Are Larger Pretrained Language Models Uniformly Better? Comparing Performance at the Instance Level

Ruiqi Zhong  Dhruba Ghosh  Dan Klein  Jacob Steinhardt
Computer Science Division, University of California, Berkeley
{ruiqi-zhong, djghosh13, klein, jsteinhardt}@berkeley.edu

Abstract
Larger language models have higher accuracy on average, but are they better on every single instance (datapoint)? Some work suggests larger models have higher out-of-distribution robustness, while other work suggests they have lower accuracy on rare subgroups. To understand these differences, we investigate these models at the level of individual instances. However, one major challenge is that individual predictions are highly sensitive to noise in the randomness in training. We develop statistically rigorous methods to address this, and after accounting for pretraining and finetuning noise, we find that our BERT-LARGE is worse than BERT-MINI on at least 1–4% of instances across MNLI, SST-2, and QQP, compared to the overall accuracy improvement of 2–10%. We also find that finetuning noise increases with model size, and that instance-level accuracy has momentum: improvement from BERT-MINI to BERT-MEDIUM correlates with improvement from BERT-MEDIUM to BERT-LARGE. Our findings suggest that instance-level predictions provide a rich source of information; we therefore recommend that researchers supplement model weights with model predictions.

1 Introduction
Historically, large deep learning models (Peters et al., 2018; Devlin et al., 2019; Lewis et al., 2020; Raffel et al., 2019) have improved the state of the art on a wide range of tasks and leaderboards (Schwartz et al., 2014; Rajpurkar et al., 2016; Wang et al., 2018), and empirical scaling laws predict that larger models will continue to increase performance (Kaplan et al., 2020). However, little is understood about such improvement at the instance (datapoint) level. Are larger models uniformly better? In other words, are larger pretrained models better at every instance, or are they better at some instances, but worse at others?

Prior works hint at differing answers. Hendrycks et al. (2020) and Desai and Durrett (2020) find that larger pretrained models consistently improve out-of-distribution performance, which implies that they might be uniformly better at a finer level. Henighan et al. (2020) claim that larger pretrained image models have lower downstream classification loss for the majority of instances, and they predict this trend to be true for other data modalities (e.g. text). On the other hand, Sagawa et al. (2020) find that larger non-pretrained models perform worse on rare subgroups; if this result generalizes to pretrained language models, larger models will not be uniformly better. Despite all the indirect evidence, it is still inconclusive how many instances larger pretrained models perform worse on.

A naïve solution is to finetune a larger model, compare it to a smaller one, and find instances where the larger model is worse. However, this approach is flawed, since model predictions are noisy at the instance level. On MNLI in-domain development set, even the same architecture with different finetuning seeds leads to different predictions on ∼8% of the instances. This is due to under-specification (D’Amour et al., 2020), where there are multiple different solutions that can minimize the training loss. Since the accuracy improvement from our BERT-BASE to BERT-LARGE is 2%, most signals across different model sizes will be dominated by noise due to random seeds.

To account for the noise in pretraining and finetuning, we define instance accuracy as “how often a model correctly predicts an instance” (Figure 1 left) in expectation across pretraining and finetuning seeds. We estimate this quantity by pretraining 10 models with different seeds, finetuning 5 times for each pretrained models (Figure 1 middle), and

---

1This is not the original release by Devlin et al. (2019); we pretrained models ourselves.
averaging across them.

However, this estimate is still inexact, and we might falsely observe smaller models to be better at some instances by chance. Hence, we propose a random baseline to estimate the fraction of \textit{false discoveries} (Section 3, Figure 1 right) and formally upper-bound the false discoveries in Section 4. Our method provides a better upper bound than the classical Benjamini-Hochberg procedure with Fisher’s exact test.

Using the 50 models for each size and our improved statistical tool, we find that, on the MNLI in-domain development set, the accuracy “decays” from BERT-LARGE to BERT-MINI on at least \(~4\)% of the instances, which is significant given that the improvement in overall accuracy is 10%. These decaying instances contain more controversial or wrong labels, but also correct ones (Section 4.2). Therefore, larger pretrained language models are not uniformly better.

We make other interesting discoveries at the instance level. Section 5 finds that instance-level accuracy has momentum: improvement from MINI to MEDIUM correlates with improvement from MEDIUM to LARGE. Additionally, Section 6 attributes variance of model predictions to pretraining and finetuning random seeds, and finds that finetuning seeds cause more variance for larger models. Our findings suggest that instance-level predictions provide a rich source of information; we therefore recommend that researchers supplement model weights with model predictions. In this spirit, we release all the pretrained models, model predictions, and code here: https://github.com/ruiqi-zhong/acl2021-instance-level.

2 Data, Models, and Predictions

To investigate model behavior, we considered different sizes of the BERT architecture and fine-tuned them on Quora Question Pairs (QQP\textsuperscript{2}), Multi-Genre Natural Language Inference (MNLI; Williams et al. (2020)), and the Stanford Sentiment Treebank (SST-2; Socher et al. (2013)). To account for pretraining and finetuning noise, we averaged over multiple random initializations and training data order, and thus needed to pretrain our own models rather than downloading off the internet. Following Turc et al. (2019) we trained 5 architectures of increasing size: MINI (L4/H256, 4 Layers with hidden dimension 256), SMALL (L4/H512), MEDIUM (L8/H512, BASE (L12/H768), and LARGE (L24/H1024). For each architecture we pre-trained models with 10 different random seeds and fine-tuned each of them 5 times (50 total) on each task; see Figure 1 middle. Since pretraining is computationally expensive, we reduced the context size during pretraining from 512 to 128 and compensated by increasing training steps from 1M to 2M. Appendix A includes more details about pretraining and finetuning and their computational cost, and Appendix B verifies that our cost-saving changes do not affect accuracy qualitatively.

\textbf{Notation.} We use $i$ to index an instance in the evaluation set, $s$ for model sizes, $P$ for pretraining seeds and $F$ for finetuning seeds. $c$ is a random variable of value 0 or 1 to indicate whether the prediction is correct. Given the pretraining seed $P$ and the finetuning seed $F$, $c^i_s = 1$ if the model of size $s$ is correct on instance $i$, 0 otherwise. To keep the notation uncluttered, we sometimes omit these superscripts or subscripts if they can be inferred from context.

Unless otherwise noted, we present results on the MNLI in-domain development set in the main paper.

3 Comparing Instance Accuracy

To find the instances where larger models are worse, a naïve approach is to finetune a larger pretrained model, compare it to a smaller one, and find instances where the larger is incorrect but the smaller is correct. Under this approach, BERT-LARGE is worse than BERT-BASE on 4.5% of the instances and better on 7%, giving an overall accuracy improvement of 2.5%.

However, this result is misleading: even if we compare two BERT-BASE model with different finetuning seeds, their predictions differ on 8% of the instances, while their accuracies differ only by 0.1%; Table 1 reports this baseline randomness across model sizes. Changing the pretraining seed also changes around 2% additional predictions beyond finetuning. Table 1 also reports the standard deviation of overall accuracy, which is about 40 times smaller. Such stability starkly contrasts with the noisiness at the instance level, which poses a unique challenge.

\textsuperscript{2}https://www.quora.com/q/quoradata/First-Quora-Dataset-Release-Question-Pairs
Figure 1: Left: Each column represents the same architecture trained with a different seed. We calculate accuracy for each instance (row) by averaging across seeds (column), while it is usually calculated for each model by averaging across instances. Middle: A visual layout of the model predictions we obtain, which is a binary-valued tensor with 4 axes: model size $s$, instance $i$, pretraining seeds $P$ and finetuning seeds $F$. Right: for each instance, we calculate the accuracy gain from MINI to LARGE and plot the histogram in blue, along with a random baseline in red. Since the blue distribution has a bigger left tail, smaller models are better at some instances.

Table 1: Larger model sizes are at the bottom rows. Diff$^\text{FTune}$: how much do the predictions differ, if two models have the same pretraining seed but different finetuning seeds $F$? Diff$^\text{PTrain}$: the difference if the pretraining seeds $P$ are different. Std$_{all}$: the standard deviation of overall accuracy, around 40 times smaller than Diff$^\text{FTune}$.

|        | Diff$^\text{FTune}$ | Diff$^\text{PTrain}$ | Std$_{all}$ |
|--------|---------------------|----------------------|------------|
| MINI   | 7.2%                | 10.7%                | 0.2%       |
| SMALL  | 7.2%                | 10.7%                | 0.3%       |
| MEDIUM | 8.0%                | 10.7%                | 0.3%       |
| BASE   | 8.5%                | 10.6%                | 0.2%       |
| LARGE  | 8.6%                | 10.1%                | 0.2%       |

Table 2: We histogram the $\hat{\Delta}$Acc$^s_i$ in Figure 2 (b). We observe a unimodal distribution centered near 0, with tails on both sides. Therefore, the estimated differences for some instances are negative. However, due to estimation noise, we might falsely observe this accuracy decay by chance. Therefore, we introduce a random baseline $s_2$ $\hat{\Delta}$Acc$^s_i$ to control for these false discoveries. Recall that we have 10 smaller pretrained models and 10 larger ones. Our baseline splits these into a group $A$ of 5 smaller + 5 larger, and another group $B$ of the remaining 5 + 5. Then the empirical accuracies $\hat{\text{Acc}}^A$ and $\hat{\text{Acc}}^B$ are identically distributed, so we take our baseline $s_2$ $\hat{\Delta}$Acc$^s_i$ to be the difference $\hat{\text{Acc}}^A - \hat{\text{Acc}}^B$. We visualize and compare how to calculate $s_1$ $\hat{\Delta}$Acc$^s_i$ to $s_2$ $\hat{\Delta}$Acc$^s_i$ in Figure 3.

|        | Diff$^\text{FTune}$ | Diff$^\text{PTrain}$ | Std$_{all}$ |
|--------|---------------------|----------------------|------------|
| MINI   | 7.2%                | 10.7%                | 0.2%       |
| SMALL  | 7.2%                | 10.7%                | 0.3%       |
| MEDIUM | 8.0%                | 10.7%                | 0.3%       |
| BASE   | 8.5%                | 10.6%                | 0.2%       |
| LARGE  | 8.6%                | 10.1%                | 0.2%       |

4 Quantifying the Decaying Instances

The left tail of $\hat{\Delta}$Acc noiselessly estimates the fraction of decaying instances, and the left tail of the random baseline $\hat{\Delta}$Acc$^s_i$ counts the false discovery fraction due to the noise. Intuitively, the true fraction of decaying instances can be captured by the difference of these left tails, and we formally quantify this below.
Suppose instance $i$ is drawn from the empirical evaluation distribution. Then we can define the true decaying fraction $\hat{\text{Decay}}$ as

$$\text{Decay} := \mathbb{P}[\Delta \text{Acc}_i < 0].$$

Since $\Delta \text{Acc}_i$ is not directly observable and $\Delta \text{Acc}_i$ is noisy, we add a buffer and only consider instances with $\Delta \text{Acc}_i \leq t$, which makes it more likely (but still uncertain) that the true $\Delta \text{Acc}_i < 0$. We denote this “discovery fraction” $\hat{\text{Decay}}(t)$ as

$$\hat{\text{Decay}}(t) := \mathbb{P}[\hat{\Delta} \text{Acc}_i \leq t].$$

Similarly, we define a baseline control (false discovery fraction) $\hat{\text{Decay}}'(t) := \mathbb{P}[\hat{\Delta} \text{Acc}'_i \leq t]$. Hence, $\hat{\text{Decay}}$ and $\hat{\text{Decay}}'$ are the cumulative distribution function of $\Delta \text{Acc}$ and $\Delta \text{Acc}'$ (Figure 4).

We have the following theorem, which we formally state and prove in Appendix D:

**Theorem 1 (Informal)** If all the random seeds are independent, then for all thresholds $t$,

$$\text{Decay} \geq \mathbb{E}[\hat{\text{Decay}}(t) - \hat{\text{Decay}}'(t)].$$

**Proof Sketch** Suppose we observe $c^{R_1}_{R_2, \ldots, R_{2k}}$ and $c^{R_{2k+1}, \ldots, R_{4k}}$, where there are $2k$ different random seeds for each model size $^3$. Then

$$\hat{\Delta} \text{Acc}_i := \frac{1}{2k} \left( \sum_{j=1}^{2k} c^{R_1}_{R_j,i} - \sum_{j=2k+1}^{4k} c^{R_{2k+1}, \ldots, R_{4k}}_{R_j,i} \right),$$

and hence the discovery rate $\hat{\text{Decay}}(t)$ is defined as

$$\hat{\text{Decay}}(t) := \frac{1}{|T|} \sum_{i=1}^{|T|} 1[\Delta \text{Acc} \leq t].$$

For the random baseline estimator, we have

$$\Delta \text{Acc}'_i := \frac{1}{2k} \left( \sum_{j=1}^{k} c^{R_1}_{R_j,i} + \sum_{j=2k+1}^{3k} c^{R_{2k+1}, \ldots, R_{4k}}_{R_j,i} \right) - \sum_{j=k+1}^{2k} c^{R_1}_{R_j,i} - \sum_{j=3k+1}^{4k} c^{R_{2k+1}, \ldots, R_{4k}}_{R_j,i},$$

and the false discovery control $\hat{\text{Decay}}'$ is defined as

$$\hat{\text{Decay}}'(t) := \frac{1}{|T|} \sum_{i=1}^{|T|} 1[\Delta \text{Acc}'_i \leq t].$$

Formally, the theorem states that

$$\text{Decay} \geq \mathbb{E}_{R_1 \ldots R_{4k}}[\hat{\text{Decay}}(t) - \hat{\text{Decay}}'(t)],$$

which is equivalent to

$$\sum_{i=1}^{|T|} (1[\Delta \text{Acc}_i < 0] - \mathbb{P}[\hat{\Delta} \text{Acc}_i \leq t]) + \mathbb{P}[\Delta \text{Acc}'_i \leq t] \geq 0.$$
This is easy to see, since $\Delta \text{Acc}$ substantial difference, as discussed in Appendix D.1. that the relative bias is at most 10% using a bootstrap method. data may incur a slight upward bias. Appendix E estimates runs for each pretraining seed; these predictions ensemble our model across 5 different finetuning bound, we slightly modify our target of interest. In- independent. share the same pretraining seeds and hence are de- is not fully rigorous, since some finetuning runs are essentially the same as individual finetuning runs, except that the finetuning randomness is av- eraged out. Hence we obtain 10 independent sets of model predictions with different random seeds, which allows us to apply Theorem 1. We compute this lower bound for other pairs of model sizes in Table 2, and the full results across other tasks and model size pairs are in Appendix C. In all of these settings we find a non-zero fraction of decaying instances, and larger model size differences usually lead to more decaying instances. Unfortunately, applying Theorem 1 as above is not fully rigorous, since some finetuning runs share the same pretraining seeds and hence are de- pendent. 6 To obtain a statistically rigorous lower bound, we slightly modify our target of interest. Instead of examining individual finetuning runs, we ensemble our model across 5 different finetuning runs for each pretraining seed; these predictions are essentially the same as individual finetuning runs, except that the finetuning randomness is averaged out. Hence we obtain 10 independent sets of model predictions with different random seeds, which allows us to apply Theorem 1. We compare MINI to LARGE using these ensembles and report the discovery decay rate $\text{Decay'}$, the false discovery control $\text{Decay''}$, and their difference (Diff) under different thresholds $t$. LARGE is worse on at least ~4% (maximum Diff) of instances.

| Threshold | Decay | Decay' | Diff  |
|----------|-------|--------|-------|
| $t = -0.4$ | 4.22% | 3.49e−3 | 3.87% |
| $t = -0.9$ | 0.91% | 1.44e−7 | 0.91% |
| $t = -1.0$ | 0.48% | 2.06e−8 | 0.48% |

Table 3: Comparing MINI vs. LARGE by calculating the discovery fraction $\text{Decay}$, the false discovery control $\text{Decay'}$, and their difference (Diff) under different thresholds $t$. LARGE is worse on at least ~4% (maximum Diff) of instances.

4.1 Fisher’s Test + Benjamini-Hochberg
Here is a more classical approach to lower-bound the decaying fraction. For each instance, we compute a significance level $\alpha$ under the null hypothesis that the larger model is better, using Fisher’s exact test. We sort the significance levels ascendingly, and call the $p$th percentile $\alpha_p$. Then we pick a false discovery rate $q$ (say, 25%), find the largest $p$ s.t. $\alpha_p < pq$, and estimate the decaying fraction to be at least $p(1-q)$. This calculation is known as the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995).

To compare our method with this classical ap-
We next manually examine the decaying instances which is usually the regime of interest. For example, across all pairs of model sizes, our approach only needs 2 random seeds (i.e. pretrained models) to provide a non-zero lower bound on the decaying fraction, while the classical approach sometimes fails to do this even with 10 seeds. Intuitively, when fewer seeds are available, the smallest possible significance level for each instance is larger than the decaying fraction, hence hurting the classical approach.

4.2 Understanding the Decaying Instances

We next manually examine the decaying instances to see whether we can find any interpretable patterns. One hypothesis is that all the decaying fractions are in fact mislabeled, and hence larger models are not in fact worse on any instances.

To investigate this hypothesis, we examined the group of instances where \( \hat{\Delta} \text{Acc}_{s_1} \leq -0.9 \). \( \text{MINI} \) is almost always correct on these instances, while \( \text{LARGE} \) is almost always wrong, and the false discovery fraction is tiny. For each instance, we manually categorize it as either: 1) Correct, if the label is correct, 2) Fine, if the label might be controversial but we could see a reason why this label is reasonable, 3) Wrong, if the label is wrong, or 4) Unsure, if we are unsure about how to label this instance. Each time we annotate, with 50% probability we randomly sample either a decaying instance or an instance from the remaining dataset as a control. We are blind to which group it comes from.

For each task of MNLI, QQP, and SST-2, the first author annotated 100 instances (decay + control group) (Table 5). We present all the annotated decaying instances in Appendix J.

**Conclusion** We find that the decaying fraction has more wrong or controversial labels, compared to the remaining instances. However, even after we adjust for the fraction of incorrect labels, the Decay fraction still exceeds the false discovery control. This implies that \( \text{MINI} \) models are better than \( \text{LARGE} \) models on some correctly labeled instances. The second author followed the same procedure and reproduced the same qualitative results.

However, we cannot find an interpretable pattern for these correctly labeled decaying instances by simply eyeballing. We discuss future directions to discover interpretable categories in Section 7.

### 5 Correlation of Instance Difference

We next investigate whether there is a momentum of instance accuracy increase: for example, if the instance accuracy improves from \( \text{MINI}(s_1) \) to \( \text{MEDIUM}(s_2) \), is it more likely to improve from \( \text{MEDIUM}(s_2) \) to \( \text{LARGE}(s_3) \)?

The na"ive approach is to calculate the Pearson correlation coefficient between \( \Delta \text{Acc}_{s_1} \) and \( \Delta \text{Acc}_{s_2} \), and we find the correlation to be zero. However, this is partly an artifact of accuracies being bounded in \([0, 1]\). If \( \text{MEDIUM} \) drastically improves over \( \text{MINI} \) from 0 to 1, there is no room for \( \text{LARGE} \) to improve over \( \text{MEDIUM} \). To remove this inherent negative correlation, we calculate the correlation conditioned on the accuracy of the middle-

| s1   | s2   | Correct | Fine | Wrong | Unsure |
|------|------|---------|------|-------|--------|
| BASE | LARGE| ours    | 1.9% | 3.1%  | 4.0%   |
| BASE | LARGE| BH      | 0.0% | 0.9%  | 1.9%   |
| MINI | LARGE| ours    | 0.4% | 0.9%  | 1.2%   |
| MINI | LARGE| BH      | 0.0% | 0.0%  | 0.0%   |

Table 4: We compare our method to the Fisher’s exact test + Benjamin-Hochberg (BH) procedure described in Section 4. For all different model size pairs and number of pretrained models available, ours always provides a higher (better) lower bound of the decaying fraction.

|                  | Correct | Fine | Wrong | Unsure |
|------------------|---------|------|-------|--------|
| MNLI\(^D\)       | 66%     | 17%  | 9%    | 5%     |
| MNLI\(^C\)       | 86%     | 5%   | 5%    | 1%     |
| SST-2\(^D\)      | 55%     | 8%   | 10%   | 25%    |
| SST-2\(^C\)      | 88%     | 4%   | 0%    | 6%     |
| QQP\(^D\)        | 60%     | 26%  | 10%   | 2%     |
| QQP\(^C\)        | 87%     | 10%  | 1%    | 0%     |

Table 5: \( \text{MINI} \) vs. \( \text{LARGE} \). We examine whether there are mislabels for the Decaying fractions (superscript \(^D\)) and the rest of the dataset (Control group \(^C\)). The decaying fraction contains more mislabels, but includes correct labels as well.
Table 6: Each row corresponds to a triplet of model sizes. Each column represents a bucket that contains instances with $\hat{\text{Acc}}_s \in [t-0.1, t]$. Within each bucket, we calculate the Pearson correlation coefficient between the estimated accuracy improvements: $\hat{s}_2 \Delta \hat{\text{Acc}}$ and $\hat{s}_3 \Delta \hat{\text{Acc}}$. These correlations are positive and become higher when model size differences are small.

| $s_1, s_2, s_3$ | Buckets | 0.10 | 0.20 | 0.30 | 0.40 | 0.50 | 0.60 | 0.70 | 0.80 | 0.90 | 1.00 |
|-----------------|---------|------|------|------|------|------|------|------|------|------|------|
| SMALL, MEDIUM, BASE | 0.07 | 0.22 | 0.29 | 0.40 | 0.35 | 0.33 | 0.38 | 0.27 | 0.24 | 0.13 |
| MINI, MEDIUM, LARGE | 0.03 | 0.15 | 0.18 | 0.33 | 0.17 | 0.16 | 0.22 | 0.20 | 0.19 | 0.09 |

Table 7: The bias, pretraining variance, and finetuning variance for each model size, averaged across all test instances. Finetuning variance is much larger than pretraining variance; larger models have larger finetuning variance.

|                | Bias$^2$ | PretVar | FineVar |
|----------------|----------|---------|---------|
| MINI           | 0.203    | 0.017   | 0.036   |
| SMALL          | 0.179    | 0.017   | 0.036   |
| MEDIUM         | 0.157    | 0.014   | 0.040   |
| BASE           | 0.134    | 0.010   | 0.043   |
| LARGE          | 0.111    | 0.007   | 0.043   |

Table 6: Each row corresponds to a triplet of model sizes. Each column represents a bucket that contains instances with $\hat{\text{Acc}}_s \in [t-0.1, t]$. Within each bucket, we calculate the Pearson correlation coefficient between the estimated accuracy improvements: $\hat{s}_2 \Delta \hat{\text{Acc}}$ and $\hat{s}_3 \Delta \hat{\text{Acc}}$. These correlations are positive and become higher when model size differences are small.

| $s_1, s_2, s_3$ | Buckets | 0.10 | 0.20 | 0.30 | 0.40 | 0.50 | 0.60 | 0.70 | 0.80 | 0.90 | 1.00 |
|-----------------|---------|------|------|------|------|------|------|------|------|------|------|
| SMALL, MEDIUM, BASE | 0.07 | 0.22 | 0.29 | 0.40 | 0.35 | 0.33 | 0.38 | 0.27 | 0.24 | 0.13 |
| MINI, MEDIUM, LARGE | 0.03 | 0.15 | 0.18 | 0.33 | 0.17 | 0.16 | 0.22 | 0.20 | 0.19 | 0.09 |

Table 7: The bias, pretraining variance, and finetuning variance for each model size, averaged across all test instances. Finetuning variance is much larger than pretraining variance; larger models have larger finetuning variance.

We may conclude from Table 7 that larger models have larger finetuning variance and smaller pretraining variance. However, lower bias also inherently implies lower variance. To see this, suppose a model has perfect accuracy and hence zero bias;
then it always predicts the same label (the correct one) and hence has zero variance. This might favor larger models and “underestimate” their variance, since they have lower bias. Therefore, we calculate and compare the variances conditioned on the bias, i.e. \( \text{PretVar}(b^2) := \mathbb{E}_i[\text{PretVar}_i | \text{Bias}^2_i = b^2] \).

We estimate \( \text{PretVar}^+ (b^2) \) using Gaussian process regression and plot it against \( b^2 \) in Figure 5. We find that larger models still have lower pretraining variance across all levels of bias on the specific task of MNLI under the 0/1 loss. To further check whether our conclusions are general, we tested them on other tasks and under the squared loss \( L_i := (1 - p_i)^2 \), where \( p_i \) is the probability assigned to the correct class. Below are the conclusions that generally hold across different tasks and loss functions.

**Conclusion** We find that 1) larger models have larger finetuning variance, 2) \( \text{LARGE} \) has smaller pretraining variance than \( \text{BASE} \); however, the ordering between other sizes varies across tasks and losses, and 3) finetuning variance is 2−8 times as large as pretraining variance, and the ratio is bigger for larger models.

7 Discussion and Future Directions

To investigate model behaviors at the instance level, we produced massive amounts of model predictions in Section 2 and treated them as raw data. To extract insights from them, we developed better metrics and statistical tools, including a new method to control the false discoveries, an unbiased estimator for the decomposed variances, and metrics that compute variance and correlation of improvements conditioned on instance accuracy. We find that larger pretrained models are indeed worse on a non-trivial fraction of instances and have higher variance due to finetuning seeds; additionally, instance accuracy improvements from \( \text{MINI} \) to \( \text{MEDIUM} \) correlate with improvements from \( \text{MEDIUM} \) to \( \text{LARGE} \).

Overall, we treated model prediction data as the central object and built analysis tools around them to obtain a finer understanding of model performance. We therefore refer to this paradigm as “instance level understanding as data mining”. We discuss three key factors for this paradigm to thrive: 1) scalability and the cost of obtaining prediction data, 2) other information to collect for each instance, and 3) better statistical tools. We analyze each of these aspects below.

**Scalability and Cost of Data** Data mining is more powerful with more data. How easy is it to obtain more model predictions? In our paper, the main bottleneck is pretraining. However, once the pretrained models are released, individual researchers can download them and only need to repeat the cheaper finetuning procedure. Furthermore, model prediction data are under-shared: while many recent research papers share their code or even model weights to help reproduce the results, it is not yet a standard practice to share all the model predictions. Since many researches follow almost the same recipe of pretraining and finetuning (McCoy et al., 2020; Desai and Durrett, 2020; Dodge et al., 2020), much computation can be saved if model predictions are shared. On the other hand, as the state of the art model size is increasing at a staggering speed\(^7\), most researchers will not be able to run inference on a single instance. The trend that models are becoming larger and more similar necessitate more prediction sharing.

**Meta-Labels and Other Predictions** Data mining is more powerful with more types of information. One way to add information to each instance is to assign “meta-labels”. In the HANS (McCoy et al., 2019) dataset, the authors tag each instance with a heuristic\(^8\) that holds for the training distribution but fails on this instance. Naik et al. (2018a) and Ribeiro et al. (2020) associate each instance

---

\(^7\)e.g. BERT (Devlin et al., 2019) has 340M parameters, while Switch-Transformer has over 1 trillion parameters (Fedus et al., 2021).

\(^8\)For example, “the label [entailment] is likely if the premise and the hypothesis have significant lexical overlap”.

---

Figure 5: The pretraining variance conditioned on \( \text{Bias}^2 \) (the level of correctness). Each curve represents a model size. Larger models have lower pretraining variance across all levels of bias.
with a particular stress test type or subgroup, for example, whether the instance requires the model to reason numerically or handle negations. Nie et al. (2020) collects multiple human responses to estimate human disagreement for each instance. This meta-information can potentially help us identify interpretable patterns for the disagreeing instances where one model is better than the other. On the flip side, identifying disagreeing instances between two models can also help us generate hypothesis and decide what subgroup information to annotate.

We can also add performance information on other tasks to each instance. For example, Pruk-sachatkun et al. (2020) studied the correlation between syntactic probing accuracy (Hewitt and Liang, 2019) and downstream task performance. Turc et al. (2019) and Kaplan et al. (2020) studied the correlation between language modelling loss and the downstream task performance. However, they did not analyze correlations at the instance level. We may investigate whether their results hold on the instance level: if an instance is easier to tag by a probe or easier to predict by a larger language model, is the accuracy likely to be higher?

Statistical Tools Data mining is more powerful with better statistical tools. Initially we used the Benjamini-Hochberg procedure with Fisher’s exact test, which required us to pretrain 10 models to formally verify that the decaying instances exist. However, we later realized that 2 is in fact enough by using our approach introduced in Section 4. We could have saved 80% of the computation for pretraining if this approach was known before we started.

Future work can explore more complicated metrics and settings. We compared at most 3 different model sizes at a time, and higher order comparisons require novel metrics. We studied two sources of randomness, pretraining and finetuning, but other sources of variation can be interesting as well, e.g. differences in pretraining corpus, different model checkpoints, etc. To deal with more sophisticated metrics, handle different sources and hierarchies of randomness, and reach conclusions that are robust to noises at the instance level, researchers need to develop new inference procedures.

To conclude, for better instance level understanding, we need to produce and share more prediction data, annotate more diverse linguistic properties, and develop better statistical tools to infer under noises. We hope our work can inform researchers about the core challenges underlying instance level understanding and inspire future work.

Acknowledgement

We thank Steven Cao, Cathy Chen, Frances Ding, David Gaddy, Colin Li, and Alex Wei for giving comments on the initial paper draft. We would also like to thank the Google Cloud TPU team for their hardware support.

References

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Alexander D’Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D Hoffman, et al. 2020. Under specification presents challenges for credibility in modern machine learning. *arXiv preprint arXiv:2011.03395*.

Shrey Desai and Greg Durrett. 2020. Calibration of pre-trained transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 295–302, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. 2020. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *arXiv preprint arXiv:2002.06305*.

William Fedus, Barret Zoph, and Noam Shazeer. 2021. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *arXiv preprint arXiv:2101.03961*.
Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedzic, Rishabh Krishnan, and Dawn Song. 2020. Pretrained transformers improve out-of-distribution robustness. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 2744–2751, Online. Association for Computational Linguistics.

Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B Brown, Pranav Raghavendra, Scott Gray, et al. 2020. Scaling laws for autoregressive generative modeling. arXiv preprint arXiv:2010.14701.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. 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 2733–2743, Hong Kong, China. Association for Computational Linguistics.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. arXiv preprint arXiv:2001.08361.

Wuwei Lan, Siyu Qiu, Hua He, and Wei Xu. 2017. A continuously growing dataset of sentential paraphrases. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 1224–1234, Copenhagen, Denmark. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7871–7880, Online. Association for Computational Linguistics.

Zhuohan Li, Eric Wallace, Sheng Shen, Kevin Lin, Kurt Keutzer, Dan Klein, and Joseph E. Gonzalez. 2020. Train large, then compress: Rethinking model size for efficient training and inference of transformers. In International Conference on Machine Learning.

R. Thomas McCoy, Junghyun Min, and Tal Linzen. 2020. BERTs of a feather do not generalize together: Large variability in generalization across models with similar test set performance. In Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, pages 217–227, Online. Association for Computational Linguistics.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Aaakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018a. Stress test evaluation for natural language inference. In The 27th International Conference on Computational Linguistics (COLING), Santa Fe, New Mexico, USA.

Aaakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018b. Stress test evaluation for natural language inference. In Proceedings of the 27th International Conference on Computational Linguistics, pages 2340–2353, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Yixin Nie, Xiang Zhou, and Mohit Bansal. 2020. What can we learn from collective human opinions on natural language inference data? In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 9131–9143, Online. Association for Computational Linguistics.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Yada Pruksachatkun, Jason Phang, Haokun Liu, Phu Mon Htut, Xiaoyi Zhang, Richard Yuanzhe Pang, Clara Vania, Katharina Kann, and Samuel R. Bowman. 2020. Intermediate-task transfer learning with pretrained language models: When and why does it work? In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 5231–5247, Online. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. arXiv preprint arXiv:1910.10683.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4902–4912, Online. Association for Computational Linguistics.
Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. 2020. An investigation of why over-parameterization exacerbates spurious correlations. In Proceedings of the 37th International Conference on Machine Learning, volume 119 of Proceedings of Machine Learning Research, pages 8346–8356. PMLR.

Lane Schwartz, Timothy Anderson, Jeremy Gwinnup, and Katherine Young. 2014. Machine translation and monolingual postediting: The AFRL WMT-14 system. In Proceedings of the Ninth Workshop on Statistical Machine Translation, pages 186–194, Baltimore, Maryland, USA. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

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 preprint arXiv:1908.08962v2.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Adina Williams, Tiago Pimentel, Hagen Blix, Arya D. McCarthy, Eleanor Chodroff, and Ryan Cotterell. 2020. Predicting declension class from form and meaning. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 6682–6695, Online. Association for Computational Linguistics.
A  Pretraining and Finetuning Details

Here we explain how to obtain the model predictions, which are analyzed in later sections. To obtain these predictions under the “pretraining and finetuning” framework (Devlin et al., 2019), we need to decide a model size, perform pretraining, finetune on a training set with a choice of hyperparameters, and test the model on an evaluation set. We discuss each bolded aspect below.

Size  Similar to Turc et al. (2019), we experimented with the following five model sizes, listed in increasing order: MINI (L4/H256) 9, SMALL (L4/H512), MEDIUM (L8/H512), BASE (L12/H768), and LARGE (L24/H1024).

Pretraining  We used the pretraining code from Devlin et al. (2019) and the pre-training corpus from Li et al. (2020). Compared to the original BERT release, we used context size 128 instead of 512, since computation cost grows quadratically with respect to context size; we also pretrained for 2M steps instead of 1M.

Training Set  We consider 3 datasets: Quora Question Pairs (QQP) 10, Multi-Genre Natural Language Inference (MNLI; Williams et al. (2020)), and the Stanford Sentiment Treebank (SST-2; Socher et al., 2013). For QQP we used the official training split. For MNLI we used 350K out of 400K instances from the original training split, and added the remaining 50K to the evaluation set, since the original in-domain development set only contains 10K examples. For SST-2, we mix the training and development set of the original split, split the instances into 5 folds, train on four of them, and evaluate on the remaining fold.

Hyperparameters  As in Turc et al. (2019), we finetune 4 epochs for each dataset. For each task and model size, we tune hyperparameters in the following way: we first randomly split our new training set into 80% and 20%; then we finetune on the 80% split with all 9 combination of batch size [16, 32, 64] and learning rate [1e-4, 5e-5, 3e-5], and choose the combination that leads to the best average accuracy on the remaining 20%.

Evaluation Set  After finetuning our pretrained models, we evaluate them on a range of in-domain, out-of-domain, or challenging datasets to obtain model predictions. Models trained on MNLI are also evaluated on Stanford Natural Language Inference (SNLI; (Bowman et al., 2015)), Heuristic Analysis for NLI Systems (HANS; (McCoy et al., 2019)), and stress test evaluations (STRESS; (Naik et al., 2018b)). Models trained on QQP are also evaluated on Twitter Paraphrase Database (TwitterPPDB; (Lan et al., 2017)).

Since pretraining introduces randomness, for each model size s, we pretrain 10 times with different random seed P; since finetuning also introduces noise, for each pretrained model we pretrain 5 times with different random seed F; besides, we also evaluate the model at the checkpoints after E epochs, where E ∈ [3, 3.117, 3.5, 4].

Pretraining 10 models for all 5 model sizes altogether takes around 3840 hours on TPU v3 with 8 cores. Finetuning all of them 5 times for all three tasks in our paper requires around 1200 hours.

B  Compare Our Models to the Original

Since we decreased the pre-training context length to save computation, these models are not exactly the same as the original BERT release by Devlin et al. (2019) and Turc et al. (2019). We need to benchmark our model against theirs to ensure that the performance of our model is still reasonable and the qualitative trend still holds. For each each size and task, we finetune the original model 5 times and calculate the average of overall accuracy.

The comparison can be seen in Table 8. We find that our model does not substantially differ from the original ones on QQP and SST-2. On MNLI, the performance of our BERT-BASE and BERT-LARGE is 2~3% below the original release, but the qualitative trend that larger models have better accuracy still holds robustly.

C  More Instance Difference Results

Similar to Figure 4, for all 10 pairs of model sizes and all in-distribution instances of MNLI, SST-2, and QQP, we plot the cumulative density of ∆Acc and ∆Acc′, or say, Decay(t) and Decay′(t) in Figure 6, 7, and 8.

Additionally, for each pair of model sizes s1 and s2, we estimate “how much instances are getting better/worse accuracy?” by taking the maximum difference between the red curve and the blue curve. We report these results for MNLI, SST-2, and QQP in Table 9. We find that larger model size gaps

---

9 Layers with hidden dimension 256
10https://www.quora.com/q/quoradata/First-Quora-Dataset-Release-Question-Pairs
lead to larger decaying fraction, but also larger improving fraction as well.

D Proof of Theorem 1

Formal Setup Our goal is to show that if all the random seeds are independent,

\[ \text{Decay} \geq \mathbb{E}[\hat{\text{Decay}}(t) - \hat{\text{Decay}}'(t)] \] (16)

More concretely, suppose each instance is indexed by \( i \), the set of all instances is \( T \), and the random seed is \( R \); then \( c_R \in \{0, 1\}^{|T|} \) is a random \( |T| \) dimensional vector, where \( c_R^{s_i} = 1 \) if the model of size \( s \) correctly predicts instance \( i \) under the random seed \( R \). We are comparing model size \( s_1 \) and \( s_2 \), where \( s_2 \) is larger; to keep notation uncluttered, we omit these indexes whenever possible.

Suppose we observe \( c_{R_1}^{s_1} \) and \( c_{R_2}^{s_2} \), where there are \( 2k \) different random seeds for each model size \(^{11}\). Then

\[ \hat{\Delta}\text{Acc}_i := \frac{1}{2k} \left( \sum_{j=1}^{2k} c_{R_j,i}^{s_1} - \sum_{j=2k+1}^{4k} c_{R_j,i}^{s_2} \right), \] (17)

and hence the discovery rate \( \hat{\text{Decay}}(t) \) is defined as

\[ \hat{\text{Decay}}(t) := \frac{1}{|T|} \sum_{i=1}^{|T|} 1[\hat{\Delta}\text{Acc} \leq t]. \] (18)

For the random baseline estimator, we have

\[ \Delta\text{Acc}_i' := \frac{1}{2k} \left( \sum_{j=1}^{k} c_{R_j,i}^{s_1} + \sum_{j=k+1}^{3k} c_{R_j,i}^{s_2} - \sum_{j=k+1}^{2k} c_{R_j,i}^{s_1} - \sum_{j=4k+1}^{6k} c_{R_j,i}^{s_2} \right), \] (19)

and the false discovery control \( \hat{\text{Decay}}' \) is defined as

\[ \hat{\text{Decay}}'(t) := \frac{1}{|T|} \sum_{i=1}^{|T|} 1[\Delta\text{Acc}_i' \leq t]. \] (20)

To reiterate, the definition of the true decay rate is

\[ \text{Decay} = \frac{1}{|T|} \sum_{i=1}^{|T|} 1[\Delta\text{Acc}_i < 0]. \] (21)

Our goal is to prove that

\[ \text{Decay} \geq \mathbb{E}_{R_1 \ldots R_4} [\hat{\text{Decay}}(t) - \hat{\text{Decay}}'(t)] \] (22)

Proof By re-arranging terms and linearity of expectation, Equation 22 is equivalent to the following

\[ \sum_{i=1}^{|T|} \{1[\Delta\text{Acc}_i < 0] - \mathbb{P}[\Delta\text{Acc}_i < t]\} + \mathbb{P}[\Delta\text{Acc}_i' \leq t] \geq 0 \] (23)

Hence, we can declare victory if we can prove that for all \( i \),

\[ 1[\Delta\text{Acc}_i < 0] - \mathbb{P}[\Delta\text{Acc}_i < t] \geq 0 \] (24)

To prove Equation 24, we observe that if \( \text{Acc}_i < 0 \), since the probabilities are bounded by 0 and 1, its left-hand side must be positive. Therefore, we only need to prove that

\[ \Delta\text{Acc}_i \geq 0 \] (25)

which will be proved in Lemma 1. \( \square \)

Lemma 1

\[ \Delta\text{Acc}_i \geq 0 \] (26)

\[ \Rightarrow \mathbb{P}[\Delta\text{Acc}_i' \leq t] \geq \mathbb{P}[\Delta\text{Acc}_i \leq t], \]
Table 9: On QQP, MNLI in domain development set and SST-2 we lowerbound the fraction of instances that improves when model size changes from $s_1$ (row) to $s_2$ (column).

For $m = 1, 2$, define

$$p_i^{s_m} := \mathbb{E}[c_i^s_m],$$  \hfill (27)$$

then

$$p_i^{s_1} \leq p_i^{s_2}$$  \hfill (28)

Since $c_i^{s_1}$ and $c_i^{s_2}$ are both Bernoulli random variables with rate $p_i^{s_1}$ and $p_i^{s_2}$ respectively, we can write down the probability distribution of $\hat{\Delta}Acc_i$ and $\Delta Acc'_i$ as the sum/difference of several binomial variables, i.e.

$$\Delta Acc_i \sim (\text{Binom}(k, p_i^{s_2}) + \text{Binom}(k, p_i^{s_1}) - \text{Binom}(k, p_i^{s_1})) / 2k,$$  \hfill (29)

and

$$\Delta Acc'^i \sim (\text{Binom}(k, p^{s_2,i}) + \text{Binom}(k, p^{s_1,i}) - \text{Binom}(k, p^{s_1,i})) / 2k,$$  \hfill (30)

Since $p_i^{s_1} \leq p_i^{s_2}$, $\text{Binom}(k, p^{s_2,i})$ first order stochastically dominates $\text{Binom}(k, p^{s_1,i})$. Therefore, $\Delta Acc'^i$ dominates $\Delta Acc_i$, hence completing the proof. \hfill \Box

### D.1 Independent Seed Assumption

We notice that Theorem 1 requires the seeds $R$ to be independent. This assumption does not hold on our data, since some finetuning runs share the same pretraining seeds. Therefore, the above proof no longer holds. Specifically, Lemma 1 fails because $\Delta Acc$ and $\Delta Acc'$ are no longer binomial variables, and the later does not necessarily dominate the first. Here is a counter-example, if the seeds are not entirely independent.

Hypothetically, suppose we are comparing a smaller model $s_1$ and a larger model $s_2$. For the smaller model, with 0.1 probability it finds a perfect pretrained model that always predict correctly across all finetuning runs and with 0.9 probability it finds a bad pretrained model that predict always incorrectly. For the larger model, with probability 1 it finds an average pretrained model that predict correctly for 0.2 fraction of finetuning runs. The larger model is on average better, because it has $0.2 > 0.1$ probability to be correct. Hence, $\Delta Acc > 0$.

Suppose we observe 2 independent pretraining seeds for each size and infinite number of finetuning seeds for each pretraining seed, and let us consider the threshold -0.8. Then

$$\mathbb{P}[\hat{\Delta}Acc_i \leq -0.8] = 0.01 \geq 0 = \mathbb{P}[\Delta Acc'_i \leq -0.8]$$  \hfill (31)

The event that $\hat{\Delta}Acc_i \leq -0.8$ happens with probability 0.01 when both of the two small pretrained models have good pretraining seeds, and $\Delta Acc'_i$ is at least -0.5 and will never be less than -0.8.
Figure 6: Similar to Figure 4, on MNLI in-distribution development set, for each pair of model sizes, we plot the cumulative density function of instance differences.

Figure 7: Similar to Figure 4, on SST-2, for each pair of model sizes, we plot the cumulative density function of instance differences.

The key idea behind this counter-example is that even if the larger model has better average, the distribution of average finetuning accuracy for different pretraining seeds might not stochastically dominate the one with lower average because of outliers. Hence, a priori, this is unlikely to happen in practice, since pretraining variance is generally small, and we have multiple pretraining seeds to average out the outliers. Nevertheless, future work is needed to make a more rigorous argument.

E Upward Bias of Adaptive Thresholds

In section 3 we picked the best threshold that can maximize the lowerbound, which can incur a slight upward bias. Here we estimate that the bias is at most 10% relative to the unbiased lowerbound with a bootstrapping method.

We use the empirical distribution of 10 pre-trained models as the ground truth distribution for bootstrapping. We first compute a best threshold with 10 sampled smaller and larger pretrained models, and then compute the lowerbound \( L \) with this threshold on another sample of 10 smaller and larger models. Intuitively, we use one bootstrap sample (which contains 10 smaller pretrained models and 10 larger pretrained models) as the development set to “tune the threshold”, and then use this threshold on a fresh bootstrap sample to compute the lowerbound. We refer to the lowerbound that uses the best threshold as \( L^* \), and compute the relative error \( \frac{E[(L^* - L)]}{E[L]} \), where the expectation is taken with respect to bootstrap samples.

We report all results in Table 10. In general, we find that the upward bias is negligible, which is at most around 10%.
Comparison with Significance Testing

We also experimented with the classical approach that calculates the significance-level for each instance and then use the Benjamini-Hochberg procedure to lowerbound the decaying fraction. To make sure that we are comparing with this approach fairly, we lend it additional power by picking the false discovery rate that can maximize the true discovery counts. We report the decaying fraction on MNLI in-domain development set found by this classical method and compare it with our method for different model size differences in Table 11; we also simulate situations when we have fewer models.

In general, we find that our method always provide a tighter (higher) lowerbound than the classical method, and 2 models are sufficient to verify the existence (i.e. lowerbound \(> 0\)) of the decaying fraction; in contrast, the classical method sometimes fails to do this even with 10 models, e.g., when comparing BASE to LARGE.

Intuitively, our approach provides a better lower-bound because it better makes use of the information that on most instances, both the smaller and the larger models agree and predict completely correctly or incorrectly. For an extreme example, suppose we only observe 2 smaller models and 2 larger models, and infinite number of datapoints, whose predictions are independent. On 99.98% datapoints, both models have instance accuracy 1; on 0.01% datapoints, smaller model is completely wrong but bigger completely correct. Setting threshold to be 2, our decay estimate \(\hat{\text{Decay}}\) is 0.01%, while \(\text{Decay}' = 0\): since the models either completely predict correct or wrongly, there is never a false discovery. Therefore, our method can provide the tightest lowerbound 0.01% in this case. On the other hand, since we only have 4 models in total, the lowest significance-level given by the fisher exact test is 17% \(\gg\) 0.1%, hence the discovery made by the Benjamini-Hochberg procedure is 0.

More Results on Momentum

We report more results on the correlation between instance differences. Specifically, for one triplet of model sizes (e.g. MINI ⇒ MEDIUM ⇒ LARGE), for each group of instances that have similar \(\hat{\text{Acc}}_{\text{MEDIUM}}\), we calculate the correlation between instance differences, i.e. the Pearson-R score between \(\Delta \text{Acc}_{\text{MINI}}\) and \(\Delta \text{Acc}_{\text{MEDIUM}}\). All results can be seen in Table 12.

We observe that

- For nearly all buckets, the improvements are positively correlated.
- When model size gap becomes larger (e.g. MINI, MEDIUM, LARGE has the largest model size differences), the correlation decreases.

Loss Decomposition and Estimation

In this section, under the bias-variance decomposition and total variance decomposition framework, we decompose loss into four components: bias, variance brought by pretraining randomness, by
|       | MNLI $s_1 \setminus s_2$ | MINI | SMALL | MEDIUM | BASE | LARGE |
|-------|--------------------------|------|-------|--------|------|-------|
| MINI  | 0.000                    | 0.031| 0.027 | 0.026  | 0.020|       |
| SMALL | 0.108                    | 0.000| 0.027 | 0.023  | 0.019|       |
| MEDIUM| 0.095                    | 0.116| 0.000 | 0.028  | 0.023|       |
| BASE  | 0.093                    | 0.100| 0.144 | 0.000  | 0.026|       |
| LARGE | 0.097                    | 0.103| 0.117 | 0.149  | 0.000|       |

|       | QQP $s_1 \setminus s_2$ | MINI | SMALL | MEDIUM | BASE | LARGE |
|-------|--------------------------|------|-------|--------|------|-------|
| MINI  | 0.000                    | 0.025| 0.022 | 0.021  | 0.020|       |
| SMALL | 0.127                    | 0.000| 0.040 | 0.020  | 0.020|       |
| MEDIUM| 0.093                    | 0.146| 0.000 | 0.032  | 0.031|       |
| BASE  | 0.087                    | 0.119| 0.123 | 0.000  | 0.049|       |
| LARGE | 0.090                    | 0.105| 0.079 | 0.106  | 0.000|       |

|       | SST-2 $s_1 \setminus s_2$ | MINI | SMALL | MEDIUM | BASE | LARGE |
|-------|----------------------------|------|-------|--------|------|-------|
| MINI  | 0.000                      | 0.028| 0.022 | 0.021  | 0.019|       |
| SMALL | 0.117                      | 0.000| 0.047 | 0.031  | 0.029|       |
| MEDIUM| 0.075                      | 0.093| 0.000 | 0.063  | 0.035|       |
| BASE  | 0.068                      | 0.067| 0.085 | 0.000  | 0.071|       |
| LARGE | 0.071                      | 0.067| 0.060 | 0.098  | 0.000|       |

Table 10: The same table as 10, except that we are now calculating the relative upward bias $\mathbb{E}[(L_* - L)]/\mathbb{E}[L]$ as described in Section E.

| s1     | s2     | method | 2   | 4   | 6   | 8   | 10  |
|--------|--------|--------|-----|-----|-----|-----|-----|
| MINI   | SMALL  | ours   | 0.004| 0.011| 0.016| 0.020| 0.023|
| MINI   | SMALL  | BH     | 0.000| 0.000| 0.000| 0.000| 0.002|
| MINI   | MEDIUM | ours   | 0.012| 0.019| 0.026| 0.032| 0.035|
| MINI   | MEDIUM | BH     | 0.000| 0.000| 0.003| 0.008| 0.011|
| MINI   | BASE   | ours   | 0.019| 0.028| 0.035| 0.040| 0.042|
| MINI   | BASE   | BH     | 0.000| 0.000| 0.008| 0.014| 0.020|
| MINI   | LARGE  | ours   | 0.019| 0.027| 0.031| 0.037| 0.040|
| MINI   | LARGE  | BH     | 0.000| 0.000| 0.009| 0.015| 0.019|
| SMALL  | MEDIUM | ours   | 0.002| 0.006| 0.010| 0.015| 0.017|
| SMALL  | MEDIUM | BH     | 0.000| 0.000| 0.000| 0.000| 0.000|
| SMALL  | BASE   | ours   | 0.013| 0.020| 0.025| 0.030| 0.033|
| SMALL  | BASE   | BH     | 0.000| 0.000| 0.002| 0.006| 0.011|
| SMALL  | LARGE  | ours   | 0.015| 0.021| 0.026| 0.031| 0.033|
| SMALL  | LARGE  | BH     | 0.000| 0.000| 0.005| 0.009| 0.013|
| MEDIUM | BASE   | ours   | 0.006| 0.010| 0.013| 0.014| 0.016|
| MEDIUM | BASE   | BH     | 0.000| 0.000| 0.000| 0.000| 0.001|
| MEDIUM | LARGE  | ours   | 0.010| 0.014| 0.019| 0.022| 0.023|
| MEDIUM | LARGE  | BH     | 0.000| 0.000| 0.002| 0.004| 0.006|
| BASE   | LARGE  | ours   | 0.004| 0.005| 0.009| 0.010| 0.012|
| BASE   | LARGE  | BH     | 0.000| 0.000| 0.000| 0.000| 0.000|

Table 11: We compare each pair of model sizes $s_1$ and $s_2$ and report the lower bound provided by our method and the Benjamin-Hochberg (BH) procedure. The numbers in column name denote how many pretrained model we used to obtain the lower bounds.
We calculate the PearsonR correlation score between Table 12: Sorted in ascending order, the model sizes are MINI, SMALL, MEDIUM, BASE, and LARGE. The three model sizes listed for each row represents the model size of interest: for example, MINI, MEDIUM, LARGE means that we are calculating the correlation between MINI bucket. seeds, and instance index can be omitted in most superscript, since we now use subscript for random move the instance index from the subscript to the on datapoint $i$.

tuning seed $F$ represents the model of size $c$ indexes each instance (datapoint) $E$ tuning seed, $P$ Recall that $P$ is the pretraining seed, $F$ is the fine-tuning seed, $E$ represents a model checkpoint, $i$ indexes each instance (datapoint). $c_{P,F,E}^{s,i} = 1$ if the model of size $s$ with pretraining seed $p$ and fine-tuning seed $F$, and trained for $E$ epochs is correct on datapoint $i$, and 0 otherwise. Notice that we move the instance index from the subscript to the superscript, since we now use subscript for random seeds, and instance index can be omitted in most of our derivations.

The expected squared loss $\mathcal{L}$ of model $s$ on instance $i$ can then be written as

$$\mathcal{L}^{s,i} = \mathbb{E}_{P,F,E}[(1 - c_{P,F,E}^{s,i})^2]$$

(33)

Since we will analyze this term at a datapoint level, we drop the subscript $s$ and $i$ to keep the notation uncluttered. By the standard bias variance decomposition and total variance decomposition, we decompose the loss $\mathcal{L}$ into four terms:

$$\mathcal{L} = \text{Bias}^2 + \text{PretVar} + \text{FineVar} + \text{CkptVar}.$$  

(34)

We will walk through the meaning and definition of these four terms one by one. $\text{Bias}^2$ captures how bad is the average prediction, defined as

$$\text{Bias}^2 = (1 - \mathbb{E}_{P,F,E}[c_{P,F,E}])^2.$$  

(35)

$\text{PretVar}$ captures the variance brought by randomness in pretraining, and is defined as

$$\text{PretVar} = \text{Var}_{P} \mathbb{E}_{E}[c_{P,F,E}].$$  

(36)

Similarly, we define the variance brought by randomness in finetuning $\text{FineVar}$

$$\text{FineVar} = \mathbb{E}_{P} \mathbb{E}_{F} \mathbb{E}_{E}[c_{P,F,E}].$$  

(37)

and that by fluctuations across checkpoints $e$

$$\text{CkptVar} = \mathbb{E}_{P,F} \mathbb{E}_{E} \text{Var}_{E}[c_{P,F,E}].$$  

(38)

| MNL1. Buckets ⇒ | 0.10 | 0.20 | 0.30 | 0.40 | 0.50 | 0.60 | 0.70 | 0.80 | 0.90 | 1.00 |
|-----------------|------|------|------|------|------|------|------|------|------|------|
| MINI, SMALL, MEDIUM | 0.00 | 0.18 | 0.19 | 0.18 | 0.23 | 0.26 | 0.24 | 0.23 | 0.20 | 0.12 |
| SMALL, MEDIUM, BASE | 0.07 | 0.22 | 0.29 | 0.40 | 0.35 | 0.33 | 0.38 | 0.27 | 0.24 | 0.13 |
| MEDIUM, BASE, LARGE | 0.05 | 0.09 | 0.17 | 0.33 | 0.20 | 0.30 | 0.12 | 0.13 | 0.16 | 0.09 |
| MINI, MEDIUM, LARGE | 0.03 | 0.15 | 0.18 | 0.33 | 0.17 | 0.16 | 0.22 | 0.20 | 0.19 | 0.09 |

| QQP. Buckets ⇒ | 0.10 | 0.20 | 0.30 | 0.40 | 0.50 | 0.60 | 0.70 | 0.80 | 0.90 | 1.00 |
|-----------------|------|------|------|------|------|------|------|------|------|------|
| MINI, SMALL, MEDIUM | 0.03 | 0.21 | 0.18 | 0.21 | 0.21 | 0.25 | 0.18 | 0.16 | 0.10 | 0.06 |
| SMALL, MEDIUM, BASE | 0.01 | 0.17 | 0.23 | 0.19 | 0.24 | 0.22 | 0.24 | 0.19 | 0.16 | 0.05 |
| MEDIUM, BASE, LARGE | -0.02 | 0.16 | 0.09 | 0.23 | 0.17 | 0.10 | 0.14 | 0.14 | 0.09 | -0.01 |
| MINI, MEDIUM, LARGE | -0.01 | 0.07 | 0.14 | 0.09 | 0.16 | 0.09 | 0.16 | 0.07 | 0.10 | 0.07 |

| SST-2. Buckets ⇒ | 0.10 | 0.20 | 0.30 | 0.40 | 0.50 | 0.60 | 0.70 | 0.80 | 0.90 | 1.00 |
|-----------------|------|------|------|------|------|------|------|------|------|------|
| MINI, SMALL, MEDIUM | 0.09 | 0.26 | 0.43 | 0.22 | 0.28 | 0.24 | 0.27 | 0.35 | 0.20 | 0.12 |
| SMALL, MEDIUM, BASE | 0.07 | 0.12 | 0.22 | 0.40 | 0.07 | 0.20 | 0.10 | 0.12 | 0.19 | 0.06 |
| MEDIUM, BASE, LARGE | 0.01 | 0.24 | 0.29 | 0.35 | 0.19 | 0.19 | 0.26 | 0.39 | 0.15 | 0.03 |
| MINI, MEDIUM, LARGE | 0.01 | 0.17 | 0.11 | 0.41 | 0.04 | 0.29 | 0.16 | 0.21 | 0.15 | 0.07 |
H.2 Unbiased Estimation

We first describe the data on which we apply our estimator. Suppose we pretrain \( P \) models with different random seeds, for each of the \( P \) pre-trained models we finetune with \( F \) different random seeds, and we evaluate at \( E \) different checkpoints. Then \( \forall j \in [P], k \in [F], l \in [E] \), we observe \( P_j, F_{jkl}, E_{jkl} \), and \( c_{P_j, F_{jkl}, E_{jkl}} \), where each observed \( P, F \) and \( E \) are i.i.d. distributed. Our goal is to estimate from \( c \) the four quantities described in the previous section.

H.2.1 Estimating \( \text{CkptVar} \)

It is straightforward to estimate \( \text{CkptVar} \). The estimator \( \text{CkptVar} \) defined below is unbiased:

\[
\text{CkptVar} := \frac{1}{|F|} \sum_{j \in [P], k \in [F]} Var_E^{P_j, F_{jkl}}, \tag{39}
\]

where

\[
Var_E^{P_j, F_{jkl}} := \frac{1}{|E| - 1} \sum_{l \in [E]} (c_{P_j, F_{jkl}, E_{jkl}} - \bar{c}_{P_j, F_{jkl}})^2, \tag{40}
\]

and

\[
\bar{c}_{P_j, F_{jkl}} := \frac{1}{|E|} \sum_{l \in [E]} c_{P_j, F_{jkl}, E_{jkl}}. \tag{41}
\]

\( \text{CkptVar} \) is unbiased, since \( Var_E^{P_j, F_{jkl}} \) is an unbiased estimation of variance of \( c \) with fixed \( P_j \) and \( F_{jkl} \), and randomness \( E \), i.e.

\[
\mathbb{E}_{E_{ijkl}}[Var_E^{P_j, F_{jkl}}] = Var_E[c_{P,F,E}|P = P_j, F = F_{jkl}]. \tag{42}
\]

Therefore, \( \forall j \in [P], k \in [F] \), we have

\[
\mathbb{E}_{P_j, F_{jkl}}[Var_E^{P_j, F_{jkl}}] = \text{CkptVar}, \tag{43}
\]

and hence by linearity of expectation

\[
\mathbb{E}_{P_j, F_{jkl}, E_{jkl}}[\text{CkptVar}] = \text{CkptVar}. \tag{44}
\]

H.2.2 Estimating \( \text{FineVar} \)

As before, by linearity of expectation, we can declare victory if we can develop an unbiased estimator for the following quantity and then average across \( P_j \):

\[
Var_F[\mathbb{E}[c_{P,F,E}|P = P_j]], \tag{45}
\]

which verbally means "variance across different finetuning seeds of the mean of \( c \) over different checkpoints \( E \), conditioned on the pretraining seed \( P_j \)."

Since \( P_j \) is fixed for this estimator, we drop the subscripts \( P \) to keep notation uncluttered. Therefore, we want to estimate

\[
Var_F := Var_F[\mathbb{E}_E[c_{F,E}]] \tag{46}
\]

A naive solution is to take first take the mean \( \bar{c}_{F_k} \) of each \( F_k \), i.e.

\[
\bar{c}_{F_k} := \frac{1}{|E|} \sum_{l \in [E]} c_{F_k, E_{jkl}}, \tag{47}
\]

and then calculate the sample variance \( Var_F \) of \( \bar{c} \) with respect to \( F \):

\[
Var_F := \frac{1}{|F| - 1} \sum_{k \in [F]} (\bar{c}_{F_k} - \bar{c})^2, \tag{48}
\]

where

\[
\bar{c} := \frac{1}{|F|} \sum_{k \in [F]} \bar{c}_{F_k}. \tag{49}
\]

However, this would create an upward bias: the empirical mean \( \bar{c}_{F_k} \) is a noisy estimate of the population mean \( E[\mathbb{E}_E[c_{F_k,E}]] \), and hence increases let \( Var_F \) over-estimate the variance. Imagine a scenario where \( Var_F \) is in fact 0; however, since \( \bar{c}_{F_k} \) is a noisy estimate, \( Var_F \) will sometimes be positive but never below 0. As a result, \( \mathbb{E}[Var_F] > 0 \), which is a biased estimator.

We introduce the following general theorem to correct this bias.

**Theorem 2** Suppose \( D_k, k \in [F] \) are independently sampled from the same distribution \( \Xi \), which is a distribution of distributions; \( \mu_k \) is an unbiased estimator of \( \mathbb{E}_{X \sim D_k}[X] \), and \( \phi_k \) to be an unbiased estimator of the variance of \( \mu_k \), then

\[
Var_F = \frac{1}{|F| - 1} \sum_{k \in [F]} (\hat{\mu}_k - \hat{\mu})^2 \tag{50}
\]

\[
\frac{1}{|F|} \sum_{k \in [F]} \hat{\mu}_k
\]

is an unbiased estimator for

\[
V = Var_{D \sim \Xi} \mathbb{E}[X \sim D][X], \tag{51}
\]

where

\[
\hat{\mu} := \frac{1}{|F|} \sum_{k \in [F]} \hat{\mu}_k. \tag{52}
\]
In this estimator, the first term “pretends” that \( \hat{\mu} \) are perfect estimator for the population mean and calculate the variance, while the second term corrects for the fact that the empirical mean estimation is not perfect. Notice the theorem only requires that \( \hat{\mu} \) and \( \hat{\phi} \) are unbiased, and is agnostic to the actual computation procedure by these estimators.

**Proof**  We define the population mean of \( D_k \) to be \( \mu_k \), i.e.

\[
\mu_k := \mathbb{E}_{X \sim D_k}[X],
\]

and the population mean of \( \mu_k \) across randomness in \( D \) to be \( \mu \), i.e.

\[
\mu := \mathbb{E}_{D \sim \Xi}[\mathbb{E}_{X \sim D}[X]].
\]

We look at the first term of the estimator in equation 50:

\[
\frac{1}{\mathcal{F}} - 1 \mathbb{E}\left[ \sum_{k \in [\mathcal{F}]} (\hat{\mu}_k - \bar{\mu})^2 \right] = \frac{1}{\mathcal{F}} - 1 \mathbb{E}\left[ \sum_{k \in [\mathcal{F}]} ((\bar{\mu}_k - \mu_k) - (\bar{\mu} - \mu) + (\mu_k - \mu))^2 \right]
\]

\[
= \frac{1}{\mathcal{F}} - 1 \mathbb{E}\left[ \sum_{k \in [\mathcal{F}]} ((\bar{\mu}_k - \mu_k)^2 + (\bar{\mu} - \mu))^2 \right]
\]

\[
+ (\mu_k - \mu)^2 - 2(\bar{\mu}_k - \mu_k)(\bar{\mu} - \mu)
\]

\[
- 2(\bar{\mu}_k - \mu_k)(\bar{\mu} - \mu)]
\]

There are 5 summands within \( \sum_{k \in [\mathcal{F}],} \), and we look at them one by one:

\[
\mathbb{E}\left[ \sum_{k \in [\mathcal{F}]} (\bar{\mu}_k - \mu_k)^2 \right] = \mathbb{E}\left[ \sum_{k \in [\mathcal{F}]} \hat{\phi}_k \right],
\]

\[
\mathbb{E}[(\bar{\mu} - \mu)^2] = \mathbb{E}\left[ \mu - \frac{1}{\mathcal{F}} \sum_{k \in [\mathcal{F}]} \mu_k \right]
\]

\[
= \frac{1}{\mathcal{F}} \mathcal{V} + \frac{1}{\mathcal{F}^2} \sum_{k \in [\mathcal{F}]} \mathbb{E}\left[ \hat{\phi}_k \right]
\]

\[
\mathbb{E}\left[ \sum_{k \in [\mathcal{F}]} (\mu_k - \mu)^2 \right] = \mathcal{F} \mathcal{V}
\]

\[
\mathbb{E}\left[ -2 \sum_{k \in [\mathcal{F}]} (\bar{\mu}_k - \mu_k)(\bar{\mu} - \mu) \right] = -2 \mathcal{V}.
\]

Putting these five terms together, we continue calculating Equation 55:

\[
\frac{1}{\mathcal{F}} - 1 \mathbb{E}\left[ \sum_{k \in [\mathcal{F}]} (\hat{\mu}_k - \bar{\mu})^2 \right] = \frac{1}{\mathcal{F}} - 1 \mathbb{E}\left[ \sum_{k \in [\mathcal{F}]} \hat{\phi}_k \right]
\]

\[
+ \mathcal{F} \mathcal{V} + \frac{1}{\mathcal{F}^2} \sum_{k \in [\mathcal{F}]} \mathbb{E}[\hat{\phi}_k]
\]

\[
- 2 \mathcal{V}
\]

\[
= \mathcal{V} + \mathbb{E}\left[ \frac{1}{\mathcal{F}} \sum_{k} \hat{\phi}_k \right]
\]

Then from Equation 50, we can tell that \( \mathcal{V} \mathcal{F} \) is unbiased. \( \square \)

Now we come back to the topic of developing an unbiased estimator for \( \mathcal{V} \mathcal{F} \) as defined in Equation 46. To utilize Theorem 2, we need two components:

- An unbiased estimator \( \hat{\mu}_{\mathcal{F}_k} \) for \( \mathbb{E} \mathcal{F} \mathcal{E}[c_{\mathcal{F}, \mathcal{F}}|F = F_k] \)
- An unbiased estimator \( \hat{\phi}_{\mathcal{F}_k} \) for the variance of \( \hat{\mu}_{\mathcal{F}_k} \), i.e. \( \mathcal{V} \mathcal{F} \mathcal{E} \mathcal{E}[\hat{\phi}_{\mathcal{F}_k}] \)

\( \hat{\phi}_{\mathcal{F}_k} \) is an unbiased estimator for \( \mathbb{E} \mathcal{F} \mathcal{E}[c_{\mathcal{F}, \mathcal{F}}|F = F_k] \), and its variance \( \mathcal{V} \mathcal{F} \mathcal{E} \mathcal{E}[\hat{\phi}_{\mathcal{F}_k}|F = F_k] \) is

\[
\mathcal{V} \mathcal{F} \mathcal{E} \mathcal{E}[\hat{\phi}_{\mathcal{F}_k}] = \frac{1}{\mathcal{E}} \mathcal{V} \mathcal{F} \mathcal{E} \mathcal{E}[c_{\mathcal{F}, \mathcal{F}}|F = F_k].
\]

Therefore, to develop an unbiased estimator for \( \mathcal{V} \mathcal{F} \mathcal{E} \mathcal{E}[\hat{\phi}_{\mathcal{F}_k}] \), it suffices to have an unbiased estimate of \( \mathcal{V} \mathcal{F} \mathcal{E} \mathcal{E}[c_{\mathcal{F}, \mathcal{F}}|F = F_k] \). We define

\[
\hat{\phi}_{\mathcal{F}_k} := \frac{1}{\mathcal{E}(\mathcal{F} - 1)} \sum_{k \in [\mathcal{F}]} (c_{\mathcal{F}_k, \mathcal{E} \mathcal{E} \mathcal{F}_k} - \bar{\phi}_{\mathcal{F}_k})^2,
\]

\[
= \frac{1}{\mathcal{E}(\mathcal{F} - 1)} \sum_{k \in [\mathcal{F}]} (c_{\mathcal{F}_k, \mathcal{E} \mathcal{E} \mathcal{F}_k} - \bar{\phi}_{\mathcal{F}_k})^2
\]

\[
= \frac{1}{\mathcal{E}(\mathcal{F} - 1)} \sum_{k \in [\mathcal{F}]} (c_{\mathcal{F}_k, \mathcal{E} \mathcal{E} \mathcal{F}_k} - \bar{\phi}_{\mathcal{F}_k})^2
\]

\[
= \frac{1}{\mathcal{E}(\mathcal{F} - 1)} \sum_{k \in [\mathcal{F}]} (c_{\mathcal{F}_k, \mathcal{E} \mathcal{E} \mathcal{F}_k} - \bar{\phi}_{\mathcal{F}_k})^2
\]

\[
= \frac{1}{\mathcal{E}(\mathcal{F} - 1)} \sum_{k \in [\mathcal{F}]} (c_{\mathcal{F}_k, \mathcal{E} \mathcal{E} \mathcal{F}_k} - \bar{\phi}_{\mathcal{F}_k})^2
\]
and we can plug in \( \hat{\phi}_F \) and \( \hat{\mu}_F = \bar{c}_F \) into Theorem 2 as an unbiased estimator to obtain an unbiased estimator for \( \Var_F[\mathbb{E}_F[c_{P,F,E}] | P = P_j] \), and we average the estimation for each \( P_j \) to obtain an unbiased estimate.

**H.2.3 Estimating \( \text{PretVar} \)**

We next estimate \( \Var_F[\mathbb{E}_F[c_{P,F,E}] \] We can still apply the idea from Theorem 2, which requires

- An unbiased estimator \( \hat{\mu}_{P_j} \) for \( \mathbb{E}_{F,E}[c_{P,F,E} | P = P_j] \)
- An unbiased estimator \( \hat{\phi}_{P_j} \) for the variance of \( \hat{\mu}_{P_j} \), i.e. \( \Var_{F,E}[\hat{\mu}_{P_j}] \).

Again, the first is easy to obtain: \( \hat{\mu}_{P_j} = \bar{c}_{P_j} \) is an unbiased estimator for \( \mathbb{E}_{F,E}[c_{P,F,E} | P = P_j] \), where

\[
\bar{c}_{P_j} := \frac{1}{F} \sum_{k \in [F], l \in [E]} c_{P_j,F_{jk},E_{jkl}} \quad (64)
\]

However, we cannot straightforwardly estimate \( \Var_{F,E}[\hat{\mu}_{P_j}] \) as before, since samples \( c_{P_j,F_{jk},E_{jkl}} \) are no longer independent. We need to use Equation 57 to develop an unbiased estimator (the LHS is exactly what we want!), i.e.

\[
\Var_{F,E}(\bar{c}_{P_j}) = \frac{1}{F} \Var_F(\mathbb{E}_F[c_{P,F,E}] | P = P_j) \]

\[
+ \frac{1}{F^2} \sum_{k \in [F]} \Var_{F,E}(\bar{c}_{P_j,F_{jk}}), \quad (65)
\]

and we already know how to estimate these two summands from the previous discussion on estimating \( \text{FineVar} \).

**H.2.4 Estimating \( \text{Bias}^2 \)**

It is easy to see that the following \( \hat{L} \) is an unbiased estimator for the loss \( L \).

\[
\hat{L} := \frac{1}{P \cdot F \cdot E} \sum_{j \in [P], k \in [F], l \in [E]} (1 - c_{P_j,F_{jk},E_{jkl}})^2, \quad (66)
\]

and

\[
\mathbb{E}[\hat{L}] = L. \quad (67)
\]

By linearity of expectation and loss decomposition in Equation 34,

\[
\text{Bias}^2 := \hat{L} - \text{PretVar} \quad (68)
\]

\[
- \text{FineVar} - \text{CkptVar}
\]
is an unbiased estimator of \( \text{Bias}^2 \).

Notice that the naïve estimator that calculates the expected bias and then squares it estimates \( \mathbb{E}[\text{Bias}]^2 \) instead of \( \mathbb{E}[\text{Bias}^2] \).

**H.3 Generalization**

We can generalize this estimation strategy to decompose variance into arbitrary number of randomness. In general, we want to estimate some quantity of the following form

\[
\mathbb{E}_{r_1 \ldots r_{n-1}}[\Var_{r_n}[\mathbb{E}_{r_{n+1} \ldots r_N}[c_{r_1 \ldots r_N}]], \quad (69)
\]

from the data that has an hierarchical tree structure of randomness.

For the goal of developing an unbiased estimator, we can get rid of the outer expectation \( r_1 \ldots r_{n-1} \) easily by linearity of expectation: simply estimate the Variance conditioned on \( r_1 \ldots r_{n-1} \) and average them together, as discussed in Section H.2.1.

To estimate

\[
\Var_{r_n}[\mathbb{E}_{r_{n+1} \ldots r_N}[c_{r_1 \ldots r_N}]], \quad (70)
\]

we make use of Theorem 2, which requires

- an unbiased estimator \( \hat{\mu}_{r_{n+1}} \) for the quantity \( \mathbb{E}_{r_{n+1} \ldots r_N}[c_{r_1 \ldots r_N}] \), which we can straightforwardly obtain by average the examples that has the same random variables \( r_1 \ldots n \) (e.g. \( \bar{c}_{P_j} \))
- an unbiased estimator for the variance of \( \hat{\mu}_{r_{n+1}} \). If \( N = n + 1 \), we can directly compute the sample variance of the \( c \) as our estimate (e.g. in Equation 63). Otherwise, we use Equation 57 to decompose the desired quantities into two, and estimate them recursively by applying Theorem 2 and Equation 57.

For readability we wrote the proof with the assumption that, in the tree of randomness, the number of branches for each node at the same depth is the same. However, our proof does not make use of this assumption and can be applied to a general tree structure of randomness as long as the the number of children is larger or equal to 2 for each non-terminal node.

**I Variance Conditioned on Bias**

Since lower bias usually implies lower variance, to tease out the latent effect, we estimate the variance “given a fixed level of bias \( \text{Bias}^2 \) of \( b^2 \in [0, 1] \)”, i.e.
Figure 9: The variance curve conditioned on $\text{Bias}^2$ for in-domain development set of MNLI, QQP and SST-2. Each curve represents a model size. Left for pretraining variance and right for finetuning variance.

$$\text{PretVar}(b^2) := E_i[\text{PretVar}_i \mid \text{Bias}^2_i = b^2]$$  \hspace{1cm} (71)

We estimate $\text{PretVar}_s(b^2)$ and $\text{FineVar}_s(b^2)$ using gaussian process and plot them against $b^2$ in Figure 9 for MNLI, QQP, and SST-2. We find that larger models usually have larger finetuning variance across all levels of biases (except for MEDIUM and MINI on SST-2), and BASE model always has larger pretraining variance than LARGE.

We also experimented with the squared loss:

$$L_i = (1 - p_i)^2,$$  \hspace{1cm} (72)

where $p_i$ is the probability assigned to the correct label for instance $i$. We plot the same curve in Figure 10 and observe the same trend.

**J Example Decaying Instances**

We manually examined the group of instances where $\text{MINI}_{\text{LARGE}} \Delta \text{Acc}_i \leq -0.9$ in Table 3. In other words, MINI is almost always correct on these instances, while LARGE is almost always wrong. For each instance in this group, we manually categorize it into one of the four categories: 1) Correct, if the label is correct, 2) Fine, if the label might be controversial but we could see a reason why this label is reasonable, 3) Wrong, if the label is wrong, and 4) Unsure, if we are unsure how to label this instance. As a control, we also examined the remaining fraction of the dataset. Each time we annotate an instance, with 50% probability it is sampled from the decaying fraction or the remaining fraction, and we do not know which group it comes from.

We show below all the annotated instances from this decaying fraction and their categories for MNLI (Section J.1), QQP, and SST-2 (Section J.3).

**J.1 MNLI**

MNLI is the abbreviation of Multi-Genre Natural Language Inference (Williams et al. (2020)). In this task, given a premise and a hypothesis, the model needs to classify whether the premise entails/contradicts the hypothesis, or otherwise. The instances can be seen below.

**Premise**: and that you’re very much right but the jury may or may not see it that way so you get a little anticipate you know anxious there and go well you know

**Hypothesis**: Jury’s operate without the benefit of
an education in law.

Premise: In fiscal year 2000, it reported estimated improper Medicare Fee-for-Service payments of $11.
Hypothesis: The payments were improper.
Label: Entailment
Category: Fine

Premise: is that what you ended up going into
Hypothesis: So that must be what you chose to do?
Label: Entailment
Category: Correct

Premise: INTEREST RATE - The price charged per unit of money borrowed per year, or other unit of time, usually expressed as a percentage.
Hypothesis: Interest rate is defined as the total amount of money borrowed.
Label: Entailment
Category: Wrong

Premise: The analyses comply with the informational requirements of the sections including the classes of small entities subject to the rule and alternatives considered to reduce the burden on the small entities.
Hypothesis: The rules place a high burden on the activities of small entities.
Label: Contradiction
Category: Correct

Premise: Isn’t a woman’s body her most personal property?
Hypothesis: Women’s bodies belong to themselves, they should decide what to do with it.
Label: Neutral
Category: Unsure

Premise: The Standard, published a few days before Deng’s death, covers similar territory.
Hypothesis: The Washington Post covers similar territory.
Label: Neutral
Category: Correct

Premise: Shoot only the ones that face us, Jon had told Adrin.
Hypothesis: Jon told Adrin and the others to only shoot the ones that face us.
Label: Entailment
Category: Wrong

Premise: But if you take it seriously, the anti-abortion position is definitive by definition.
Hypothesis: If you decide to be serious about supporting anti-abortion, it’s a very run of the mill belief to hold.
Label: Neutral
Category: Unsure

Premise: yeah well that’s the other thing you know they talk about women leaving the home and going out to work well still taking care of the children is a very important job and and someone’s got to do it and be able to do it right and
Hypothesis: It is not acceptable for anybody to refuse work in order to take care of children.
Label: Contradiction
Category: Correct

Premise: The researchers found expected stresses like the loss of a check in the mail and the illness of loved ones.
Hypothesis: The stresses affected people much differently than the researchers expected.
Label: Contradiction
Category: Correct

Premise: so you know it’s something we we have tried to help but yeah
Hypothesis: We did what we could to help.
Label: Entailment
Category: Correct

Premise: Czarek was welcomed enthusiastically, even though the poultry brotherhood was paying a lot of sudden attention to the newcomers - a strong group of young and talented managers from an egzemo-exotic chicken farm in Fodder Band nearby Podunkowice.
Hypothesis: Czarek was welcomed into the group by the farmers.
Label: Entailment
Category: Correct

Premise: ’I don’t suppose you could forget
I ever said that?'
**Hypothesis**: I hope that you can remember that forever.
**Label**: Contradiction
**Category**: Wrong

**Premise**: Oh, my friend, have I not said to you all along that I have no proofs.
**Hypothesis**: I told you from the start that I had no evidence.
**Label**: Entailment
**Category**: Correct

**Premise**: I should put it this way.
**Hypothesis**: I should phrase it differently.
**Label**: Entailment
**Category**: Correct

**Premise**: An organization’s activities, core processes, and resources must be aligned to support its mission and help it achieve its goals.
**Hypothesis**: An organization is successful if its activities, resources, and goals align.
**Label**: Entailment
**Category**: Correct

**Premise**: A more unusual dish is azure, a kind of sweet porridge made with cereals, nuts, and fruit sprinkled with rosewater.
**Hypothesis**: Azure is a common and delicious food made with cereals, nuts and fruit.
**Label**: Entailment
**Category**: Wrong

**Premise**: once you have something and it’s like i was watching this program on TV yesterday in nineteen seventy six NASA came up with Three D graphics right
**Hypothesis**: I was watching a program about gardening.
**Label**: Contradiction
**Category**: Correct

**Premise**: , First-Class Mail used by households to pay their bills) and the household bill mail (i.e.
**Hypothesis**: Second-Class Mail used by households to pay their bills
**Label**: Contradiction
**Category**: Unsure

**Premise**: Rightly or wrongly, America is seen as globalization's prime mover and head cheerleader and will be blamed for its excesses until we start paying official attention to them.
**Hypothesis**: America’s role in the globalization movement is important whether we agree with it or not.
**Label**: Entailment
**Category**: Correct

**Premise**: After being diagnosed with cancer, Carrey’s Kaufman decides to do a show at Carnegie Hall.
**Hypothesis**: Carrey’s Kaufman is only diagnosed with cancer after doing a show at Carnegie Hall.
**Label**: Contradiction
**Category**: Correct

**Premise**: Several pro-life Dems are mounting serious campaigns at the state level, often against pro-choice Republicans.
**Hypothesis**: Serious campaigns are being run by a few pro-life Democrats.
**Label**: Entailment
**Category**: Correct

**Premise**: On the northwestern Alpine frontier, a new state had appeared on the scene, destined to lead the movement to a united Italy.
**Hypothesis**: The unite Italy movement was waiting for a leader.
**Label**: Neutral
**Category**: Fine

**Premise**: well we bought this with credit too well we found it with a clearance uh down in Memphis i guess and uh
**Hypothesis**: We bought non-sale items in Memphis on credit.
**Label**: Contradiction
**Category**: Correct

**Premise**: He slowed.
**Hypothesis**: He stopped moving so quickly.
**Label**: Entailment
**Category**: Correct

**Premise**: As legal scholar Randall Kennedy wrote in his book Race, Crime, and the Law, Even if race is only one of several factors behind a decision, tolerating it at all means tolerating it as
potentially the decisive factor.

**Hypothesis**: Race is one of several factors in some judicial decisions  
**Label**: Entailment  
**Category**: Correct

**Premise**: Although all four categories of emissions are down substantially, they only achieve 50-75% of the proposed cap by 2007 (shown as the dotted horizontal line in each of the above figures).  
**Hypothesis**: All of the emission categories experienced a downturn except for one.  
**Label**: Contradiction  
**Category**: Correct

**Premise**: He sat up, trying to free himself.  
**Hypothesis**: He was trying to take a nap.  
**Label**: Contradiction  
**Category**: Correct

**Premise**: Impossible.  
**Hypothesis**: Cannot be done.  
**Label**: Entailment  
**Category**: Correct

**Premise**: But, as the last problem I’ll outline suggests, neither of the previous two objections matters.  
**Hypothesis**: I will not continue to outline any more problems.  
**Label**: Entailment  
**Category**: Correct

**Premise**: As the Tokugawa shoguns had feared, this opening of the floodgates of Western culture after such prolonged isolation had a traumatic effect on Japanese society.  
**Hypothesis**: The Tokugawa shoguns had feared that, because they understood the Japanese society very well.  
**Label**: Neutral  
**Category**: Fine

**Premise**: In the ancestral environment a man would be likely to have more offspring if he got his pick of the most fertile-seeming women.  
**Hypothesis**: Only a man who stayed with one female spread his genes most efficiently.  
**Label**: Contradiction  
**Category**: Fine

**Premise**: Tommy was suddenly galvanized into life.  
**Hypothesis**: Tommy had been downcast for days.  
**Label**: Neutral  
**Category**: Correct

**Premise**: Improved products and services initiate actions and manage risks to develop new products and services within or outside the organization.  
**Hypothesis**: Managed risks lead to new products  
**Label**: Entailment  
**Category**: Fine

**Premise**: Coast Guard rules establishing bridge opening schedules.  
**Hypothesis**: The Coast Guard is in charge of opening bridges.  
**Label**: Entailment  
**Category**: Correct

**Premise**: The anthropologist Napoleon Chagnon has shown that Yanomamo men who have killed other men have more wives and more offspring than average guys.  
**Hypothesis**: Yanomamo men who kill other men have better chances at getting more wives.  
**Label**: Entailment  
**Category**: Fine

**Premise**: The Varanasi Hindu University has an Art Museum with a superb collection of 16th-century Mughal miniatures, considered superior to the national collection in Delhi.  
**Hypothesis**: The Varanasi Hindu University has an art museum on its campus which may be superior objectively to the national collection in Delhi.  
**Label**: Entailment  
**Category**: Correct

**Premise**: Part of the reason for the difference in pieces per possible delivery may be due to the fact that five percent of possible residential deliveries are businesses, and it is thought, but not known, that a lesser percentage of possible deliveries on rural routes are businesses.  
**Hypothesis**: We all know that the reason for a lesser percentage of possible deliveries on rural routes being businesses, is because of the fact that
people prefer living in cities rather than rural areas.

**Label**: Neutral  
**Category**: Correct

**Premise**: right oh they’ve really done uh good job of keeping everybody informed of what’s going on sometimes i’ve wondered if it wasn’t almost more than we needed to know  
**Hypothesis**: I don’t think I have shared enough information with everyone.  
**Label**: Contradiction  
**Category**: Correct

**Premise**: To reach any of the three Carbet falls, you must continue walking after the roads come to an end for 20 minutes, 30 minutes, or two hours respectively.  
**Hypothesis**: There are three routes to the three Carbet falls, each a different length and all continue after the road seemingly ends.  
**Label**: Entailment  
**Category**: Correct

**Premise**: But when the cushion is spent in a year or two, or when the next recession arrives, the disintermediating voters will find themselves playing the roles of budget analysts and tax wonks.  
**Hypothesis**: The cushion will likely be spent in under two years.  
**Label**: Entailment  
**Category**: Correct

**Premise**: But, Slate protests, it was [Gates’] byline that appeared on the cover.  
**Hypothesis**: Slate was one hundred percent positive it was Gates’ byline on the cover.  
**Label**: Neutral  
**Category**: Correct

**Premise**: But it’s for us to get busy and do something.”  
**Hypothesis**: ”We don’t do much, so maybe this would be good for us to bond and be together for the first time in a while.”.  
**Label**: Neutral  
**Category**: Fine

**Premise**: Pearl Jam detractors still can’t stand singer Eddie They say he’s unbearably self-important and limits the group’s appeal by refusing to sell out and make videos.  
**Hypothesis**: A lot of people consider Eddie to be a bad singer.  
**Label**: Neutral  
**Category**: Correct

**Premise**: it’s the very same type of paint and everything  
**Hypothesis**: It’s the same paint formula, it’s great!  
**Label**: Entailment  
**Category**: Fine

**Premise**: Exhibit 3 presents total national emissions of NOx and SO2 from all sectors, including power.  
**Hypothesis**: In Exhibit 3 there are the total regional emissions od NOx and SO2 from all sectors.  
**Label**: Entailment  
**Category**: Correct

**Premise**: uh-huh and is it true i mean is it um  
**Hypothesis**: It’s true.  
**Label**: Entailment  
**Category**: Wrong

**Premise**: When a GAGAS attestation engagement is the basis for an auditor’s subsequent report under the AICPA standards, it would be advantageous to users of the subsequent report for the auditor’s report to include the information on compliance with laws and regulations and internal control that is required by GAGAS but not required by AICPA standards.  
**Hypothesis**: The report is required by GAGAS but not AICPA.  
**Label**: Entailment  
**Category**: Correct

**Premise**: i’m on i’m in the Plano school system and living in Richardson and there is a real dichotomy in terms of educational and economic background of the kids that are going to be attending this school  
**Hypothesis**: The Plano school system only has children with poor intelligence.  
**Label**: Contradiction  
**Category**: Correct
J.2  QQP

QQP is the abbreviation of Quora Question Pairs. Given two questions, the model needs to tell whether they have the same meaning (i.e. Paraphrase/Non-paraphrase).

Question 1: Which universities for MS in CS should I apply to?
Question 2: Which universities should I apply to for an MS in CS?
Label: Paraphrase
Category: Correct

Question 1: What should I do to make life worth living?
Question 2: What makes life worth living?
Label: Paraphrase
Category: Fine

Question 1: Why did Quora remove my question?
Question 2: Why does Quora remove questions?
Label: Paraphrase
Category: Correct

Question 1: How do I get thousands of followers on Instagram?
Question 2: How can I get free 10k real Instagram followers fast?
Label: Paraphrase
Category: Correct

Question 1: What is the basic knowledge of computer science engineers?
Question 2: What is basic syllabus of computer science engineering?
Label: Non-paraphrase
Category: Fine

Question 1: How many mosquito bites does it take to kill a human being?
Question 2: How many times can a single mosquito bite a human within 8 hours?
Label: Non-paraphrase
Category: Correct

Question 1: How does it feel to become attractive from unattractive?
Question 2: What does it feel like to go from physically unattractive to physically attractive?

---

https://www.quora.com/q/quoradata/First-Quora-Dataset-Release-Question-Pairs

Label: Paraphrase
Category: Correct

Question 1: Who is answering the questions asked on Quora?
Question 2: Who can answer the questions asked on Quora?
Label: Paraphrase
Category: Correct

Question 1: What machine learning theory do I need to know in order to be a successful machine learning practitioner?
Question 2: What do I need to know to learn machine learning?
Label: Paraphrase
Category: Wrong

Question 1: If you could go back in time and change one thing, what would it be and why?
Question 2: If you could go back in time and do one thing, what would it be?
Label: Paraphrase
Category: Correct

Question 1: Will there be a civil war if Trump doesn’t become president?
Question 2: Will there be a second civil war if Trump becomes president?
Label: Paraphrase
Category: Correct

Question 1: Do Quora contributors get paid?
Question 2: How do contributors get paid by Quora?
Label: Paraphrase
Category: Correct

Question 1: Did India meet Abdul Kalam’s 2020 vision so far?
Question 2: How far do you think India has reached on President APJ Kalam’s vision in the book India 2020?
Label: Non-paraphrase
Category: Correct

Question 1: How do I stop my dog from whining after getting spayed?
Question 2: How do I stop my dog from whining?
Label: Paraphrase
Category : Wrong

Question 1 : What difference are exactly between Euclidean space and non-Euclidean space?
Question 2 : What is the difference between Euclidean and non-Euclidean?
Label : Non-paraphrase
Category : Wrong

Question 1 : Why doesn’t Hillary Clinton win the White House if she won the popular vote?
Question 2 : How did Hillary Clinton win the popular vote but Donald Trump win the election?
Label : Paraphrase
Category : Correct

Question 1 : How is public breastfeeding seen where you live?
Question 2 : How is breastfeeding in public seen in your country?
Label : Paraphrase
Category : Correct

Question 1 : What are some ways to change your Netflix password?
Question 2 : How do you change your Netflix password and email?
Label : Paraphrase
Category : Fine

Question 1 : What do you think, is your best answer on Quora?
Question 2 : What is your best answer on Quora?
Label : Non-paraphrase
Category : Correct

Question 1 : How can I travel to Mexico without a passport?
Question 2 : Can I travel to Mexico without a passport?
Label : Paraphrase
Category : Correct

Question 1 : How do modern Congolese people view Mobutu in retrospect?
Question 2 : How do Congolese currently view Mobutu Sese Seko?
Label : Non-paraphrase
Category : Correct

Question 1 : How is Tanmay Bhat losing weight?
Question 2 : Tanmay Bhat: How did you manage to reduce your fat?
Label : Non-paraphrase
Category : Unsure

Question 1 : Is Xiaomi a brand to trust (comparing it with brands like Samsung and HTC)? What is better: Xiaomi MI3 or HTC Desire 816?
Question 2 : Is xiaomi a trusted brand?
Label : Non-paraphrase
Category : Correct

Question 1 : Why did Buddhism spread in East Asia and not in its native land India?
Question 2 : How was Buddhism spread in Asia?
Label : Non-paraphrase
Category : Correct

Question 1 : Can I become a multi billion-aire betting on horses?
Question 2 : How much money can I make betting on horses? A month? Can I make 20,000 a month?
Label : Paraphrase
Category : Fine

Question 1 : What is a diet for gaining weight?
Question 2 : What is a way to gain weight?
Label : Non-paraphrase
Category : Correct

Question 1 : How do I use Instagram on my computer?
Question 2 : How can I get Instagram on my computer?
Label : Paraphrase
Category : Fine

Question 1 : What is the legal basis of a "you break it, you buy it" policy?
Question 2 : Is a "you break it you buy it" policy actually legal?
Label : Paraphrase
Category : Correct

Question 1 : How I should fix my computer while it is showing no boot device found?
Question 2 : How do I fix the "Boot device not
found” problem?
Label : Paraphrase
Category : Correct

Question 1 : What innovative name can I use for an interior designing firm?
Question 2 : What can i name my interior designing firm?
Label : Paraphrase
Category : Correct

Question 1 : What would it realistically cost to go to Tomorrowland?
Question 2 : How much is a ticket to Tomorrowland?
Label : Non-paraphrase
Category : Fine

Question 1 : Is there a gender pay gap? If so why?
Question 2 : Is the gender pay gap a myth?
Label : Paraphrase
Category : Correct

Question 1 : How can I get rid of a canker sore on the bottom of my tongue?
Question 2 : How can I get rid or a canker sore on the tip of my tongue?
Label : Paraphrase
Category : Correct

Question 1 : How can I sleep better and early in night?
Question 2 : How can I sleep better at night?
Label : Paraphrase
Category : Correct

Question 1 : Why did DC Change Captain Marvel’s name?
Question 2 : Why did DC have to change Captain Marvel’s name but Marvel didn’t have to change Scarecrow’s name?
Label : Paraphrase
Category : Fine

Question 1 : Should be there any difference between IIT and non IIT students in terms of placement package from a company if both of them are equally talented?
Question 2 : Should be there any difference between IIT and non IIT students in terms of placement package from a company if both of them are equally capable?
Label : Paraphrase
Category : Correct

Question 1 : What happened to The Joker after The end of The Dark Knight?
Question 2 : What happens to the Joker at the end of The Dark Knight (2008 movie)?
Label : Non-paraphrase
Category : Wrong

Question 1 : I love my wife more then anything. Why do I fantasize about her with other men?
Question 2 : Why do I fantasize about other men having sex with my wife?
Label : Paraphrase
Category : Fine

Question 1 : What is your opinion on the new MacBook Pro Touch Bar?
Question 2 : What do you think about the OLED touch bar on the new MacBook Pro?
Label : Paraphrase
Category : Correct

Question 1 : How do I get rid of my negative alter ego?
Question 2 : How do you get rid of your negative alter ego?
Label : Paraphrase
Category : Correct

Question 1 : How can I get wifi driver for my hp laptop with windows 7 os?
Question 2 : How can I get wifi driver for my laptop with windows 7 os?
Label : Paraphrase
Category : Correct

Question 1 : What’s your attitude towards life?
Question 2 : What should be your attitude towards life?
Label : Paraphrase
Category : Wrong

Question 1 : What books would I like if I loved A Song of Ice and Fire?
Question 2 : Are there books which are similar to
A Song of Ice and Fire?
**Label**: Paraphrase
**Category**: Fine

**Question 1**: Why do Muslims think they will conquer the whole world?
**Question 2**: Do you think Muslims will take over the world?
**Label**: Non-paraphrase
**Category**: Correct

**Question 1**: Is dark matter a sea of massive dark photons that ripple when galaxy clusters collide and wave in a double slit experiment?
**Question 2**: Does a superfluid dark matter which ripples when Galaxy clusters collide and waves in a double slit experiment relate GR and QM?
**Label**: Paraphrase
**Category**: Correct

**Question 1**: What is Batman like?
**Question 2**: What is Batman’s personality like?
**Label**: Non-paraphrase
**Category**: Correct

J.3 SST-2

SST-2 is the abbreviation of Stanford Sentiment Treebank (Socher et al., 2013). In this task, the model needs to recognize whether the phrases or sentences reflect positive or negative sentiments.

**Input**: predictability is the only winner
**Label**: Negative
**Category**: Correct

**Input**: abandon their scripts and go where the moment takes them
**Label**: Negative
**Category**: Correct

**Input**: chases for an hour and then
**Label**: Positive
**Category**: Unsure

**Input**: provide much more insight than the inside column of a torn book jacket
**Label**: Negative
**Category**: Unsure

**Input**: a children’s party clown
**Label**: Negative
**Category**: Fine

**Input**: perhaps even the sle high command found writer-director mitch davis ’s wall of kitsch hard going.
**Label**: Negative
**Category**: Correct

**Input**: own placid way
**Label**: Negative
**Category**: Correct

**Input**: get on a board and , uh , shred ,
**Label**: Negative
**Category**: Correct

**Input**: asks what truth can be discerned from non-firsthand experience , and specifically questions cinema ‘s capability for recording truth .
**Label**: Positive
**Category**: Correct

**Input**: puts the dutiful efforts of more disciplined grade-grubbers
**Label**: Positive
**Category**: Correct

**Input**: filter out the complexity
**Label**: Positive
**Category**: Correct

**Input**: told what actually happened as if it were the third ending of clue'
**Label**: Negative
**Category**: Correct

**Input**: is more in love with strangeness than excellence .
**Label**: Positive
**Category**: Wrong

**Input**: i found myself howling more than cringing
**Label**: Positive
**Category**: Correct

**Input**: goldbacher draws on an elegant visual sense and a talent for easy , seductive pacing
... but she and writing partner laurence coriat do
n’t manage an equally assured narrative coinage
Input : for a thirteen-year-old ‘s book re-
port
Label : Negative
Category : Correct

Input : a problem hollywood too long has
ignored
Label : Negative
Category : Correct

Input : twisted sense
Label : Negative
Category : Correct

Input : a stab at soccer hooliganism
Label : Negative
Category : Correct

Input : sinuously plotted
Label : Negative
Category : Correct

Input : shiner can certainly go the distance
, but is n’t world championship material
Label : Positive
Category : Correct

Input : holding equilibrium up
Label : Negative
Category : Wrong

Input : i am highly amused by the idea that
we have come to a point in society where it has
been deemed important enough to make a film in
which someone has to be hired to portray richard
dawson.
Label : Positive
Category : Wrong

Input : waters
Label : Negative
Category : Wrong

Input : what might have emerged as hilari-
ous lunacy in the hands of woody allen or
Label : Positive
Category : Correct

Input : of those airy cinematic bon bons
whose aims – and by extension , accomplishments
– seem deceptively slight on the surface
Label : Positive
Category : Correct

Input : do n’t blame eddie murphy but
Label : Negative
Category : Correct

Input : melodramatic paranormal romance
Label : Negative
Category : Correct

Input : could possibly be more contemptu-
ous of the single female population
Label : Negative
Category : Correct

Input : cremaster 3 ” should come with the
warning “ for serious film buffs only ! ”
Label : Negative
Category : Correct

Input : softheaded metaphysical claptrap
Label : Negative
Category : Correct

Input : owed to benigni
Label : Negative
Category : Unsure

Input : to be a suspenseful horror movie or
a weepy melodrama
Label : Positive
Category : Correct

Input : genuinely unnerving .
Label : Positive
Category : Correct

Input : gaping enough to pilot an entire
olympic swim team through
Label : Negative
Category : Correct

Input : this is popcorn movie fun with equal doses
of action , cheese , ham and cheek ( as well as a
serious debt to the road warrior ) , but it feels like
unrealized potential
Label: Positive
Category: Fine

Input: feeling like it was worth your seven bucks, even though it does turn out to be a bit of a cheat in the end
Label: Negative
Category: Correct

Input: pull it back
Label: Negative
Category: Correct

Input: , this is more appetizing than a side dish of asparagus.
Label: Negative
Category: Correct

Input: crime drama
Label: Negative
Category: Unsure

Input: like most movies about the pitfalls of bad behavior
Label: Negative
Category: Fine

Input: befallen every other carmen before her
Label: Positive
Category: Unsure

Input: appeal to those without much interest in the elizabethans (as well as rank frustration from those in the know about rubbo ’s dumbed-down tactics)
Label: Negative
Category: Unsure

Input: about existential suffering
Label: Negative
Category: Fine

Input: , if uneven ,
Label: Negative
Category: Unsure

Input: succumbs to sensationalism
Label: Positive
Category: Wrong

Input: that turns me into that annoying specimen of humanity that i usually dread encountering the most
Label: Negative
Category: Fine

Input: at least a minimal appreciation
Label: Positive
Category: Unsure

Input: underlines even the duldest tangents
Label: Negative
Category: Correct

Input: heard before
Label: Positive
Category: Unsure

Input: i like my christmas movies with more elves and snow and less pimps and ho ’s.
Label: Negative
Category: Unsure

Input: can aspire but none can equal
Label: Negative
Category: Unsure

Input: fathom
Label: Negative
Category: Unsure

Input: attempt to bring cohesion to pamela ’s emotional roller coaster life
Label: Negative
Category: Unsure

Input: movie version
Label: Positive
Category: Wrong

Input: of spontaneity in its execution and a dearth of real poignancy
Label: Positive
Category: Correct