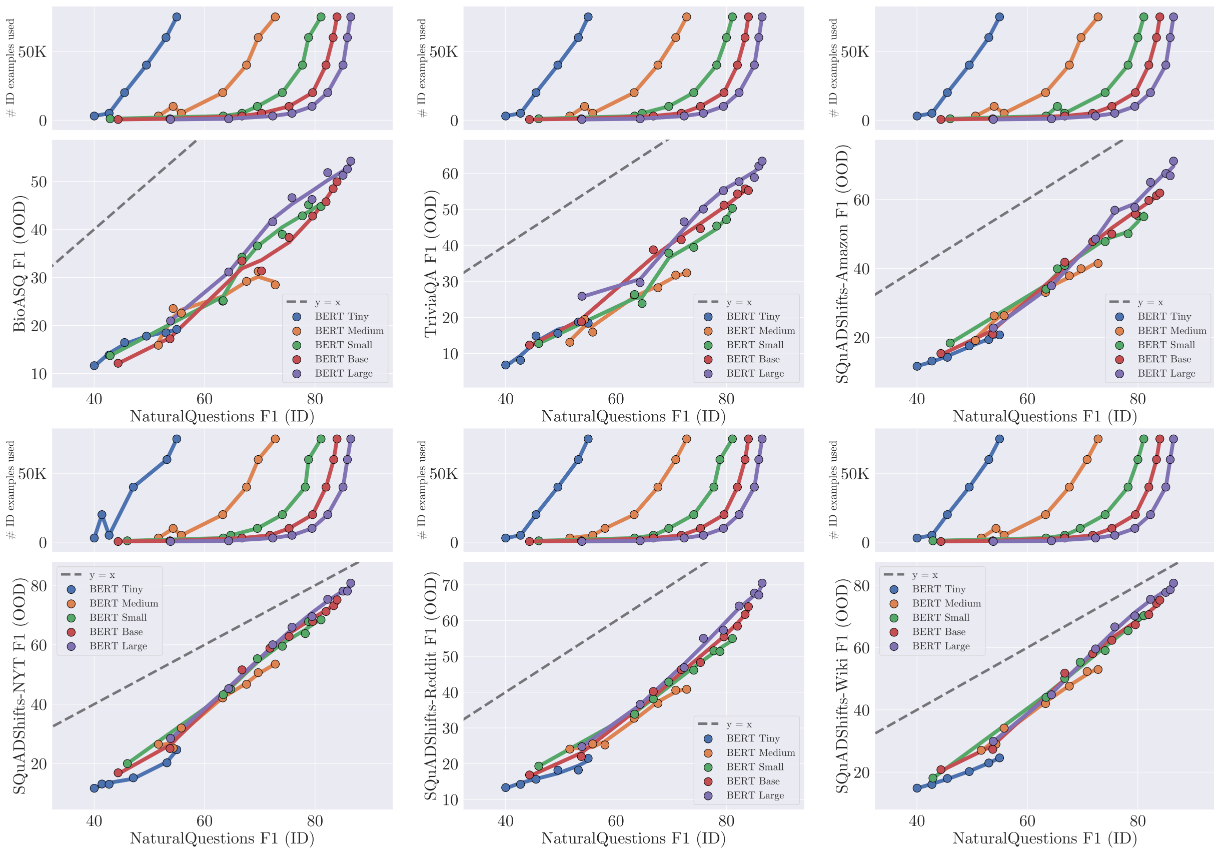


Figure 15: Results on all extractive QA OOD settings when training on NaturalQuestions with pre-trained models of increasing size.

B.4 Pre-Training on More Data

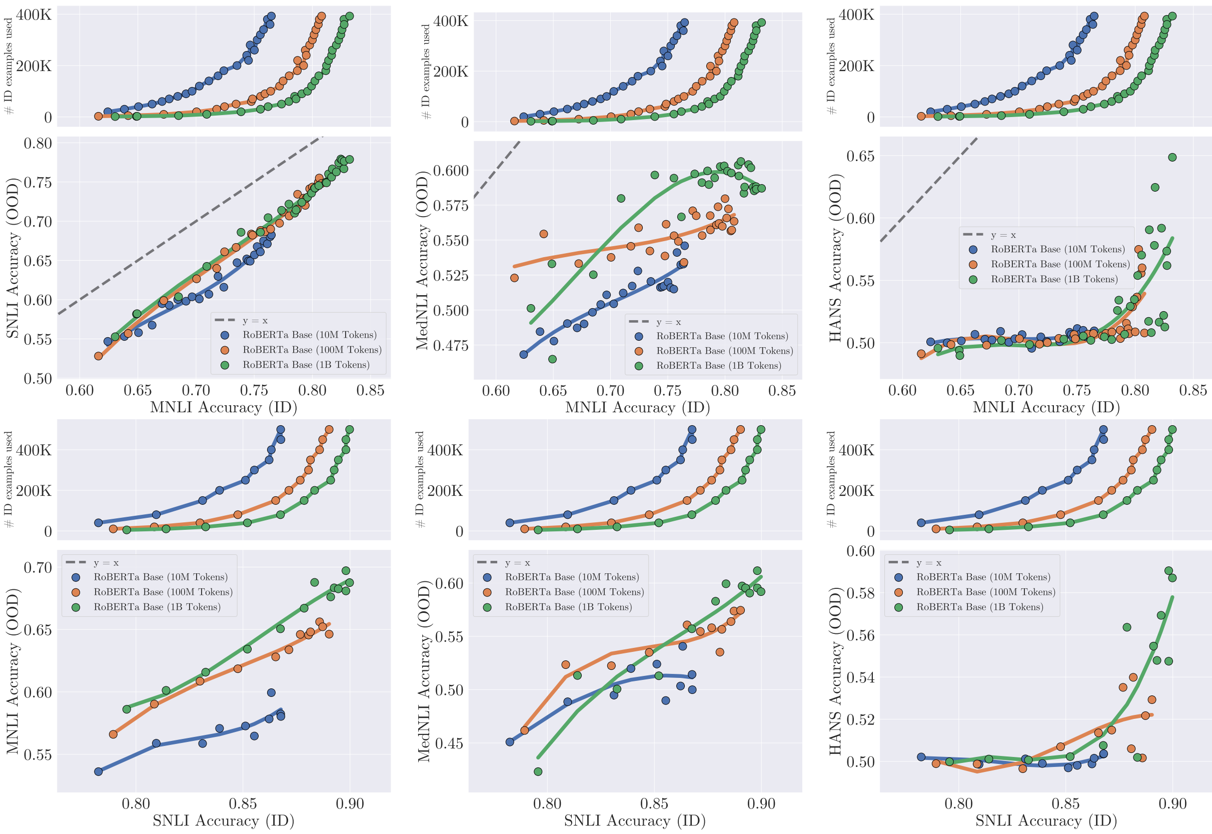


Figure 16: Results on all NLI ID-OOD settings when increasing the amount of pre-training data.

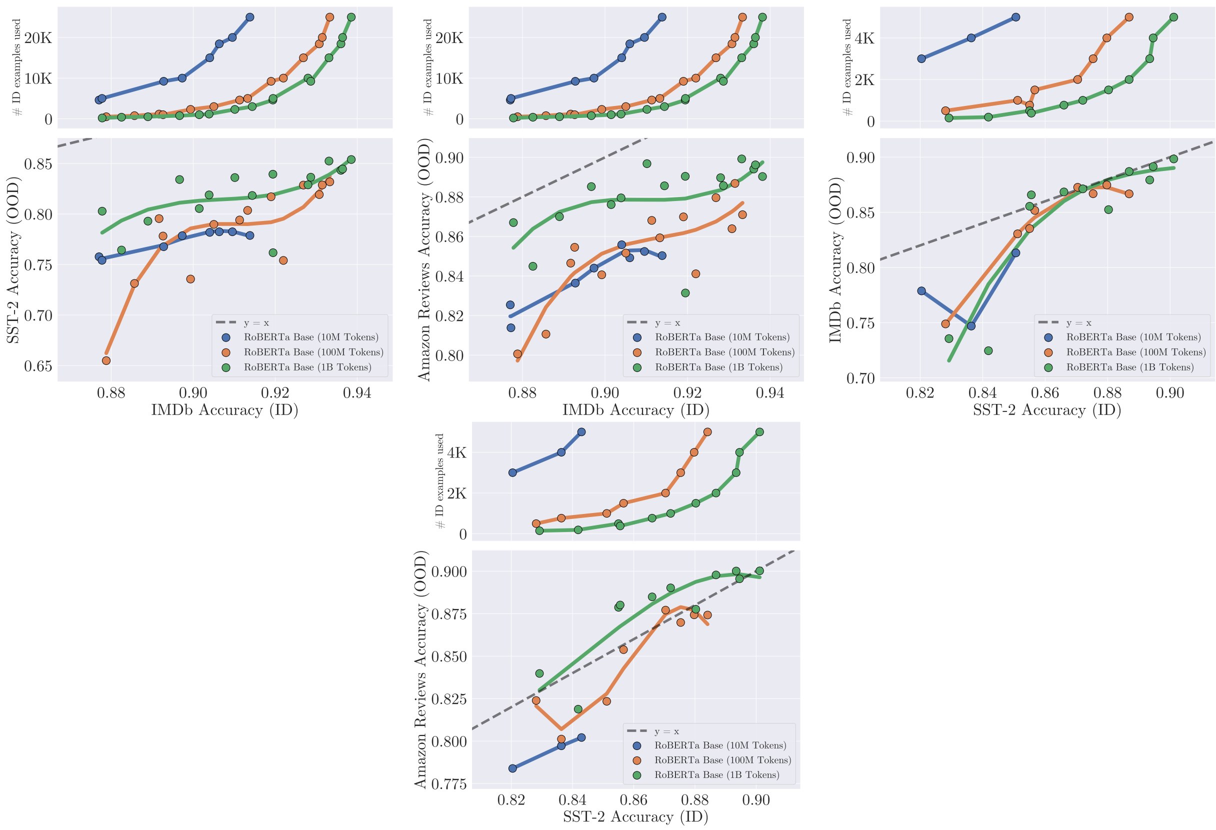


Figure 17: Results on all sentiment ID-OOD settings when increasing the amount of pre-training data.

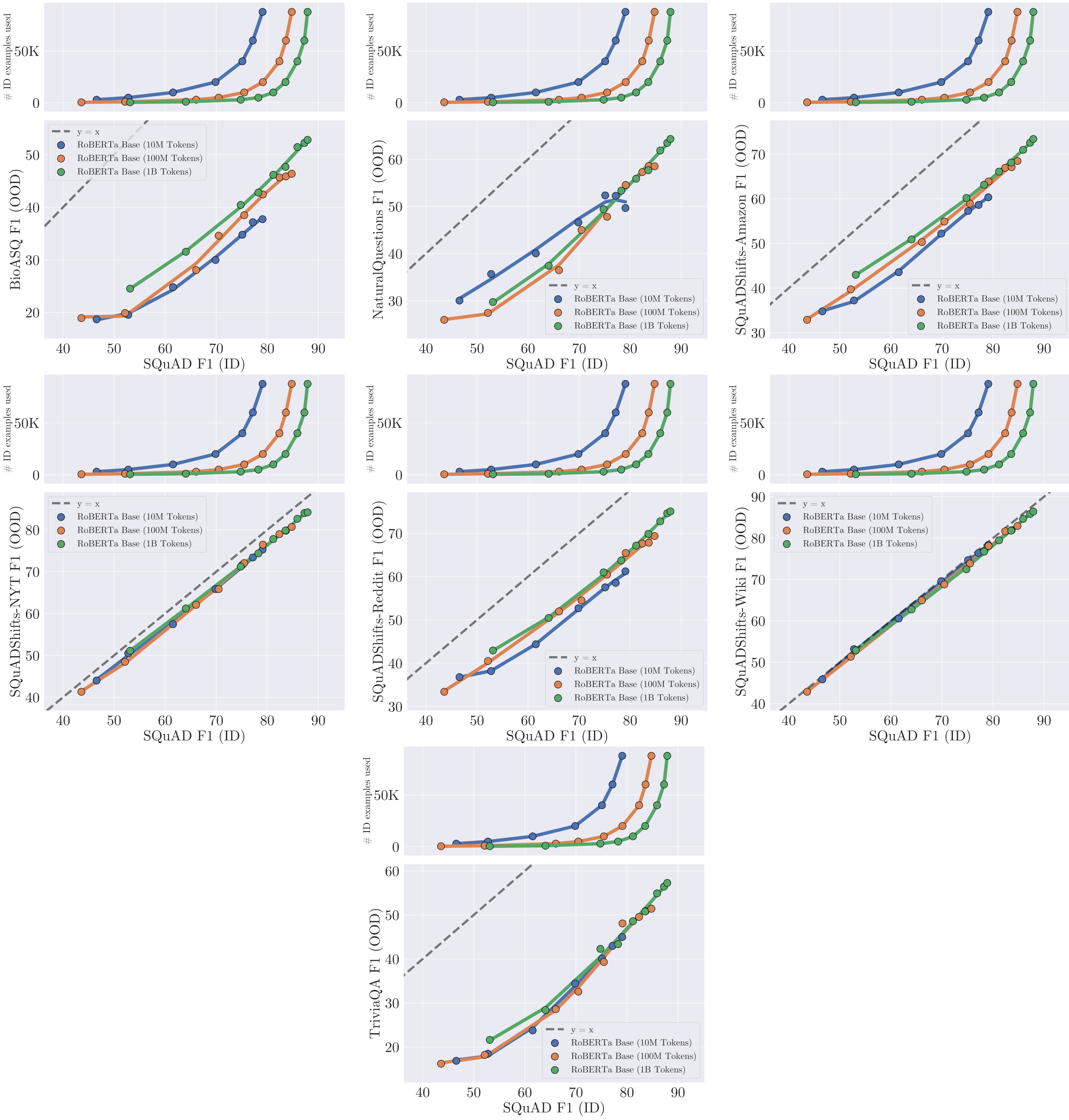


Figure 18: Results on all extractive QA OOD settings when training on SQuAD with models pre-trained on varying amounts of data.

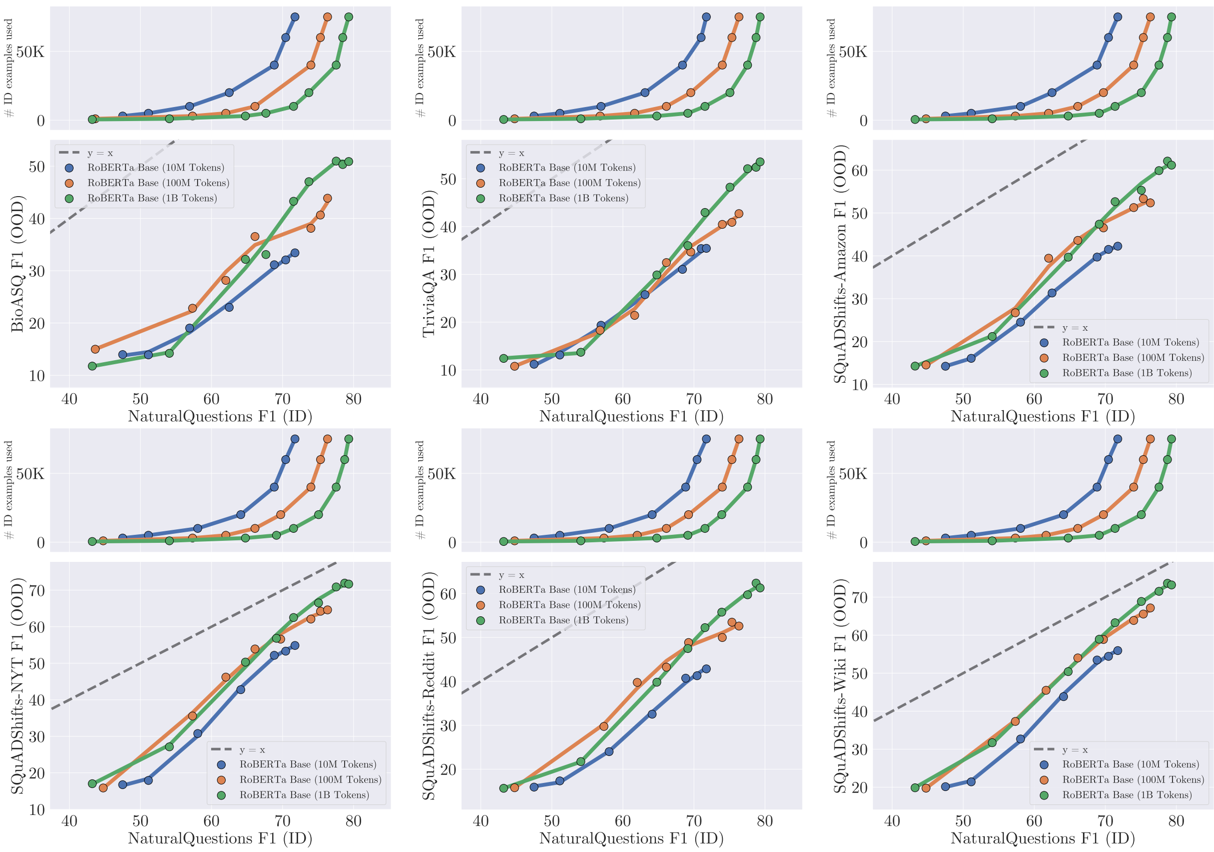


Figure 19: Results on all extractive QA OOD settings when training on NaturalQuestions with models pre-trained on varying amounts of data.