

# Evaluating the fairness of task-adaptive pretraining on unlabeled test data before text classification

Anonymous ACL submission

## Abstract

Few-shot learning benchmarks are critical for evaluating modern NLP techniques. It's possible, however, that benchmarks favor methods which easily make use of unlabeled text, because researchers can use unlabeled text from the test set to pretrain their models. Given the dearth of research on this potential problem, we run experiments to quantify the bias caused by pretraining on unlabeled test set text instead of on unlabeled, independently drawn text. Controlled few-shot and zero-shot experiments on 25 classification tasks and 3 language models—BERT, GPT-2, and Mistral 7B—do not find evidence of overoptimism. Furthermore, we demonstrate the importance of repeated subsampling when studying few-shot text classification, and recommend that few-shot learning benchmarks include multiple training folds. Code and data are available here: <https://github.com> (currently omitted for anonymity).

## 1 Introduction

It's standard for NLP benchmarks to release text from the test set. This allows researchers to submit a file of predictions instead of submitting code. A potential concern is that researchers can use this text during training. Consider the Real-world Annotated Few-shot Tasks (RAFT) benchmark (Alex et al., 2021), which contains "few-shot" text classification tasks—tasks where the training set contains a relatively small number of labeled examples. Below is an excerpt from the RAFT paper (emphasis added):

For each task, we release a public training set with 50 examples and a larger unlabeled test set. *We encourage unsupervised pre-training on the unlabelled examples* and open-domain information retrieval.

In the RAFT competition, a model is evaluated by scoring its predictions on the same set of unlabeled text which the model may have been trained on (using an unsupervised training procedure).

It's wrong to train a model on test set features with their labels and then evaluate on the test set when one needs to estimate performance on out-of-sample data. Test set performance would be overoptimistic (Hastie et al., 2009). This fact is widely known. But what if, as encouraged by Alex et al. (2021), a model is trained on test set features *without* test set labels? This paper studies this question for the domain of text classification.

## 2 Motivation

NLP benchmarks for few-shot learning are prevalent, as having only a handful of labeled examples is more realistic. One consideration when designing these benchmarks is that some few-shot approaches can—at least theoretically—use unlabeled text from the test set. With Pattern-Exploiting Training (Schick and Schütze, 2021), for example, one can train the final classifier on test set text with soft labels predicted by an ensemble of supervised models. With Pre-trained Prompt Tuning (Gu et al., 2022), one can pretrain the language model (LM) on unlabeled test set text before prompt-tuning on the labeled training set. A more classical approach would be to train a word2vec model (Mikolov et al., 2013) on unlabeled test set text, run this model on training text to get embeddings, and finally train a classifier on these embeddings with labels from the training set.

For other few-shot approaches, such as SetFit (Tunstall et al., 2022) and in-context learning with LLMs (as popularized by Brown et al., 2020), it's more common to only use labeled text.

While the ability to exploit unlabeled text is useful, applying this ability to test set text could be substantively different than applying it to text which is

statistically independent of the test set. This difference in methodology may be more concerning in the few-shot setting than in the many-shot setting. It’s conceivable that differences between few-shot methods are due just as much to how unlabeled text is used as they are to how the few, labeled examples are used. This raises the question: does pretraining a model on a benchmark’s unlabeled test set text inflate the model’s performance on that benchmark?

### 3 Related work

As indicated by the quote in §1, the RAFT benchmark implicitly assumes that the answer is no. The validity of using test set features is not a fringe opinion. The popular textbook by [Hastie et al. \(2009\)](#) contains the following passage without a reference or evidence (emphasis added):

There is one qualification: *initial unsupervised screening steps can be done before samples are left out*. For example, we could select the 1000 predictors with highest variance across all 50 samples, before starting cross-validation. *Since this filtering does not involve the class labels, it does not give the predictors an unfair advantage*.

The opposite opinion—that exploiting unlabeled test set features is unfair—may align more closely with best practices. For example, [Gururangan et al. \(2020\)](#) contains the following criticism of another study when comparing performances on a popular text classification benchmark:

[Thongtan and Phienthrakul \(2019\)](#) report a higher number (97.42) on IMDB, but they train their word vectors on the test set.

[Jacovi et al. \(2023\)](#) argue that benchmarks which release unlabeled test set text can be compromised in two ways. First, the text can be discretely labeled and trained on by model developers. Second, if test set texts were scraped from websites which also host their labels, e.g., movie reviews and ratings on IMDB, then an LLM may have already been pretrained on these labeled texts (Scenario 1 in [Jacovi et al., 2023](#)). They do not discuss potential problems with using unlabeled test set text by itself.

[Moscovich and Rosset \(2022\)](#) contains experiments and theory for unsupervised methods which

are common to tasks involving tabular data. They find that estimators of out-of-sample performance which were subject to these methods may be biased positively or negatively, depending on the parameters of the problem. They recommend further research on this bias in more domains, particularly when dealing with small sample sizes and high-dimensional data.

## 4 Experimental design

In the absence of answers to this question in NLP, we study whether pretraining on unlabeled test set text biases test set performance for 25 diverse text classification tasks and two types of LMs: BERT ([Devlin et al., 2019](#)), and GPT-2 ([Radford et al., 2019](#)). Appendix A describes each task.

The goal of the experiment is to first establish that pretraining is beneficial, in line with [Gururangan et al. \(2020\)](#). Second, given that pretraining has a detectable benefit, the experiment measures the accuracy difference between using test set text for the pretraining stage—an arguably unfair methodology—and using text which is independent of the test set—an inarguably fair methodology.

In more detail, the experiment starts by subsampling three separate sets of data from the full sample of data for a given text classification task:

- extra:  $n$  (either 50, 100, 200 or 500) unlabeled texts which are optionally used for pre-training
- train:  $m$  (either 50 or 100) labeled texts for classification training
- test:  $n$  labeled texts to report accuracy.

Next, three accuracy estimators are computed. The procedures used to obtain them are described below.

### 4.1 $\text{acc}_{\text{extra}}$

1. Train a freshly loaded, pretrained LM on the  $n$  unlabeled texts in extra using the LM’s pretraining objective—masked language modeling loss for BERT, or autoregressive/causal language modeling loss for GPT-2.
2. Add a linear layer to this model and finetune all of the LM’s weights to minimize classification cross entropy loss on train.
3. Compute the classification accuracy of this model on test.

Step 1 is task-adaptive pretraining—a procedure broadly recommended by Gururangan et al. (2020). Step 2 is a canonical way to train a transformer-based LM for a classification task, according to Section 2 of Zhang et al. (2021).

$\text{acc}_{\text{extra}}$  is clearly an unbiased estimator of out-of-sample accuracy because it never trains on test.

## 4.2 $\text{acc}_{\text{test}}$

$\text{acc}_{\text{test}}$  is identical to  $\text{acc}_{\text{extra}}$ , except that task-adaptive pretraining is done on test instead of extra in step 1.

$\text{acc}_{\text{test}}$  represents what one might see in a competition like RAFT, where pretraining on unlabeled text from test is encouraged. It’s unclear whether this accuracy estimator is unbiased, because it came from pretraining and evaluating on the same set of test set text. A reasonable hypothesis is that it’s overoptimistic, i.e.,  $E[\text{acc}_{\text{test}}] > E[\text{acc}_{\text{extra}}]$ .

## 4.3 $\text{acc}_{\text{base}}$

$\text{acc}_{\text{base}}$  doesn’t do task-adaptive pretraining; it doesn’t make any use of unlabeled text. It trains a pretrained LM on train to do classification, and then computes this model’s accuracy on test.

This score provides a sanity check. If there’s no boost from  $\text{acc}_{\text{base}}$  to  $\text{acc}_{\text{extra}}$ , then it may not be surprising to observe no difference between  $\text{acc}_{\text{extra}}$  and  $\text{acc}_{\text{test}}$ . A boost from  $\text{acc}_{\text{base}}$  to  $\text{acc}_{\text{extra}}$  rules out undertraining as the cause of a null difference between  $\text{acc}_{\text{extra}}$  and  $\text{acc}_{\text{test}}$  due to insufficient pre-training epochs or too low a learning rate.

## 4.4 Repeated subsampling

The three accuracy estimators are paired, because their classification training and test sets are identical. The only difference is the source of unlabeled text for pretraining. For  $\text{acc}_{\text{extra}}$ , the source is independent of test data. For  $\text{acc}_{\text{test}}$ , the source is exactly the test set text. For  $\text{acc}_{\text{base}}$ , no unlabeled text is used.

A potentially important source of variation in this experiment is the particular subsamples, i.e., the particular realizations of extra, train, and test for a given classification task. To expose this variation, the experiment procedure is repeated tens of times for each classification task.<sup>1</sup> For example, for  $n = 500$ , and for each of the 25 classification

<sup>1</sup>For  $n = 50$  and  $n = 100$ , the experiment is repeated 100 times. For  $n = 200$ , the experiment is repeated 50 times. For  $n = 500$ , the experiment is repeated 20 times. In total, 81,000 finetuned BERT and GPT-2 models were evaluated.

tasks, 20 ( $\text{acc}_{\text{extra}}$ ,  $\text{acc}_{\text{test}}$ ,  $\text{acc}_{\text{base}}$ ) triples are computed.

Appendix B explains more experiment choices.

## 5 Results

Appendix D.2 visualizes the distributions of  $\text{acc}_{\text{extra}} - \text{acc}_{\text{base}}$  and  $\text{acc}_{\text{test}} - \text{acc}_{\text{extra}}$ .  $\text{acc}_{\text{extra}} - \text{acc}_{\text{base}}$  is a control—it’s the accuracy boost from pretraining on unlabeled independent text versus not pretraining at all.  $\text{acc}_{\text{test}} - \text{acc}_{\text{extra}}$  is the main quantity of interest: it’s the evaluation bias from pretraining on unlabeled test set text instead of on unlabeled independent text.

Table 1 contains means of these differences for each configuration of the experiment. It roughly suggests that while pretraining is consistently beneficial, pretraining on unlabeled test set text does not bias test set performance one way or the other.

A more complete analysis of this data is motivated and performed in the next section.

	BERT	GPT-2
$n = 50$	4.1% 0.19%	3.8% 0.18%
$n = 100$	3.9% 0.18%	4.1% 0.11%
$n = 200$	3.9% -0.39%	4.4% -0.05%
$n = 500$	3.5% 0.48%	4.6% -0.08%

(a)  $m = 50$

	BERT	GPT-2
$n = 50$	6.2% -0.08%	2.2% -0.05%
$n = 100$	6.1% -0.37%	2.5% 0.03%
$n = 200$	4.1% 0.33%	6.3% -0.01%
$n = 500$	6.1% -0.16%	3.9% -0.21%

(b)  $m = 100$

Table 1: Means of accuracy differences taken across all subsamples of all 25 classification tasks. For each cell, the upper-left of the diagonal corresponds to the sample mean of  $\text{acc}_{\text{extra}} - \text{acc}_{\text{base}}$ , and the lower-right corresponds to the sample mean of  $\text{acc}_{\text{test}} - \text{acc}_{\text{extra}}$ .

## 6 Analysis

Reporting means is not enough, especially when studying few-shot learning. Appendix D.2 demonstrates that there’s considerable variance, despite

pairing the accuracy estimators.<sup>2</sup> While these visualizations tell us about how raw accuracy differences vary, they do not tell us how the mean accuracy difference varies. We seek a neat answer to the core questions: on this benchmark of 25 classification tasks, how much does the overall accuracy differ between two modeling techniques, and how much does this difference vary?

One way to communicate the variance is to estimate the standard error of the mean difference across classification tasks. But the standard error statistic can be difficult to interpret (Morey et al., 2016). Furthermore, its computation is not completely trivial due to the data’s hierarchical dependency structure: each triple, ( $\text{acc}_{\text{extra}}$ ,  $\text{acc}_{\text{test}}$ ,  $\text{acc}_{\text{base}}$ ), is drawn from ( $\text{train}$ ,  $\text{test}$ ), which is itself drawn from the given classification dataset.

## 6.1 Model

This analysis does not aim to estimate standard errors. Instead, a hierarchical model is fit (\* is short for  $ijkl$ —the lowest-level observation):

$$Y_* \sim \text{Binomial}(n, \lambda_*) \quad (1)$$

$$\text{logit}(\lambda_*) = \mu + \alpha z_i + U_j + V_{jk} + W_{jl} + \beta x_l \quad (2)$$

$$\mu \sim \text{Normal}(0, 1) \quad (3)$$

$$\alpha \sim \text{Normal}(0, 5) \quad (4)$$

$$U_j \sim \text{Normal}(0, \sigma_U) \quad (5)$$

$$V_{jk} \sim \text{Normal}(0, \sigma_V) \quad (6)$$

$$W_{jl} \sim \text{Normal}(0, \sigma_W) \quad (7)$$

$$\beta \sim \text{Normal}(0, 1) \quad (8)$$

$$\sigma_U, \sigma_V \sim \text{HalfNormal}(0, 1) \quad (9)$$

$$\sigma_W \sim \text{HalfNormal}(0, 3.5355) \quad (10)$$

- (1) number of correct predictions
- (2) logit link for accuracy rate, additive effects
- (3) prior for the global intercept
- (4) prior for the effect of the type of LM (BERT or GPT-2)—a control variable
- (5) prior for the effect of the classification task (partial-pooled to reduce overfitting)
- (6) prior for the nested effect of the task’s subsampled dataset

<sup>2</sup>One source of variance is intentionally introduced: the subsamples/splits, as explained in §4.4. The other source of variance is inherent: the added linear layer to perform classification is initialized with random weights.

- (7) prior for the interaction effect of the task and the effect of interest (to reduce underfitting)

- (8) prior for the effect of interest ( $x_l = 1$  indicates the modeling intervention,  $x_l = 0$  indicates no intervention)

- (9) prior for standard deviations

- (10) prior for standard deviation.

The model is fit using Markov Chain Monte Carlo, using the interface provided by the `bambi` package (Capretto et al., 2022).

## 6.2 Overall effects

In NLP benchmarks, methods are assessed by taking their average performance across tasks. To place the analysis results in this context, samples from the posterior predictive distribution of  $Y_{ijk1} - Y_{ijk0}$  (6.1) are taken, then averaged across  $i$  (the 2 LM types—BERT and GPT-2),  $j$  (the 25 classification tasks), and  $k$  (their subsamples), and divided by  $n$  to obtain the distribution of the average accuracy difference:

$$\frac{\bar{Y}_{\dots 1} - \bar{Y}_{\dots 0}}{n} \quad (11)$$

These distributions are plotted in Figure 1. Each distribution is that of the marginal effect of the modeling intervention: pretraining versus not pretraining before classification training (the pretraining boost), or pretraining on unlabeled test set text instead of on unlabeled independent text before classification training (the evaluation bias).

## 6.3 Task-level effects

While taking an average across tasks provides a concise summary, it cannot be used to rule out the existence of an evaluation bias. If the direction of the bias depends on latent properties of the task, averaging may cancel out real, positive biases with real, negative ones. Alternatively, it may dilute the few real, positive biases with many null ones.

Jin et al. (2021) argue and demonstrate that the benefit of task-adaptive pretraining depends on the task’s causal direction. If the principle of independent causal mechanisms is also relevant to the fairness of pretraining on test set features, then our accuracy data may contain (for the sake of argument) positive evaluation biases for anti-causal tasks, and null biases for causal tasks.<sup>3</sup>

<sup>3</sup>We won’t assess any particular hypothesis about the role of causality. We are only motivating task-level analysis.



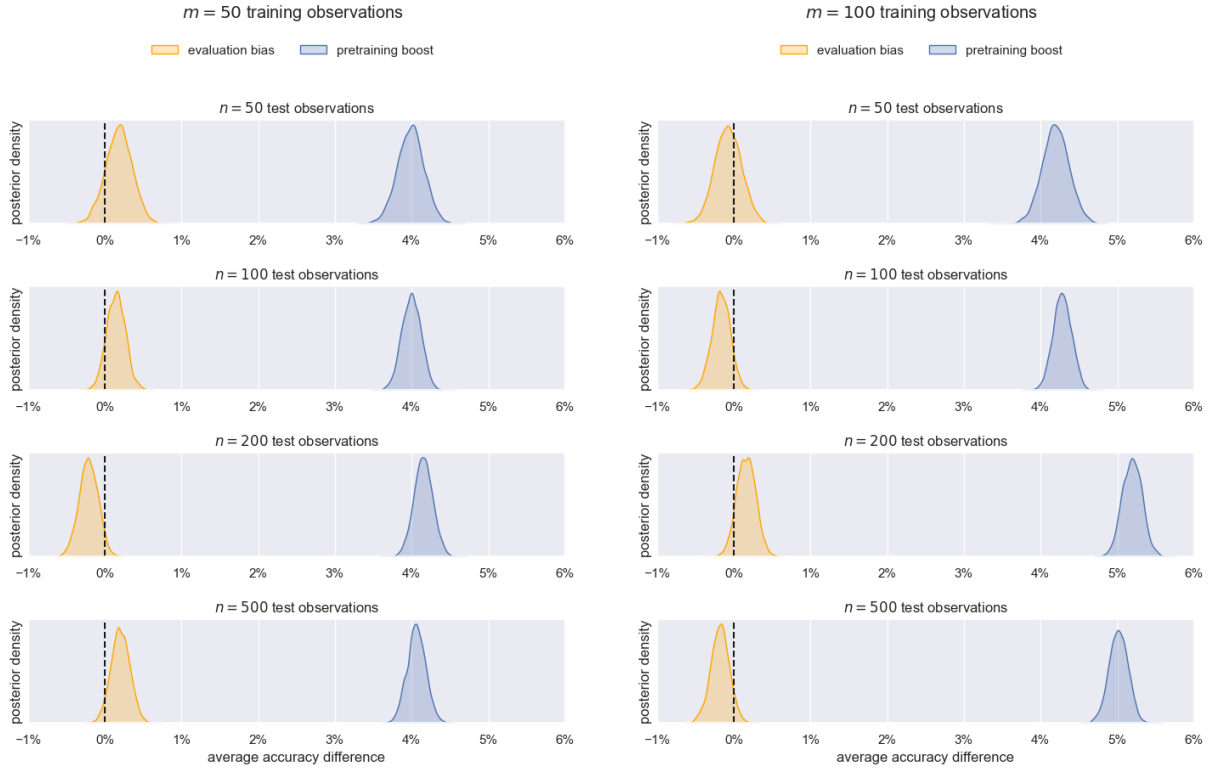


Figure 1: Distributions of average accuracy differences (11). The evaluation bias is akin to  $\text{acc}_{\text{test}} - \text{acc}_{\text{extra}}$ . The pretraining boost is akin to  $\text{acc}_{\text{extra}} - \text{acc}_{\text{base}}$ .

One way to analyze tasks is to sample from the posterior predictive distribution of the accuracy difference, and only average across subsamples:

$$\frac{\bar{Y}_{ij \cdot 1} - \bar{Y}_{ij \cdot 0}}{n}. \quad (12)$$

A more concise way is to perform a hypothesis test for each setting of  $m, n$ , and the LM type:

$$H_0 : E[\text{acc}_{\text{test}} - \text{acc}_{\text{extra}}] = 0 \quad (13)$$

$$H_1 : E[\text{acc}_{\text{test}} - \text{acc}_{\text{extra}}] > 0. \quad (14)$$

The  $p$ -value is estimated via permutation testing. It's then adjusted to control the false discovery rate (Benjamini and Hochberg, 1995).

## 7 Discussion

Figure 1 demonstrates that the average pretraining boost is significant in every configuration of the experiment, ranging from 2% to 6%. This finding replicates that from Gururangan et al. (2020). After averaging across settings for  $m, n$ , and the 2 LM types, only two of the 25 classification tasks had a pretraining boost less than 0, and both were greater

than -1%.<sup>4</sup> Overall, task-adaptive pretraining had the intended effect.

As shown in Figure 1, the evaluation bias bounces inconsistently and insignificantly around 0. After averaging, 12 of the 25 classification tasks had a positive evaluation bias, 13 had a negative evaluation bias, and all tasks had an average evaluation bias less than 1% in absolute value.

To provide a detailed picture of the evaluation bias, and to avoid excessive averaging, we lemon-picked tasks which reported a bias of at least +3% in any configuration of the experiment. All 6 tasks matching this criterion were from experiments with BERT, simply because BERT had greater training variance. If there were a task-dependent evaluation bias, one could expect that the magnitude of the bias is consistent across  $m$  or  $n$  for a task, or there's a consistent pattern with how the bias changes with  $m$  or  $n$  across tasks. Figure 2 does not support either of these hypotheses.

Given the lack of evidence for an evaluation bias in either direction, it's unlikely that a benchmark which releases unlabeled test set text can systematically promote models pretrained on it over equally

<sup>4</sup>The tasks were `blog_authorship_corpus` and `movie_rationales`.

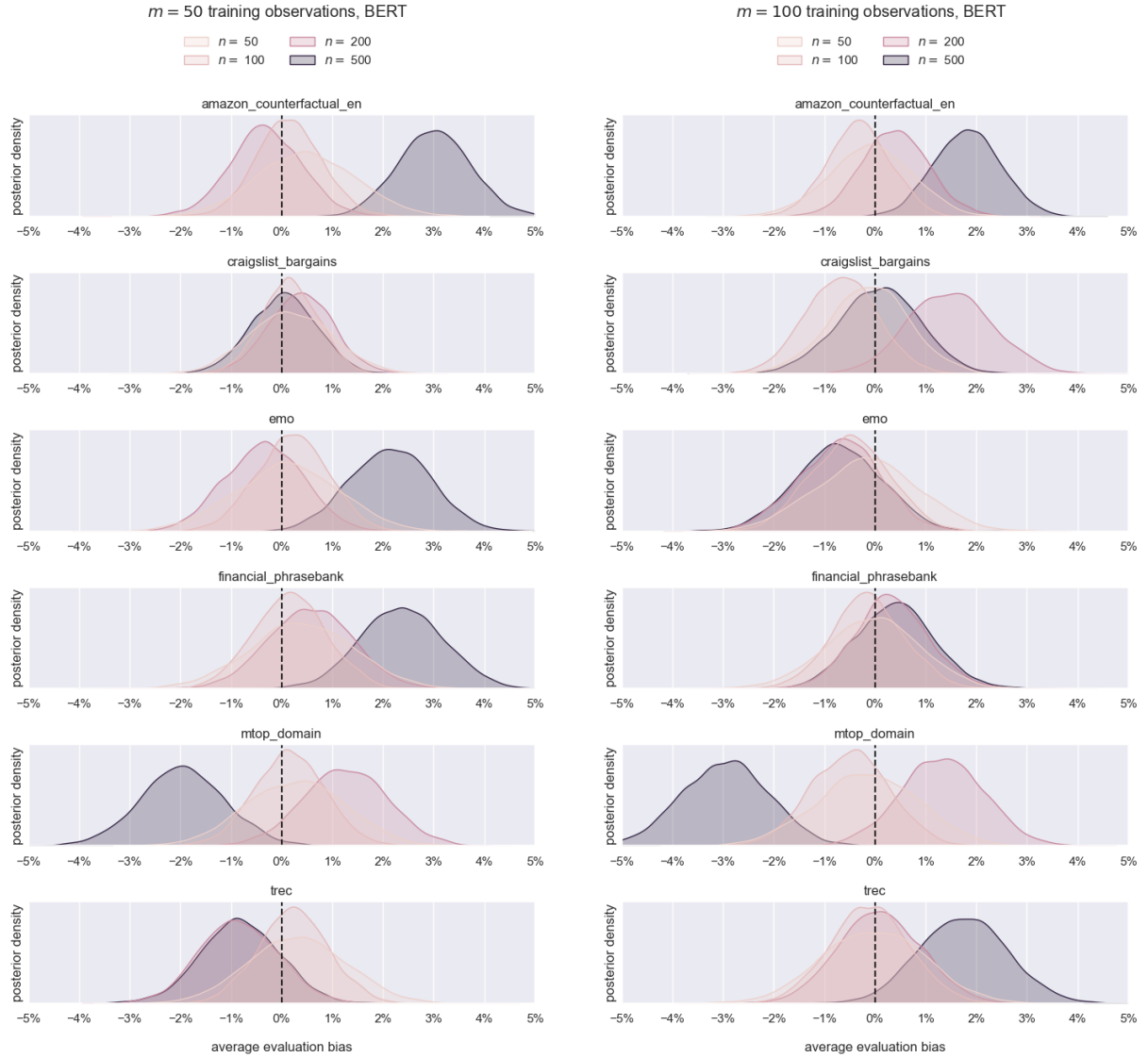


Figure 2: Distributions of average evaluation biases (12) for the subset of tasks which reported an average evaluation bias of at least +3% accuracy in any configuration of the experiment.

performant models which pretrained on unlabeled independent text.

Moscovich and Rosset (2022) found that the evaluation bias caused by unsupervised methods for tabular data converges to 0 as  $n$  increases. This finding is not confirmed by this experiment. Figure 1 shows that within  $m = 50$  and  $m = 100$ , distributions of the evaluation bias consistently hover around 0 across  $n$ . Figure 2 also does not support a relationship between  $n$  and the evaluation bias for lemon-picked tasks. More experiments varying  $n$  are needed to thoroughly assess this insensitivity.

## 8 Overtraining

§7 rules out undertraining on unlabeled text as the cause of a null evaluation bias. What if we over-

train? Overtraining on labeled test data trivially increases test set performance. Perhaps overtraining on unlabeled test set text has a similar effect. To test this hypothesis for text classification, GPT-2 is intentionally overtrained on unlabeled text for 2 epochs instead of 1.

For each of the 25 classification tasks and their subsamples, pretraining for 2 epochs instead of 1 resulted in a lower pretraining loss. The average pretraining loss is 20% lower, and the pretraining boost is negative, which indicates overfitting, as intended. Figure 3 shows that, despite overtraining, the evaluation bias hovers around 0. All  $p$ -values from the test in (13) are greater than 0.5.<sup>5</sup> Over-

<sup>5</sup>For the test that  $E[\text{acc}_{\text{extra}} - \text{acc}_{\text{base}}] < 0$ , 16 of the 25  $p$ -values for  $m = 50, n = 50$  in §8 were less than 0.05,

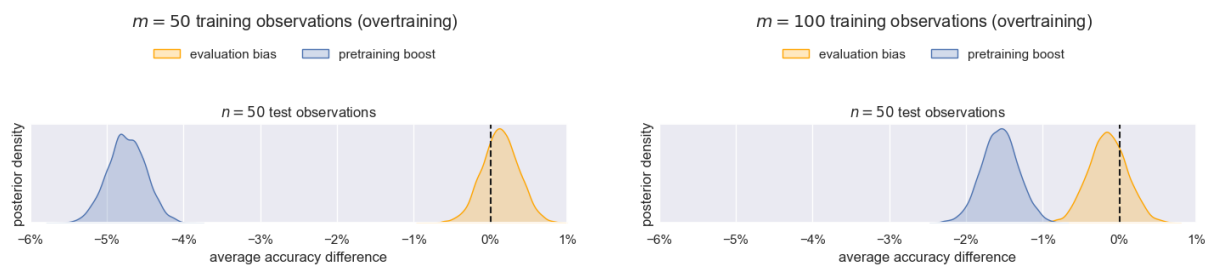


Figure 3: Distributions of average accuracy differences (11) after pretraining GPT-2 for 2 epochs instead of 1 (§8).

training on unlabeled test set text causes test set performance to degrade to the same degree that overfitting on unlabeled independent text does.

## 9 Zero-shot text classification

Prompting an LLM is a popular choice for solving NLP problems. These prompts can be pretrained on. For example, Gemma 2 (Gemma Team, 2024) is intentionally pretrained on prompts from the LM-SYS benchmark (Zheng et al., 2023).

To study a modern prompting approach, the experiment in §4 is repeated with two modifications. First, task-adaptive pretraining is done on prompts—unlabeled texts with instructions for solving the task. Second, classification training is not performed. The further-pretrained LLM is immediately prompted to do the task on test data.

More specifically, pretraining is performed by adding a QLoRA adapter layer (Dettmers et al., 2024) to every linear layer in Mistral 7B<sup>6</sup> (Jiang et al., 2023). Perhaps notably, instructions mention the set of possible answers—the class names. Here is an example of a prompt for the `ag_news` task (Zhang et al., 2015):

```
Your task is to classify a given text as
one of these categories:
World
Sports
Business
Sci/Tech
```

```
The text is a news article. Answer with its topic.
```

```
### Text: Bombardier CEO Quits, Shares
Dive Paul Tellier stepped down on Monday
as president and chief executive of
Bombardier Inc. (BBDsvb.TO: Quote,
Profile, Research) (BBDsb.
### Answer:
```

Figure 4 shows that, while pretraining on prompts improves accuracy, pretraining on test

which demonstrates that the testing procedure may have some statistical power.

<sup>6</sup>Mistral-7B-v0.3, the non-instruction-trained model.

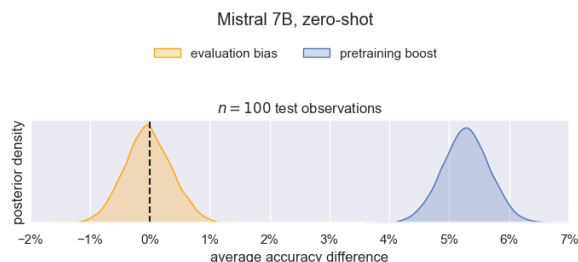


Figure 4: Distributions of average accuracy differences (11) for zero-shot classification (§9). For each of the 25 classification tasks, 20 subsamples were taken.

set prompts does not increase test set accuracy compared to pretraining on independently drawn prompts.

12 of the 25 tasks had a positive evaluation bias and 13 had a negative evaluation bias. All 25  $p$ -values from the hypothesis test in (13) were greater than 0.5. In other words, at the task level, there is still no evidence of evaluation bias.

A limitation of this experiment is that it doesn’t account for data contamination. If Mistral 7B’s pretraining data included labeled or unlabeled parts of the classification datasets used here, the pretraining boost and evaluation bias may be diluted.

## 10 Meta-analysis

§4.4 briefly argues for subsampling multiple datasets from the full classification dataset. To assess this argument, the analysis was repeated on 500 random slices of the  $m = 100, n = 500$  dataset of accuracies such that exactly 1 ( $\text{acc}_{\text{extra}}, \text{acc}_{\text{test}}, \text{acc}_{\text{base}}$ ) triple per classification task (instead of 20 triples) is included. This unreplicated data is often all one gets from benchmarks.

Figure 5 (left) displays the cumulative distribution of the posterior mean of the evaluation bias for  $m = 100, n = 500$  under this unreplicated experimental design. The distribution is quite variant. There’s a 47% chance that the posterior mean of

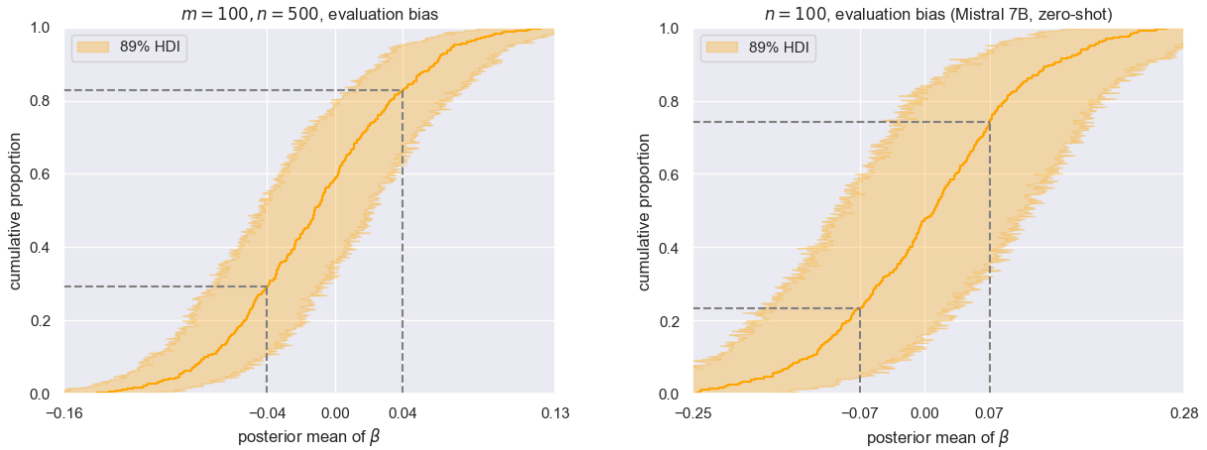


Figure 5: Distributions of conclusions had there been no technical replication (§10).

$\beta$ —the average increase in the log-odds of a correct prediction by pretraining on unlabeled test set text instead of on unlabeled independent text—is outside the interval  $(-0.04, 0.04)$ , which would indicate a significant negative or positive bias.<sup>7</sup> For the zero-shot experiment in §9, there’s a 49% chance that the posterior mean of  $\beta$  is outside  $(-0.07, 0.07)$ . Without repeated subsampling, one may as well flip a coin to decide whether pretraining on unlabeled test set text is fair.

## 11 Conclusion

Across combinations for the number of classification training examples ( $m = 50, 100$ ), the number of pretraining or evaluation examples ( $n = 50, 100, 200, 500$ ), and for zero-shot prompting with an LLM, task-adaptive pretraining on unlabeled test set text—instead of on unlabeled independent text—did not result in a consistent or significant evaluation bias. This appears to be the case when pretraining helps (§7) and when it hurts (§8).

For benchmarks which only release unlabeled text from the test set, this finding doesn’t completely absolve LLM evaluations from scrutiny. The reason is that the boost from pretraining on unlabeled text—which is often significant—could be viewed as a type of evaluation bias, depending on how LLMs generalize. More concretely, suppose there’s a benchmark and two LLMs,  $A$  and  $B$ .  $A$  was *not* pretrained on unlabeled text from the test set of the benchmark, while  $B$  was. With the

perspective that LLM benchmarks supply scores which are correlates of performance on real-world tasks—instead of indicators of performance solely on the benchmark’s tasks—then  $B$  scoring higher on the benchmark than  $A$  may be a misleading signal. If pretraining on the benchmark’s unlabeled text causes  $B$  to generalize better only *within* the distribution of the benchmark, then  $B$ ’s edge on this benchmark does not signal an edge in real-world tasks. This argument implies that knowing whether or not an LLM was pretrained on unlabeled text from the test set is still important.

One recommendation for designing few-shot benchmarks, which expands on the principle about robustness from Bragg et al. (2021), is based on the meta-analysis in §10: empirical studies of few-shot learning should consider including multiple, independent subsamples of training data. While a single training set combined with a large test set is sufficient for precise, unbiased estimation of out-of-sample performance, this estimator is conditional on the training set. In few-shot learning, the training set is, by definition, minimal. The estimator hides two sources of variance—that from the randomly drawn training set, and that from randomness inherent in the training procedure. Figure 5 shows that this variance is large-enough to turn a methodology into a coin flip for two different training procedures. In-context learning with LLMs is also sensitive to the selection of few-shot examples (Lu et al., 2022, Alzahrani et al., 2024). Benchmarks which require training on multiple, independent subsamples would expose training variance.

<sup>7</sup>For 0.04, the odds ratio is  $e^{0.04} \approx 1.04$ . For context, the average odds ratio between adjacent submissions in the RAFT leaderboard is 1.03. For posterior means outside  $(-0.04, 0.04)$ , all of their 89% credible intervals exclude 0, which evidences a non-null effect.



## Limitations

This paper does not analyze semi-supervised methods like Pattern-Exploiting Training, or hand-inspecting the test set text and targeting interventions accordingly. We also don't study what happens when unlabeled test set texts are included as part of the initial pretraining stage of an LLM. This paper's conclusions are limited to task-adaptive pretraining.

It's possible that there are tasks where an evaluation bias exists, and these were not part of the 25 classification tasks we collected.

## Acknowledgements

Currently omitted for anonymity.

## References

Neel Alex, Eli Lifland, Lewis Tunstall, Abhishek Thakur, Pegah Maham, C. Riedel, Emmie Hine, Carolyn Ashurst, Paul Sedille, Alexis Carlier, Michael Noetel, and Andreas Stuhlmüller. 2021. [Raft: A real-world few-shot text classification benchmark](#). In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

Norah Alzahrani, Hisham Abdullah Alyahya, Yazeed Alnumay, Sultan Alrashed, Shaykhah Alsubaie, Yusef Almushaykeh, Faisal Mirza, Nouf Alotaibi, Nora Altwairesh, Areeb Alowisheq, et al. 2024. [When benchmarks are targets: Revealing the sensitivity of large language model leaderboards](#). *arXiv preprint arXiv:2402.01781*.

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.

Jonathan Bragg, Arman Cohan, Kyle Lo, and Iz Beltagy. 2021. [Flex: Unifying evaluation for few-shot nlp](#). *Advances in Neural Information Processing Systems*, 34:15787–15800.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. [Language models are few-shot learners](#). *Advances in neural information processing systems*, 33:1877–1901.

Tomás Capretto, Camen Pihó, Ravin Kumar, Jacob Westfall, Tal Yarkoni, and Osvaldo A Martin. 2022. [Bambi: A simple interface for fitting bayesian linear models in python](#). *Journal of Statistical Software*, 103(15):1–29.

Emile Chapuis, Pierre Colombo, Matteo Manica, Matthieu Labeau, and Chloé Clavel. 2020. [Hierarchical pre-training for sequence labelling in spoken dialog](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2636–2648, Online. Association for Computational Linguistics.

Ankush Chatterjee, Kedhar Nath Narahari, Meghana Joshi, and Puneet Agrawal. 2019. [SemEval-2019 task 3: EmoContext contextual emotion detection in text](#). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 39–48, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. [Qlora: Efficient finetuning of quantized llms](#). *Advances in Neural Information Processing Systems*, 36.

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.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. [ERASER: A benchmark to evaluate rationalized NLP models](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online. Association for Computational Linguistics.

Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2020. [Climate-fever: A dataset for verification of real-world climate claims](#).

Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, Laurie Crist, Misha Britan, Wouter Leeuwis, Gokhan Tur, and Prem Nataraajan. 2023. [MASSIVE: A 1M-example multilingual natural language understanding dataset with 51 typologically-diverse languages](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4277–4302, Toronto, Canada. Association for Computational Linguistics.

Gemma Team. 2024. [Gemma 2: Improving open language models at a practical size](#).

Giovanni Grano, Andrea Di Sorbo, Francesco Mercaldo, Corrado A Visaggio, Gerardo Canfora, and Sebastiano Panichella. 2017. [Android apps and user feedback: a dataset for software evolution and quality improvement](#). In *Proceedings of the 2nd ACM SIGSOFT international workshop on app market analytics*, pages 8–11.

637	Yuxian Gu, Xu Han, Zhiyuan Liu, and Minlie Huang.	Johannes Kiesel, Maria Mestre, Rishabh Shukla, Em-	693
638	2022. <a href="#">PPT: Pre-trained prompt tuning for few-shot</a>	manuel Vincent, Payam Adineh, David Corney,	694
639	<a href="#">learning</a> . In <i>Proceedings of the 60th Annual Meet-</i>	Benno Stein, and Martin Potthast. 2019. <a href="#">SemEval-</a>	695
640	<i>ing of the Association for Computational Linguistics</i>	<a href="#">2019 task 4: Hyperpartisan news detection</a> . In	696
641	(Volume 1: Long Papers), pages 8410–8423, Dublin,	<i>Proceedings of the 13th International Workshop on</i>	697
642	Ireland. Association for Computational Linguistics.	<i>Semantic Evaluation</i> , pages 829–839, Minneapolis,	698
		Minnesota, USA. Association for Computational Lin-	699
		guistics.	700
643	Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré,	Ravin Kumar, Colin Carroll, Ari Hartikainen, and Os-	701
644	Adam Chilton, Alex Chohlas-Wood, Austin Peters,	valdo Martin. 2019. <a href="#">Arviz a unified library for</a>	702
645	Brandon Waldon, Daniel Rockmore, Diego Zam-	<a href="#">exploratory analysis of bayesian models in python</a> .	703
646	brano, et al. 2024. <a href="#">Legalbench: A collaboratively</a>	<i>Journal of Open Source Software</i> , 4(33):1143.	704
647	<a href="#">built benchmark for measuring legal reasoning in</a>		
648	<a href="#">large language models</a> . <i>Advances in Neural Informa-</i>		
649	<i>tion Processing Systems</i> , 36.		
650	Suchin Gururangan, Ana Marasović, Swabha	Nikola Ljubešić, Darja Fišer, and Tomaž Erjavec. 2019.	705
651	Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey,	<a href="#">The frenk datasets of socially unacceptable discourse</a>	706
652	and Noah A. Smith. 2020. <a href="#">Don’t stop pretraining:</a>	<a href="#">in slovene and english</a> .	707
653	<a href="#">Adapt language models to domains and tasks</a> . In		
654	<i>Proceedings of the 58th Annual Meeting of the</i>	Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel,	708
655	<i>Association for Computational Linguistics</i> , pages	and Pontus Stenetorp. 2022. <a href="#">Fantastically ordered</a>	709
656	8342–8360, Online. Association for Computational	<a href="#">prompts and where to find them: Overcoming few-</a>	710
657	Linguistics.	<a href="#">shot prompt order sensitivity</a> . In <i>Proceedings of the</i>	711
		<i>60th Annual Meeting of the Association for Compu-</i>	712
658	Trevor Hastie, Robert Tibshirani, Jerome H Friedman,	<i>tational Linguistics (Volume 1: Long Papers)</i> , pages	713
659	and Jerome H Friedman. 2009. <a href="#">The elements of statis-</a>	8086–8098, Dublin, Ireland. Association for Compu-	714
660	<a href="#">tical learning: data mining, inference, and prediction</a> ,	tational Linguistics.	715
661	volume 2. Springer.		
662	He He, Derek Chen, Anusha Balakrishnan, and Percy	P. Malo, A. Sinha, P. Korhonen, J. Wallenius, and	716
663	Liang. 2018. <a href="#">Decoupling strategy and generation in</a>	P. Takala. 2014. <a href="#">Good debt or bad debt: Detecting se-</a>	717
664	<a href="#">negotiation dialogues</a> . In <i>Proceedings of the 2018</i>	<a href="#">mantic orientations in economic texts</a> . <i>Journal of the</i>	718
665	<i>Conference on Empirical Methods in Natural Lan-</i>	<i>Association for Information Science and Technology</i> ,	719
666	<i>guage Processing</i> , pages 2333–2343, Brussels, Bel-	65.	720
667	gium. Association for Computational Linguistics.		
668	Zhang Huangzhao. 2018. Yahoo-	Irene Manotas, Ngoc Phuoc An Vo, and Vadim Sheinin.	721
669	answers-topic-classification-dataset.	2020. <a href="#">LiMiT: The literal motion in text dataset</a> . In	722
670	<a href="https://github.com/LC-John/">https://github.com/LC-John/</a>	<i>Findings of the Association for Computational Lin-</i>	723
671	<a href="#">Yahoo-Answers-Topic-Classification-Dataset</a> .	<i>guistics: EMNLP 2020</i> , pages 991–1000, Online.	724
		Association for Computational Linguistics.	725
672	Alon Jacovi, Avi Caciularu, Omer Goldman, and Yoav	Richard McElreath. 2018. <a href="#">Statistical rethinking: A</a>	726
673	Goldberg. 2023. <a href="#">Stop uploading test data in plain</a>	<a href="#">Bayesian course with examples in R and Stan</a> . Chap-	727
674	<a href="#">text: Practical strategies for mitigating data contami-</a>	man and Hall/CRC.	728
675	<a href="#">nation by evaluation benchmarks</a> . In <i>Proceedings of</i>		
676	<i>the 2023 Conference on Empirical Methods in Natu-</i>	Vangelis Metsis, Ion Androutsopoulos, and Georgios	729
677	<i>ral Language Processing</i> , pages 5075–5084, Singa-	Paliouras. 2006. <a href="#">Spam filtering with naive bayes-</a>	730
678	apore. Association for Computational Linguistics.	<a href="#">which naive bayes?</a> In <i>CEAS</i> , volume 17, pages	731
		28–69. Mountain View, CA.	732
679	Albert Q Jiang, Alexandre Sablayrolles, Arthur Men-	Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Cor-	733
680	sch, Chris Bamford, Devendra Singh Chaplot, Diego	rado, and Jeff Dean. 2013. <a href="#">Distributed representa-</a>	734
681	de las Casas, Florian Bressand, Gianna Lengyel, Guil-	<a href="#">tions of words and phrases and their compositionality</a> .	735
682	laume Lample, Lucile Saulnier, et al. 2023. <a href="#">Mistral</a>	<i>Advances in neural information processing systems</i> ,	736
683	<a href="#">7b</a> . <i>arXiv preprint arXiv:2310.06825</i> .	26.	737
684	Zhijing Jin, Julius von Kügelgen, Jingwei Ni, Tejas	Richard D Morey, Rink Hoekstra, Jeffrey N Rouder,	738
685	Vaidhya, Ayush Kaushal, Mrinmaya Sachan, and	Michael D Lee, and Eric-Jan Wagenmakers. 2016.	739
686	Bernhard Schoelkopf. 2021. <a href="#">Causal direction of data</a>	<a href="#">The fallacy of placing confidence in confidence inter-</a>	740
687	<a href="#">collection matters: Implications of causal and an-</a>	<a href="#">vals</a> . <i>Psychonomic bulletin &amp; review</i> , 23:103–123.	741
688	<a href="#">ticausal learning for NLP</a> . In <i>Proceedings of the</i>		
689	<i>2021 Conference on Empirical Methods in Natural</i>	Amit Moscovich and Saharon Rosset. 2022. <a href="#">On the</a>	742
690	<i>Language Processing</i> , pages 9499–9513, Online and	<a href="#">cross-validation bias due to unsupervised preprocess-</a>	743
691	Punta Cana, Dominican Republic. Association for	<a href="#">ing</a> . <i>Journal of the Royal Statistical Society Series B:</i>	744
692	Computational Linguistics.	<i>Statistical Methodology</i> , 84(4):1474–1502.	745

746	Niklas Muennighoff, Nouamane Tazi, Loic Magne, and	<i>the 57th Annual Meeting of the Association for Com-</i>	803
747	Nils Reimers. 2023. <a href="#">MTEB: Massive text embedding</a>	<i>putational Linguistics: Student Research Workshop,</i>	804
748	<a href="#">benchmark</a> . In <i>Proceedings of the 17th Conference</i>	pages 407–414, Florence, Italy. Association for Com-	805
749	<i>of the European Chapter of the Association for Com-</i>	putational Linguistics.	806
750	<i>putational Linguistics</i> , pages 2014–2037, Dubrovnik,		
751	Croatia. Association for Computational Linguistics.		
752	James O’Neill, Polina Rozenshtein, Ryuichi Kiryo, Mo-	Lewis Tunstall, Nils Reimers, Unso Eun Seo Jo, Luke	807
753	toko Kubota, and Danushka Bollegala. 2021. <a href="#">I wish</a>	Bates, Daniel Korat, Moshe Wasserblat, and Oren	808
754	<a href="#">I would have loved this one, but I didn’t – a multilin-</a>	Pereg. 2022. <a href="#">Efficient few-shot learning without</a>	809
755	<a href="#">gual dataset for counterfactual detection in product</a>	<a href="#">prompts</a> . <i>arXiv preprint arXiv:2209.11055</i> .	810
756	<a href="#">review</a> . In <i>Proceedings of the 2021 Conference on</i>	Mengqiu Wang, Noah A. Smith, and Teruko Mita-	811
757	<i>Empirical Methods in Natural Language Processing</i> ,	mura. 2007. <a href="#">What is the Jeopardy model? a quasi-</a>	812
758	pages 7092–7108, Online and Punta Cana, Domini-	<a href="#">synchronous grammar for QA</a> . In <i>Proceedings of the</i>	813
759	can Republic. Association for Computational Lin-	<i>2007 Joint Conference on Empirical Methods in Natu-</i>	814
760	guistics.	<i>ral Language Processing and Computational Natural</i>	815
761	Bo Pang and Lillian Lee. 2005. <a href="#">Seeing stars: Exploit-</a>	<i>Language Learning (EMNLP-CoNLL)</i> , pages 22–32,	816
762	<a href="#">ing class relationships for sentiment categorization</a>	Prague, Czech Republic. Association for Computa-	817
763	<a href="#">with respect to rating scales</a> . In <i>Proceedings of the</i>	tional Linguistics.	818
764	<i>43rd Annual Meeting of the Association for Compu-</i>	Thomas Wolf, Lysandre Debut, Victor Sanh, Julien	819
765	<i>tational Linguistics (ACL’05)</i> , pages 115–124, Ann	Chaumond, Clement Delangue, Anthony Moi, Pier-	820
766	Arbor, Michigan. Association for Computational Lin-	ric Cistac, Tim Rault, Remi Louf, Morgan Funtow-	821
767	guistics.	icz, Joe Davison, Sam Shleifer, Patrick von Platen,	822
768	Alec Radford, Jeffrey Wu, Rewon Child, David Luan,	Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu,	823
769	Dario Amodei, Ilya Sutskever, et al. 2019. <a href="#">Language</a>	Teven Le Scao, Sylvain Gugger, Mariama Drame,	824
770	<a href="#">models are unsupervised multitask learners</a> . <i>OpenAI</i>	Quentin Lhoest, and Alexander Rush. 2020. <a href="#">Trans-</a>	825
771	<i>blog</i> , 1(8):9.	<a href="#">formers: State-of-the-art natural language processing</a> .	826
772	Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang,	In <i>Proceedings of the 2020 Conference on Empirical</i>	827
773	Junlin Wu, and Yi-Shin Chen. 2018. <a href="#">CARER: Con-</a>	<i>Methods in Natural Language Processing: System</i>	828
774	<a href="#">textualized affect representations for emotion recog-</a>	<i>Demonstrations</i> , pages 38–45, Online. Association	829
775	<a href="#">nition</a> . In <i>Proceedings of the 2018 Conference on</i>	for Computational Linguistics.	830
776	<i>Empirical Methods in Natural Language Processing</i> ,	Tianyi Zhang, Felix Wu, Arzo Katiyar, Kilian Q Wein-	831
777	pages 3687–3697, Brussels, Belgium. Association	berger, and Yoav Artzi. 2021. <a href="#">Revisiting few-sample</a>	832
778	for Computational Linguistics.	<a href="#">bert fine-tuning</a> . In <i>International Conference on</i>	833
779	Timo Schick and Hinrich Schütze. 2021. <a href="#">Exploiting</a>	<i>Learning Representations</i> .	834
780	<a href="#">cloze-questions for few-shot text classification and</a>	Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015.	835
781	<a href="#">natural language inference</a> . In <i>Proceedings of the</i>	<a href="#">Character-level convolutional networks for text classi-</a>	836
782	<i>16th Conference of the European Chapter of the Asso-</i>	<a href="#">fication</a> . <i>Advances in neural information processing</i>	837
783	<i>ciation for Computational Linguistics: Main Volume</i> ,	<i>systems</i> , 28.	838
784	pages 255–269, Online. Association for Computa-	Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle	839
785	tional Linguistics.	Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang,	840
786	Jonathan Schler, Moshe Koppel, Shlomo Argamon, and	Zhuohan Li, Zi Lin, Eric Xing, et al. 2023. <a href="#">Lmsys-</a>	841
787	James W Pennebaker. 2006. <a href="#">Effects of age and gen-</a>	<a href="#">chat-1m: A large-scale real-world llm conversation</a>	842
788	<a href="#">der on blogging</a> . In <i>AAAI spring symposium: Compu-</i>	<a href="#">dataset</a> . <i>arXiv preprint arXiv:2309.11998</i> .	843
789	<i>tational approaches to analyzing weblogs</i> , volume 6,	<b>A Classification tasks</b>	844
790	pages 199–205.	Inclusion criteria:	845
791	Eva Sharma, Chen Li, and Lu Wang. 2019. <a href="#">BIG-</a>	1. All text is in English.	846
792	<a href="#">PATENT: A large-scale dataset for abstractive and</a>	2. The number of classes is not greater than 25.	847
793	<a href="#">coherent summarization</a> . In <i>Proceedings of the 57th</i>	3. The task is to classify one text, not a pair as	848
794	<i>Annual Meeting of the Association for Computational</i>	in, e.g., textual entailment tasks.	849
795	<i>Linguistics</i> , pages 2204–2213, Florence, Italy. Asso-	4. Texts aren’t so long that too much useful sig-	850
796	ciation for Computational Linguistics.	nal is dropped when text is truncated to fit in	851
797	Roshan Sharma. 2019. Twitter-sentiment-	BERT/GPT-2’s context window, which is set	852
798	analysis. <a href="https://github.com/sharmaroshan/">https://github.com/sharmaroshan/</a>	to 256 tokens.	853
799	<a href="#">Twitter-Sentiment-Analysis</a> .	5. Based on our best judgment, it’s likely that	854
800	Tan Thongtan and Tanasanee Phientrakul. 2019. <a href="#">Sen-</a>	BERT/GPT-2 can do better than guessing.	855
801	<a href="#">timent classification using document embeddings</a>	Table 2 lists the exact tasks.	856
802	<a href="#">trained with cosine similarity</a> . In <i>Proceedings of</i>		

<b>Hugging Face dataset</b>	<b>Author(s)</b>	<b>Number of classes</b>	<b>Text length (25, 75) percentiles</b>
<a href="#">ag_news</a>	<a href="#">Zhang et al. (2015)</a>	4	(196, 266)
<a href="#">SetFit/amazon_counterfactual_en</a>	<a href="#">O’Neill et al. (2021)</a>	2	(60, 125)
<a href="#">app_reviews</a>	<a href="#">Grano et al. (2017)</a>	5	(10, 77)
<a href="#">blog_authorship_corpus</a>	<a href="#">Schler et al. (2006)</a>	2	(92, 556)
<a href="#">christinacdl/clickbait_notclickbait_dataset</a>		2	(46, 69)
<a href="#">climate_fever</a>	<a href="#">Diggelmann et al. (2020)</a>	4	(80, 156)
<a href="#">aladar/craigslist_bargains</a>	<a href="#">He et al. (2018)</a>	6	(346, 713)
<a href="#">disaster_response_messages</a>		3	(74, 178)
<a href="#">emo</a>	<a href="#">Chatterjee et al. (2019)</a>	4	(44, 83)
<a href="#">dair-ai/emotion</a>	<a href="#">Saravia et al. (2018)</a>	6	(53, 129)
<a href="#">SetFit/enron_spam</a>	<a href="#">Metsis et al. (2006)</a>	2	(342, 1553)
<a href="#">financial_phrasebank</a>	<a href="#">Malo et al. (2014)</a>	3	(79, 157)
<a href="#">classla/FRENK-hate-en</a>	<a href="#">Ljubešić et al. (2019)</a>	2	(34, 160)
<a href="#">hyperpartisan_news_detection</a>	<a href="#">Kiesel et al. (2019)</a>	2	(39, 63)
<a href="#">limit</a>	<a href="#">Manotas et al. (2020)</a>	2	(53, 123)
<a href="#">AmazonScience/massive</a>	<a href="#">FitzGerald et al. (2023)</a>	18	(24, 44)
<a href="#">movie_rationales</a>	<a href="#">DeYoung et al. (2020)</a>	2	(2721, 4659)
<a href="#">mteb/mtop_domain</a>	<a href="#">Muennighoff et al. (2023)</a>	11	(26, 44)
<a href="#">ccdvp/patent-classification</a>	<a href="#">Sharma et al. (2019)</a>	9	(441, 775)
<a href="#">rotten_tomatoes</a>	<a href="#">Pang and Lee (2005)</a>	2	(76, 149)
<a href="#">silicone</a>	<a href="#">Chapuis et al. (2020)</a>	4	(29, 75)
<a href="#">trec</a>	<a href="#">Wang et al. (2007)</a>	6	(36, 61)
<a href="#">tweets_hate_speech_detection</a>	<a href="#">Sharma (2019)</a>	2	(62, 107)
<a href="#">yahoo_answers_topics</a>	<a href="#">Huangzhao (2018)</a>	10	(58, 213)
<a href="#">yelp_review_full</a>	<a href="#">Zhang et al. (2015)</a>	5	(287, 957)

Table 2: Brief descriptions of the 25 classification tasks used in this experiment. Click the link in the cell to be taken to the dataset homepage in <https://huggingface.co/datasets>. The dataset subset (or config) and the chosen prediction task are specified in code in `src/pretrain_on_test/data.py`.



## B Other experiment choices

This section expands on §4.

First, we clarify how classification training is performed. For BERT, the linear layer transforms the [CLS] token embedding. For GPT-2, the linear layer transforms the last token’s embedding. The output dimension of the linear layer is the number of classes in the classification task. This layer, along with the rest of the weights in the LM, are finetuned to minimize classification cross entropy loss on train.

The BERT model used here is `bert-base-uncased`. The GPT-2 model used here is `gpt2 (small)`, with 124M parameters.

`train` is stratify-sampled by the class to ensure every class is represented, and to reduce the variance of accuracy estimators. `test` is not stratify-sampled. We’re only interested in the *difference* between accuracies, which is a function of the difference between model likelihoods because the priors are uniform. So even if accuracies are worse than the majority vote, differences are still meaningful for the purposes of this experiment.

`train` text is not included during pretraining to eliminate the overlap of pretraining data between  $\text{acc}_{\text{extra}}$  and  $\text{acc}_{\text{test}}$ . This choice was made in an effort to widen any gap between them. The experiment tries to go out of its way to provide evidence of a bias.

`train` contains  $m = 50$  or  $m = 100$  observations.  $m = 50$  is inspired by the RAFT benchmark.  $m = 100$  stretches the intention of “few” in few-shot learning, but was tested in an attempt to make lower-variance comparisons. BERT is quite sensitive—see Appendix D.2.

The discussion focuses on the BERT and GPT-2 results because their pretraining data is (likely) not already contaminated with text from the 25 text classification tasks. While modern NLP solutions usually involve LLMs, these models’ pretraining data are opaque and more likely to include text from the 25 classification tasks (for example, from web-crawling the Dataset Viewer in HuggingFace’s datasets web pages, which hosts the experiment’s data) (Jacovi et al., 2023). As a result, the comparisons— $\text{acc}_{\text{extra}}$  versus  $\text{acc}_{\text{base}}$  and  $\text{acc}_{\text{test}}$  versus  $\text{acc}_{\text{extra}}$ —would be less valid.

## C Hyperparameters and reproducibility

This paper’s experiment and analysis code, and data, is available here: <https://github.com>.

`experiment.sh` lists hyperparameters used for each classification task and experiment configuration. Hyperparameters were pre-specified based on Zhang et al. (2021), with batch sizes set to avoid out-of-memory errors. Run the script on a GPU with at least 15 GB RAM to reproduce results in §5. It takes about 5 days on a T4 GPU. Training is performed using the transformers package (Wolf et al., 2020).

## D Results

### D.1 Individual analysis

The notebook `analysis/dataset.ipynb` can be run to (1) produce visualizations of the distributions of  $\text{acc}_{\text{extra}}$ ,  $\text{acc}_{\text{test}}$ , and  $\text{acc}_{\text{base}}$  (for each classification task and experiment configuration), and (2) compute  $p$ -values for the hypothesis test specified in (13). For all settings of  $m$  and  $n$ , no  $p$ -values were statistically significant at the 0.05 level.

Care has to be taken when attempting to analyze or interpret  $\text{acc}_{\text{extra}} - \text{acc}_{\text{base}}$  and  $\text{acc}_{\text{test}} - \text{acc}_{\text{extra}}$  together. That’s because these differences are not independent: if  $\text{acc}_{\text{extra}}$  is high, then  $\text{acc}_{\text{extra}} - \text{acc}_{\text{base}}$  increases and  $\text{acc}_{\text{test}} - \text{acc}_{\text{extra}}$  decreases. This paper does not analyze the scores together, per se. We care about  $\text{acc}_{\text{test}} - \text{acc}_{\text{extra}}$ .  $\text{acc}_{\text{extra}} - \text{acc}_{\text{base}}$  only exists to sanity check that the pretraining code works; there may be an effect to detect.

### D.2 Difference distributions

Figures 12 - 19 visualize the distributions of the paired differences— $\text{acc}_{\text{extra}} - \text{acc}_{\text{base}}$  and  $\text{acc}_{\text{test}} - \text{acc}_{\text{extra}}$ —for each configuration of the experiment.

## E Analysis

The analysis in §6 can be reproduced by running all of the notebooks in `analysis/fit_posteriors/`. Figure 1 can be reproduced by running the notebook `analysis/results/posterior_pred.ipynb`. Figure 2 can be reproduced by running the notebook `analysis/results/posterior_pred_conditional.ipynb`. Changing the threshold for the bias to +2% accuracy instead of +3% did not change conclusions.

Posterior samples of  $\beta$  (which were used to draw posterior predictive samples) were taken from four chains with 1,000 draws each, after 500 steps of tuning.



## E.1 Hierarchical model checks

Hierarchical models require some basic checks to have faith in their results (McElreath, 2018).

For each of the 22 hierarchical models (16 in §6, 4 in §8, and 2 in §9), no divergences were observed during the fitting procedure. All trace plots were healthy.

Figure 10 contains prior predictive distributions for  $m = 100, n = 200$ , demonstrating that priors are not unreasonable. Using default priors from the `bambi` package (Capretto et al., 2022), while scientifically unreasonable (because they result in wide, basin-like accuracy distributions), did not change the conclusions of this paper.

Figure 11 contains posterior distributions of  $\beta$  for  $m = 100, n = 200$ , demonstrating the hierarchical model’s ability to recover both null and non-null effects. This test can be reproduced by running the notebook `analysis/test.ipynb`.

Figure 8 checks that posterior predictions for the average task accuracies are calibrated. Figure 9 demonstrates the importance of including the  $W_{jl}$  term. These figures can be reproduced by running the notebook `analysis/results/posterior_pred_conditional.ipynb`.

## F Meta-analysis

The meta-analysis in §10 can be reproduced by running the script, `analysis/meta/meta.py`, and then the notebook `analysis/meta/meta.ipynb`. No divergences were observed.

Another question is whether the subsample causes a consistent evaluation bias. §10 establishes that picking a single subsample causes the comparison between  $\text{acc}_{\text{test}}$  and  $\text{acc}_{\text{extra}}$  to be a coin flip. But is the result of the coin flip explained by the specific subsample? If so, comparing models on a single subsample is not as problematic, because the effect of pretraining on unlabeled test set text would be consistent across models.

One way to answer this question is to measure the correlation between the evaluation bias of BERT and GPT-2 for each setting of  $m$  and  $n$ , and each of the 25 tasks. A positive correlation suggests that the subsample causes the evaluation bias.

The observed distribution of correlations across  $m, n$ , and the tasks is plotted in Figure 6 (a). For context, 10 distributions of randomly permuted pairs of subsample-level biases are plotted in Figure 6 (b) and (c). By design, these correlations

are theoretically 0, and are positive or negative by chance. The observed distributions are qualitatively indistinguishable from the null distributions. Notably, the variance is consistent with the null distributions. A deeper dive into the correlations did not reveal any consistently positive correlations at the task level. This result further evidences the importance of repeated subsampling. Taking a single subsample does not result in a consistent pretraining boost or evaluation bias between BERT and GPT-2.

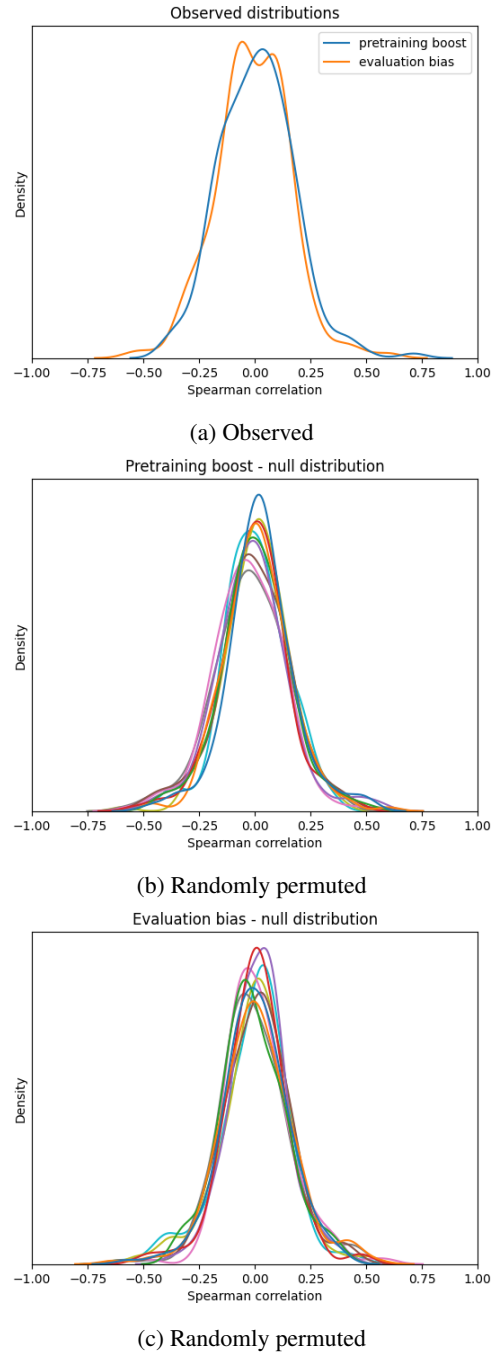


Figure 6: Distribution of correlation between BERT and GPT-2 across all  $m, n$ , and the 25 classification tasks.

## G Zero-shot text classification

The zero-shot experiment files are in [cloud\\_scripts/gcp/experiments/zero\\_shot/](#). Batch sizes are set to run on a GPU with at least 20 GB RAM. The GPU must support the data types needed for QLoRA, e.g., an L4 GPU. Figure 4 can be reproduced by running the notebook in [analysis/fit\\_posteriors/zero\\_shot](#) and then the notebook, [analysis/results/posterior\\_pred.ipynb](#).

We only study  $n = 100$  in an initial effort to provide evidence of an evaluation bias (due to the relatively small test set), and take 20 repeated subsamples instead of 50. While  $n = 100$  is quite small, benchmarks such as LegalBench (Guha et al., 2024) have test data in this range. And the analysis transparently exposes variance.

QLoRA hyperparameters were pre-specified: every adapter has rank 16 with  $\alpha = 32$  (LoRA scaling factor), a 0.05 dropout rate, and no bias parameters. The adapter layers introduce 41,943,040 new, trainable parameters to Mistral 7B, whose parameters are frozen.

## H Experiment with PCA

This is a tangential experiment which served as part of the inspiration of the paper.

This experiment’s design is akin to §4 except that the unsupervised pretraining procedure is PCA, the supervised training procedure is linear regression on synthetic data, and 2,000 subsamples are taken instead of 20-100. 20 features are generated with known effective ranks. The dependent variable is a linear transformation of these features plus independent, normally distributed noise with a standard deviation of 1. Performance is measured using  $R^2$ .

Figure 7 shows that the evaluation bias increases with the effective rank of the features. In other words, fitting a PCA on test set features artificially improves test set performance particularly when fitting a PCA on independent features is unhelpful.

Figure 7 can be reproduced by running the notebook [analysis/pca.ipynb](#)



Figure 7: The shaded regions are 95% confidence intervals.

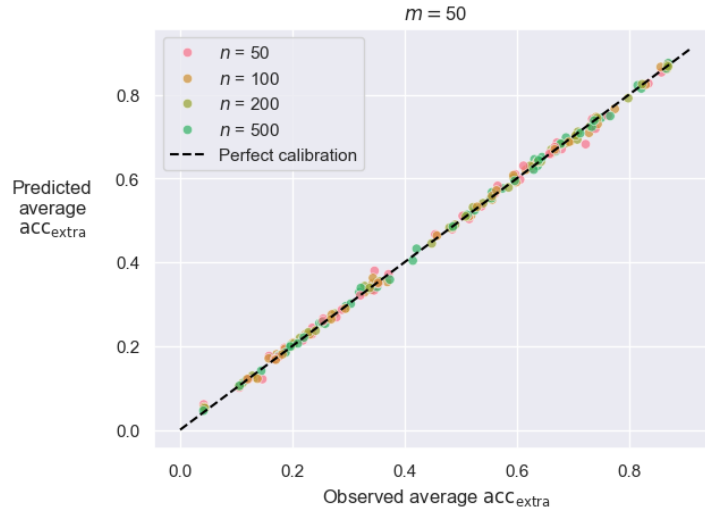


Figure 8: Each of the points represents a task and an LM type.

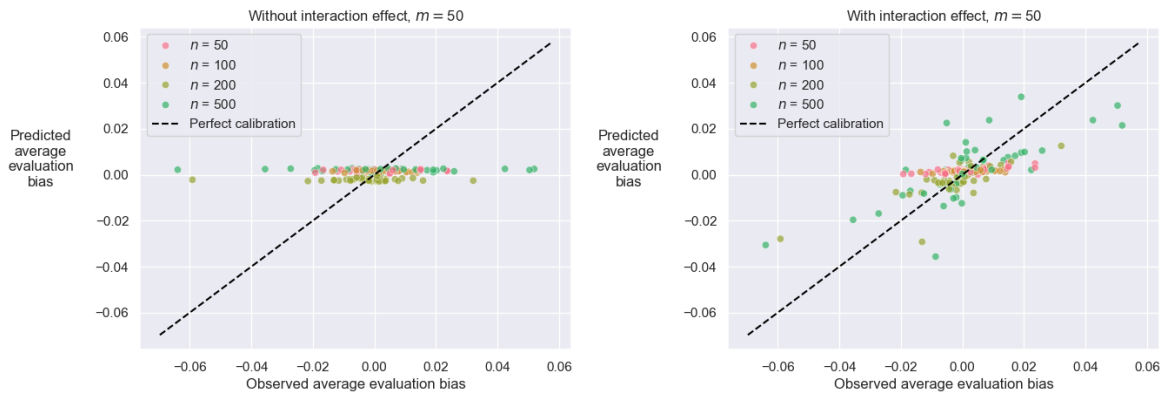


Figure 9: Omitting the interaction effect causes underfitting. Note that the prior causes effects to shrink towards 0.

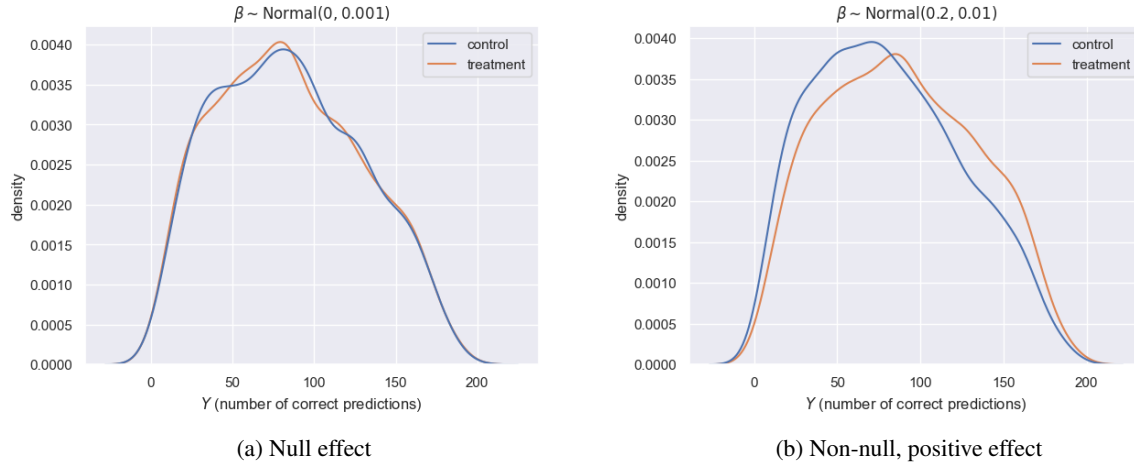


Figure 10: Prior predictive distributions for  $m = 100, n = 200$  from two different priors for  $\beta$ —the expected increase in the log-odds of a correct prediction resulting from an intervention/treatment.

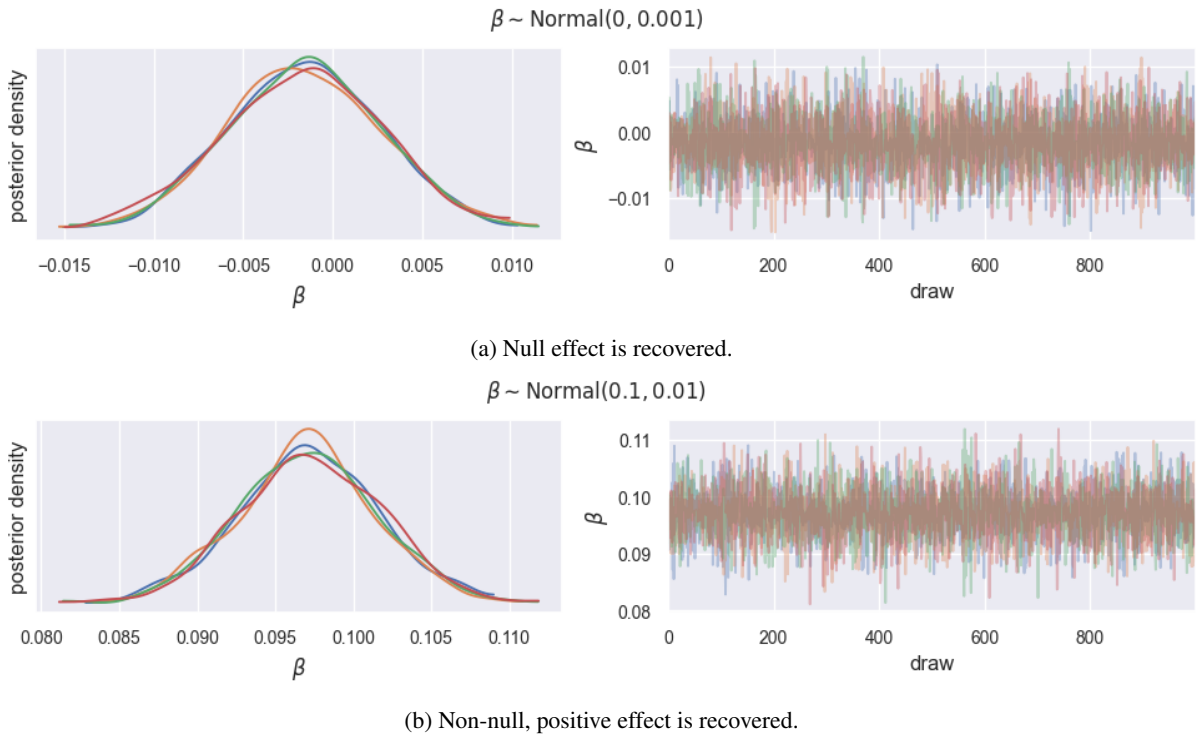


Figure 11: Posterior distributions and trace plots for null and non-null effects **from simulated data** where  $m = 100, n = 200$ , approximated by four chains with 1,000 draws each, after 500 steps of tuning. For each model, no divergences were observed during the fitting procedure. Visualizations were produced by the arviz package (Kumar et al., 2019).

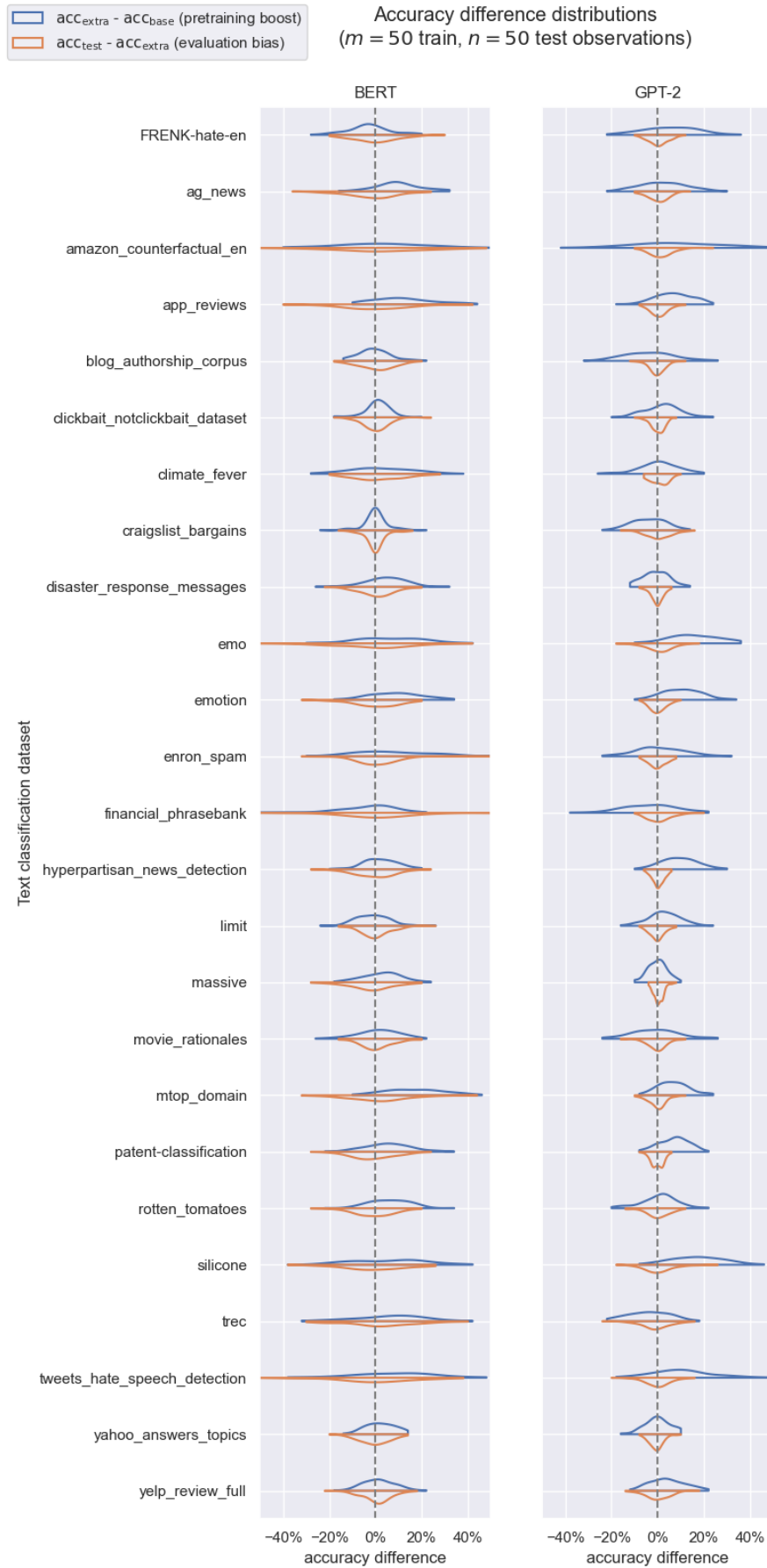


Figure 12



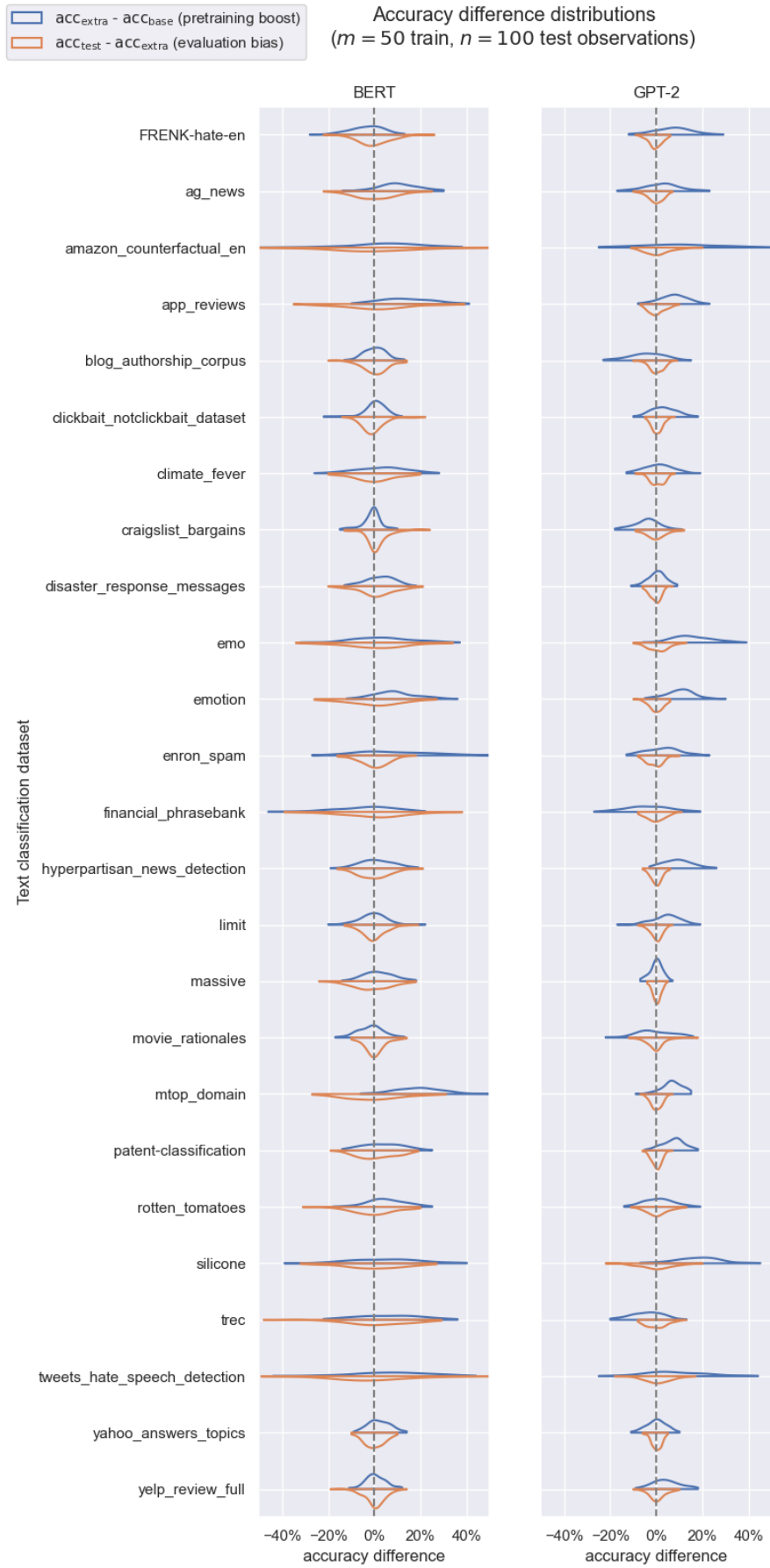


Figure 13

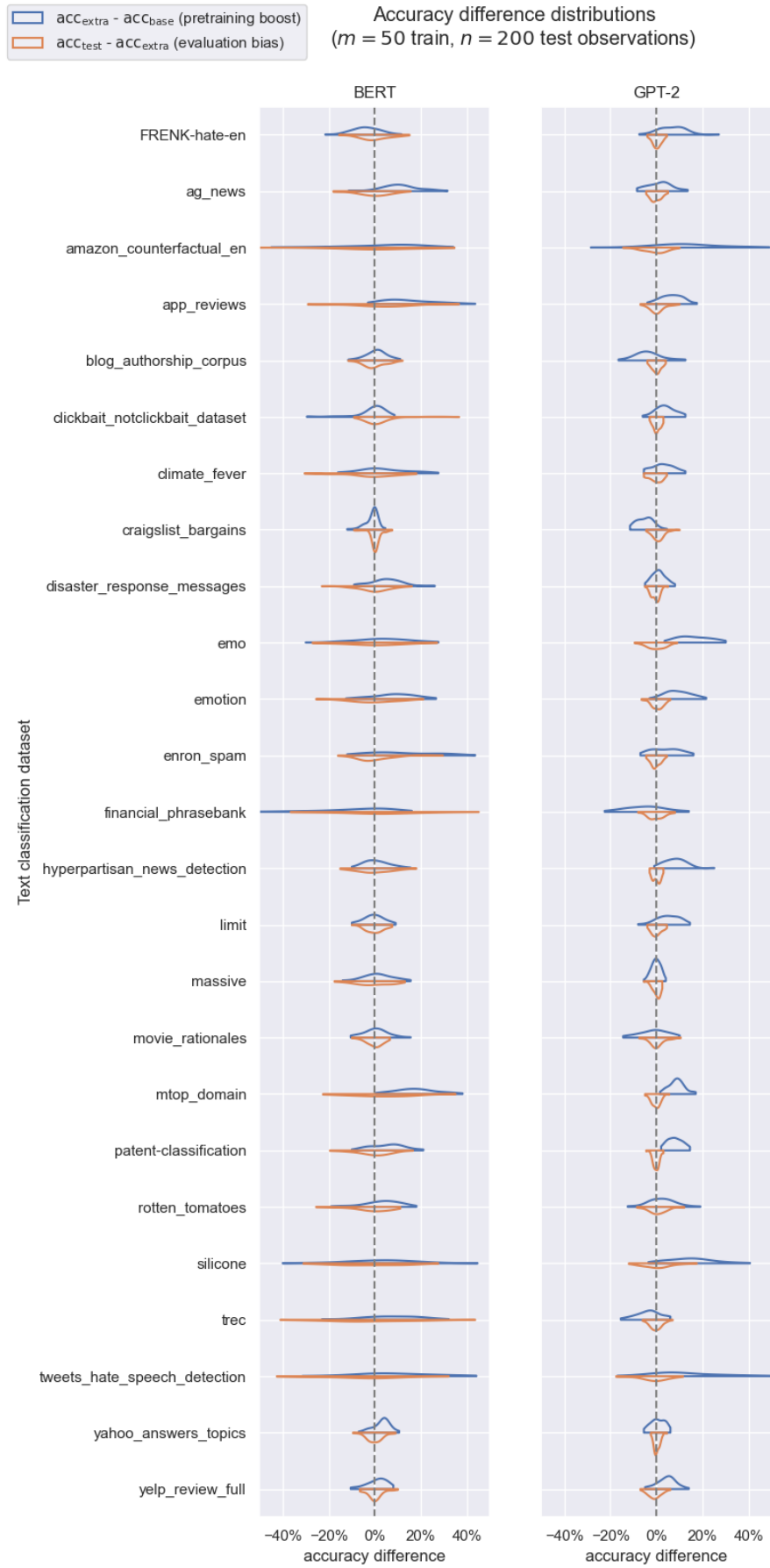


Figure 14

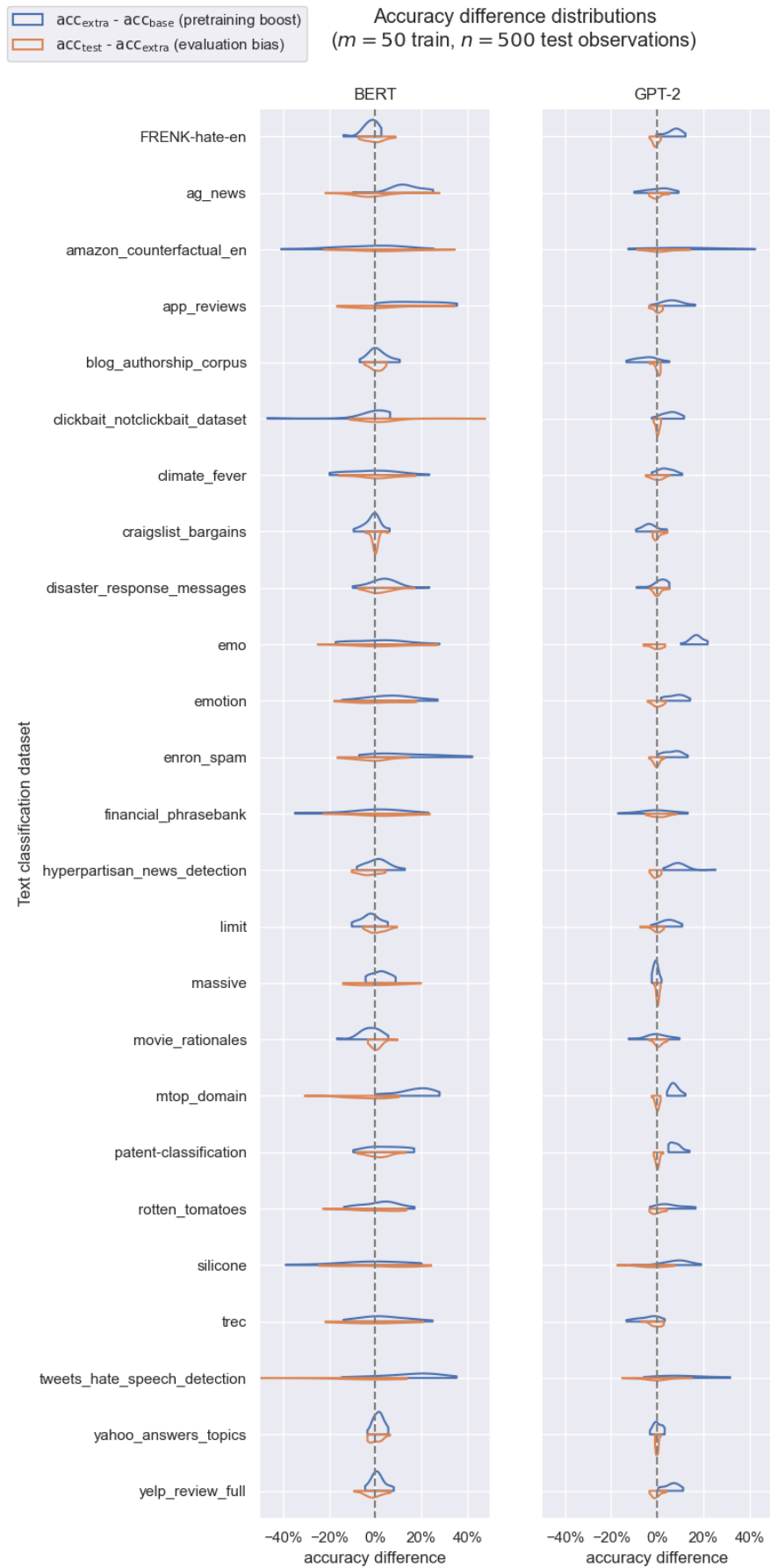


Figure 15

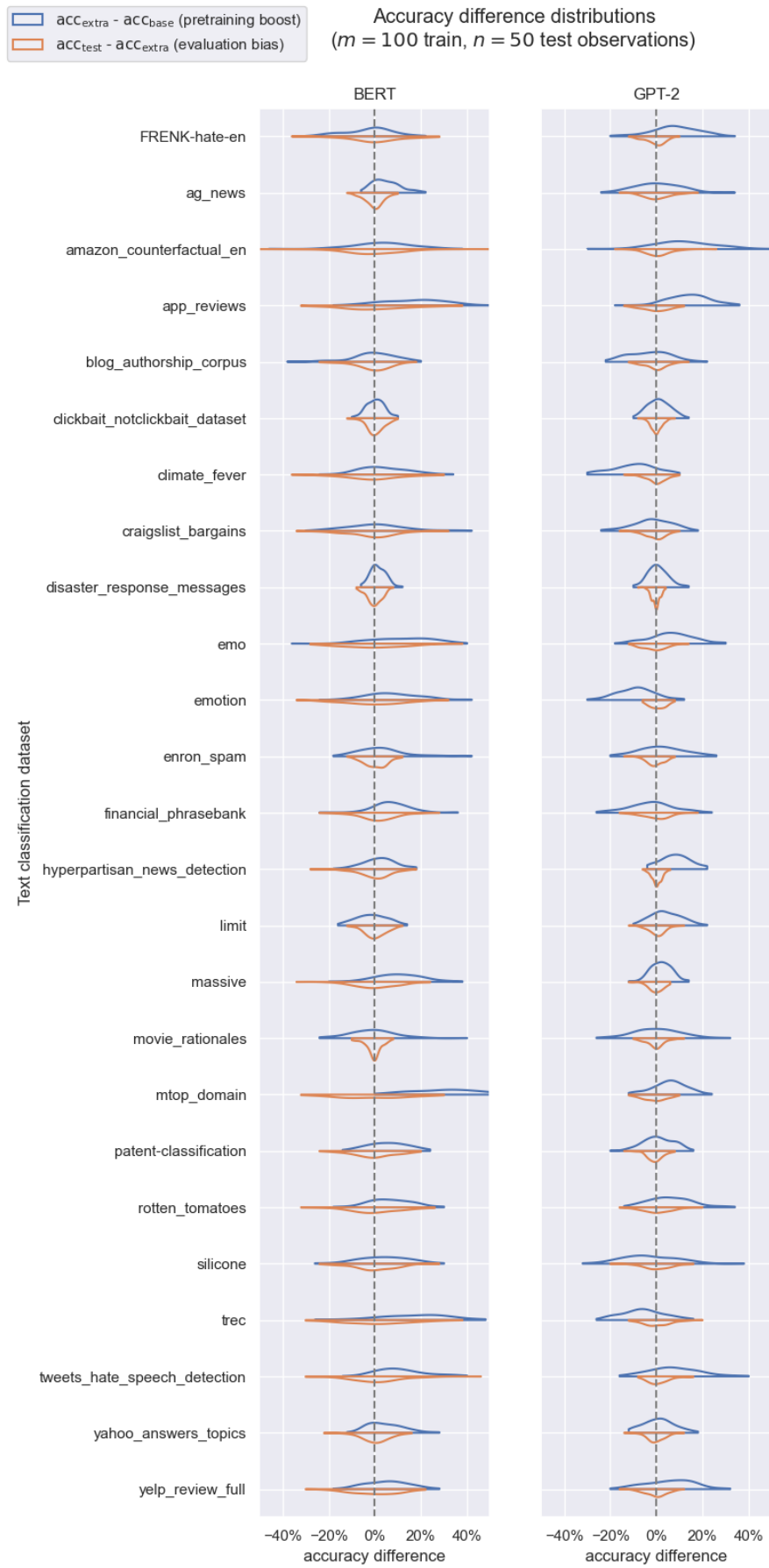


Figure 16

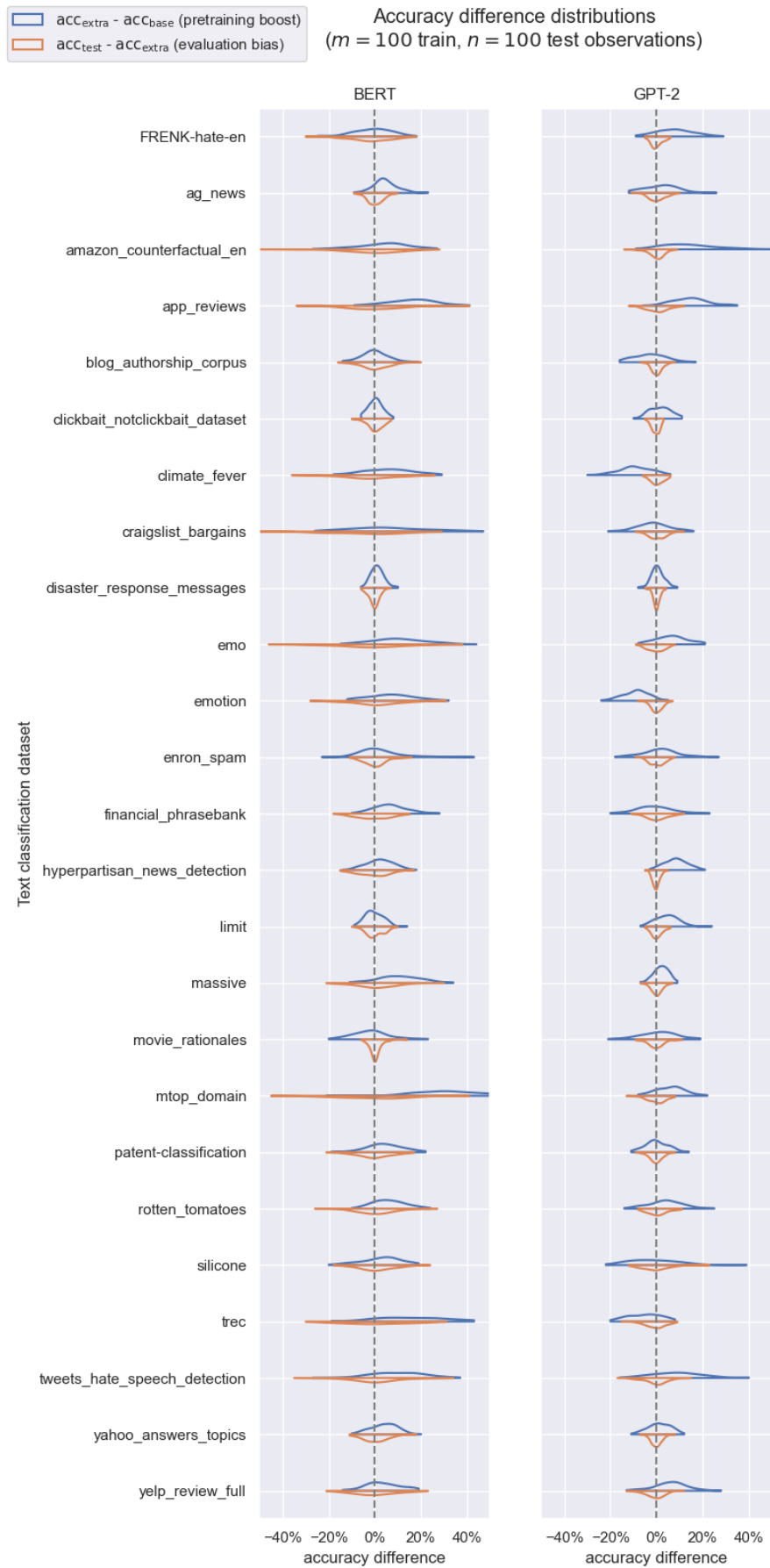
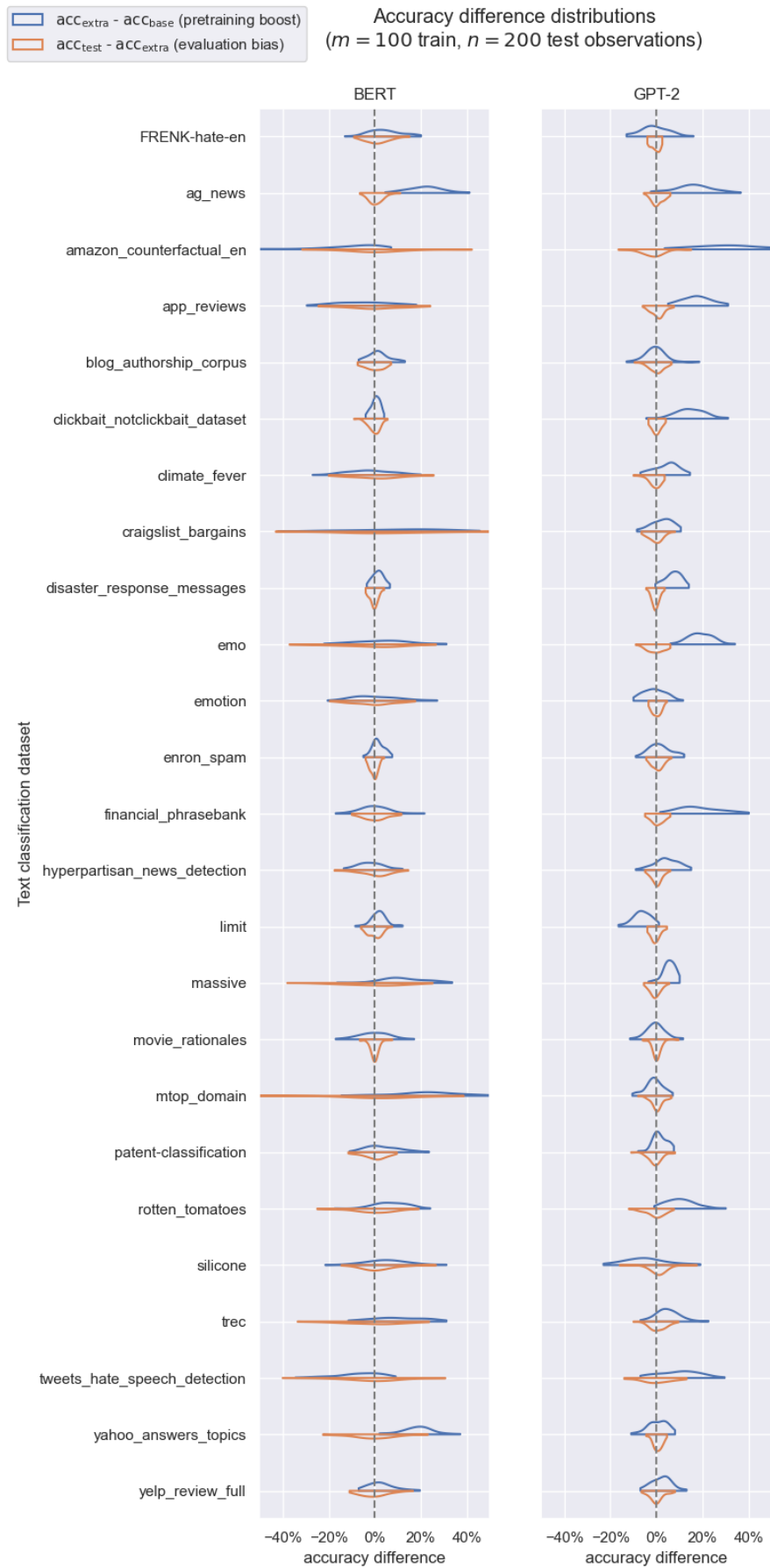


Figure 17





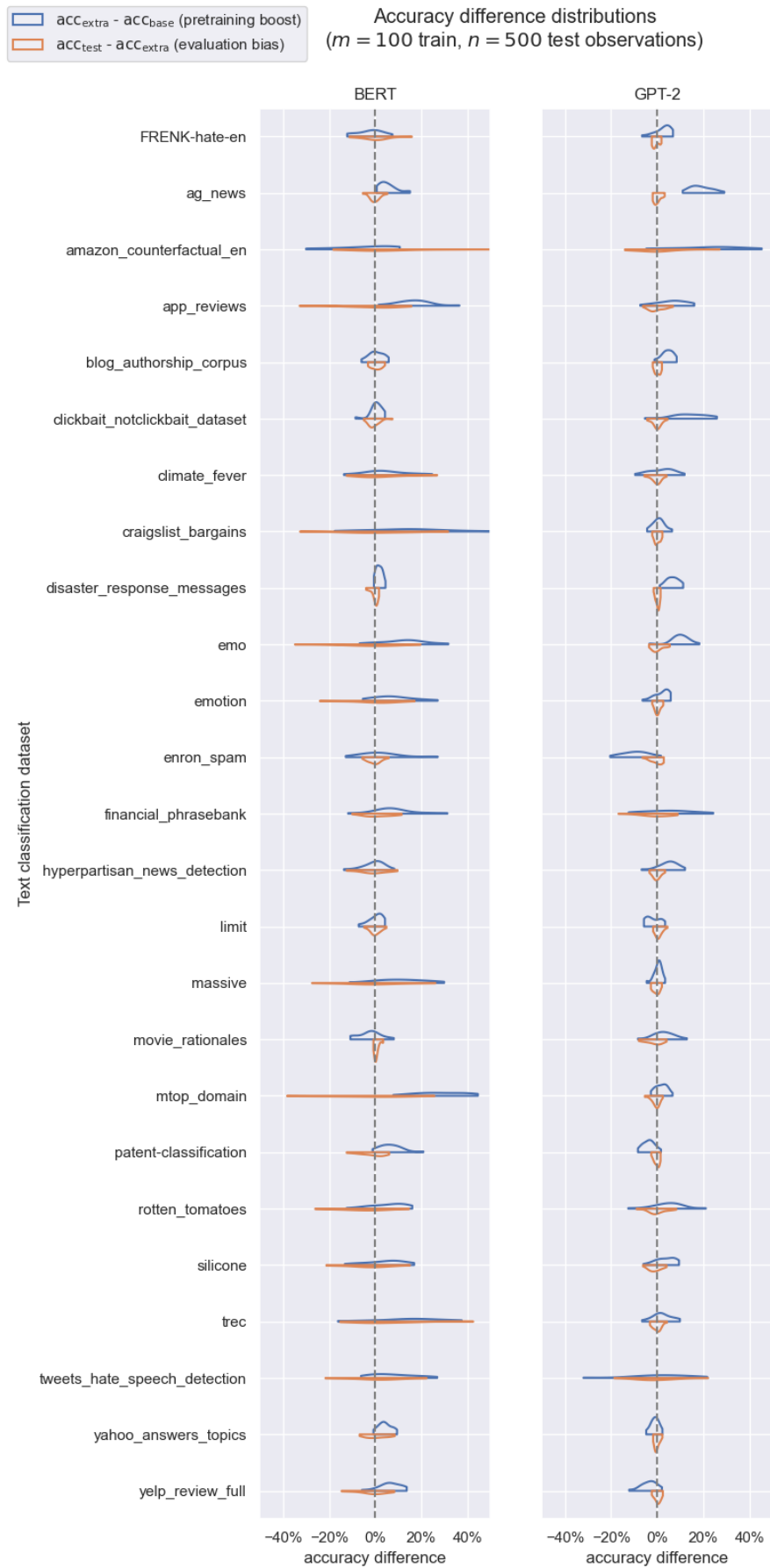


Figure 19