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 aspects 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 $E \in [3, 3\frac{1}{3}, 3\frac{2}{3}, 4]$.

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 $\Delta\hat{\text{Acc}}$ and $\Delta\text{Acc}'$, or say, $\hat{\text{Decay}}(t)$ and $\text{Decay}'(t)$ in Figure 6, 7, and 8.

Additionally, for each pair of model sizes s_1 and s_2, 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] 4 Layers with hidden dimension 256

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

	QQP	MNLI	SST-2
MINI ^{orig}	88.2%	74.6%	92.8%
MINI ^{ours}	87.3%	74.3%	92.8%
SMALL ^{orig}	89.1%	77.3%	93.9%
SMALL ^{ours}	88.7%	76.7%	93.9%
MEDIUM ^{orig}	89.8%	79.6%	94.2%
MEDIUM ^{ours}	89.5%	78.9%	94.2%
BASE ^{orig}	90.8%	83.8%	95.0%
BASE ^{ours}	90.6%	81.2%	94.6%
LARGE ^{orig}	91.3%	86.8%	95.2%
LARGE ^{ours}	91.0%	83.8%	94.8%

Table 8: Comparing our pretrained model (superscript ^{orig}) to the original release by Devlin et al. (2019) and Turc et al. (2019) (superscript ^{ours}). All pretrained models are finetuned with the training set and tested on the in-distribution evaluation set described in Appendix A.

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,

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

More concretely, suppose each instance is indexed by i, the set of all instances is $\mathcal{T}$, and the random seed is R; then $c^s_R \in \{0,1\}^{|\mathcal{T}|}$ is a random $|\mathcal{T}|$ dimensional vector, where $c^s_{R,i} = 1$ 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 $c^{s_1}_{R_{1\ldots 2k}}$ and $c^{s_2}_{R_{2k+1\ldots 4k}}$, where there are $2k$ different random seeds for each model size [11]. Then

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

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

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

[11] We assumed even number of random seeds since we will mix half of the models from each size to compute the random baseline

For the random baseline estimator, we have

$$\Delta\text{Acc}'_i := \frac{1}{2k}\Big(\sum_{j=1}^{k} c^{s_1}_{R_j,i} + \sum_{j=2k+1}^{3k} c^{s_2}_{R_j,i} - \sum_{j=k+1}^{2k} c^{s_1}_{R_j,i} - \sum_{j=3k+1}^{4k} c^{s_2}_{R_j,i}\Big), \tag{19}$$

and the false discovery control Decay' is defined as

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

To reiterate, the definition of the true decay rate is

$$\text{Decay} = \frac{1}{|\mathcal{T}|}\sum_{i=1}^{|\mathcal{T}|} \mathbf{1}[\Delta\text{Acc}_i < 0]. \tag{21}$$

Our goal is to prove that

$$\text{Decay} \geq \mathbb{E}_{R_{1\ldots R_{4k}}}[\hat{\text{Decay}}(t) - \text{Decay}'(t)] \tag{22}$$

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

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

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

$$\mathbf{1}[\Delta\text{Acc}_i < 0] - \mathbb{P}[\Delta\hat{\text{Acc}}_i \leq t] + \mathbb{P}[\Delta\text{Acc}'_i \leq t] \geq 0 \tag{24}$$

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

$$\Delta\text{Acc}_i \geq 0 \Rightarrow \mathbb{P}[\Delta\text{Acc}'_i \leq t] \geq \mathbb{P}[\Delta\hat{\text{Acc}}_i \leq t], \tag{25}$$

which will be proved in Lemma 1.□

Lemma 1

$$\Delta\text{Acc}_i \geq 0 \Rightarrow \mathbb{P}[\Delta\text{Acc}'_i \leq t] \geq \mathbb{P}[\Delta\hat{\text{Acc}}_i \leq t], \tag{26}$$

MNLI $s_1 \setminus s_2$	MINI	SMALL	MEDIUM	BASE	LARGE
MINI	0.000	0.087	0.136	0.179	0.214
SMALL	0.033	0.000	0.089	0.139	0.180
MEDIUM	0.050	0.028	0.000	0.090	0.143
BASE	0.060	0.048	0.026	0.000	0.101
LARGE	0.059	0.052	0.040	0.021	0.000
QQP $s_1 \setminus s_2$	MINI	SMALL	MEDIUM	BASE	LARGE
MINI	0.000	0.057	0.076	0.100	0.107
SMALL	0.019	0.000	0.039	0.073	0.084
MEDIUM	0.029	0.014	0.000	0.044	0.063
BASE	0.034	0.027	0.016	0.000	0.032
LARGE	0.036	0.031	0.027	0.016	0.000
SST-2 $s_1 \setminus s_2$	MINI	SMALL	MEDIUM	BASE	LARGE
MINI	0.000	0.037	0.043	0.052	0.057
SMALL	0.010	0.000	0.015	0.031	0.036
MEDIUM	0.016	0.008	0.000	0.020	0.028
BASE	0.019	0.014	0.009	0.000	0.014
LARGE	0.020	0.017	0.015	0.008	0.000

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

$$p_i^{s_m} := \mathbb{E}_R[c_i^{s_m}], \tag{27}$$

then

$$p_i^{s_1} \leq p_i^{s_2} \tag{28}$$

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 $\Delta\hat{\text{Acc}}_i$ and $\Delta\text{Acc}'_i$ as the sum/difference of several binomial variables, i.e.

$$\Delta\hat{\text{Acc}}_i \sim (\text{Binom}(k, p_i^{s_2}) + \text{Binom}(k, p_i^{s_2}) - \text{Binom}(k, p_i^{s_1}) - \text{Binom}(k, p_i^{s_1}))/2k, \tag{29}$$

and

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

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

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 $\Delta\hat{\text{Acc}}$ and $\Delta\text{Acc}'$ 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.

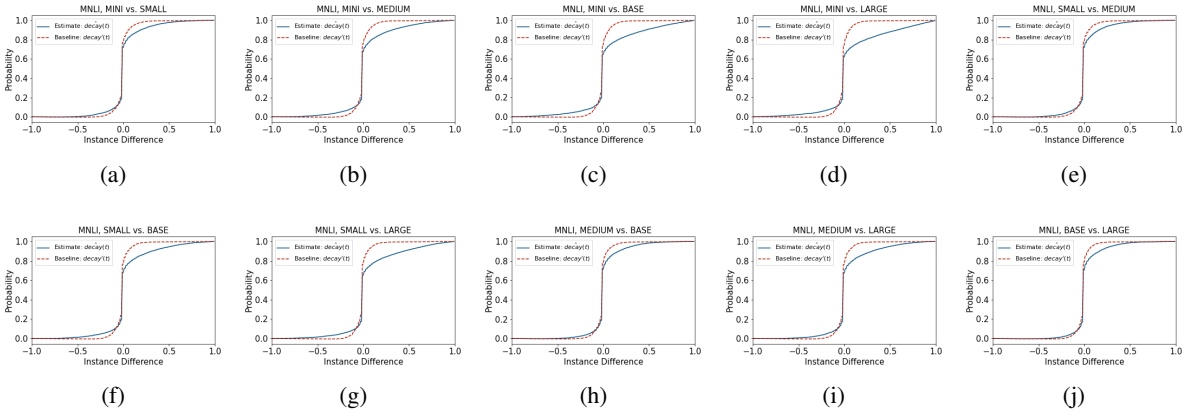
Hypothetically, suppose we are comparing a smaller model s_1 and a larger model s_2. For the smaller model, with .1 probability it finds a perfect pretrained model that always predict correctly across all finetuning runs and with .9 probability it finds a bad pretrained model that predict always incorrectly. For the larger model, with probability 1 it finds an average pretrained model that predict correctly for .2 fraction of finetuning runs. The larger model is on average better, because it has $.2 > .1$ probability to be correct. Hence, $\Delta\text{Acc} > 0$

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

$$\mathbb{P}[\Delta\hat{\text{Acc}}_i \leq -0.8] \tag{31}$$

$$= 0.01 \geq 0 = \mathbb{P}[\Delta\text{Acc}'_i \leq -0.8] \tag{32}$$

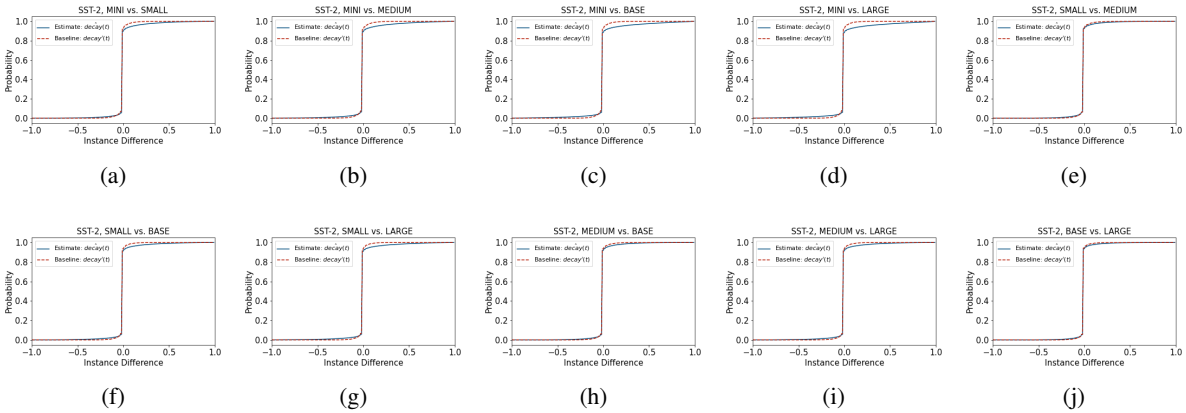
The event that $\Delta\hat{\text{Acc}}_i \leq -0.8$ happens with probability 0.01 when both of the two small pretrained models have good pretraining seeds, and $\Delta\text{Acc}'_i$ is at least -0.5 and will never be less than -0.8.



(a) (b) (c) (d) (e)

(f) (g) (h) (i) (j)

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.



(a) (b) (c) (d) (e)

(f) (g) (h) (i) (j)

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

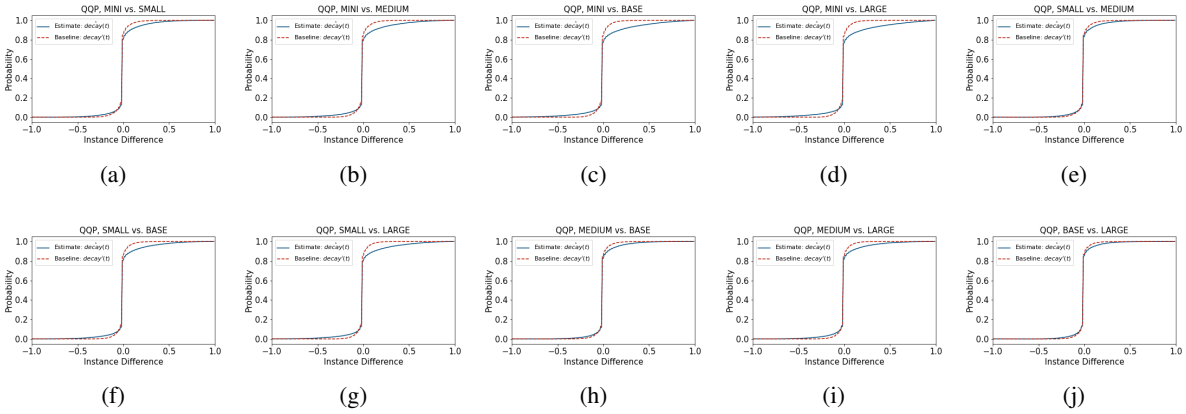
The key idea behind this counter-example is that even if the larger model has better average, the distribution of average finetuning accuracy for different pretraining seeds might not stochastically dominate the one with lower average because of outliers. Hence, a priori, this is unlikely to happen in 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 pretrained 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 $\mathbb{E}[(L^* - L)]/\mathbb{E}[L)]$, 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%.



(a) (b) (c) (d) (e)

(f) (g) (h) (i) (j)

Figure 8: Similar to Figure 4, on QQP in-domain development set, for each pair of model sizes, we plot the cumulative density function of instance differences.

F 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 lowerbound 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[12]. 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 correct while bigger completely wrong, while on the rest 0.01% smaller 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 Benjamin-Hochberg procedure is 0.

[12]This is for intuition, though, and we do not need any assumption on the prior of instance accuracy, which requires a Bayes interpretation.

G More Results on Momentum

We report more results on the correlation between instance differences. Specifically, for one triplet of model sizes (e.g. MINI $\Rightarrow$ MEDIUM $\Rightarrow$ 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 ${}^{\text{MINI}}_{\text{MEDIUM}}\Delta\text{Acc}$ and ${}^{\text{MEDIUM}}_{\text{LARGE}}\Delta\text{Acc}$. 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.

H 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.

MNLI. 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

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 ${}^{\text{MINI}}_{\text{MEDIUM}}\Delta\text{Acc}$ and ${}^{\text{MEDIUM}}_{\text{LARGE}}\Delta\text{Acc}$. Each column t represents a bucket that contains instances with middle size accuracy in $[t-0.1, t]$. For example, if the row name is MINI, MEDIUM, LARGE, then the column 0.2 corresponds to a bucket where $\hat{\text{Acc}}_i^{\text{MEDIUM}}$ is between 0.1 and 0.2. We calculate the PearsonR correlation score between ${}^{\text{MINI}}_{\text{MEDIUM}}\Delta\hat{\text{Acc}}$ and ${}^{\text{MEDIUM}}_{\text{LARGE}}\Delta\hat{\text{Acc}}$ across all instances in the bucket.

finetuning randomness, and across different checkpoints throughout training. We formally define the quantities we want to estimate in Appendix H.1, present an unbiased estimator for these quantities in Appendix H.2, and show that our method can be generalized to arbitrary number of source of randomness in Appendix H.3.

Specifically, the main paper focused on scenarios with 2 sources of randomness: pretraining and finetuning. We discuss the case with 3 sources of randomness in the appendix, rather than 2 as in the main paper, because it is easier to understand the general estimation strategy in the case of 3.

H.1 Formalizing Decomposition

Recall that P is the pretraining seed, F is the finetuning seed, E represents a model checkpoint, i indexes each instance (datapoint). $c^{s,i}_{P,F,E} = 1$ if the model of size s with pretraining seed p and finetuning 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

$$\mathcal{L}^{s,i} = \mathbb{E}_{P,F,E}[(1 - c^{s,i}_{P,F,E})^2] \tag{33}$$

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:

$$\begin{aligned}\mathcal{L} =& \text{Bias}^2 + \text{PretVar} \\ &+ \text{FineVar} + \text{CkptVar}.\end{aligned} \tag{34}$$

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

$$\text{Bias}^2 = (1 - \mathbb{E}_{P,F,E}[c_{P,F,E}])^2. \tag{35}$$

PretVar captures the variance brought by randomness in pretraining, and is defined as

$$\text{PretVar} = Var_P[\mathbb{E}_{F,E}[c_{P,F,E}]]. \tag{36}$$

Similarly, we define the variance brought by randomness in finetuning FineVar

$$\text{FineVar} = \mathbb{E}_P[Var_F[\mathbb{E}_E[c_{P,F,E}]]], \tag{37}$$

and that by fluctuations across checkpoints e

$$\text{CkptVar} = \mathbb{E}_{P,F}[Var_E[c_{P,F,E}]]. \tag{38}$$

H.2 Unbiased Estimation

We first describe the data on which we apply our estimator. Suppose we pretrain $\mathcal{P}$ models with different random seeds, for each of the $\mathcal{P}$ pretrained models we finetune with $\mathcal{F}$ different random seeds, and we evaluate at $\mathcal{E}$ different checkpoints. Then $\forall j \in [\mathcal{P}], k \in [\mathcal{F}], l \in [\mathcal{E}]$ [13] , we observe P_j, F_{jk}, E_{jkl}, and $c_{P_j,F_{jk},E_{jkl}}$, 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 CkptVar

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

$$\hat{\text{CkptVar}} := \frac{1}{\mathcal{P}\mathcal{F}} \sum_{j\in[\mathcal{P}],k\in[\mathcal{F}]} \hat{Var}_E^{P_j,F_{jk}}, \quad (39)$$

where

$$\hat{Var}_E^{P_j,F_{jk}} := \frac{1}{\mathcal{E}-1} \sum_{l\in\mathcal{E}} (c_{P_j,F_{jk},E_{jkl}} - \bar{c}_{P_j,F_{jk}})^2, \quad (40)$$

and

$$\bar{c}_{P_j,F_{jk}} := \frac{1}{\mathcal{E}} \sum_{l\in[\mathcal{E}]} c_{P_j,F_{jk},E_{jkl}}. \quad (41)$$

$\hat{\text{CkptVar}}$ is unbiased, since $\hat{Var}_E^{P_j,F_{jk}}$ is an unbiased estimation of variance of c with fixed P_j and F_{jk}, and randomness E, i.e.

$$\mathbb{E}_{E_{ij(\cdot)}}[\hat{Var}_E^{P_j,F_{jk}}] = Var_E[c_{P,F,E}|P = P_j, F = F_{jk}]. \quad (42)$$

Therefore, $\forall j \in [\mathcal{P}], k \in [\mathcal{F}]$, we have

$$\mathbb{E}_{P_j,F_{jk}}[\hat{Var}_E^{P_j,F_{jk}}] = \text{CkptVar}, \quad (43)$$

and hence by linearity of expectation

$$\mathbb{E}_{P_{(\cdot)},F_{(\cdot)},E_{(\cdot)}}[\hat{\text{CkptVar}}] = \text{CkptVar}. \quad (44)$$

H.2.2 Estimating 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:

$$Var_F[\mathbb{E}_E[c_{P,F,E}]|P = P_j], \quad (45)$$

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

$$Var_F := Var_F[\mathbb{E}_E[c_{F,E}]] \quad (46)$$

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

$$\bar{c}_{F_{jk}} := \frac{1}{\mathcal{E}} \sum_{l\in[\mathcal{E}]} c_{F_k,E_{kl}}, \quad (47)$$

and then calculate the sample variance $\tilde{Var}_F$ of $\bar{c}$ with respect to F:

$$\tilde{Var}_F := \frac{1}{\mathcal{F}-1} \sum_{k\in[\mathcal{F}]} (\bar{c}_{F_k} - \bar{c})^2, \quad (48)$$

where

$$\bar{c} := \frac{1}{\mathcal{F}} \sum_{k\in[\mathcal{F}]} \bar{c}_{F_k} \quad (49)$$

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

We introduce the following general theorem to correct this bias.

Theorem 2 *Suppose $\mathcal{D}_k, k \in [\mathcal{F}]$ are independently sampled from the same distribution Ξ, which is a distribution of distributions; $\hat{\mu}_k$ is an unbiased estimator of $\mathbb{E}_{X\in\mathcal{D}_k}[X]$, and $\hat{\phi}_k$ to be an unbiased estimator of the variance of $\hat{\mu}_k$, then*

$$\hat{Var}_F = \frac{1}{\mathcal{F}-1} \sum_{k\in[\mathcal{F}]} (\hat{\mu}_k - \hat{\mu})^2 - \frac{1}{\mathcal{F}} \sum_{k\in[\mathcal{F}]} \hat{\phi}_k \quad (50)$$

is an unbiased estimator for

$$V = Var_{\mathcal{D}\sim\Xi}[\mathbb{E}_{X\sim\mathcal{D}}[X]], \quad (51)$$

where

$$\hat{\mu} := \frac{1}{\mathcal{F}} \sum_{k\in[\mathcal{F}]} \hat{\mu}_k \quad (52)$$

[13] $[L] := \{l : l \in N, l \in [0, L-1]\}$

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 $\mathcal{D}_k$ to be μ_k, i.e.

$$\mu_k := \mathbb{E}_{X\sim\mathcal{D}_k}[X], \tag{53}$$

and the population mean of μ_k across randomness in $\mathcal{D}$ to be μ, i.e.

$$\mu := \mathbb{E}_{\mathcal{D}\sim\Xi}[\mathbb{E}_{X\sim\mathcal{D}}[X]] \tag{54}$$

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

$$\begin{aligned}
&\frac{1}{\mathcal{F}-1}\mathbb{E}[\sum_{k\in[\mathcal{F}]}(\hat{\mu}_k-\hat{\mu})^2] \\
&=\frac{1}{\mathcal{F}-1}\mathbb{E}[\sum_{k\in[\mathcal{F}]}((\hat{\mu}_k-\mu_k)-(\hat{\mu}-\mu) \\
&+(\mu_k-\mu))^2] \\
&=\frac{1}{\mathcal{F}-1}\mathbb{E}[\sum_{k\in[\mathcal{F}]}[(\hat{\mu}_k-\mu_k)^2+(\hat{\mu}-\mu))^2 \\
&+(\mu_k-\mu)^2-2(\hat{\mu}_k-\mu_k)(\hat{\mu}-\mu) \\
&-2(\hat{\mu_k}-\mu_k)(\hat{\mu}-\mu)]]
\end{aligned} \tag{55}$$

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

$$\mathbb{E}[\sum_{k\in[\mathcal{F}]}(\hat{\mu}_k-\mu_k)^2]=\mathbb{E}[\sum_{k\in[\mathcal{F}]}\hat{\phi}_k], \tag{56}$$

$$\begin{aligned}
\mathbb{E}[(\hat{\mu}-\mu)^2] &= \mathbb{E}[(\mu-\frac{1}{\mathcal{F}}\sum_{k\in[\mathcal{F}]}\mu_k) \\
&+\frac{1}{\mathcal{F}}\sum_{k\in[\mathcal{F}]}(\mu_k-\hat{\mu}_k))^2] \\
&=\frac{1}{\mathcal{F}}V+\frac{1}{\mathcal{F}^2}\sum_{k\in[\mathcal{F}]}\mathbb{E}[\hat{\phi}_k]
\end{aligned} \tag{57}$$

$$\mathbb{E}[\sum_{k\in[\mathcal{F}]}(\mu_k-\mu)^2]=\mathcal{F}V \tag{58}$$

$$\begin{aligned}
&\mathbb{E}[-2\sum_{k\in[\mathcal{F}]}(\hat{\mu}_k-\mu_k)(\hat{\mu}-\mu)] \\
&=-\frac{2}{\mathcal{F}}\mathbb{E}[\sum_{k\in[\mathcal{F}]}\hat{\phi}_k].
\end{aligned} \tag{59}$$

$$\begin{aligned}
&\mathbb{E}[-2\sum_{k\in[\mathcal{F}]}(\hat{\mu}_k-\mu_k)(\hat{\mu}-\mu)] \\
&=-2V.
\end{aligned} \tag{60}$$

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

$$\begin{aligned}
&\frac{1}{\mathcal{F}-1}\mathbb{E}[\sum_{k\in[\mathcal{F}]}(\hat{\mu}_k-\hat{\mu})^2] \\
&=\frac{1}{\mathcal{F}-1}\mathbb{E}[\sum_{k\in[\mathcal{F}]}\hat{\phi}_k \\
&+\mathcal{F}(\frac{1}{\mathcal{F}}V+\frac{1}{\mathcal{F}^2}\sum_{k\in[\mathcal{F}]}\mathbb{E}[\hat{\phi}_k]) \\
&+\mathcal{F}V \\
&-\frac{2}{\mathcal{F}}\sum_{k\in[\mathcal{F}]}\hat{\phi}_k \\
&-2V] \\
&=V+\mathbb{E}[\frac{1}{\mathcal{F}}\sum_{k}\hat{\phi}_k]
\end{aligned} \tag{61}$$

Then from Equation 50, we can tell that $\hat{Var}_F$ is unbiased. □

Now we come back to the topic of developing an unbiased estimator for Var_F as defined in Equation 46. To utilize Theorem 2, we need two components:

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

$\bar{c}_{F_k}$ is an unbiased estimator for $\mathbb{E}_E[c_{F,E}|F=F_k]$, and its variance $Var_E[\bar{c}_F|F=F_k]$ is

$$Var_{E_{(.)}}(\bar{c}_{F_k})=\frac{1}{\mathcal{E}}Var_E[c_{F,E}|F=F_k]. \tag{62}$$

Therefore, to develop an unbiased estimator for $Var_E(\bar{c}_{F_{jk}})$, it suffices to have an unbiased estimate of $Var_E[c_{F,E}|F=F_k]$. We define

$$\hat{\phi}_{F_k}:=\frac{1}{\mathcal{E}(\mathcal{F}-1)}\sum_{k\in[\mathcal{L}]}(c_{F_k,E_{kl}}-\bar{c}_{F_k})^2, \tag{63}$$

and we can plug in $\hat{\phi}_{F_k}$ and $\hat{\mu}_{F_k} = \bar{c}_{F_k}$ into Theorem 2 as an unbiased estimator to obtain an unbiased estimator for $Var_F[\mathbb{E}_E[c_{P,F,E}]|P = P_j]$, and we average the estimation for each P_j to obtain an unbiased estimate.

H.2.3 Estimating $\mathrm{PretVar}$

We next estimate $Var_P[\mathbb{E}_{F,E}[c_{P,F,E}]]$ 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

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

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.

$$\begin{aligned} Var_{F,E}(\bar{c}_{P_j}) = \frac{1}{\mathcal{F}} Var_F(\mathbb{E}_E[c_{P,F,E}]|P = P_j) \\ + \frac{1}{\mathcal{F}^2} \sum_{k\in[\mathcal{F}]} Var_E(\bar{c}_{P_j,F_{jk}}), \end{aligned} \tag{65}$$

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

H.2.4 Estimating Bias^2

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

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

and

$$\mathbb{E}[\hat{\mathcal{L}}] = \mathcal{L}. \tag{67}$$

By linearity of expectation and loss decomposition in Equation 34,

$$\begin{aligned} \hat{\mathrm{Bias}^2} := \hat{\mathcal{L}} - \hat{\mathrm{PretVar}} \\ - \hat{\mathrm{FineVar}} - \hat{\mathrm{CkptVar}} \end{aligned} \tag{68}$$

is an unbiased estimator of Bias^2.

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

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

$$\mathbb{E}_{r_1\ldots,r_{n-1}}[Var_{r_n}[\mathbb{E}_{r_{n+1}\ldots r_N}[c_{r_1\ldots,c_N}]]], \tag{69}$$

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 $r_{1\ldots n-1}$ and average them together, as discussed in Section H.2.1.

To estimate

$$Var_{r_n}[\mathbb{E}_{r_{n+1}\ldots r_N}[c_{r_1\ldots,c_N}]], \tag{70}$$

we make use of Theorem 2, which requires

- an unbiased estimator $\hat{\mu}_{r_{n+1}}$ for the quantity $\mathbb{E}_{r_{n+1}\ldots r_N}[c_{r_1\ldots,c_N}]$, 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 Bias^2 of $b^2 \in [0, 1]$", i.e.

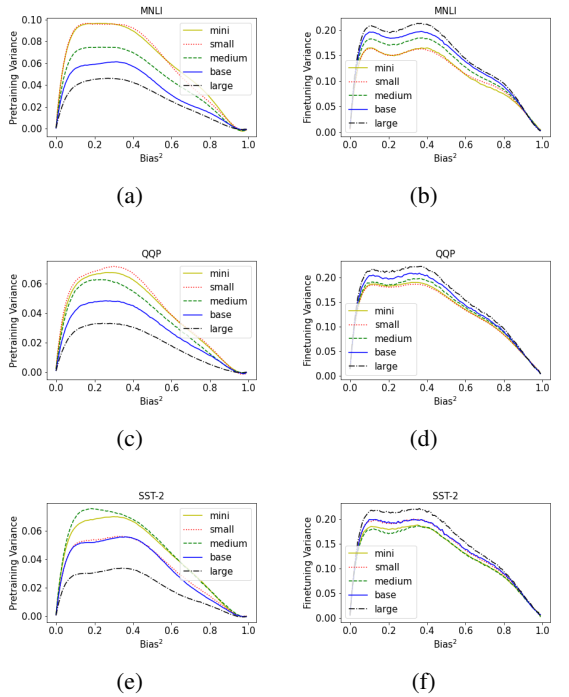


Figure 9: The variance curve conditioned on Bias^2 for in-domain development set of MNLI, QQP and SST-2. Each curve represents a model size. Left for pretraining variance and right for finetuning variance.

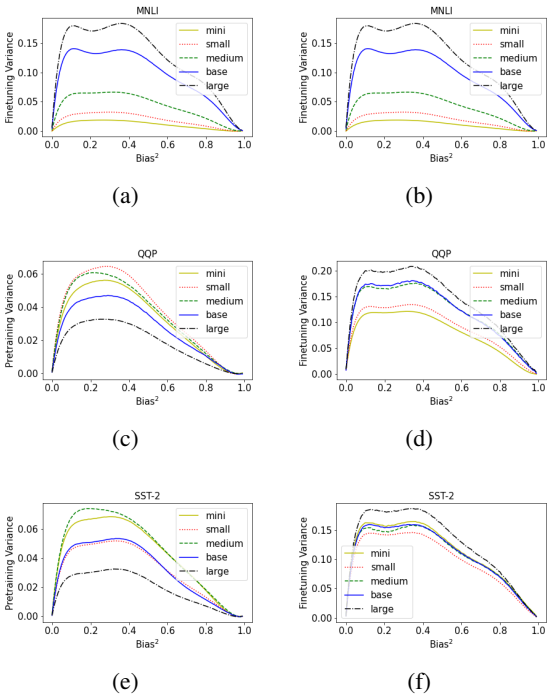


Figure 10: The same figure as 9, except for using the squared loss function $\mathcal{L} = (1-p)^2$, where p is the probability assigned to the correct label, instead of 0/1 loss.

$$\text{PretVar}(b^2) := \mathbb{E}_i[\text{PretVar}_i | \text{Bias}^2{}_i = b^2] \quad (71)$$

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:

$$\mathcal{L}_i = (1-p_i)^2, \quad (72)$$

where p_i is the probability the 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 $\overset{\text{MINI}}{\text{LARGE}}\Delta\hat{\text{Acc}}_i \leq -0.9$ 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.
Label : Neutral
Category : Correct

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 diffferently 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 : Fine

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 bridgeopening 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[14]. 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 : Fine

Question 1 : What is the basic knowledge of computer science engineers?
Question 2 : What is basic syallbus 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?
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

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

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 : 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 billionaire 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 : Fine

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 : Fine

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 slc 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
Label : Positive
Category : Unsure

Input : for a thirteen-year-old 's book report
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 hilarious 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 contemptuous 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 dullest 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