Debiasing Language Models for In-Context Learning by Adapting Causal Inference

Hao Zou[1], **Karin de Langis**[1], **Dongyeop Kang**[1], **and Yohan Jo**[2]

[1]Computer Science & Engineering, University of Minnesota
{zou00080,dento019,dongyeop}@umn.edu
[2]Alexa AI, Amazon
jyoha@amazon.com

Abstract

Large pre-trained language models have shown remarkable prediction performance in in-context learning, i.e. using only a few task demonstrations in input prompts. But in-context learning performs poorly when input text and output labels are confounded by confounding variables (e.g. a movie being horror affects both review texts and likely review sentiment), and how to control for them without fine-tuning a model is not straightforward. In our view, the problem of common prompting practice is that it is eliciting a summary of observational data (i.e. pre-training corpora) rather than measuring the true causal effect of the input text on possible labels. To measure the latter more accurately, we present a method inspired by causal inference on observational data. Specifically, our method proposes (1) prompting conditionally on confounding variables and (2) accounting for control-group effects. Our method substantially increases the accuracy of GPT-2 in three classification tasks (by 6-24% points) and reduces accuracy variance across different class distributions in prompts (by up to 14% points).[1]

Text: "horror"
GPT-2 prediction: *Positive* (87.92%)

Text: "If this were marketed as something other than a horror movie, i might have enjoyed it more."
Ground Truth: *Negative*
GPT-2 prediction: *Positive* (51.37%)
Our method: *Negative* (98.22%)

(a) IMDb Task.

Text: "never"
GPT-2 prediction: *Contradiction* (56.27%)

"yeah well that's really neat"
Hypothesis: "That's amazing, I've never seen anything like it.."
Ground Truth: *Neutral*
GPT-2 prediction: *Contradiction* (53.31%)
Our method: *Neutral* (66.11%)

(b) MultiNLI Task.

Figure 1: GPT-2 is biased to predict the positive class for texts containing "horror" (IMDb task) and to predict the contradiction class when the Hypothesis contains "never" (MultiNLI task). The score in each bracket indicates the model-predicted probability for the class. Labels marked as red indicate wrong predictions.

1 Introduction

In-context learning, a new paradigm in NLP for language models, uses pre-trained language models (PLMs) to conduct prediction tasks by providing only a few examples of input and output, leading to an extremely lightweight learning pipeline. It is critical to consider not only the task performance of this paradigm but also its robustness to the shift of data distributions. Prior work has shown the presence of confounding variables, influencing both the text features and the prediction labels, can degrade text classifier performance upon confounding shift (Landeiro and Culotta, 2016). However, the question of how to alleviate the influence of confounding variables in in-context learning is still under-explored. This paper investigates the negative impact of confounding variables in in-context learning and presents effective methods for controlling them by adapting a causal inference framework.

Confounding variables can deteriorate a prediction model upon confounding shift, that is, the probability distribution of output labels given a confounding variable differs between training data and test data. One of the main sources of such confounding shift in in-context learning is pre-training corpora. Many PLMs are pre-trained on large corpora, from which they implicitly pick up on the distribution of output labels in the presence of a confounding variable. For example, horror movies tend to receive higher ratings than non-

[1]We will release our code upon publication.

horror movies according to a large pool of reviews like IMDb (Landeiro and Culotta, 2016), and unsurprisingly, GPT-2 (Radford et al., 2019a) without any fine-tuning is biased to predict the positive class for reviews of horror movies as illustrated in Figure 1. Another potential source of confounding shift is prompts used in in-context learning. As we will see in our experiments (§5), the in-context learning accuracy of GPT-2 drops substantially as the confounding shift becomes larger between the examples in a prompt and a test set.

To alleviate the negative impact of confounding variables, we present a method inspired by causal inference (Pearl and Mackenzie, 2018). The key perspective of our work is that the common prompting practice is eliciting a summary of observational data (i.e. pre-training corpora) rather than measuring the true causal effect of an input text on output labels. We borrow the framework of Average Treatment Effect based on observational data to better measure the causal effect of input texts (§3). Our formulation identifies two important components: (1) prompting conditioned on confounding variables (§3.2) and (2) accounting for the effect of "control" group (§3.3). The two components together help alleviate biases derived from prompts and pre-training corpora.

We evaluate our method on three classification tasks that have known confounding variables (§4–5): (a) sentiment classification of IMDb reviews confounded by movie genre (horror vs. non-horror), (b) sentiment classification of IMDb reviews confounded by movie rating (R vs. non-R), and (c) natural language inference with use of negation words as a confounding variable. Our experiments with GPT-2 demonstrate that a standard prompting method is highly vulnerable to confounding shift in prompts. In contrast, our method achieves lower variance in accuracy across confounding ratios. Further, our method outperforms baseline models' accuracy by large margins (up to 24%) on average. We also demonstrate the effectiveness of our method in handling multiple confounding variables together. Our method was found to be not as effective for a larger model, GPT-3, and we discuss potential reasons and implications.

To summarize our contributions, our work is the first, to our knowledge, that uses a causal inference framework to alleviate the problem of confounding variables in in-context learning. Our method effectively increases the prediction accuracy of a moderately large PLM (GPT-2), making the model a more useful tool for lightweight prediction tasks.

2 Related Work

In-context Learning with Language Models PLMs can perform tasks in a zero- or few-shot manner using in-context learning (Radford et al., 2019a; Brown et al., 2020a). To do so, a natural language prompt is fed into the model, which contains three main components: a format, a set of demonstrations, and an ordering of those demonstrations (Zhao et al., 2021). GPT-2 (Radford et al., 2019b) and GPT-3 (Brown et al., 2020b) have demonstrated astonishing few-shot capabilities via in-context learning on various downstream tasks. Given only a natural language prompt and a few demonstrations of the task, they can make accurate predictions without updating any model parameters. Despite this remarkable performance, in-context learning has also shown some instability. For instance, GPT-3 yields high variance in accuracy depending on the selection and permutation of input examples in a prompt, as well as on the prompt format (Zhao et al., 2021), while input-label mappings in prompts have little impact on model performance (Min et al., 2022). Our paper addresses another aspect of their instability: in-context learning can yield high variance in accuracy upon confounding shift.

Confounding Bias Prior work (Zemel et al., 2013) has analyzed the model bias and algorithmic bias in modern machine learning to tackle the problems produced by data bias. Systems trained to make decisions based on historical data will naturally inherit the past biases, and these bias can blind the system to some attributes that account for predicting target class labels. Thus, the system might fail to make fair decisions; e.g., the system might be biased for some subgroups in a population (Zemel et al., 2013). The data bias exists due to the fact that supervised machine learning models are trained on historical data that itself might contain biases, including confounding bias, where there exists a confounding variable Z that influences both X and Y through distributions $P(X|Z)$ and $P(Y|Z)$. The presence of confounding bias is problematic since it affects the relationship between input features and output variables and can lead to spurious correlations (Elazar and Goldberg, 2018; Landeiro and Culotta, 2018; Pryzant et al., 2018; Garg et al., 2018).

Dataset Shift Dataset shift occurs when the joint distribution of features and labels differ between the training and testing sets (i.e. $P_{train}(X,Y) \neq P_{test}(X,Y)$) (Quionero-Candela et al., 2009). Prior work has investigated different dataset shift issues over the last few decades, including Covariate Shift where $P_{train}(X) \neq P_{test}(X)$ (Sugiyama et al., 2007; Bickel et al., 2009; Chen et al., 2016); Prior Probability Shift where $P_{train}(Y) \neq P_{test}(Y)$ (Webb and Ting, 2005); and similarly, Concept Drift where $P_{train}(Y|X) \neq P_{test}(Y|X)$ (Widmer and Kubat, 1994). In this paper, we specifically focus a type of dataset shift called Confounding Shift, where there exists a confounding variable Z that has correlations with both X and Y through distributions $P(X|Z)$ and $P(Y|Z)$, and there is a shift from $P_{train}(Y|Z)$ to $P_{test}(Y|Z)$ (Landeiro and Culotta, 2018). Previous work builds robust classifier under confounding shift (Landeiro and Culotta, 2016, 2018) for traditional machine learning algorithms. However, little attention has been paid to confounding shift in in-context learning via PLMs.

3 Proposed Methods

When predicting the label of an input text using in-context learning, we see it as rather eliciting a summary of observational data, i.e. pre-training corpora, because the model is not trained or fine-tuned to measure the causal effect of the input text on possible labels. To measure the causal relationship between the input and labels more accurately, we propose a method inspired by causal inference on observational data.

3.1 Problem Formulation

To motivate our method, let's use a simplified analogy of measuring the effect of a target drug on outcomes via observational data. The true causal effect may be measured by taking the effect observed from people who took the target drug, subtracted by the effect observed from people who took fake drugs. Further, since it is not a randomized test, each of the effects is measured conditionally on confounding variables so as to identify the causal effect with the resulting statistics. In our case, the target drug is an input text, and the outcome effects are the logits of possible labels. The fake drugs correspond to "control" texts (defined later). Considering the fake effect is important for addressing a human or model's inherent bias toward certain outcomes.

In causal inference (VanderWeele, 2013, 2016), one way to calculate causal effects is Average Treatment Effect (ATE). For in-context learning, let X denote the space of input texts (e.g., review texts), Y the outcome variable that represents the logit for an output label (e.g., 'positive' in sentiment classification), and Z a confounding variable (e.g., movie genre). For a certain text x, we want to predict its label by measuring its true causal effect on Y given Z. The ATE with the do-notation (Pearl and Mackenzie, 2018) can be computed as follows:

$$\begin{aligned} ATE &= \mathbb{E}\left[Y|do(x)\right] - \mathbb{E}\left[Y|do(x')\right] \\ &= \mathbb{E}_{z\sim Z}\left[\mathbb{E}\left[Y|x,z\right] - \mathbb{E}\left[Y|x',z\right]|z\right] \\ &= \mathbb{E}\left[Y|x,z^{test}\right] - \mathbb{E}\left[Y|x',z^{test}\right] \end{aligned} \quad (1)$$

where x' is a "control" text, and z^{test} is the confounder value of the test instance. The last equation holds because we compute the ATE for each test instance and only one instance at a time. We compute ATE for each output class, each of which is equated with the final logit for that class (§3.4). We choose logits rather than actual probability values as the outcome variable because logits can take any real number, whereas probability values require an additional constraint that they range between 0 and 1. In addition, using logits allows us to handle each class separately, whereas probability values incurs a dependency among the classes to form a simplex.

This equation is different from the common prompting method that feeds only texts (and possibly labels for demonstration), which can be considered $P(Y|x)$. The first main difference is that the two terms here are conditioned on a confounder value in addition to the input text. We will discuss one conditioning method based on confounder-aware prompts in Section 3.2. The second main difference is that the second term calibrates the observed effect of the input text by subtracting the observed effect of a control text. Subtracting this "control" effect alleviates the confounding shift between pre-training corpora and test data (i.e. the model's inherent bias). We will present several methods for obtaining control texts in Section 3.3.

3.2 Confounder-Aware Prompts

A straightforward way to estimate $\mathbb{E}\left[Y|x,z\right]$ in Equation 1 is to directly incorporate the confounder value z as a token into prompts because the probability of an output is conditioned on prompt tokens

for most generative models. An example format of the confounder-aware prompt is as follows:

> **Review:** I cannot recommend enough if you love horror!
> **Genre:** Horror
> **Sentiment:** Positive

"Genre: Horror" indicates the confounder value of this instance. More examples are in Figure 2.

3.3 Control Texts

To calculate $\mathbb{E}[Y|x', z]$ in Equation 1, we explore several methods for obtaining control texts x'. Most importantly, control texts should be *neutral* in that they should not have a causal effect on possible labels. Otherwise, it is difficult to measure the true effect of the input text. Further, we may compute $\mathbb{E}[Y|x', z]$ by averaging the observed effects of *multiple* control texts (rather than one) so that we can obtain the control effect more reliably against outlier texts.

Following this guidance, we propose four simple but effective methods. The first two methods sample the actual texts of instances in data, and the other two methods use tokens as control texts.

Neutral Class Random Sampling (NeutralClsRandSamp) When a neutral class is well-defined, we randomly select the same number of texts in the neutral class for each possible value of the confounding variable. For instance, for review sentiment classification, we select review texts that have a rating of 5 out of 10 for both horror movies and non-horror movies.[2]

Cross-Class Random Sampling (CrossClsRandSamp) We randomly select the same number of texts for every class and for every confounder value. The rationale is that the expected value of the effect would be neutral on average. A benefit of this method is that it works even if a neutral class is not well-defined (e.g. classification of age groups).

PMI(conf↑, class↓) For the remaining two methods, we use tokens obtained via Pointwise Mutual Information (PMI) (Church and Hanks, 1990) as control texts. Specifically, neutral tokens for a certain confounding value z would satisfy close-to-zero PMI(class) and high PMI(z). Hence, we extract 20 tokens that maximize: PMI(class)/|PMI(z)|, where z is the confounding value of each test instance. We provide the token lists selected by PMI in Table 13 in Appendix E.

PMI(conf↑) According to our observation, many confounder-related tokens (i.e. high PMI(conf)) that are supposed to be neutral (e.g. "HORROR" for movie reviews, "NOBODY" for natural language inference) indeed have high PMI(class). Hence, in this method, we ignore PMI(class) and extract four tokens that minimize: |PMI(z)| where z is the representative value of the confounding variable[3]. The rationale is that tokens that are most representative of a confounding variable are expected to be neutral. The tokens obtained via this method are shown in Table 4 for different classification tasks and confounding variables, and it shows that these tokens are reliably neutral. We verify that these tokens are useful as control texts in Appendix A.1. We also considered setting z the same as that of each test instance, but the neutrality of those tokens are questionable and they did not perform as well.

3.4 Final Logits

Here we describe how to compute the final logit of each class (i.e. ATE) using the confounder-aware prompts and control texts. The first term $\mathbb{E}\left[Y|x, z^{test}\right]$ in Equation 1 is simply the logit for the target class predicted by a language model (e.g. GPT-2) using a prompt that includes both the input text x and confounder value z^{test}. The second term $\mathbb{E}\left[Y|x', z^{test}\right]$ is the average of the logits that the model predicts the same way, except that each prompt now includes a control text instead of x. The difference between the two terms becomes the final logit for the target class.

Note that this can be seen as one way of calibrating model predictions (Zhao et al., 2021; Guo et al., 2017; Platt, 1999). The main difference, however, is that our work focuses on confounding variables (i.e. alleviating their negative impact) whereas prior work on calibration did not pay attention to confounding variables. Further, we interpret calibration from a causal inference perspective, proposing confounder-aware prompts for causal identification and suggesting new ways to obtain control texts for

[2] We also considered selecting texts only from the confounder value of each test instance, but it performed poorly. This is possibly because our method already conditions on the confounder value of each test instance via confounder-aware prompts or because selecting texts from one confounder value is not robust.

[3] In our experiments, these values are 'Horror' (genre) and 'R' (rating) for review sentiment classification, 'Negation' for NLI.

a calibration factor. As a result, our method outperforms a prior calibration method by large margins (Section 5).

4 Experimental Setups

Using three real-world datasets, we conducted experiments in which the relationship between the confounder Z and the class variable Y varies between training data and test data.

4.1 Train-Test Bias Ratio

To sample train/test sets with different confounding ratios $P(Y|Z)$, we assume we have labeled repository D_{train} and D_{test} where each instance (x_i, y_i, z_i) is a tuple of the input text, its label, and the confounder value. In k-shot learning cases, k instances are sampled from D_{train} without replacement. To simulate confounding shift, we follow a similar pipeline to Landeiro and Culotta (2016); we use different confounding ratio b for train and test sets and sample according to the following constraints:

- $P_{train}(Y|Z) = b_{train}$
- $P_{test}(Y|Z) = b_{test}$
- $P_{train}(Y) = P_{test}(Y)$
- $P_{train}(Z) = P_{test}(Z)$

We adjust b_{train} and b_{test} to simulate confounding shift between training and testing sets. The last two constraints ensure that we can interpret results as an effect derived purely by confounding shift rather than any other distributional changes.

4.2 Datasets

We evaluate our method on two datasets that have known confounding variables: IMDb, and MultiNLI.

IMDb The IMDb dataset (Pal et al., 2020) contains movie reviews labeled with ratings. It contains 17 genres, and each genre has a list of 100 movies. Following previous work on polarity classification (Maas et al., 2011), we consider reviews with a score ≤ 3 out of 10 as negative, and reviews with a score ≥ 7 out of 10 as positive.

We focus on two confounding variables. The first is **movie genre** (horror vs. non-horror) as tagged in the dataset. We set $Z = 1$ for horror movies and $Z = 0$ otherwise. Horror movies have been shown to be positively correlated with positive reviews (Landeiro and Culotta, 2016). Roughly 60.12% of reviews from horror movies are positive. The second confounding variable is **movie rating** (R vs. others), and 75.09% of reviews from R-rated movies are positive. We set $Z = 1$ for R-rated movies and $Z = 0$ otherwise. The statistics of data are shown in Table 10 in Appendix E.

MultiNLI The Multi-Genre Natural Language Inference (MultiNLI) corpus (Williams et al., 2018) is a collection of 433K sentence pairs annotated to determine whether a *hypothesis* is true (entailment), false (contradiction), or undetermined (neutral) given a *premise*. Prior work has shown the spurious correlation between contradiction labels and the presence of the negation words like *nobody, no, never*, and *nothing* in the hypothesis as significant annotation artifacts (Gururangan et al., 2018). Hence, our target confounding variable is whether these negation words are present in the hypothesis ($Z = 1$) or not present ($Z = 0$). We use the MATCHED test set as our validation (Gururangan et al., 2018). The statistics of data are in Table 11 in Appendix E.

4.3 Prompt Format and Details

The prompt format of each evaluation task is shown in Figure 2, and more examples are in Table 7 in Appendix B.

We vary the confounding ratio $b = P(Y|Z)$ from 0 to 1 for training but fix b_{test} to 0.5, i.e., $P_{test}(Y|Z) = 0.5$, to investigate how prediction accuracy is affected by the discrepancy of confounding ratios between training data and test data, i.e., across different values of b_{train}.

For each task, we use five random seeds for shuffling the order of training examples in a prompt to average out the ordering influence of prompting examples (Zhao et al., 2021). In addition, we prepare five different splits for each task and report the average accuracy.

We fix the number of examples in each prompt to ensure fair comparison throughout different tasks; each prompt uses 16 examples that encompass all possible output labels and confounding values. Under each setting, we fix $P(Y) = 0.5$ and $P(z) = 0.5$.

4.4 Comparison and Model Details

We compare our method to three baselines: Standard, CalibrateBeforeUse and GrIPS. **Standard** means prompts without confounder tokens and without subtracting the effect of control texts. **CalibrateBeforeUse** (Zhao et al., 2021) is a calibra-

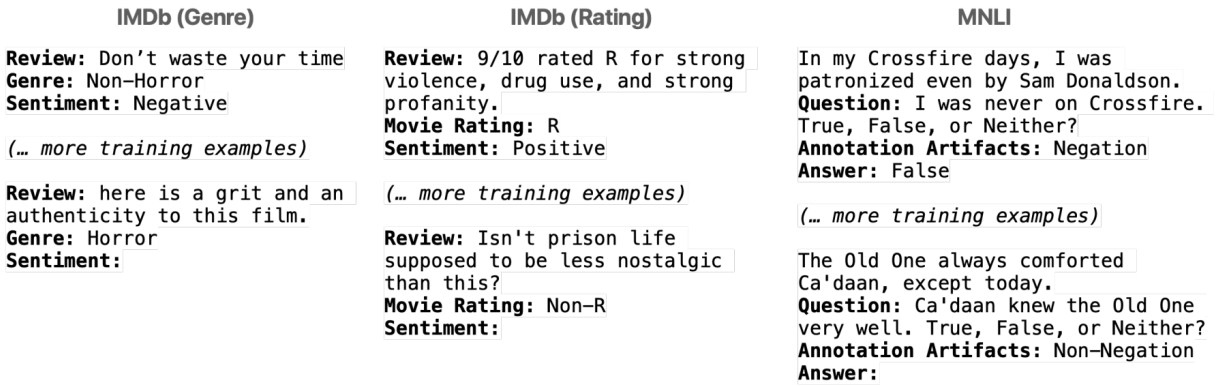


Figure 2: Prompt formats across evaluation tasks.

tion method where the probability distribution predicted by the language model using the standard prompt is scaled by the probability distribution via context-free tokens (i.e. "N/A", the empty string and "[MASK]"). **GrIPS** (Prasad et al., 2022) is a gradient-free, edit-based prompt search method, which takes in manually designed instructions and automatically returns an improved, edited prompt. The prompts in the search are iteratively edited using four operations (delete, add, swap, paraphrase) on phrase-level text. Our method, abbreviated as ***ConfAware***, is the "Confounder-Aware Prompts" defined in Section 3.2. We evaluate various combinations of the confounder-aware prompt and the four methods for obtaining control texts. As seed models, we use GPT-2 XL (1.5B parameters) and GPT-3 (Davinci Engines).

5 Results

5.1 Performance Under Confounding Shift

We investigate how different confounding ratios affect in-context learning performance, given a fixed number of demonstrations.

Table 1 shows that when b_{train} shifts, both *Standard* and *CalibrateBeforeUse* baselines have high variance across the shifts for all tasks. For instance, in IMDb review sentiment classification, the accuracy score of the *Standard* system drops from nearly 70% to 52% (row 1). In the MultiNLI task, the largest gap between the maximum score (37%) and the minimum score (26%) across b_{train} is around 11% (row 2). A large variance in accuracy across b_{train} from baseline models clearly indicates the potential for negative effects of confounders in in-context learning.

On the other hand, our methods demonstrate a clear improvement over the baselines. First, the variance in accuracy across different confounding ratios is substantially lower for IMDb Genre and IMDb Rating tasks. For instance, in the IMDb Genre task, the *ConfAware+CrossClsRandSamp* method (row 6) the highest and lowest accuracy scores have only a difference of 4.0 points, compared to 17.4 points for the *Standard* system, 11.4 points for the *CalibrateBeforeUse* system and 7.1 points for the *GrIPS* system. In addition, for the IMDb and MultiNLI tasks, our methods outperform the baseline systems by a large margin for most values of b_{train}. Among our methods, *ConfAware+CrossClsRandSamp* performs best overall for the IMDb tasks, and token-based control texts perform best for the MultiNLI.

5.2 Combination of Confounding Variables

Our method can naturally be extended to cases where multiple confounding variables exist. In this case, we add all confounders in prompts, and we extract control texts by considering all confounders. Table 2 presents the classification results of the IMDb task with both genre and rating as confounders. Considering both confounders (row 7) performs significantly better than the baselines across confounding ratios. However, it does not substantially improve over using only one of the confounders (row 5 to row 6). One potential reason is a strong correlation between the confounders; horror movies are likely to be R-rated. This correlation may not provide additive benefit to using both confounders together. We leave a more thorough analysis to future work.

5.3 Performance with GPT-3

To investigate the effectiveness of our model on larger models, we present classification results based on GPT-3 in Table 3. The IMDb task is relatively easy, and GPT-3 already achieves near-

	$B_{train} \rightarrow$	0.0	0.25	0.50	0.75	1.0
				IMDb Genre Confounding		
Baseline	*Standard*	$69.7_{2.4}$	$64.2_{4.1}$	$54.4_{2.8}$	$54.5_{2.5}$	$52.3_{1.4}$
	CalibrateBeforeUse	$72.7_{2.3}$	$71.0_{1.8}$	$61.3_{2.8}$	$67.7_{1.4}$	$69.8_{3.1}$
	GrIPS	$64.8_{4.4}$	$66.0_{2.6}$	$66.7_{3.2}$	$62.3_{4.8}$	$59.6_{3.3}$
Ours	ConfAware	$72.4_{2.0}$	$67.8_{5.6}$	$57.1_{2.8}$	$59.2_{3.8}$	$54.5_{1.8}$
	+NeutralClsRandSamp	$76.3_{1.1}$	$75.1_{1.6}$	$\mathbf{76.5}_{1.1}$	$69.9_{0.3}$	$\mathbf{76.1}_{1.2}$
	+CrossClsRandSamp	$\mathbf{77.4}_{0.9}$	$\mathbf{77.3}_{1.5}$	$\mathbf{76.5}_{1.6}$	$73.4_{1.1}$	$75.7_{0.9}$
	+PMI(conf↑, class↓)	$74.0_{1.1}$	$76.5_{1.3}$	$69.0_{2.5}$	$\mathbf{74.9}_{1.0}$	$70.8_{1.5}$
	+PMI(conf↑)	$72.6_{1.2}$	$74.2_{1.3}$	$70.5_{0.9}$	$70.2_{1.3}$	$68.4_{1.5}$
				IMDb Rating Confounding		
Baseline	*Standard*	$66.1_{3.5}$	$68.5_{4.2}$	$61.7_{4.7}$	$59.4_{4.2}$	$50.5_{0.5}$
	CalibrateBeforeUse	$67.3_{2.1}$	$71.6_{1.6}$	$72.4_{2.4}$	$66.0_{3.4}$	$60.2_{3.9}$
	GrIPS	$64.6_{7.5}$	$68.0_{2.7}$	$66.0_{9.3}$	$65.4_{5.6}$	$61.4_{5.9}$
Ours	ConfAware	$69.6_{3.2}$	$72.9_{2.3}$	$71.0_{5.1}$	$65.9_{3.7}$	$55.2_{3.8}$
	+NeutralClsRandSamp	$67.9_{1.3}$	$74.8_{0.7}$	$68.6_{1.2}$	$67.5_{2.3}$	$\mathbf{74.5}_{2.3}$
	+CrossClsRandSamp	$\mathbf{76.3}_{1.3}$	$\mathbf{76.5}_{0.9}$	$76.5_{1.3}$	$71.3_{2.3}$	$68.9_{1.6}$
	+PMI(conf↑, class↓)	$78.46_{1.4}$	$71.13_{2.8}$	$\mathbf{77.86}_{1.1}$	$\mathbf{74.4}_{2.7}$	$62.93_{1.9}$
	+PMI(conf↑)	$71.10_{2.4}$	$65.86_{3.6}$	$52.83_{0.9}$	$70.33_{2.1}$	$55.06_{1.3}$
				MultiNLI		
Baseline	*Standard*	$41.1_{2.1}$	$42.6_{2.2}$	$38.1_{3.7}$	$37.8_{2.9}$	$41.4_{3.9}$
	CalibrateBeforeUse	$29.3_{1.8}$	$35.0_{2.5}$	$26.1_{2.0}$	$30.7_{1.0}$	$36.8_{3.4}$
	GrIPS	$42.3_{4.2}$	$41.3_{5.5}$	$39.8_{5.9}$	$33.7_{4.4}$	$37.5_{5.2}$
Ours	ConfAware	$39.1_{0.4}$	$48.6_{0.4}$	$37.5_{3.1}$	$44.0_{2.2}$	$47.8_{0.3}$
	+NeutralClsRandSamp	–	–	–	–	–
	+CrossClsRandSamp	$40.5_{0.3}$	$47.8_{0.3}$	$40.7_{0.9}$	$37.3_{1.9}$	$47.0_{0.4}$
	+PMI(confounder↑, class↓)	$38.1_{0.7}$	$\mathbf{48.7}_{0.4}$	$33.6_{1.3}$	$37.1_{1.8}$	$47.1_{0.4}$
	+PMI(confounder↑)	$\mathbf{45.3}_{1.8}$	$48.0_{0.6}$	$\mathbf{42.3}_{1.8}$	$\mathbf{44.2}_{1.3}$	$\mathbf{48.7}_{0.3}$

Table 1: Averaged accuracies among five splits of different models across different confounding bias ratios (B_{train}) in prompts. The subscript following each accuracy score is the standard deviation. The best (highest) accuracy scores for a given B_{train} and task are in *bold*. The NeutralClsRandSamp method is computed only for the IMDb task because a neutral class is not well-defined for the other tasks.

$B_{train} \rightarrow$	0.0	0.50	1.0
Standard	$66.1_{3.5}$	$61.6_{4.7}$	$50.4_{0.4}$
CalibrateBeforeUse	$57.7_{2.8}$	$56.2_{1.2}$	$54.5_{2.1}$
GrIPS	$63.3_{3.6}$	$66.1_{4.5}$	$64.5_{5.0}$
CrossClsRandSamp			
with Genre	$\mathbf{77.4}_{0.9}$	$\mathbf{76.5}_{1.6}$	$\mathbf{75.7}_{0.9}$
with Rating	$76.3_{1.3}$	$\mathbf{76.5}_{1.3}$	$68.9_{1.6}$
with Genre & Rating	$74.8_{1.4}$	$73.5_{0.8}$	$66.7_{1.9}$

Table 2: Multiple confounding variables.

$B_{train} \rightarrow$	0.0	0.50	1.0
	IMDb Genre Confounding		
Standard	$92.5_{1.4}$	$94.16_{0.4}$	$97.5_{1.4}$
Ours	$91.7_{2.0}$	$90.3_{1.2}$	$96.3_{1.2}$
	MultiNLI		
Standard	$51.1_{2.2}$	$40.0_{0.1}$	$36.7_{3.3}$
Ours	$50.0_{4.9}$	$\mathbf{44.9}_{5.1}$	$34.0_{4.4}$

Table 3: GPT-3 vs. *ConfAware+PMI(Conf↑)*.

perfect accuracy scores across confounding shifts. For this task, there is little room for our method to improve the model further. For MultiNLI tasks, our method slightly improves the average accuracy but not by a large margin as in GPT-2. There could be a fundamental limit on the model's ability to solve these problems with only a few examples. In that case, it is difficult to expect that our method would increase the model performance drastically.

Furthermore, GPT-3 has been found to be less sensitive to the class distribution in the prompt (Min et al., 2022). Our experiments also show that GPT-3 has a lower variance in accuracy across confounding ratios; our intervention's lessened impact on GPT-3 may be attributed to this pre-existing low variance. We leave more investigation to future work.

While our method is less effective on GPT-3 than GPT-2 overall, it is important to note that GPT-3 requires much more computational resources, making it not as practical and accessible as smaller models like GPT-2. Thus, our method can provide significant benefits in many applications.

6 Discussion

The role of confounder values for in-context learning Prior work (Min et al., 2022) observed that for in-context learning, the distribution or order of output labels in a prompt has only a small impact on prediction accuracy. Our work further raises the question: does the distribution of confounding variables affect prediction accuracy in in-context learning, and if so, can this effect be mitigated? According to our experiments, surprisingly, some confounding values do have spurious correlations between inputs and inferences. For example, consider sentiment classification of movie reviews. The term 'horror' may be correlated with the positive class if the majority of horror movies have positive reviews. However, the term itself does not indicate a positive review. In other words, the positive prediction should not be attributed to the term 'horror.' Such spurious correlations can harm the PLMs' performance when there are distribution discrepancies between inputs and inferences. E.g., a PLM that has observed a spurious correlation between horror and positive reviews may see a new negative review of a horror movie and incorrectly classify it as positive. Our findings indicate that in-context learning with PLMs may suffer from degraded performance if confounding variables are not taken into consideration during prompting. We show that the degraded performance that comes with confounding shift can be mitigated by confounder-aware prompts.

Controlling confounding effects under in-context learning We want to study whether PLMs can learn the invariant features that attribute to the predictions while not be influenced much by confounder values under in-context learning. Our experiments indicate that some confounding values do influence prediction performance. To investigate whether this effect can be ameliorated through careful prompting, we redesign the prompts to model the relevant causal relationships with *ConfAware* prompts to distinguish spurious and genuine correlations between inputs and predictions. Table 1 shows that these *ConfAware* prompts can control the effects of bias in the confounding variables for in-context learning, resulting in more robust predictions.

7 Conclusion and Future Work

In this paper, we proposed a method to alleviate the negative impact of confounding variables in in-context learning. Our method is inspired by causal inference and has two important components—confounder-aware prompts and the effect of control texts. While confounding shift deteriorates the in-context learning accuracy of GPT-2 under standard prompting practice, our method is robust to changes in confounding ratios and outperforms baseline models often by a large margin. On the other hand, our model does not show a great improvement when a model is nearly perfect at a task (GPT-3 on sentiment classification), and when a model is insensitive to class distributions in prompts (GPT-3).

It is somewhat surprising that the existence of confounder values in a prompt has such a high impact on prediction accuracy, given that prior research found no strong relationship between prediction accuracy and class distribution in a prompt or true mapping of example instances to their classes. Our results imply that confounding values could be further investigated as a potential breakthrough to improve in-context learning once the benefit provided by the class labels of example instances plateaus.

There are various potential methods for selecting control texts. We randomly sampled texts but k-means clustering may be used instead so that control texts are sampled from diverse and representative clusters of input space. In addition, it would be an interesting direction to study evaluation metrics for control texts and the associations between the quality of control texts and prediction accuracy.

8 Limitation and Ethics

In this paper, we experimented only with binary confounding variables. Since our method uses confounder values as input tokens, it can be applied to both multi-class and continuous values of confounding variables in theory. However, the effective of the method in these cases (especially continuous confounding values) should be investigated empirically.

Unknown confounders are a well-known limitation of causal inference in general, and most theoretical and empirical work on causal inference assumes that all confounders are observed through domain knowledge. Our work focuses on a

new method for applying a causal inference framework to resolve negative confounding effects in in-context learning; finding confounders is a whole new problem and we leave it to future work.

Large PLMs have been found to have toxic biases against specific groups of people. As a result, the application of such models, including predictions tasks addressed in this paper, have the risk of reflecting the same biases. Our method would be beneficial for alleviating such biases.

References

Shiladitya Bhattacharya, Siddharth Singh, Ritesh Kumar, Akanksha Bansal, Akash Bhagat, Yogesh Dawer, Bornini Lahiri, and Atul Kr. Ojha. 2020. Developing a multilingual annotated corpus of misogyny and aggression.

Steffen Bickel, Michael Brückner, and Tobias Scheffer. 2009. Discriminative learning under covariate shift. *Journal of Machine Learning Research*, 10(75):2137–2155.

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

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

Xiangli Chen, Mathew Monfort, Anqi Liu, and Brian D. Ziebart. 2016. Robust covariate shift regression. In *AISTATS*.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

Yanai Elazar and Yoav Goldberg. 2018. Adversarial removal of demographic attributes from text data.

Sahaj Garg, Vincent Perot, Nicole Limtiaco, Ankur Taly, Ed H. Chi, and Alex Beutel. 2018. Counterfactual fairness in text classification through robustness.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Virgile Landeiro and Aron Culotta. 2016. Robust text classification in the presence of confounding bias. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, page 186–193. AAAI Press.

Virgile Landeiro and Aron Culotta. 2018. Robust text classification under confounding shift. *J. Artif. Int. Res.*, 63(1):391–419.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work?

Aditya Pal, Abhilash Barigidad, and Abhijit Mustafi. 2020. Imdb movie reviews dataset.

Judea Pearl and Dana Mackenzie. 2018. *The Book of Why: The New Science of Cause and Effect*, 1st edition. Basic Books, Inc., USA.

John Platt. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods.

Archiki Prasad, Peter Hase, Xiang Zhou, and Mohit Bansal. 2022. Grips: Gradient-free, edit-based instruction search for prompting large language models.

Reid Pryzant, Kelly Shen, Dan Jurafsky, and Stefan Wagner. 2018. Deconfounded lexicon induction for interpretable social science. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1615–1625, New Orleans, Louisiana. Association for Computational Linguistics.

Joaquin Quionero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence. 2009. Dataset shift in machine learning.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019a. Language models are unsupervised multitask learners.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019b. Language models are unsupervised multitask learners.

Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Müller. 2007. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(35):985–1005.

T. J. VanderWeele. 2013. A three-way decomposition of a total effect into direct, indirect, and interactive effects. *Epidemiology (Cambridge, Mass.)*, pages 224–32.

T. J. VanderWeele. 2016. Explanation in causal inference: developments in mediation and interaction. *International journal of epidemiology*, pages 1904–1908.

Geoffrey I. Webb and Kai Ming Ting. 2005. On the application of roc analysis to predict classification performance under varying class distributions. *Machine Learning*, 58:25–32.

Gerhard Widmer and M. Kubat. 1994. Learning in the presence of concept drift and hidden contexts. *Machine Learning*, 23.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning fair representations. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 325–333, Atlanta, Georgia, USA. PMLR.

Tony Z. Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. *arXiv preprint arXiv:2102.09690*.

A Bias Score

A.1 Justification of our Confounding Tokens

This section justifies that our confounding token list selected by PMI is valid enough given the change of bias scores we designed.

We vary the confounding ratio $b = P(Y|Z)$ for demonstration examples to study how distribution of confounding labels influence the system's prediction behavior on our PMI(confounder↑) tokens.

We set up three scenarios based on confounding bias ratio we defined previously:

- all samples are from non confounding group in the demonstration, we treat it as the baseline
- $b_{train} = 0$, i.e., $P_{train}(Y = 1|Z = 1) = 0$, indicating that none of the samples with class label $Y = 1$ are from the confounding group
- $b_{train} = 1$, i.e., $P_{train}(Y = 1|Z = 1) = 1$, indicating that all the samples with class label $Y = 1$ are from the confounding group

Under each setting, we query GPT-2 the prediction class labels for our PMI(confounder↑) tokens. Ideally, the confounder-related tokens should be neutral to the class prediction labels, hence GPT-2 and GPT-3 would score those tokens as 50% Positive and 50% Negative for binary classification, which would yield nearly 0 Bias Score (defined below)

Bias Score We measure the model's bias for a certain class using the following bias score:

$$\Theta = \frac{\text{label\_prob}(Y = 1)}{1 - \text{label\_prob}(Y = 1)} - 1 \qquad (2)$$

We put the -1 term here since for neutral inputs, ideally the system should score 50% for label_prob$(Y = 1)$ for binary class labels case, hence the first term in Equation 2 should be as close to 1 as possible, then for the most neutral case, $\Theta = 0$, indicating the smallest bias for the class y=1 prediction. The more the model is biased to the certain class, the larger the gap between label_prob$(Y = 1)$ and $1 -$ label_prob$(Y = 1)$ is and the higher the bias scores are.

We compare the bias scores under three setting we state above. Bias Scores are presented in Table 4. When all the training samples with class label $Y = 1$ are from the confounding group $Z = 1$, the bias scores for IMDB genre confounder tokens are increased in column $b_{train} = 1$ compared to the *Baseline*, again, the *Baseline* means there are no samples are from confounding group in the demonstration. The increase in the Bias Score indicates higher class label ($Y = 1$) probabilities are assigned, which contradicts the 'neutral' attributes of those tokens. That's the reason why calibration needed for the effects of those tokens.

After our ConfAware prompts, the bias scores drop, which means the probabilities of prediction class labels become much more unified, then it can mitigate the bias effect for those tokens to make them neutral to the prediction class labels.

	Baseline	$b_{train} = 0$	$b_{train} = 1$
	bias scores	bias scores	bias scores
HORROR	0.18	0.25 ↑	4.22 ↑
ZOMBIE	0.07	0.20 ↑	1.48 ↑
CREEPY	0.05	0.31 ↑	3.38 ↑
SCARE	0.28	0.09	4.98 ↑
NOBODY	1.20	3.56 ↑	2.39 ↑
NOTHING	0.55	0.42	0.14
NEVER	0.76	0.18	0.52
NO	0.47	0.05	0.44
KUTIYA	0.33	1.27 ↑	0.12
HOMOSEXUALS	0.63	1.38 ↑	0.16
HOMOSEXUALITY	0.69	1.34 ↑	0.28
SUCHITRA	0.22	0.86 ↑	0.12

Table 4: Bias Scores for PMI(confounder↑) Tokens before ConfAware. ↑ means the Bias Score increase due to the b_{train} setting demonstration compared to the baseline.

	$b_{train} = 0$	$b_{train} = 1$
	bias scores	bias scores
HORROR	0.29	1.66 ↓
ZOMBIE	0.32	1.04 ↓
CREEPY	0.46	1.86 ↓
SCARE	0.47	1.89 ↓
NOBODY	3.64	1.28 ↓
NOTHING	0.74	0.06 ↓
NEVER	0.23	0.16 ↓
NO	0.52	0.82
KUTIYA	0.08 ↓	0.04 ↓
HOMOSEXUALS	0.08 ↓	0.37
HOMOSEXUALITY	0.08 ↓	0.28
SUCHITRA	0.15 ↓	0.28

Table 5: Bias Scores for PMI(confounder↑) Tokens after ConfAware. ↓ means the Bias Score decrease after ConfAware conditioning on the confounding labels in the demonstration.

B Default and ConfAware Prompt Formats

Tables 6 shows the default prompt format used for all tasks. Table 7 shows the format for adding confounder-related prefix.

Task	Prompt	Prediction Label Names
IMDb +Genre	Review: Definitely a must-see for any horror movie fanatic. Sentiment: Positive Review: I can't believe disney accepted this as the final draft. Sentiment:	Positive, Negative
Task	**Prompt**	**Prediction Label Names**
IMDb +Rating	Review: 9/10 rated R for strong violence, drug use, and strong profanity. Sentiment: Positive Review: I don't think it's good for children seeing how it's a relatively violent movie. Sentiment:	Positive, Negative
Task	**Prompt**	**Prediction Label Names**
IMDb +Genre +Rating	Review: There's plenty of killing and blood but also some nice humorous bits and a good plot line drives this film through and i highly recommend watching it even if you're not a horror fan. Sentiment: Positive Review: To recapitulate, my family and i agree this movie deserves a thumbs down. Sentiment:	Positive, Negative
Task	**Prompt**	**Prediction Label Names**
MNLI +Negation	For instance, when Clinton cited executive privilege as a reason for holding back a memo from FBI Director Louis Freeh criticizing his drug policies, Bob Dole asserted that the president had no basis for refusing to divulge it. question: Bob Dole stated that Clinton had no right to privilege for actions not involving the presidency. True, False, or Neither? answer: Neither And not only is it you know trouble to have to drive but it takes time away from your home and your family when you're out driving. question: Driving is a fast experience with no downfalls. True, False, or Neither? answer:	True, False, Neither
Task	**Prompt**	**Prediction Label Names**
HateSpeech +misogynous	Sentence: Ever wondered why so many homosexual pedophilia takes place in hollywood? Aggressive: True Sentence: Judicious stap by Guy,May God keep him safe. Aggressive:	True, False

Table 6: The standard prompts used for our datasets. We show only one input example here for illustration purpose. We query the PLMs and check the probability for the corresponding label names.

Task	Prompt	Prediction Label Names
IMDb +Genre	Review: Definitely a must-see for any horror movie fanatic. Genre: Horror Sentiment: Positive Review: I can't believe disney accepted this as the final draft. Genre: Non-Horror Sentiment:	Positive, Negative
Task	**Prompt**	**Prediction Label Names**
IMDb +Rating	Review: 9/10 rated R for strong violence, drug use, and strong profanity. Movie Rating: R Sentiment: Positive Review: I don't think it's good for children seeing how it's a relatively violent movie. Movie Rating: Non-R Sentiment:	Positive, Negative
Task	**Prompt**	**Prediction Label Names**
IMDb +Genre +Rating	Review: all shrieks aside, this film made me feel like a vampire in a blood bank. Genre: Horror Movie Rating: R Sentiment: Positive Review: To recapitulate, my family and i agree this movie deserves a thumbs down. Genre: Non-Horror Movie Rating: Others Sentiment:	Positive, Negative
Task	**Prompt**	**Prediction Label Names**
MNLI +Negation	In my Crossfire days, I was patronized even by Sam Donaldson. question: I was never on Crossfire. True, False, or Neither? Annotation Artifacts: Negation answer: Neither 8. Jury Nullification. question: Annulment by the jury. True, False, or Neither? Annotation Artifacts: Non-Negation answer:	True, False, Neither
Task	**Prompt**	**Prediction Label Names**
HateSpeech +Misogynous	Sentence: Ever wondered why so many homosexual pedophilia takes place in hollywood? Gendered: True Aggressive: True Sentence: Judicious stap by Guy,May God keep him safe. Gendered: False Aggressive:	True, False

Table 7: ConfAware prompts used for our datasets, which condition on Confounding class labels to force in-context learning consider both the prediction labels and confounding labels.

C Effect of Number of Demonstrations

This section studies how the number of demonstrations affects GPT-2's performance under $b_{test} = 0.5$. We plot the accuracy score across different numbers of training examples in Figure 3. For IMDb and MNLI tasks, there are huge drops from 0-shot to 8-shot after we add more confounding labeled samples to the prompt demonstrations, which might reflect that the labels do matter in demonstrations and can influence the in-context learning performance. We speculate that the seemingly decreasing performance with small sample sizes is because in-context learning is highly sensitive to prompts and test sets when examples are only a few (Zhao et al., 2021). In Figure 3, the accuracy of the models is rather fluctuating than monotonically decreasing for small sample sizes, which probably represents the models' high sensitivity to the examples in prompts and the specific test sets chosen.

When there are zero demonstration samples, the accuracy of the *ConfAware+PMI* method is significantly higher than the accuracy of the GPT-2 baseline for IMDb Genre task. Even if there are zero example instances in the prompt, there are still the effects of confounder-aware prompting (i.e., specifying the confounder value of a test instance) and adjustment of the control effect (through a PMI-based control text). Both seem to benefit *ConfAware+PMI*.

From Figure 3 we see that conditioning on confounding variables alone not enough for good performance, and we should further address the effect of control texts.

D HateSpeech Task

HateSpeech The HateSpeech task is to predict the aggression level of speeches from social media (Bhattacharya et al., 2020), where each speech is labeled with either Overly Aggression (OAG), Covertly Aggression (CAG), or Non-Aggression (NAG). We combine OAG and CAG groups into Aggression and treat this task as a binary classification. The texts are also labeled as Misogyny or not. Misogyny indicates that aggression is aimed at gender roles, one's sexuality, or sexual orientation. Our confounding variable is whether a speech is misogynous ($Z = 1$) or not ($Z = 0$). The statistics of data are in Table 12 in Appendix E.

Table 8 and Figure 3 indicates that our methods are not as effective for the HateSpeech task. Given that the accuracy of the baseline systems are near random, GPT-2 seems to have no ability to solve this task and accordingly our methods do not add much benefit.

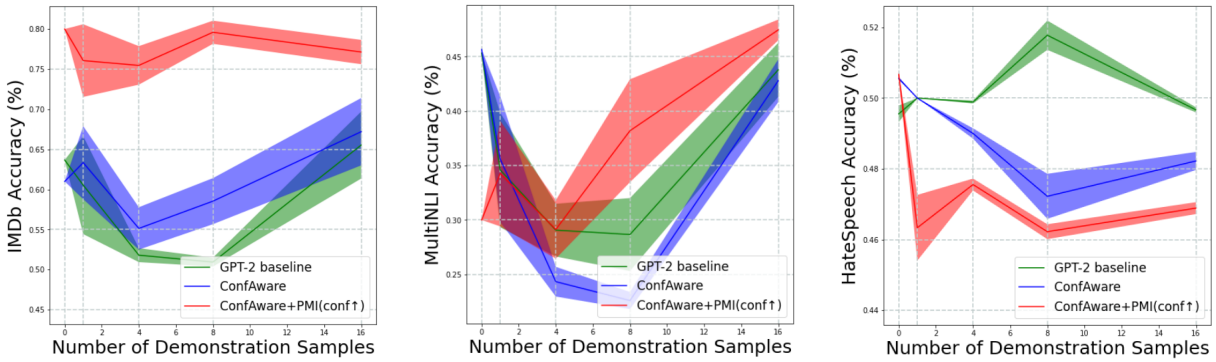
$B_{train} \rightarrow$	0.0	0.50	1.0
HateSpeech			
Standard	$50.2_{0.2}$	$47.9_{0.8}$	$48.7_{0.3}$
CalibrateBeforeUse	$48.6_{0.6}$	$\mathbf{51.2}_{1.1}$	$\mathbf{52.1}_{0.7}$
ConfAware	$50.4_{1.2}$	$49.6_{0.8}$	$51.5_{0.8}$
+NeutralClsRandSamp	–	–	–
+CrossClsRandSamp	$48.1_{1.2}$	$47.2_{0.6}$	$45.2_{0.7}$
+PMI(conf↑, class↓)	$\mathbf{51.3}_{0.5}$	$49.1_{0.5}$	$47.8_{0.8}$
+PMI(conf↑)	$50.0_{0.6}$	$48.4_{0.4}$	$47.6_{0.9}$

Table 8: HateSpeech Task.

$B_{train} \rightarrow$	0.0	0.50	1.0
HateSpeech			
Standard	$57.5_{3.2}$	$60.0_{5.0}$	$55.0_{2.8}$
Ours	$\mathbf{62.5}_{2.5}$	$56.3_{7.1}$	$\mathbf{57.5}_{1.4}$

Table 9: GPT-3 vs. *ConfAware+PMI(Conf↑)*.

According to our experiments so far, our method is effective at least when the standard prompting method achieves a 15% relative gain over random. For example, in HateSpeech, our method was not effective with GPT-2, where the standard method has an average accuracy of 50%, but it shows effectiveness with GPT-3, where the standard method has an average accuracy of 57.5%.



(a) IMDb with Genre as confounder

(b) MultiNLI with negation words in hypothesis as confounder

(c) HateSpeech with misogyny as confounder

Figure 3: Accuracy scores across different numbers of training examples.

reviews(positive; Horror)	34546
reviews(negative; Horror)	17950
reviews(neutral; Horror)	4960
reviews(positive; Non Horror)	715978
reviews(negative; Non Horror)	174308
reviews(neutral; Non Horror)	60321
reviews(positive; R)	328605
reviews(negative; R)	83920
reviews(neutral; R)	25062
reviews(positive; non R ratings)	421922
reviews(negative; non R ratings)	108338
reviews(neutral; non R ratings)	40219

Table 10: Number of reviews for each corresponding confounding group for positive, negative and neutral sentiment labels.

samples(contradiction; Negation)	21399
samples(neutral; Negation)	3765
samples(entailment; Negation)	2857
samples(contradiction; Non Negation)	109504
samples(neutral; Non Negation)	127135
samples(entailment; Non Negation)	128042

Table 11: Number of reviews for each corresponding confounding group for contradiction, neutral and entailment labels.

E Dataset Statistics and PMI Tokens

samples(Aggressive; Gendered)	1071
samples(Non-Aggressive; Gendered)	537
samples(Aggressive; Non Gendered)	3050
samples(Non-Aggressive; Non Gendered)	5378

Table 12: Number of samples for each corresponding confounding group for Aggressive, Non-Aggressive labels.

F Samples Illustrating the Efficiency of Our Method

Task	PMI(↑, class↓)	PMI(↑)
IMDb +Genre	"shock", "scary", "scared", "genuinely", "creepy", "flicks", "zombie", "flick", "street", "blood", "shocked", "jump", "evil", "monster", "fans", "tension"	"horror", "zombie", "creepy", "scare"
IMDb +Rating	"kingsman", "debauchery", "fishburn", "hmong", "iggy", "prides", "penitentiary", "stieg", "stutter", "lili", "masse", "builder", "gopher", "pollack", "scrubs", "mistrust", "boong", "skype", "foxy", "herzog"	"well", "higher", "failed", "admit"
MNLI +Negation	"longer", "religion", "statute", "books", "wear", "taxes", "damage", "cycle", "lay", "difference", "administration", "homes", "knows", "allowed"	"nobody", "no", "never", "nothing"
HateSpeech +Misogynous	"nature", "kids", "allowed", "c", "e", "opposite", "ye", "h", "sexual", "tu", "dangerous", "created", "brothers", "according", "kung", "fine", "relationships", "understanding", "culture", "earth"	"kutiya", "homosexuals", "homosexuality", "suchitra"

Table 13: Top 20 words with highest PMI(hor) and closest-to-zero PMI(pos).

Task	Instances	Ground Truth	Orig Pred Prob	Orig Pred Label	New Pred Prob	New Pred Label
IMDb +Genre	If you want horror then look no further than the excorcist now that film even by today's standards is up there forget the blair stichers project.	*Negative*	75.01%	*Positive*	52.11%	*Negative*
	Waste of time, poor storyline, the creatures were funny not scary.	*Negative*	81.39%	*Positive*	50.03%	*Negative*
	But if you are a more hardcore star wars fan then i don't see how you can really enjoy this movie.	*Negative*	64.42%	*Positive*	58.01%	*Negative*
	Fairly original story based on what other movies i've seen overall- 7.4/10	*Positive*	56.74%	*Negative*	62.21%	*Positive*
	I wish more horror movies were this clever.	*Positive*	50.44%	*Negative*	53.30%	*Positive*
IMDb +Rating	This movie is nothing near the other classic stanely kubrick movies.	*Negative*	79.03%	*Positive*	52.02%	*Negative*
	Mr. woody allen, i'll go to your movies, but you'll have to seduce me.	*Negative*	73.38%	*Positive*	64.81%	*Negative*
	Enjoyed it tremendously because it was different and reasonably original.	*Positive*	54.94%	*Negative*	59.92%	*Positive*
	It is unbearable to watch at times, but that is as it should be.	*Positive*	52.48%	*Negative*	56.74%	*Positive*
	Strip it from the terrorising narative and violence, the only thing left is a use of mental disturbances.	*Negative*	83.27%	*Positive*	71.22%	*Negative*
MultiNLI +Negation	Nonetheless, the rationality of service tiers remains. **Hypothesis:** Rationality of service tiers continues on..	*Entailment*	57.01%	*Neutral*	37.69%	*Entailment*
	I knew him and liked and respected him. **Hypothesis:** I never knew the man..	*Contradiction*	62.21%	*Neutral*	46.52%	*Contradiction*
	Oh thank God i've never been to Midland **Hypothesis:** Midland is a crap hole so I am glad that I have never been there..	*Neutral*	60.40%	*Contradiction*	100.00%	*Neutral*

MultiNLI +Negation	The interior of the palace is very dark, and the use of flash is forbidden, so photographers should think twice before paying the extra fee for bringing in a camera or video equipment. **Hypothesis:** Think hard about whether or not you want to bring a camera, there is an extra fee and no flash allowed..	*Entailment*	48.87%	*Contradiction*	35.67%	*Entailment*
	It was deserved. **Hypothesis:** It was not deserved at all.	*Contradiction*	60.42%	*Entailment*	100.00%	*Contradiction*

Table 14: Examples showing the efficiency of our methods. Column 4 & 5 are the results from GPT-2 while Column 5 &6 shows the outputs from our methods. All the results are contextual based. Labels marked as red indicate wrong predictions.