

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

Anonymous ACL submission

Abstract

Few-shot learning benchmarks are critical for evaluating modern NLP techniques. It is 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 is common for NLP benchmarks to release text from the test set, as researchers can 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 is 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 few-shot 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 is 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 is 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 text classification task:

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

[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

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 effect, 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 pretraining
- train: m (either 50 or 100) labeled texts for classification training
- test: n labeled texts to report accuracy.

Next, three accuracy estimators are computed. 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 causal language modeling loss for GPT-2. Texts are passed independently, and padded to form batches.
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. In other words, the expected value of $\text{acc}_{\text{extra}}$ is the accuracy one would observe on independent, identically distributed data.

4.2 acc_{test}

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

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

4.3 acc_{base}

acc_{base} does not do task-adaptive pretraining; it does not 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 is no boost from acc_{base} to $\text{acc}_{\text{extra}}$, then it may not be surprising to observe no difference between $\text{acc}_{\text{extra}}$ and acc_{test} . A boost from acc_{base} to $\text{acc}_{\text{extra}}$ would rule out undertraining as the cause of a null difference between $\text{acc}_{\text{extra}}$ and acc_{test} due to insufficient pretraining 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 acc_{test} , the source is exactly the test set text. For acc_{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 task.¹ For example, for $n = 500$, and for each of the 25 tasks, 20 ($\text{acc}_{\text{extra}}$, acc_{test} , acc_{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 is 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 is the evaluation bias from pretraining on unlabeled test set text instead of on unlabeled independent text.

¹For $n = 50$ 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.

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.

	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 is considerable variance, despite pairing the accuracy estimators.² 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

²One source of variance is intentionally introduced: the subsample 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.

et al., 2016). Furthermore, its computation is not completely trivial due to the data’s hierarchical dependency structure: each triple, ($\text{acc}_{\text{extra}}$, acc_{test} , acc_{base}), is drawn from (train, 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
- (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

Benchmarks assess methods by taking their average performance across tasks. To place the 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:

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

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.³

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 is then adjusted to control the false discovery rate (Benjamini and Hochberg, 1995).

³We will not assess any particular hypothesis about the role of causality. We are only motivating task-level analysis.

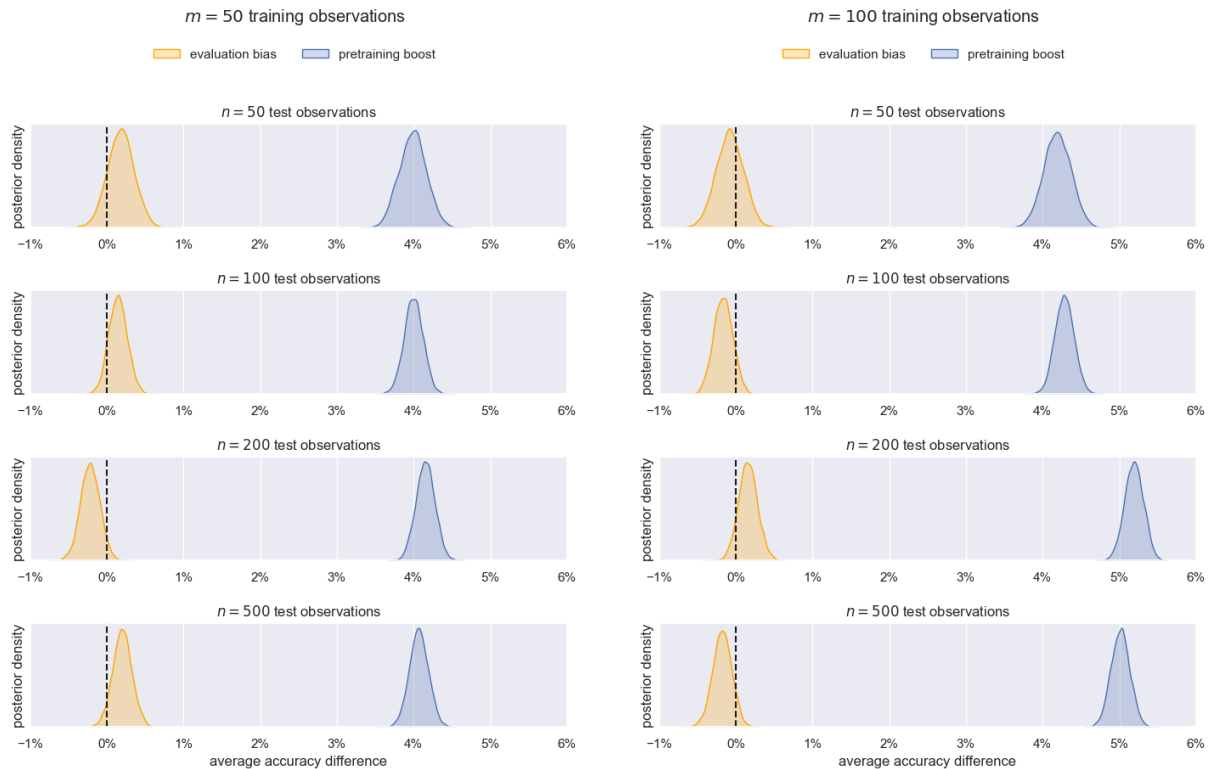


Figure 1: Distributions of average accuracy differences (11). The evaluation bias is akin to $\text{acc}_{\text{test}} - \text{acc}_{\text{extra}}$. The pretraining boost is akin to $\text{acc}_{\text{extra}} - \text{acc}_{\text{base}}$.

7 Discussion

Figure 1 demonstrates that the average pretraining boost is significant in every configuration of the experiment. 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%. 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 avoid excessive averaging, we lemon-picked tasks which reported a bias of at least +3% in any experiment configuration. All tasks matching this criterion were from experiments with BERT, as BERT had greater training variance. If there were a task-dependent evaluation bias, one could expect that the bias is consistent across m or n within a task, or there is a consistent pattern with how the bias changes with m or n across tasks. Figure 2 does not clearly support either of these hypotheses.

[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 hover around 0 across n . Figure 2 also does not support a relationship between n and the evaluation bias for lemon-picked tasks. But far 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 overtrain? 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

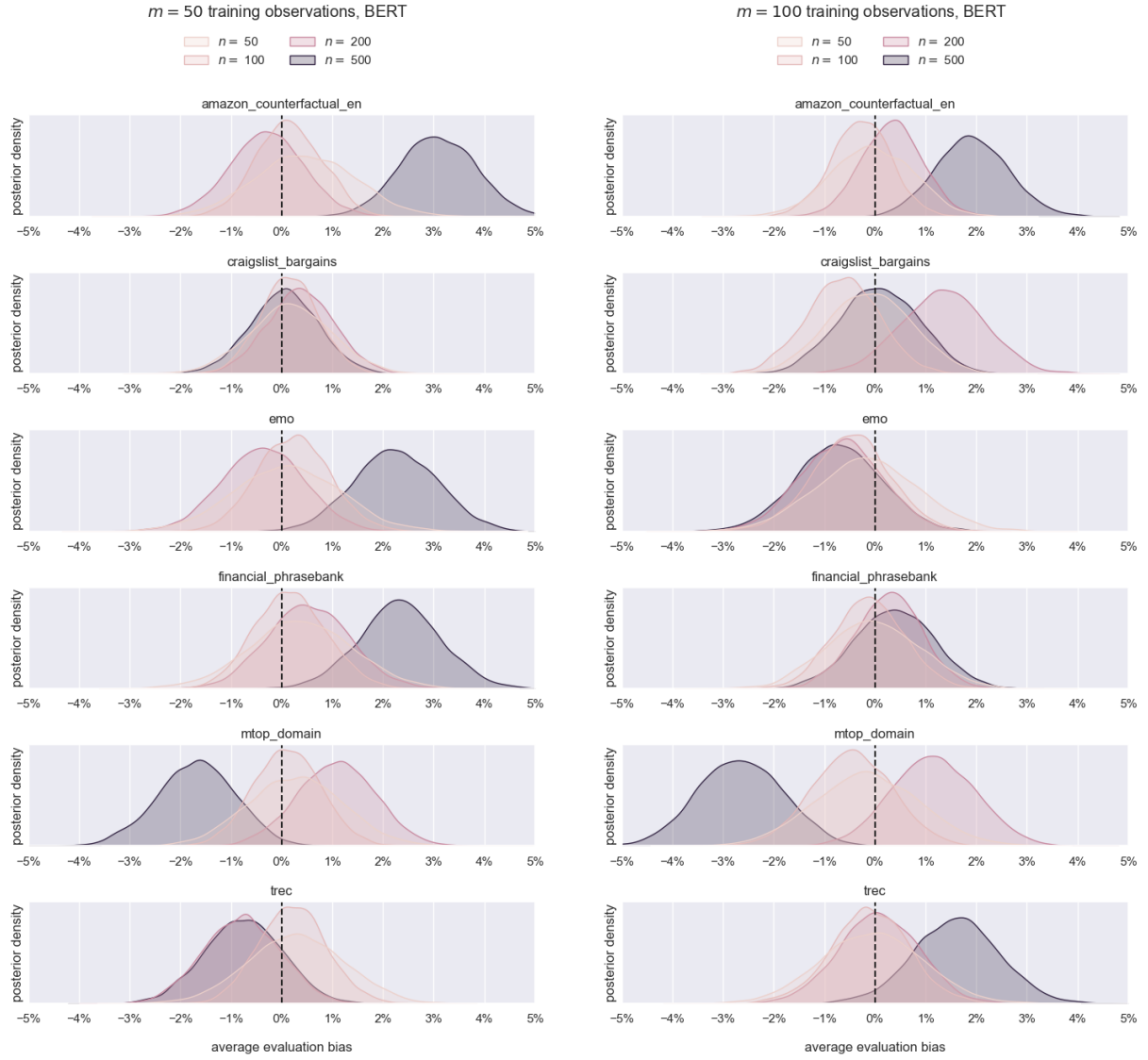


Figure 2: Distributions of average evaluation biases (12) for the subset of tasks which reported an average evaluation bias of at least +3% accuracy in any configuration of the experiment.

from the test in (13) are greater than 0.5.⁴ Over-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 (Team et al., 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 modifi-

cations. 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 (Jiang et al., 2023). Perhaps notably, instructions mention the set of possible answers—the class names.

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

⁴For 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, which demonstrates that the test has some statistical power.

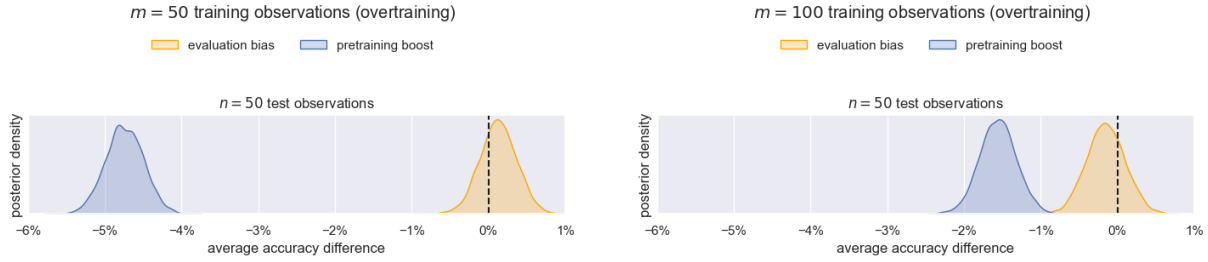


Figure 3: Average accuracy differences (11) after pretraining GPT-2 for 2 epochs instead of 1 (§8).

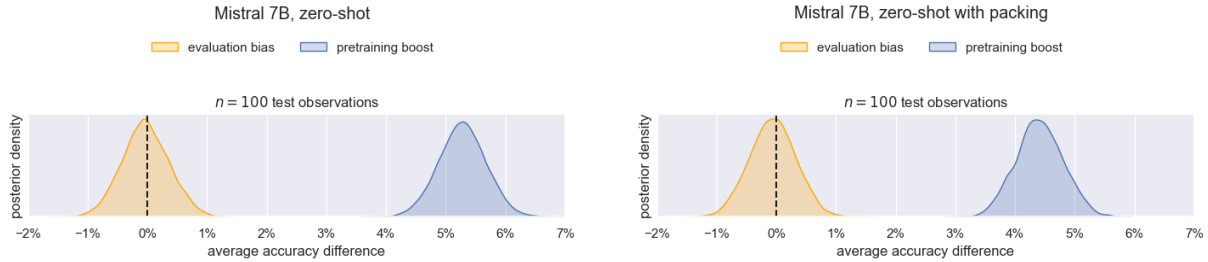


Figure 4: Average accuracy differences (11) for zero-shot classification (§9) with padding (left) and packing (right). For each of the 25 classification tasks, 20 subsamples were taken.

25 p -values from (13) are greater than 0.5; there is no evidence of a task-level evaluation bias.

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

9.1 Packing instead of padding

Experiments so far passed pretraining texts independently, adding and masking pad tokens to enable batching. Packing instead combines texts into a single sequence of tokens whose length is the model’s context length. Packing is often used during the initial pretraining of an LLM, where the model is trained on continuous streams of text to increase throughput (Brown et al., 2020).

Does packing impact evaluation bias differently than padding? One hypothesis is that packing causes the model to attend to previous texts, so the transformer has greater flexibility in modeling unlabeled text. To study the effects of packing, the zero-shot experiment in §9 is repeated with packing instead of padding. Figure 4 (right) shows that there is a pretraining boost, but no evaluation bias. All 25 p -values from (13) are greater than 0.5.

9.2 On testing test set contamination

Contamination detectors aim to flag overoptimistic LLM evaluations. An LLM is contaminated if it

was pretrained and evaluated on the same set of labeled data, as this procedure results in an evaluation bias. In contrast, the result from §9.1 implies that contamination of *unlabeled* test set text does not result in an evaluation bias. Do contamination detectors pick up this nuance?

The experiment in §9.1 is run for the `ag_news` task and $n = 500$. Next, text-label pairs from test are passed to the contamination hypothesis test in Oren et al. (2024). The p -value for the model pretrained on unlabeled text from `extra` is 0.33. The p -value for the model pretrained on unlabeled text from `test` is 0.015, which indicates contamination. However, the observed evaluation bias for this task is statistically indistinguishable from 0.

Detectors need to be able to differentiate the contamination of labeled text from the contamination of unlabeled text. For those that do not, contamination flags should be interpreted with care. Even if such a detector never raises false flags, a contamination flag may not indicate an overoptimistic evaluation.

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

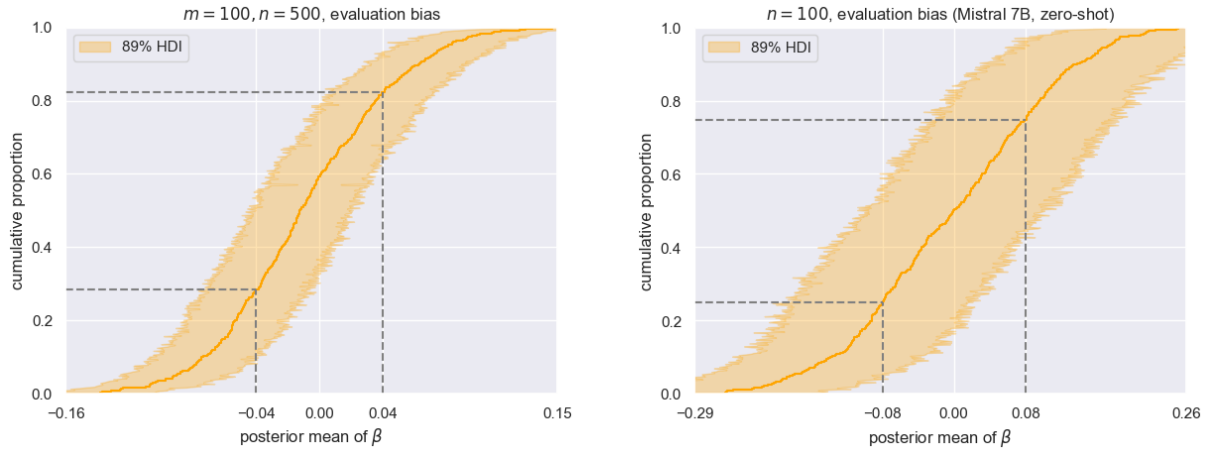


Figure 5: Distributions of conclusions had there been no technical replication (§10).

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 is a 47% chance that the posterior mean of β —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.⁵ For the zero-shot experiment in §9, there is a 50% chance that the posterior mean of β is outside $(-0.08, 0.08)$. Without repeated subsampling, one may as well flip a coin to decide whether pretraining on unlabeled test set text is fair.

11 Conclusion

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, when it hurts, and when pretraining is done on texts with instructions.

For benchmarks which release unlabeled text from the test set, this finding does not 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 is a benchmark and two LLMs, A and B . A was *not* pretrained on the benchmark’s unlabeled test set text, 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. Knowing whether an LLM was pretrained on unlabeled test set text 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.

⁵For 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 study semi-supervised methods like Pattern-Exploiting Training, or hand-inspecting the test set text and targeting interventions accordingly. We also do not study the effect of including unlabeled test set texts in the initial pretraining stage of an LLM.

The results are empirical. It is possible that there are tasks where an evaluation bias exists, and these were not part of the 25 classification tasks we collected. The results do not theoretically or universally establish that pretraining on unlabeled test set text is fair.

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 Piho, 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.
- 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.
- Yuxian Gu, Xu Han, Zhiyuan Liu, and Minlie Huang. 2022. [PPT: Pre-trained prompt tuning for few-shot](#)

633	learning. In <i>Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , pages 8410–8423, Dublin, Ireland. Association for Computational Linguistics.	
634		
635		
636		
637	Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré,	
638	Adam Chilton, Alex Chohlas-Wood, Austin Peters,	
639	Brandon Waldon, Daniel Rockmore, Diego Zam-	
640	brano, et al. 2024. Legalbench: A collaboratively	
641	built benchmark for measuring legal reasoning in	
642	large language models . <i>Advances in Neural Informa-</i>	
643	<i>tion Processing Systems</i> , 36.	
644	Suchin Gururangan, Ana Marasović, Swabha	
645	Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey,	
646	and Noah A. Smith. 2020. Don't stop pretraining:	
647	Adapt language models to domains and tasks . In	
648	<i>Proceedings of the 58th Annual Meeting of the</i>	
649	<i>Association for Computational Linguistics</i> , pages	
650	8342–8360, Online. Association for Computational	
651	Linguistics.	
652	Trevor Hastie, Robert Tibshirani, Jerome H Friedman,	
653	and Jerome H Friedman. 2009. <i>The elements of statis-</i>	
654	<i>tical learning: data mining, inference, and prediction</i> ,	
655	volume 2. Springer.	
656	He He, Derek Chen, Anusha Balakrishnan, and Percy	
657	Liang. 2018. Decoupling strategy and generation in	
658	negotiation dialogues . In <i>Proceedings of the 2018</i>	
659	<i>Conference on Empirical Methods in Natural Lan-</i>	
660	<i>guage Processing</i> , pages 2333–2343, Brussels, Bel-	
661	gium. Association for Computational Linguistics.	
662	Zhang Huangzhao. 2018. Yahoo-	
663	answers-topic-classification-dataset.	
664	https://github.com/LC-John/	
665	Yahoo-Answers-Topic-Classification-Dataset .	
666	Alon Jacovi, Avi Caciularu, Omer Goldman, and Yoav	
667	Goldberg. 2023. Stop uploading test data in plain	
668	text: Practical strategies for mitigating data contami-	
669	nation by evaluation benchmarks . In <i>Proceedings of</i>	
670	<i>the 2023 Conference on Empirical Methods in Natu-</i>	
671	<i>ral Language Processing</i> , pages 5075–5084, Singa-	
672	pore. Association for Computational Linguistics.	
673	Albert Q Jiang, Alexandre Sablayrolles, Arthur Men-	
674	sch, Chris Bamford, Devendra Singh Chaplot, Diego	
675	de las Casas, Florian Bressand, Gianna Lengyel, Guil-	
676	laume Lample, Lucile Saulnier, et al. 2023. Mistral	
677	7b . <i>arXiv preprint arXiv:2310.06825</i> .	
678	Zhijing Jin, Julius von Kügelgen, Jingwei Ni, Tejas	
679	Vaidhya, Ayush Kaushal, Mrinmaya Sachan, and	
680	Bernhard Schoelkopf. 2021. Causal direction of data	
681	collection matters: Implications of causal and an-	
682	ticausal learning for NLP . In <i>Proceedings of the</i>	
683	<i>2021 Conference on Empirical Methods in Natural</i>	
684	<i>Language Processing</i> , pages 9499–9513, Online and	
685	Punta Cana, Dominican Republic. Association for	
686	Computational Linguistics.	
687	Johannes Kiesel, Maria Mestre, Rishabh Shukla, Em-	
688	manuel Vincent, Payam Adineh, David Corney,	
	Benno Stein, and Martin Potthast. 2019. SemEval-	689
	2019 task 4: Hyperpartisan news detection . In	690
	<i>Proceedings of the 13th International Workshop on</i>	691
	<i>Semantic Evaluation</i> , pages 829–839, Minneapolis,	692
	Minnesota, USA. Association for Computational Lin-	693
	guistics.	694
	Ravin Kumar, Colin Carroll, Ari Hartikainen, and Os-	695
	valdo Martin. 2019. Arviz a unified library for	696
	exploratory analysis of bayesian models in python .	697
	<i>Journal of Open Source Software</i> , 4(33):1143.	698
	Nikola Ljubešić, Darja Fišer, and Tomaž Erjavec. 2019.	699
	The frenk datasets of socially unacceptable discourse	700
	in slovene and english .	701
	Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel,	702
	and Pontus Stenetorp. 2022. Fantastically ordered	703
	prompts and where to find them: Overcoming few-	704
	shot prompt order sensitivity . In <i>Proceedings of the</i>	705
	<i>60th Annual Meeting of the Association for Compu-</i>	706
	<i>tational Linguistics (Volume 1: Long Papers)</i> , pages	707
	8086–8098, Dublin, Ireland. Association for Compu-	708
	tational Linguistics.	709
	P. Malo, A. Sinha, P. Korhonen, J. Wallenius, and	710
	P. Takala. 2014. Good debt or bad debt: Detecting se-	711
	mantic orientations in economic texts . <i>Journal of the</i>	712
	<i>Association for Information Science and Technology</i> ,	713
	65.	714
	Irene Manotas, Ngoc Phuoc An Vo, and Vadim Sheinin.	715
	2020. LiMiT: The literal motion in text dataset . In	716
	<i>Findings of the Association for Computational Lin-</i>	717
	<i>guistics: EMNLP 2020</i> , pages 991–1000, Online.	718
	Association for Computational Linguistics.	719
	Richard McElreath. 2018. <i>Statistical rethinking: A</i>	720
	<i>Bayesian course with examples in R and Stan</i> . Chap-	721
	man and Hall/CRC.	722
	Vangelis Metsis, Ion Androutsopoulos, and Georgios	723
	Paliouras. 2006. Spam filtering with naive bayes-	724
	which naive bayes? In <i>CEAS</i> , volume 17, pages	725
	28–69. Mountain View, CA.	726
	Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Cor-	727
	rado, and Jeff Dean. 2013. Distributed representa-	728
	tions of words and phrases and their compositionality .	729
	<i>Advances in neural information processing systems</i> ,	730
	26.	731
	Richard D Morey, Rink Hoekstra, Jeffrey N Rouder,	732
	Michael D Lee, and Eric-Jan Wagenmakers. 2016.	733
	The fallacy of placing confidence in confidence inter-	734
	vals . <i>Psychonomic bulletin & review</i> , 23:103–123.	735
	Amit Moscovich and Saharon Rosset. 2022. On the	736
	cross-validation bias due to unsupervised preprocess-	737
	ing . <i>Journal of the Royal Statistical Society Series B:</i>	738
	<i>Statistical Methodology</i> , 84(4):1474–1502.	739
	Niklas Muennighoff, Nouamane Tazi, Loic Magne, and	740
	Nils Reimers. 2023. MTEB: Massive text embedding	741
	benchmark . In <i>Proceedings of the 17th Conference</i>	742

743	<i>of the European Chapter of the Association for Computational Linguistics</i> , pages 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics.	800
744		801
745		802
746	James O’Neill, Polina Rozenshtein, Ryuichi Kiryo, Motoko Kubota, and Danushka Bollegala. 2021. I wish I would have loved this one, but I didn’t – a multilingual dataset for counterfactual detection in product review . In <i>Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing</i> , pages 7092–7108, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.	803
747		804
748		805
749		806
750		807
751		808
752		809
753		810
754		811
755	Yonatan Oren, Nicole Meister, Niladri S. Chatterji, Faisal Ladhak, and Tatsunori Hashimoto. 2024. Proving test set contamination in black-box language models . In <i>The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024</i> . OpenReview.net.	812
756		813
757		814
758		815
759		816
760		817
761	Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales . In <i>Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL’05)</i> , pages 115–124, Ann Arbor, Michigan. Association for Computational Linguistics.	818
762		819
763		820
764		821
765		822
766		823
767		824
768	Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners . <i>OpenAI blog</i> , 1(8):9.	825
769		826
770		827
771		828
772	Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. 2018. CARER: Contextualized affect representations for emotion recognition . In <i>Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing</i> , pages 3687–3697, Brussels, Belgium. Association for Computational Linguistics.	829
773		830
774		831
775		832
776		833
777		834
778		835
779		836
780	Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference . In <i>Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume</i> , pages 255–269, Online. Association for Computational Linguistics.	837
781		838
782		839
783		840
784		841
785		842
786	Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W Pennebaker. 2006. Effects of age and gender on blogging . In <i>AAAI spring symposium: Computational approaches to analyzing weblogs</i> , volume 6, pages 199–205.	843
787		844
788		845
789		846
790		847
791	Eva Sharma, Chen Li, and Lu Wang. 2019. BIG-PATENT: A large-scale dataset for abstractive and coherent summarization . In <i>Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics</i> , pages 2204–2213, Florence, Italy. Association for Computational Linguistics.	848
792		849
793		850
794		851
795		852
796		853
797	Roshan Sharma. 2019. Twitter-sentiment-analysis. https://github.com/sharmaroshan/Twitter-Sentiment-Analysis .	854
798		
799		
	Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size . <i>arXiv preprint arXiv:2408.00118</i> .	
	Tan Thongtan and Tanasanee Phienthrakul. 2019. Sentiment classification using document embeddings trained with cosine similarity . In <i>Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop</i> , pages 407–414, Florence, Italy. Association for Computational Linguistics.	
	Lewis Tunstall, Nils Reimers, Unso Eun Seo Jo, Luke Bates, Daniel Korat, Moshe Wasserblat, and Oren Pereg. 2022. Efficient few-shot learning without prompts . <i>arXiv preprint arXiv:2209.11055</i> .	
	Mengqiu Wang, Noah A. Smith, and Teruko Mitamura. 2007. What is the Jeopardy model? a quasi-synchronous grammar for QA . In <i>Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)</i> , pages 22–32, Prague, Czech Republic. Association for Computational Linguistics.	
	Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing . In <i>Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations</i> , pages 38–45, Online. Association for Computational Linguistics.	
	Tianyi Zhang, Felix Wu, Arzo Katiyar, Kilian Q Weinberger, and Yoav Artzi. 2021. Revisiting few-sample bert fine-tuning . In <i>International Conference on Learning Representations</i> .	
	Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification . <i>Advances in neural information processing systems</i> , 28.	
	Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhonghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric Xing, et al. 2023. Lmsys-chat-1m: A large-scale real-world llm conversation dataset . <i>arXiv preprint arXiv:2309.11998</i> .	
	A Classification tasks	
	The experiment was ran on 25 publicly available text classification tasks found in https://huggingface.co/datasets . Inclusion criteria:	
	1. All text is in English.	

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

2. The number of classes is not greater than 25, because only 50 or 100 observations are used for training the classifier.
3. The task is to classify one text, not a pair as in, e.g., textual entailment tasks.
4. Texts are not so long that too much useful signal is dropped when text is truncated to fit in BERT/GPT-2’s context window, which is set to 256 tokens.
5. Based on our best judgment, it is likely that BERT/GPT-2 can do better than guessing.

Table 2 lists the exact tasks.

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 are 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 `accextra` and `acctest`. This choice was made in an effort to widen any gap between them.

`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’ pre-training 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—`accextra` versus `accbase` and `acctest` versus `accextra`—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. For the experiment in §4, BERT was pre-trained for 2 epochs, and GPT-2 was pretrained for 1 epoch. Classification 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 `accextra`, `acctest`, and `accbase` (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 `accextra − accbase` and `acctest − accextra` together. That’s because these differences are not independent: if `accextra` is high, then `accextra − accbase` increases and `acctest − accextra` decreases. This paper does not analyze the scores together, per se. We care about `acctest − accextra`. `accextra − accbase` only exists to sanity check that the pretraining code works; there may be an effect to detect.

D.2 Difference distributions

Figures 11 - 18 visualize the distributions of the paired differences—`accextra − accbase` and `acctest − accextra`—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 β (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 24 hierarchical models (16 in §7, 4 in §8, and 4 in §9), no divergences were observed during the fitting procedure. All trace plots were healthy.

Figure 9 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 10 contains posterior distributions of β 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 7 checks that posterior predictions for the average task accuracies are calibrated. Figure 8 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 acc_{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. Spearman’s rank correlation coefficient is used because we are only interested in the consistency of the relationship, not its linearity.

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.

G Zero-shot text classification

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:

For packing (§9.1), the prompts at inference are in the same format as above. For training, 8 texts were packed. Here is an example of an input sequence for `ag_news`, where 4 texts are packed:

Your task is to classify a given text as one of these categories:
World

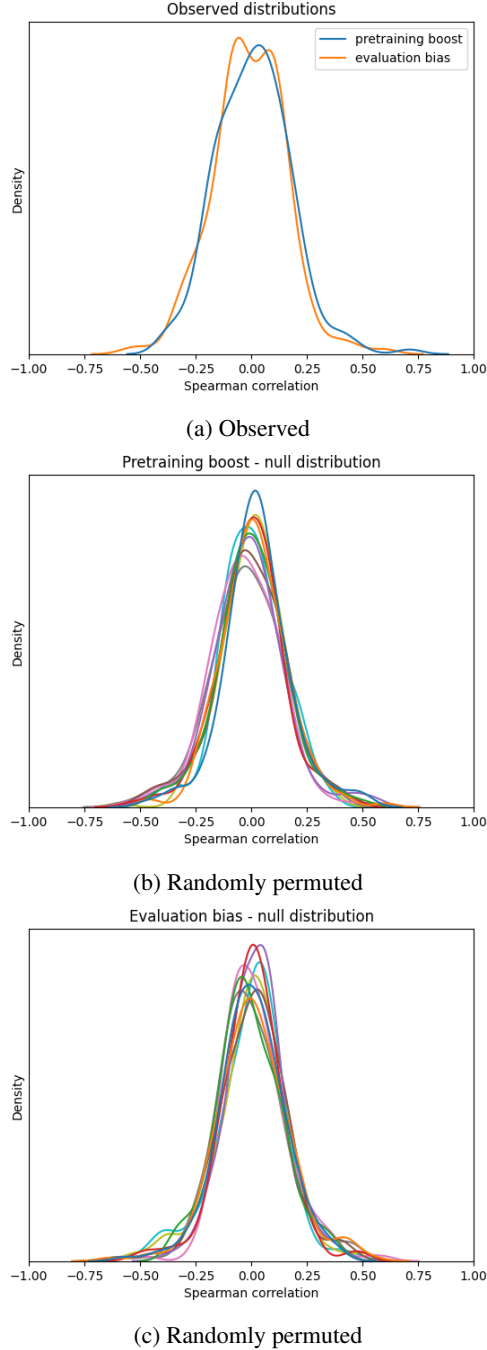


Figure 6: Distribution of correlation between BERT and GPT-2 across all m , n , and the 25 classification tasks.

Sports 1051
Business 1052
Sci/Tech 1053

The text is a news article. Answer with its topic. 1054

Text: US Electoral College withstands critics 1055
... so far (AFP) AFP - Lambasted as antiquated 1056
and anti-democratic, the Electoral College that 1057
decides the US presidency has survived for 1058
centuries as an unmovable albeit creaky pillar 1059
of the American political system. 1060

Text: Voters in Hungary decide referenda 1061
Voters in Hungary went to the polls Sunday to 1062
decide a double referendum on citizenship 1063
rights and their nation's health care 1064
system. 1065

Text: White House: Trying to Confirm Terror 1066
Group's Allegiance to bin 1067
... The Bush administration 1068
says it's trying to confirm the latest 1069
declaration from the most feared militant group 1070
in Iraq. In a statement posted on a Web site 1071
Sunday, the group led by terror mastermind Abu 1072
Musab 1073

Text: Fans rush to create mods for 1074
long-awaited 'Doom 3' Activision's 1075
Doom 3, which launched earlier this month, 1076
wasn't on store shelves for three days 1077
before players started creating their own 1078
modifications - known as mods - o the game. 1079

The zero-shot experiment files are in 1080
[cloud_scripts/gcp/experiments/zero_shot/](#) 1081
and [cloud_scripts/gcp/experiments/](#) 1082
[zero_shot_packing/](#). Batch sizes are set 1083
to run on a GPU with at least 20 GB RAM. 1084
The GPU must support the data types needed 1085
for QLoRA, e.g., an L4 GPU. Figure 4 can 1086
be reproduced by running the notebook 1087
in [analysis/fit_posteriors/zero_shot](#) 1088
and then the notebook, 1089
[analysis/results/posterior_pred.ipynb](#). 1090

The Mistral 7B model is [Mistral-7B-v0.3](#), the 1091
non-instruction-trained model. 1092

We only study $n = 100$ in an initial effort to 1093
provide evidence of an evaluation bias (due to the 1094
relatively small test set), and take 20 repeated sub- 1095
samples instead of 50. While $n = 100$ is quite 1096
small, benchmarks such as LegalBench (Guha et al., 1097
2024) have test data in this range. And the analysis 1098
transparently exposes variance. 1099

QLoRA hyperparameters were pre-specified: ev- 1100
ery adapter has rank 16 with $\alpha = 32$ (LoRA scaling 1101
factor), a 0.05 dropout rate, and no bias param- 1102
eters. The adapter layers introduce 41, 943, 040 new, 1103

trainable parameters to Mistral 7B, whose parameters are frozen. Pretraining was done for 1 epoch.

To increase the power of the contamination hypothesis test run in §9.2, shards were formed to be similar to the sequences passed in during pretraining. Here is an example of what the first 2 text-label pairs in the dataset passed to the contamination test looks like:

```
### Text: Customers bemoan changes in Quicken  
2005 The new version of the personal finance  
program drops support for a widely used file  
format.
```

```
### Answer: Sci/Tech
```

```
### Text: Blair gives partial Iraq apology Tony  
Blair has offered his Labour party a partial  
apology for waging war in Iraq, striving to pull  
angry supporters behind him ahead of an election  
next year.
```

```
### Answer: World
```

The 2 p -values in §9.2 can be obtained by running the notebook [analysis/contamination/test.ipynb](#) on an L4 GPU.

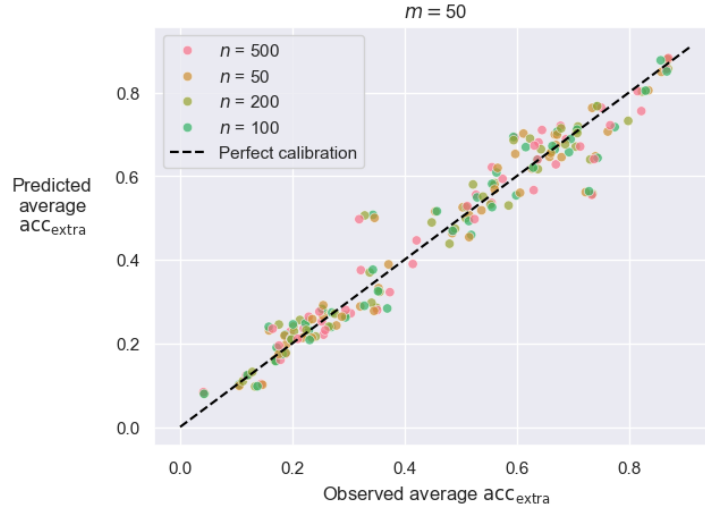


Figure 7: Each of the points represents a task and an LM type.

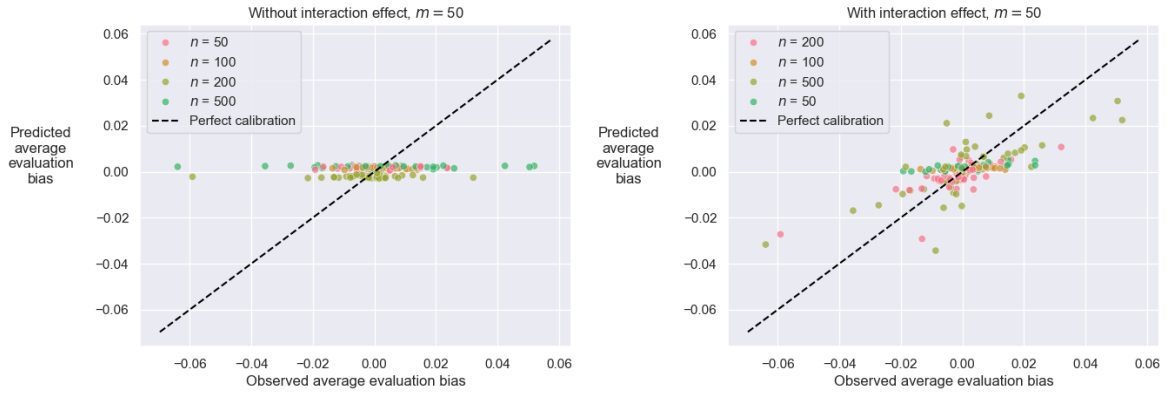


Figure 8: Omitting the interaction effect causes underfitting. Note that the prior causes effects to shrink towards 0.

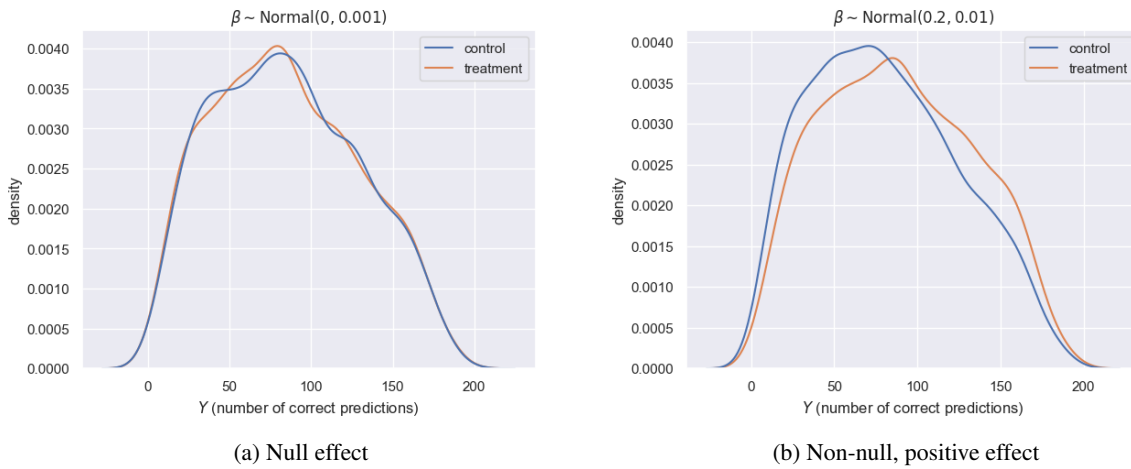


Figure 9: Prior predictive distributions for $m = 100, n = 200$ from two different priors for β —the expected increase in the log-odds of a correct prediction resulting from an intervention/treatment.

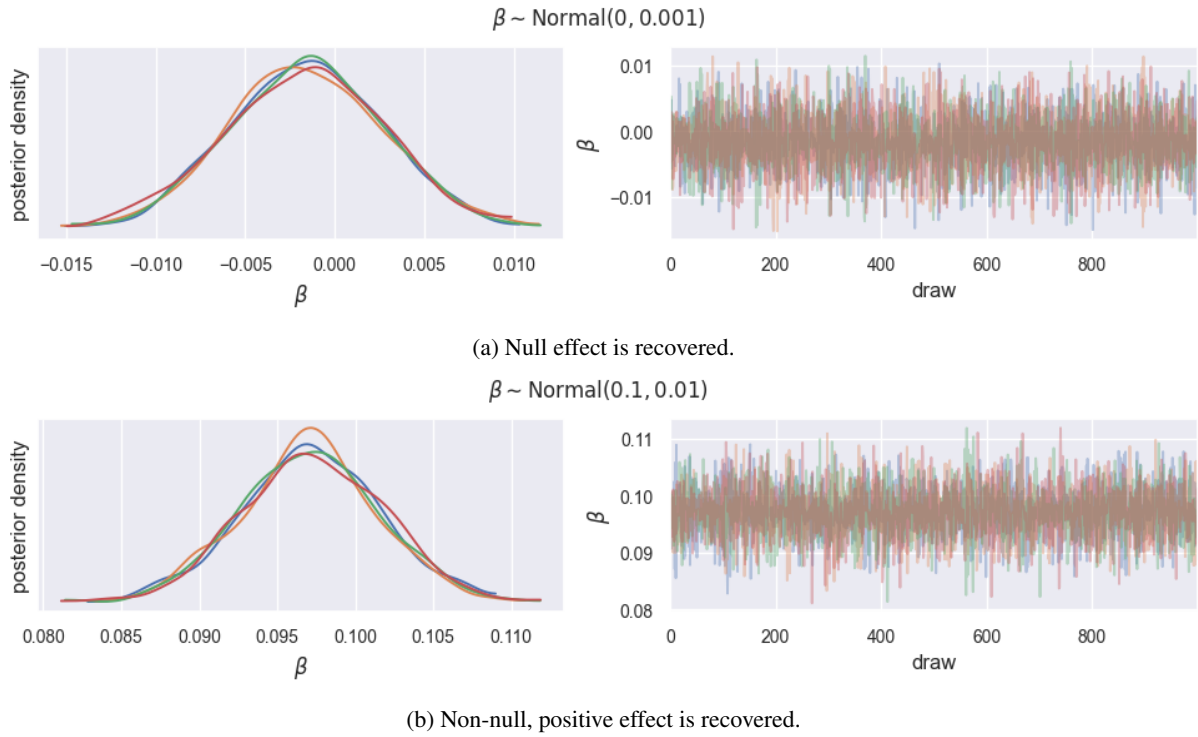


Figure 10: Posterior distributions and trace plots for null and non-null effects **from simulated data** where $m = 100, n = 200$, approximated by four chains with 1,000 draws each, after 500 steps of tuning. For each model, no divergences were observed during the fitting procedure. Visualizations were produced by the arviz package (Kumar et al., 2019).

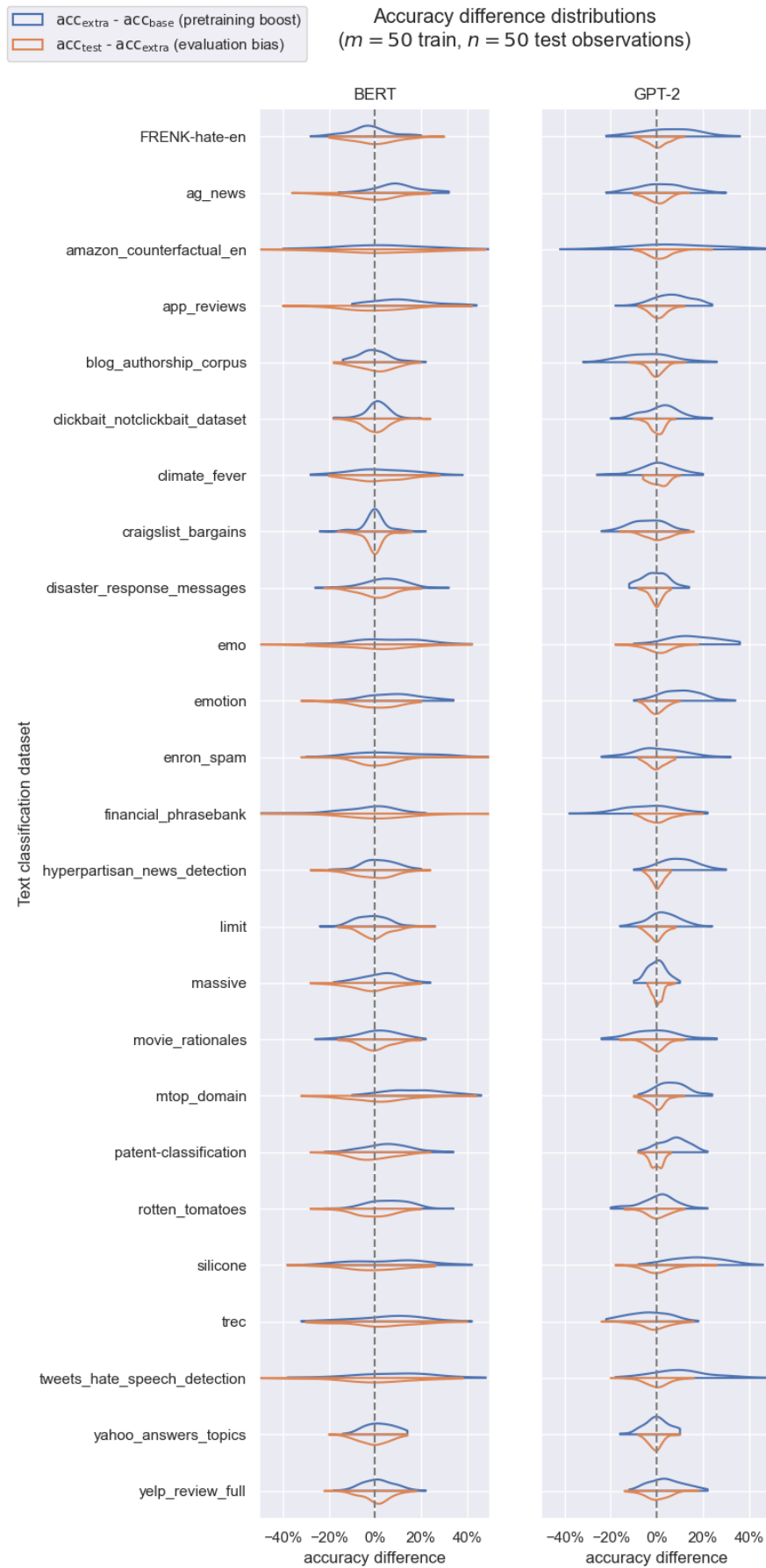


Figure 11

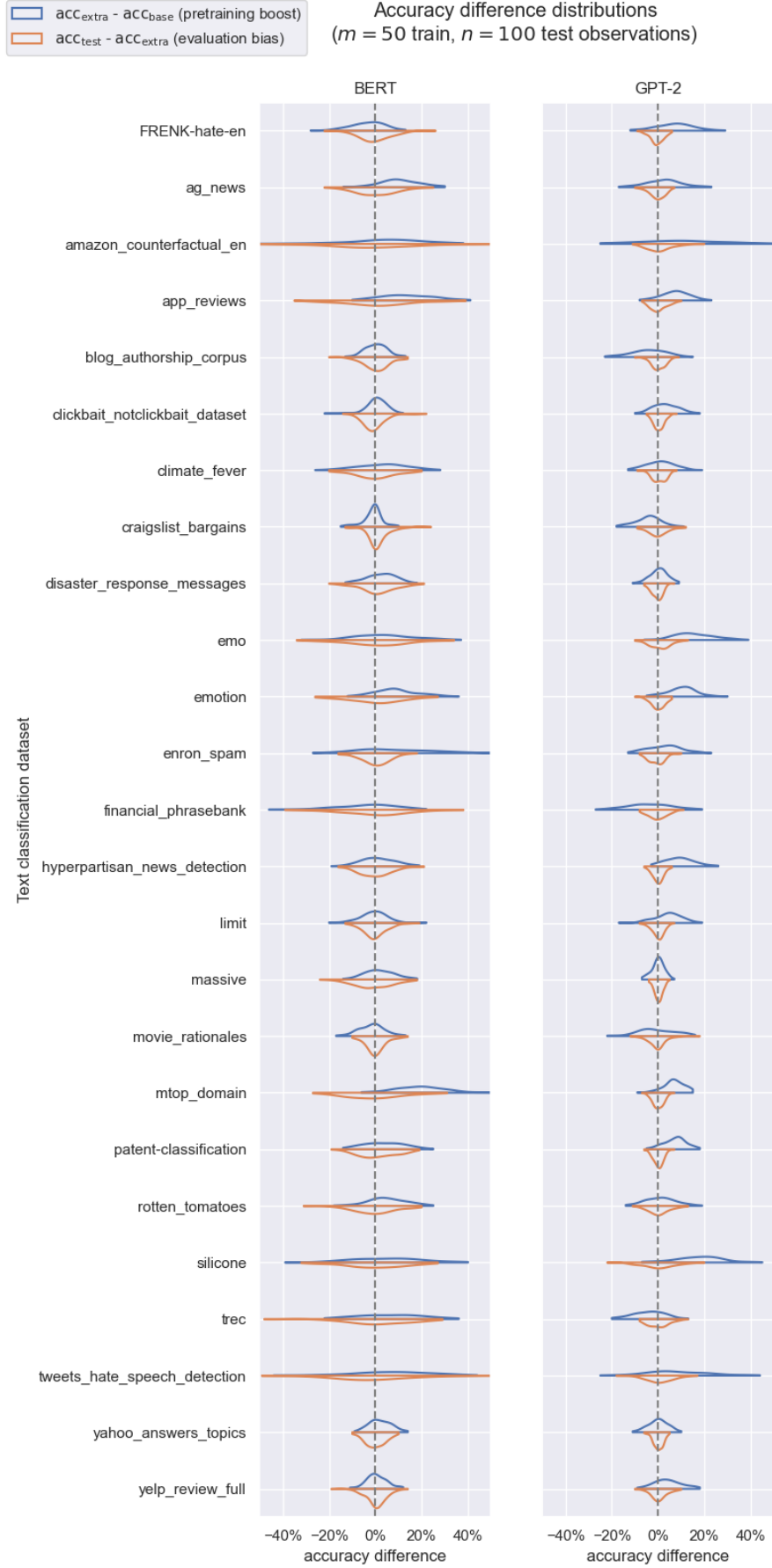


Figure 12

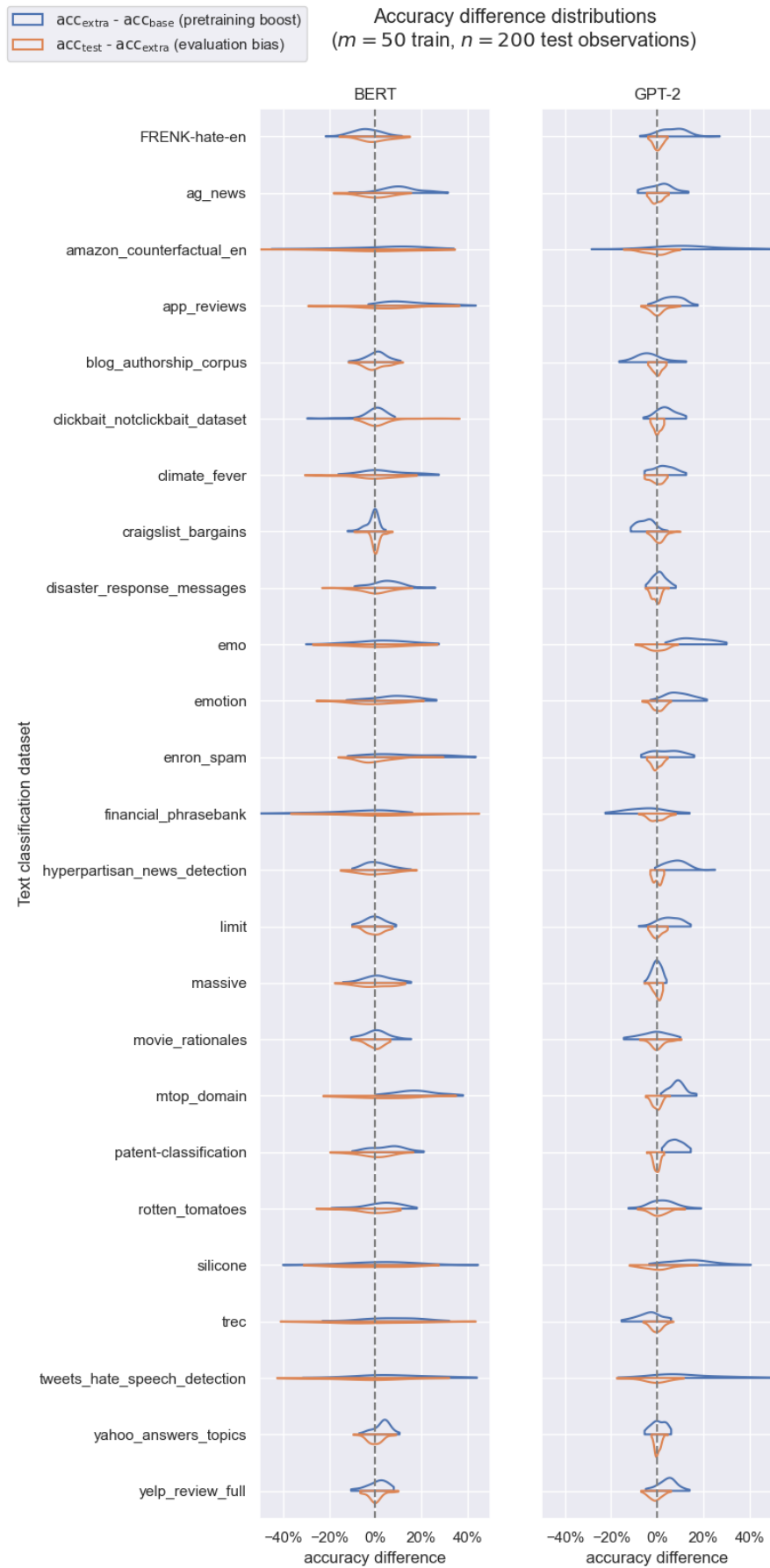


Figure 13

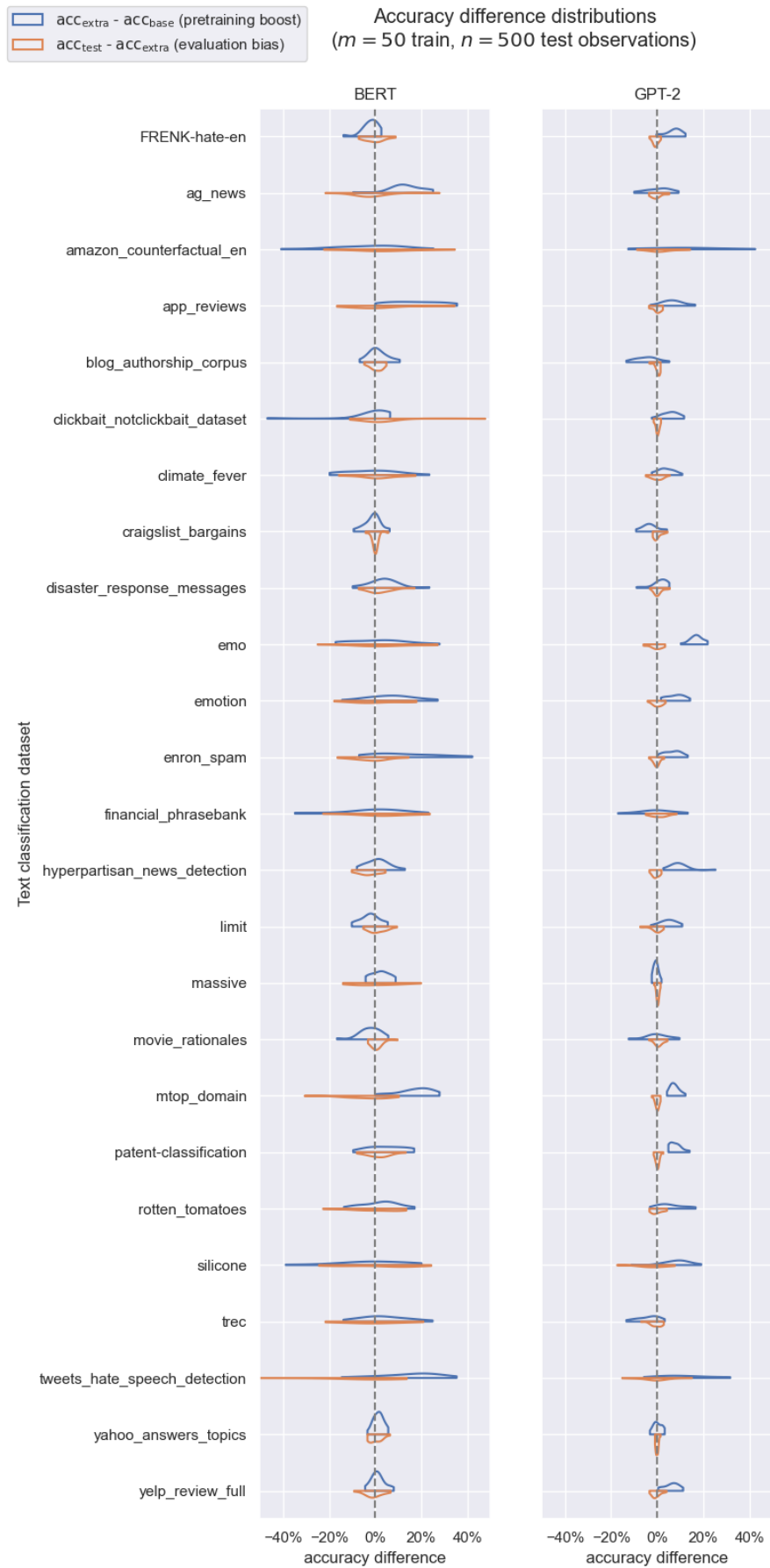


Figure 14

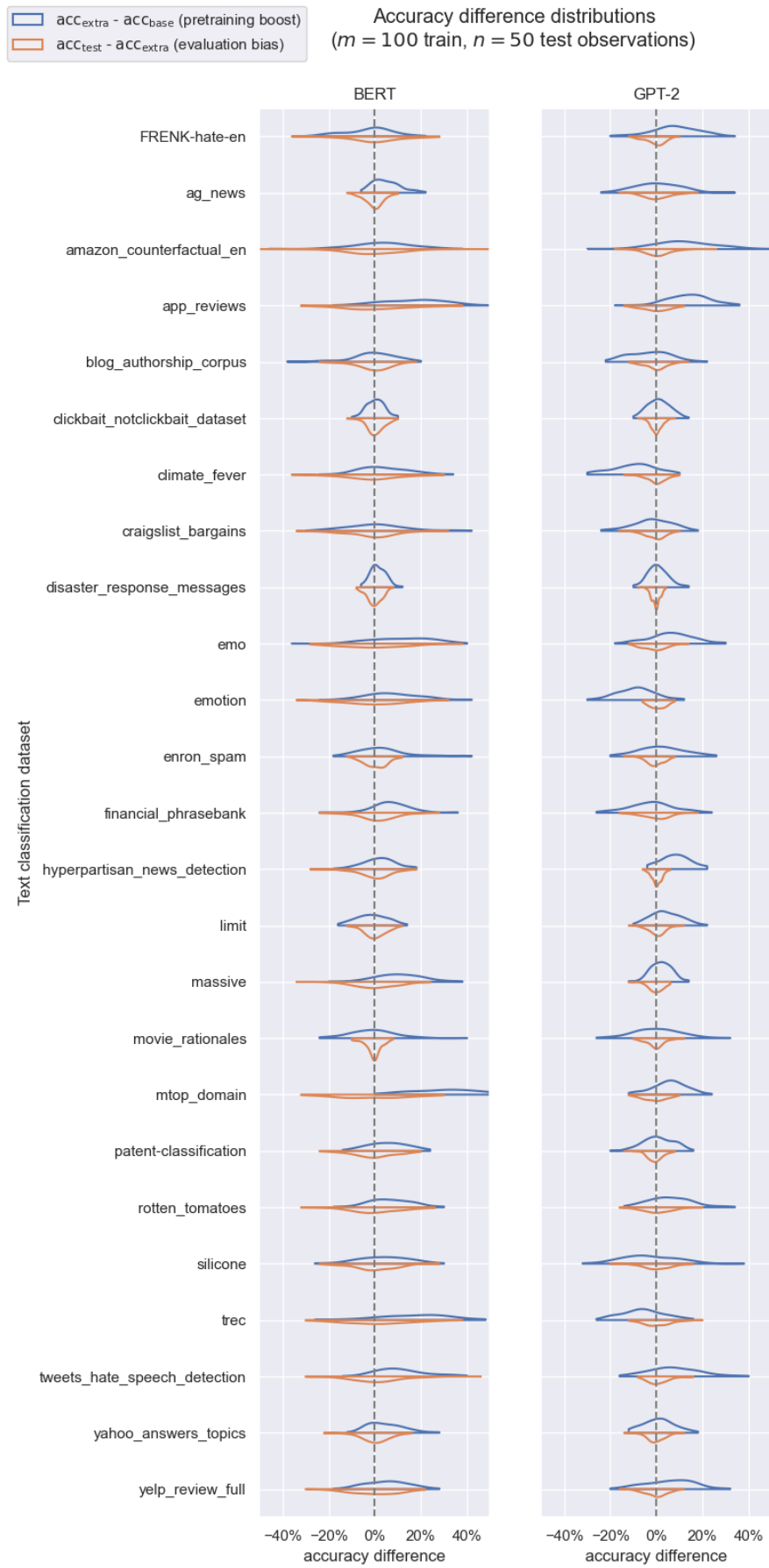


Figure 15

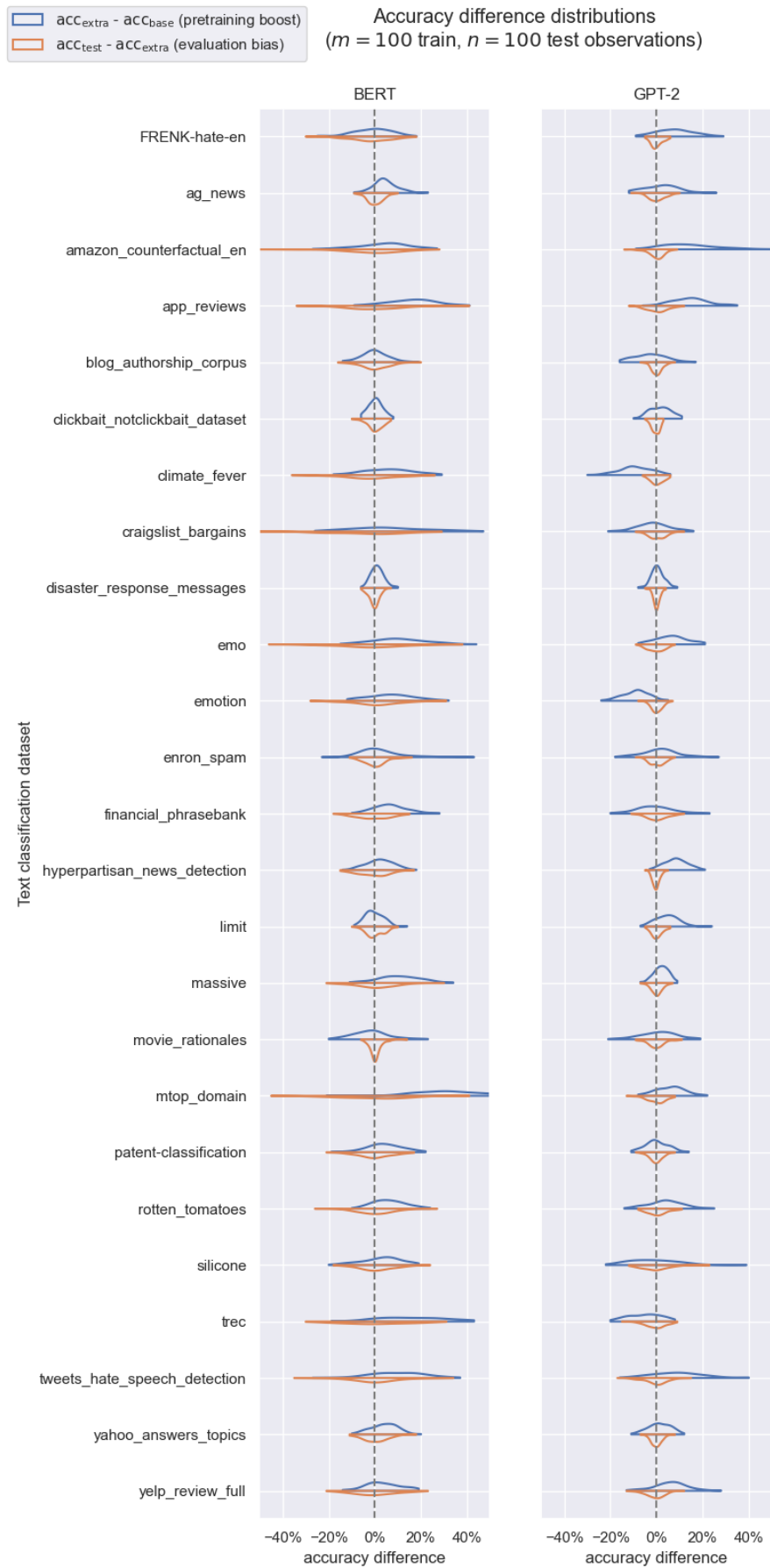


Figure 16

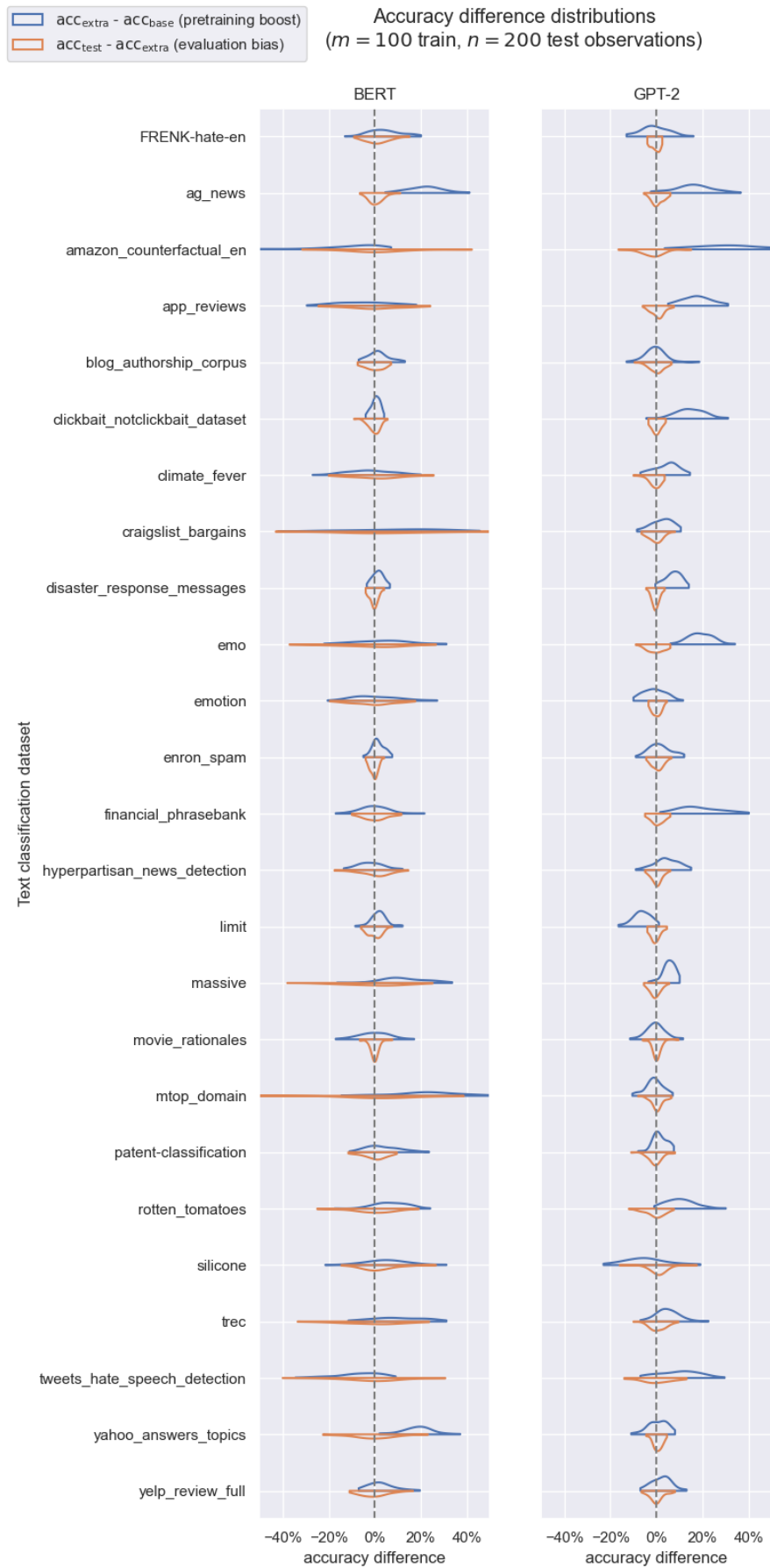


Figure 17

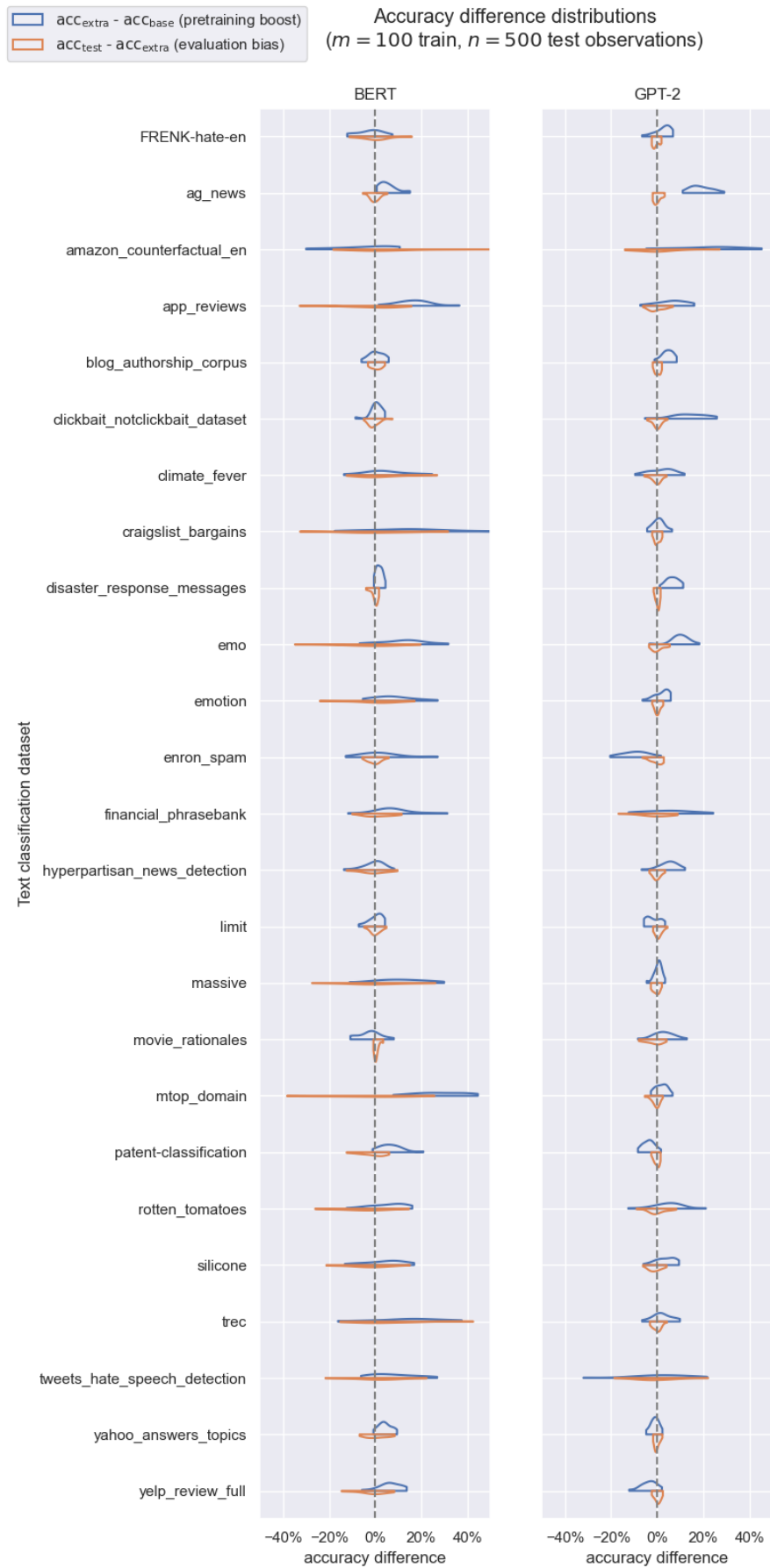


Figure 18