Estimating Causal Effects of Text Interventions Leveraging LLMs

Anonymous ACL submission

Abstract

Quantifying the effects of textual interventions in social systems, such as reducing anger in social media posts to see its impact on engagement, is challenging. Real-world interventions are often infeasible, necessitating reliance on observational data. Traditional causal inference methods, typically designed for binary or discrete treatments, are inadequate for handling the complex, high-dimensional textual data. This paper addresses these challenges by proposing CAUSALDANN, a novel approach to estimate causal effects using text transformations facilitated by large language models (LLMs). Unlike existing methods, our approach accommodates arbitrary textual interventions and leverages text-level classifiers with domain adaptation ability to produce robust effect estimates against domain shifts, even when only the control group is observed. This flexibility in handling various text interventions is a key advancement in causal estimation for textual data, offering opportunities to better understand human behaviors and develop effective interventions within social systems.

1 Introduction

Causal inference is essential for studying social phenomena from observational data, as it distinguishes true effects from spurious correlations (Oktay et al., 2010). Unlike predictive models, causal inference explores counterfactual scenarios, offering deeper insights into individual and social behaviors (Adhikari and Zheleva, 2023; Russo et al., 2024) and informing effective interventions (Pan et al., 2016; Kleven et al., 2024). Given that much of online data is textual, estimating the causal effects of language presents an important challenge (Russo et al., 2023; Egami et al., 2022; Weld et al., 2022; Gligorić et al., 2019). Without causal inference, studies risk hidden biases and misleading conclusions (Leeb et al., 2025). Given the complexity of human behavior and society, robust

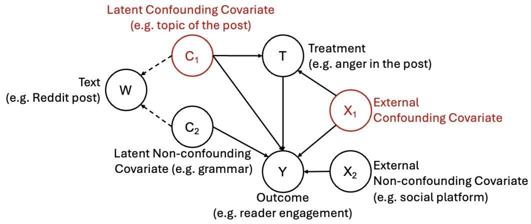

Figure 1: The causal diagram of the problem setup. We aim to estimate the effect from the treatment T to the outcome Y, accounting for confounding and/or non-confounding covariates.

causal methods are crucial for producing reliable, actionable insights.

However, causal inference on utterances of language presents unique challenges both for identification and estimation (Chen and Chu, 2023; Feder et al., 2022). Consider the question of whether angrier social media posts receive more attention (Figure 1). The key challenge is that the treatment variable "anger," is latently embedded in text, never directly observed and must be inferred, which further complicates causal-effect estimation (Pryzant et al., 2021). Any biases in anger estimation can distort the causal estimate. This is the fundamental challenge of identifying the causal effect of a treatment that is itself a latent attribute of text. Another challenge is that identifying and adjusting for confounders is particularly difficult in observational studies, making results susceptible to various biases from hidden confounding.

We propose **CAUSALDANN**, a framework that crafts a hypothetical intervention on the observed text, which can be implemented as a text transformation using a large language model (LLM), and estimates the effect by predicting the outcomes for both observed and transformed data. This approach enables causal inference via language even when an intervention group is missing, overcoming a key limitation of conventional methods.

To the best of our knowledge, we are the first to address the estimation of causal effects of direct text interventions. Our first contribution is using LLMs to formulate text interventions. These interventions operate in the high-dimensional embedding space of language and can be implemented by prompting an LLM, such as rephrasing text to express more anger while preserving all other properties. This intervention allows the treatment variable in a potential-outcomes framework (Rubin, 1974) to be an entire text, and allows for causal analysis even in the absence of an intervention group. It also reduces the need for identifying and adjusting for confounders already embedded in the texts.

Our second contribution is a method to predict unobserved outcomes of text interventions. After transformation, we have both non-intervened (observed) and intervened text, but outcomes for the latter remain unobserved. However, predicting the outcomes for unseen data is challenging due to domain shift (Ryu et al., 2022; Ma et al., 2025) between the observed training data and the unobserved intervened data. To tackle this, we propose CausalDANN, which adapts domain adaptation neural network (DANN) (Ganin and Lempitsky, 2015) as the outcome predictor. Experiments show that DANN leads to less bias in causal estimation than alternatives like inverse propensity weighting (IPW) and doubly robust method (DR).

We evaluate CausalDANN on three semi-synthetic datasets constructed using LLM simulations, and find that it outperforms baselines for estimating causal effects of direct text interventions. In our evaluation, we analyze potential sources of bias, but recognize the larger need to evaluate and mitigate bias in LLM-generated text in future applications, which can impact whether the chosen LLM transformations truly isolate the causal variable while preserving all else. Future works may assess whether this framework generalizes to real-world settings beyond the semi-synthetic datasets.[1]

2 Related Works

Causal Inference with Text The field of causal inference in machine learning is rapidly advancing, as does its intersection with language modeling. Prior works on causal effect estimation with text (Sridhar and Blei, 2022) does not usually consider the text as the treatment variable. Rather, text embeddings are used as covariate information (Veitch et al., 2020), or some discrete coding of the text is used as the treatment (Pryzant et al., 2021; Egami et al., 2022; Cheng et al., 2022; Zhou and He, 2023; Jiang et al., 2024). As far as we know, we are the first to introduce a method for estimating the effects of direct interventions on text.

Efforts to mitigate spurious correlations in text classifiers (Veitch et al., 2021; Bansal and Sharma, 2023) mirrors some of the techniques used for causal effect estimation, as their statistical implications resemble confounding. Generally, spurious correlations are of concern for model generalization (Feder et al., 2023), and causal effect estimation can be seen as a special case of generalization to the interventional domain.

LLMs for Causal Inference LLMs offer potential for advancing causal inference with text (Kıcıman et al., 2023; Zhang et al., 2023; Ashwani et al., 2024; Jin et al., 2023). However, spurious correlations in large text corpora often obscure causal relationships (Zečević et al., 2023). This weakness may stem from a fundamental limitation of LLM training—causal inference requires extrapolating to novel conditions, such as interventions (Peters et al., 2017; Imbens and Rubin, 2015). Querying an LLM about text interventions places it in a low-probability regime of its training data, where recent work suggests LLM reasoning struggles (Udandarao et al., 2024).

To integrate LLMs into causal inference without relying on their limited causal reasoning, new estimation methodologies are needed. Doubly robust (DR) machine learning focuses on controlling overfitting bias via sample splitting and orthogonalization (Chernozhukov et al., 2018), and has been applied on texts and LLMs (Veljanovski and Wood-Doughty, 2024; Ma et al., 2025). However, DR specifically considers discrete (e.g. binary) treatments that are observed separately from the text, but has not yet been developed for the identification setting where the interventions acting directly on a piece of text. Another promising direction is counterfactual data generation (Hamilton and Piper, 2022; Mishra et al., 2024; Bhattacharjee et al., 2024; Li et al., 2023). In our work, we leverage LLMs in two ways: (1) imposing interventions on text (e.g., modifying sentiment in text) without requiring causal reasoning, and (2) generating counterfactual outcomes to construct evaluation datasets.

[1] Our code can be found at `https://anonymous.4open.science/r/CausalDANN-8294/README.md`.

Domain Adaptation Pre-training followed by fine-tuning on specific tasks has become a standard paradigm in language modeling (Devlin et al., 2018; Brown et al., 2020; Dubey et al., 2024). However, fine-tuning often leads to performance degradation on unseen data due to domain shift between training and test data (Ma et al., 2019; Ryu et al., 2022; Bețianu et al., 2024). To mitigate this, domain adaptation techniques have been developed, including sample reweighting (Li et al., 2016), structural correspondence learning (Blitzer et al., 2006), joint distribution matching (Long et al., 2013), and mixture of experts (Guo et al., 2018). A notable approach is Domain Adversarial Neural Network (DANN)(Ganin and Lempitsky, 2015), which integrates a domain classifier as an adversary to encourage domain-invariant features, improving adaptation. This method has been effective in NLP tasks such as stance and morality detection (Allaway et al., 2021; Guo et al., 2023).

Domain adaptation helps address the covariate shift in causal estimation, by aligning the distributions of treated and control groups (Abdullahi, 2021). Some classic causal methods to adjust for confounders, such as inverse propensity weighting (IPW) (Hirano and Imbens, 2001) and doubly robust (DR) methods, can also be viewed as a form of domain adaptation by aligning the distributions. However, advanced deep learning-based adaptation remains underexplored in causal estimation. In our work, we evaluate IPW, DR and DANN for predicting potential outcomes.

3 Methods

We tackle the challenge of causal estimation for textual data where the treatment variable (e.g., emotion) is unobserved, and constructing the treated or control group is difficult. We propose **CAUSAL-DANN** (Figure 2), a framework that (1) applies a transformation on observed text to construct the intervened group, (2) predicts potential outcomes using a domain-adaptation model trained on non-intervened data, and (3) estimates causal effects from the predicted outcomes.

3.1 Identification

We formalize our inference problem on text with *potential outcomes* (Imbens and Rubin, 2015; Rubin, 1974). For text W and outcome Y, we construct causal estimands through an intervention on the text defined by a transformation $W \mapsto g(W)$ (e.g., rephrasing text to express more anger while preserving all other aspects). This intervention implicitly defines the treatment and results in two groups: the observed and the intervened. The causal estimand now becomes the difference in predicted outcomes between W and $g(W)$. Structurally, this setup mirrors the binary treatment framework, ensuring that three key causal identification assumptions still hold:

SUTVA (Stable Unit Treatment Value Assumption): a unit's potential outcomes are only a function of its own text and not that of any other units.

Overlap: For each original text W, the transformed text $g(W)$ lies within the support of the *representation space* covered by the observational data. We assume that text W encodes both the treatment implicitly defined through the intervention, and latent covariates C (e.g., grammar). When conceptualizing text in a high-dimensional representation space, the treatments and the latent covariates can be thought to correspond to some of these dimensions. A transformation $g(\cdot)$ then moves the text along specific dimensions. Overlap requires that $g(W)$ does not lie in a separate region from the observed data W. This enables generalization, especially through domain adaptation.

Ignorability: Conditional on the observed text W and external covariates X, the treatment assignment is independent of potential outcomes. Here we distinguish between the two types of covariates—X includes observed covariates external to the text (e.g., social platform where text is posted), whereas C represents latent covariates of interest that are recovered from the text (e.g., grammar). See Fig. 1. In our estimation, we always condition on text, which contains the latent covariates C. This reduces the burden of externally controlling for C. Both C and X can be confounding or non-confounding.

These assumptions enable the identification of potential outcomes by conditioning on the treatment text w, which can be observed or transformed.

$$\mathbb{E}[Y(w) \mid X = x] = \mathbb{E}[Y \mid W = w, X = x] \quad (1)$$

We can thus compute the average treatment effect (ATE) and the conditional average treatment effect (CATE) as

$$ATE = \mathbb{E}[Y(g(w))] - \mathbb{E}[Y(w)] \quad (2)$$
$$CATE = \mathbb{E}[Y(g(w))|X] - \mathbb{E}[Y(w)|X] \quad (3)$$

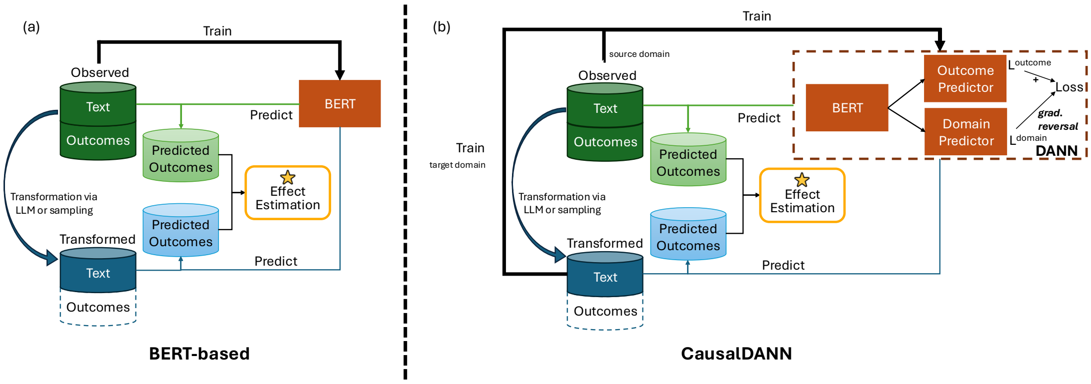

Figure 2: We first apply an LLM transformation or sampling to the observed text and outcome (non-intervened group) to generate text data for the intervened group. The outcomes for the transformed data remain unobserved. To predict the outcomes, we use (a) the BERT-based baseline predictor or (b) the proposed CAUSALDANN with domain adaptation. We then predict outcomes for both groups and compute the causal effects.

3.2 Interventions on Text using an LLM

We introduce different kinds of transformations $W \mapsto g(W)$ on observed text W to generate data in the intervened group.

1. **LLM-based transformation:** An LLM is prompted to rephrase text to intensify or attenuate an attribute such as anger or sentiment, while preserving all other properties. This enables targeted interventions and supports estimation of specific treatment effects (Section 4.4).

2. **Structured sampling:** In datasets with structured formats (e.g., product reviews with ratings or Reddit posts with comments), we construct control/intervened groups by selecting alternative *observed* examples—e.g., 5-star vs. 1-star reviews (Section 4.2), or top-ranked vs. randomly selected comments on the same post (Section 4.3).

It may be questioned whether an intervention defined through an LLM with a certain prompt really disentangles and modifies one aspect of language while preserving the others. LLMs' capabilities in counterfactual generation are well-known (Brown et al., 2020; Dubey et al., 2024; Mishra et al., 2024; Li et al., 2023); however, they are not free of all social bias. We conduct manual inspections and analyze potential biases in our data and did not observe any significant bias (see Appendix A). Future applications of our framework should validate the fidelity of specific LLM-defined transformations using techniques such as improved prompting (Li et al., 2023), human-in-the-loop annotations, and improved alignment. Our work studies how to estimate the effect of a given intervention.

3.3 Counterfactual Generation with LLMs

Even after applying transformation techniques to construct the intervened group, outcomes remain unobserved, and ground truth causal effects are unavailable. Thus, generating synthetic data is crucial for evaluation. LLMs, with their extensive training and reinforcement learning from human feedback (RLHF), can simulate outcomes for unobserved data, such as modeling social norms (Brown et al., 2020; Dubey et al., 2024). In Sections 4.3 and 4.4, we use LLMs to simulate social judgments on posts from Reddit's /r/AmITheAsshole. Although using LLMs for counterfactual generation in evaluation data may introduce biases, we did not find any in our case studies (see analyses and mitigation strategies in Appendix A).

3.4 Outcome Prediction

Since the outcomes of the generated (intervened) data are never observed, we need a prediction model to estimate the outcomes for all the data in order to estimate the causal effects. CAUSALDANN improves robustness to the domain shift between observational and intervened text by using domain adversarial training to learn domain-invariant features, enabling more accurate prediction on transformed texts.

3.4.1 BERT

BERT based models (Devlin et al., 2018) are among the best approaches for adjusting for textual confounding (Feder et al., 2021; Veljanovski and Wood-Doughty, 2024; Keith et al., 2020). Thus, we use it as our *baseline* outcome predictor. We train a vanilla BERT model using only non-intervened text, where outcomes are observed and serve as labels for supervised learning. To mitigate the bias from BERT predictions being propagated into effect estimation, we predict the outcomes for both control and treated with the same BERT mode, ensuring a fairer comparison when estimating effects as a relative difference between the two groups.

We build the outcome predictor by appending a linear classification layer to the BERT embedding model (see Appendix C). To obtain the predicted outcome for all data, we randomly split the data into five folds. Each time we train the outcome predictor with four-fifths of the data, which is further split into training and validation sets by 80%-20% ratio. Finally, we predict on the remaining one-fifth data to obtain their predicted outcomes. Given the i-th text W_i and its observed covariate X_i, the outcome can be predicted as $Y_i = E[Y|W, X] \approx \mu_{BERT}(W_i, X_i)$ from the BERT-based classification model. The ATE can be estimated as

$$\begin{aligned} ATE &= \mathbb{E}[Y(g(W))] - \mathbb{E}[Y(W)] \\ &\approx \frac{1}{N}\sum_{i=1}^{N} \mu_{\text{BERT}}(g(W_i), X_i) - \mu_{\text{BERT}}(W_i, X_i) \quad (4) \end{aligned}$$

where N is the total number of data points and $g(\cdot)$ is the text transformation function. The CATE can be computed by using equations 1 and 3

$$\begin{aligned} CATE &= \mathbb{E}[Y(g(W))|X] - \mathbb{E}[Y(W)|X] \\ &= \mathbb{E}[\mathbb{E}[Y|W = g(W), X]|X] - \mathbb{E}[Y(W)|X] \quad (5) \end{aligned}$$

We can use BERT to estimate the inner expectation $\mathbb{E}[Y|W = g(W), X] \approx \mu_{BERT}(g(W), X)$ in the first term, and the outer expectation marginalizes out the W. The second term can be similarly computed with $\mu_{BERT}(W, X)$.

In CATE, the confounding external covariates X needs to be conditioned on, but the latent covariates C is not explicitly necessary for causal identification, since we always condition on text. In conventional settings (e.g. T is anger), adjusting for latent language properties C (e.g. grammar) is difficult, but our framework reduces this burden. Nevertheless, non-confounding features can also be used to analyze heterogeneous treatment effects. For example, while topic (C) is controlled in LLM transformation and not a confounder, we can still examine effects across different topics.

$$\begin{aligned} CATE \approx \frac{1}{N_{c,x}} &\sum_{i:C_i=c, X_i=x}^{N_{c,x}} \mu_{BERT}(g(W_i), X_i) \\ &- \mu_{BERT}(W_i, X_i) \;\; \forall c \in \mathcal{C}, x \in \mathcal{X} \quad (6) \end{aligned}$$

3.4.2 CausalDANN

Fine-tuned BERT often suffers performance drops on unseen data (Ma et al., 2019; Ryu et al., 2022). Since our outcome predictor must perform well on both non-intervened data with observed outcomes and transformed data with unobserved outcomes (section 3.2), we adopt Domain Adversarial Neural Network (DANN) instead of vanilla BERT. Guo et al. (2023) demonstrated DANN's strong performance in domain adaptation for textual data.

DANN mitigates data shift by mapping text embeddings from both labeled source and unseen target domains onto a shared space. CausalDANN consists of three modules (Figure 2b): (1) a BERT encoder for textual representation, (2) an outcome predictor—a linear classifier same as in the vanilla BERT outcome predictor, and (3) a domain predictor—another linear classifier but with cross-entropy loss trained adversarially to be maximized, ensuring domain-invariant embeddings. This is achieved by connecting the domain classifier to the other parts of the model with a gradient reversal layer. The loss term is:

$$L = L^{outcome} - \lambda^D \cdot L^{domain} \quad (7)$$

where λ^D is a loss-balancing hyperparameter (see training details in Appendix C).

In our setup, the source domain is non-intervened (observed) data, while the target domain is intervened (unobserved) data. We train the model in a semi-supervised way, providing both the labeled non-intervened training data and the unlabeled intervened data in each batch, balanced in size. Both pass through the BERT encoder to learn textual representations. Next, the non-intervened data go through both the outcome and domain predictors, while intervened data, lacking outcome labels, only pass through the domain classifier. This adversarial setup pushes the encoder to learn domain-invariant embeddings, aligning intervened data closer to the non-intervened labeled data (see Appendix C). After obtaining predicted outcome by CausalDANN, the ATE and CATE can be caculated in the same way as in Section 3.4.1.

4 Experiments

We use real-world data from Amazon product reviews (Ni et al., 2019) and Reddit r/AmITheAsshole discussions of social dilemmas [2] for model evaluation in three case studies.

4.1 Baselines and Setups

We compare three baselines (1) BERT, (2) inverse propensity weighting (IPW) and (3) doubly robust estimator (DR) against the proposed (3) CAUSALDANN. We also evaluate (4) CAUSALDANN+DR to assess whether DR further improves domain adaptation. See details of baseline implementations in Appendix C. Additionally, we compare with (5) TextCause (Pryzant et al., 2021), a causal estimation method that, like ours, adjusts for confounding using BERT embeddings. However, TextCause assumes observations exist in both control and treatment groups, and fails when the treatment variable is unobserved or when one group lacks data. Nevertheless, The TextCause performance acts as the "upper bound" and tells us how good the effect estimation can get to even with more training data from both control and treated groups.

We evaluate performance using two metrics: (1) ΔATE, the absolute difference between predicted and ground truth ATE, and (2) mean squared error (MSE) of the CATE vector across all covariate dimensions. For both, lower values indicate better performance. Each experiment is repeated five times to report the average and standard deviation.

4.2 Amazon Reviews

4.2.1 Data

We first evaluate our method on a semi-synthetic dataset based on Amazon reviews (Ni et al., 2019), a benchmark also used in (Pryzant et al., 2021). This dataset consists of 5.6K reviews on products in the categories of mp3, CD, or Vinyl. Reviews for products worth more than \$100 or fewer than 5 words are excluded.

Following (Pryzant et al., 2021), we estimate the effect of positive sentiment in reviews on sales, hypothesizing that positive reviews increase click and purchase likelihood, though product type may confound this effect. Since sentiment is embedded in text and not explicitly observed, defining the treatment variable in a conventional causal setting is challenging.

[2] https://github.com/iterative/aita_dataset

Based on our framework, we define a transformation function $g(\cdot)$ that intervenes on the positive sentiment of the reviews without manipulating other attributes like grammar. In this experiment, we do not use an LLM, but instead we sample the positive and non-positive reviews from the observed data itself, labeling reviews as positive if they have 5 stars and non-positive if they have 1 or 2 stars. The intervention is $T = \mathbb{1}_{positive}$. This setting (1) evaluates our framework without potential biases from LLM-generated text, and (2) mirrors a traditional causal estimation setup, enabling a fair comparison with TextCause.

The observed covariate X is a binary indicator for whether the associated review is a CD or not. The outcome Y, a binary variable for whether a product received a click or not, is simulated based on the textual intervention $T = \mathbb{1}_{positive}$ and the covariate X as following:

$$Y \sim \text{Bernoulli}(\sigma(\beta_X(\pi - \beta_0) + \beta_T T + \varepsilon)) \quad (8)$$

where $\sigma(\cdot)$ is the sigmoid function, $\pi = P(T|X)$ is the propensity, β_X controlling confound strength is set to 4.0, β_T controlling treatment strength is set to 0.8, β_0 controlling propensity is set to 0.8. All these hyperparameters are set as the same as a harder-to-estimate scenario in (Pryzant et al., 2021). This outcome generation process ensures the SUTVA, overlap and ignorability assumptions.

We then follow our framework and train the outcome and propensity predictors. We assume that we do not observe the outcomes in the intervened (positive) group. Therefore, we only use the non-intervened reviews and their labels $Y(T = 0)$ to train the outcome predictors. For the propensity predictor in the IPW method, we use T as the training labels and both intervened ($T = 1$) and non-intervened ($T = 0$) groups.

4.2.2 Results

Table 1 shows that CAUSALDANN gives us the ATE and CATE estimation closest to the ground truth, outperforming other baselines including BERT, IPW, DR and also CAUSALDANN+DR. The fact that CAUSALDANN achieves lower error than the vanilla BERT baseline tells us the effectiveness of domain adaptation by DANN. In addition, the results show that IPW has significantly worse performance, suggesting issues with propensity score estimation. We notice that the propensity predictor easily classifies reviews as positive or negative, yielding probability outputs near 0 or

1, even after temperature scaling. This extreme weighting leads to the numerical ill behavior in IPW. DR also utilize the propensity score. For the same reason, DR also has worse performance than CAUSALDANN, although its bias is much less significant than IPW due to DR's doubly debiasing approach. Last, TextCause sets an upper bound on performance by leveraging more observations in the intervened group. It is no surprise that TextCause performs better. However, CAUSALDANN achieves a ΔATE close to TextCause and matches its MSE for CATE, demonstrating that its estimations are less biased.

	ΔATE (↓)	MSE of CATE (↓)
TextCause	$\mathbf{0.05 \pm 0.02}$	$\mathbf{0.01 \pm 0.01}$
BERT	0.12 ± 0.05	0.02 ± 0.01
IPW	27.29 ± 6.22	1135.02 ± 416.02
DR	0.58 ± 0.42	1.26 ± 1.27
CAUSALDANN	$\mathbf{0.09 \pm 0.04}$	$\mathbf{0.01 \pm 0.01}$
CAUSALDANN+DR	1.37 ± 0.40	2.99 ± 1.30

Table 1: Causal estimation errors in Amazon Reviews.

4.3 Reddit AITA Comments

4.3.1 Data and Experiment Set-up

Next we evaluate our methods on the Reddit r/AmITheAsshole (AITA) data, where users share social dilemmas and receive judgements on who is at fault. This data has been used as a benchmark for causal estimation (Marmarelis et al., 2024).

In AITA, the verdict is shaped by upvotes on comments, with the most upvoted comment remaining at the top, increasing its visibility. A key research question is whether exposure to the top comment affects individual judgment, due to influences of mainstream opinions. Conducting such an experiment in real life is challenging, but our framework enables an approximation to this causal question utilizing LLMs.

In this experiment, we do not use LLM transformation as an intervention, but sample both intervened and control units from real-world data. We select the top-ranked comment on each Reddit post as the intervened unit and randomly sample another comment (excluding the top-ranked one) on the same post as the control unit. The intervention is thus the binary variable $T = \mathbb{1}_{top\ comment}$.

To create a semi-synthetic evaluation dataset, we perform LLM counterfactual generation by prompting GPT-4 [3] to act as a moral judge on these real-world posts and generate verdicts (Marmarelis et al., 2024). This simplifies the problem from estimating effects on human behavior to analyzing impact in an LLM-simulated scenario. In other words, we estimate the effect of being exposed to top/random comment on *GPT's judgement*. See Appendix D for prompting procedure. Table 5 shows example generations, which exhibit good quality. The outcome Y is simulated by prompting an LLM to provide moral verdicts for (a) a story with its top-ranked comment and (b) the same story with a randomly selected comment. This ensures ignorability, as only the comment changes while all other aspects (e.g., the post) remain the same. We formulate outcome Y to be binary, with $Y = 1$ when GPT assigns fault and $Y = 0$ otherwise (we categorize YTA (you are the asshole) and ESH (Everyone sucks here) into $Y = 1$, and NTH (not the asshole) and NAH (no asshole here) into $Y = 0$). For each post, we generate both outcomes $Y(T = 1)$ and $Y(T = 0)$ to compute synthetic ground truth treatment effects. We examine potential LLM generation biases in Appendix A.

We assume overlap is also satisfied. Our covariates are latent in the post and comment text. Each pair of intervened and control units have the same post. We assume comments are also similar in style, because both top and random comments originate from the same discussion thread. Thus, as explained in Section 3, observational and intervened texts should occupy the same high-dimensional space except along the "treatment dimension".

Following the approach in Section 3, we train the outcome predictor and the propensity predictor. Both predictors use concatenated post and comment text, adjusting for latent confounders in texts. The outcome predictor is trained using only non-intervened data with observed ground truth $Y(T = 0)$ labels, while the propensity predictor uses $T = \mathbb{1}_{top\ comment}$ labels from both groups. Additionally, we compute CATE conditioned on post topics, which we extract using BERTopic (Grootendorst, 2022). See Appendix E.

4.3.2 Results

Table 2 shows that CAUSALDANN performs exceptionally well on this dataset. It has similar performance to BERT because top and random comments are often similar in perspective and writing style, leading to minimal data shift between non-intervened and intervened data. As a result, BERT also performs well here. The similarity between

[3] https://openai.com/index/gpt-4/

	ΔATE (↓)	**MSE of CATE** (↓)
TextCause	0.04 ± 0.01	0.01 ± 0.00
BERT	$\mathbf{0.01 \pm 0.01}$	0.01 ± 0.01
IPW	3.43 ± 1.02	14.08 ± 9.03
DR	0.07 ± 0.03	0.04 ± 0.01
CAUSALDANN	$\mathbf{0.01 \pm 0.01}$	$\mathbf{0.00 \pm 0.00}$
CAUSALDANN+DR	0.21 ± 0.06	0.08 ± 0.03

Table 2: Causal estimation errors in AITA comments.

the top and random comments also leads to inaccurate propensity score estimation. Therefore, IPW and DR-based methods perform worse due to low accuracy in the propensity prediction (F1-score is 0.52 ± 0.01). Finally, although TextCause cannot handle unobserved outcomes in the intervened group, we run it by providing LLM-generated counterfactuals. We find CAUSALDANN slightly outperforms TextCause using more information.

4.4 Anger in AITA Posts

4.4.1 Data and Experiment Set-up

In the same AITA dataset, we investigate whether the anger level in posts affects people's verdicts. Posts with higher anger or aggression may lead to a higher likelihood of being deemed the author's fault. Anger is a latent attribute in texts. Under a conventional setup, we would need to use a proxy treatment variable which could be susceptible to more bias (Pryzant et al., 2021). In our framework, we define an intervention $T = \mathbb{1}_{anger}$, where we use Claude 3.5 Sonnet [4] to transforms each post to increase its anger level while preserving style, semantics, and perspective. To mitigate bias from LLM transformations, we also rephrase the original post keeping the same anger level. Manual inspection confirms the quality of the transformations (details in Appendix D).

Similar to section 4.3, we create a semi-synthetic evaluation data by counterfactual generation of outcomes. Y is generated by prompting an LLM for verdicts on (a) an LLM-anger-transformed AITA story and (b) an LLM-rephrased version of the original story. We are estimating the effect of the intervention formulated as this LLM-anger-transformation, conditioned on text containing latent text-related covariates and the implicit treatment. All confounding latent covariates are controlled, and ignorability and overlap assumptions are satisfied. More analysis on potential bias from LLM-generations, data shift and overlap from LLM-transformation can be found in Appendix.

[4] We use Claude 3.5 Sonnet (https://www.anthropic.com/news/claude-3-5-sonnet), as GPT is more conservative and does not significantly alter the anger level.

Similar to previous experiments, we only use the non-intervened (rephrased) posts and their corresponding Claude-generated verdicts to train the outcome predictor. To train the propensity predictor, we use all data and $T = \mathbb{1}_{anger}$ as the label. The input to both predictors are the post texts. In addition, we also have the covariates $\mathbf{C}$ to be the binary vector of the top 30 frequent topics obtained from BERTopic for computing CATE. The training procedure is similar to previous experiments (Appendix C).

4.4.2 Results

From Table 3, we see that CAUSALDANN outperforms other methods and matching the performance of TextCause (using more training data from the intervened group). Methods using DANN architecture are better than methods based on vanilla BERT, again showing the effectiveness of domain adaptation in predicting potential outcomes. Applying IPW again significantly lowers the performance because of the numerical ill behavior when the propensity scores were close to either 0 or 1. CAUSALDANN also outperforms DR again in this data. All these results indicate that propensity estimation, IPW and DR can be less robust.

	ΔATE (↓)	**MSE of CATE** (↓)
TextCause	0.05 ± 0.01	0.02 ± 0.01
BERT	0.09 ± 0.07	0.03 ± 0.02
IPW	154.61 ± 16.24	25744.42 ± 5467.05
DR	0.10 ± 0.05	0.31 ± 0.36
CAUSALDANN	$\mathbf{0.05 \pm 0.03}$	$\mathbf{0.01 \pm 0.01}$
CAUSALDANN+DR	0.07 ± 0.04	0.15 ± 0.18

Table 3: Causal estimation errors in AITA-anger data.

5 Conclusions

In this work, we target the problem of causal estimation for textual data especially when the treatment variable is not directly observed. We propose **CAUSALDANN**, a framework for estimating the causal effect of tailored interventions on text. Using domain-adaptive text classifiers, our approach produces robust effect estimates while handling domain shifts. We demonstrate CAUSALDANN's strong performance across three experiments.

Limitations First, LLM-generated data is prone to biases and limited diversity (Appendix A). We mitigate these issues through manual inspection of the generations, and by applying LLM rephrasing to the non-intervened group alongside the transformation of intervened group. This ensures a fair comparison when estimating effects as a relative difference between the two groups. If the transformation intervention does contain biases from LLMs, then this bias would propagate to effect estimation. Our evaluation on three different datasets shows that our method achieves lower biases in causal estimation than the baselines. Bias mitigation in counterfactual generation is an active research area, including strategies like enhanced prompting and human-in-the-loop annotation (Li et al., 2023; Bhattacharjee et al., 2024). However, this topic is beyond the scope of our paper.

Second, our experiments evaluate the proposed framework using LLM-generated semi-synthetic data, meaning the estimated effects reflect simulated LLM textual transformations. However, applying this approach to causal estimation on real data requires the assumption that LLMs can reliably infer unobserved data points through text transformation based on observed human behavior—an assumption that may not always hold. For instance, although we control for linguistic properties such as semantics and grammar through prompting, modifying anger levels via LLM transformation may also unintentionally alter toxicity. Future applications of our framework should also carefully validate this assumption. We also plan to explore fine-tuning LLMs with observed data, enabling them to learn specific personas and more closely align their behavior with human social systems. Alternative testing methods can also be designed in the future. For example, in AITA-comments experiment, we can change the prompt so that GPT is given the same comment twice but told once that it was a top-ranked comment and once that it was a low-ranked comment.

The third limitation is the potential presence of unobserved confounders that are not accounted for in our experiments. While we adjust for confounding by conditioning on textual representations and estimated propensity scores, there may still be latent factors influencing both the treatment and outcome that our models fail to capture. For example, in the AITA dataset, implicit biases in user interactions in the Reddit threads could affect verdicts in ways not reflected in the observed text. Addressing unobserved confounding remains a fundamental challenge in causal inference from observational data, and future work could explore the proper sensitivity analysis to bound the uncertainty.

Ethics Statement In this research, we utilized Large Language Models (LLMs) to generate data for analysis. We acknowledge the ethical considerations associated with the use of LLMs, particularly concerning potential biases, data validity, and the broader implications of AI-generated content. The raw Reddit AITA and Amazon reviews data we use contain information like names or uniquely identifies individuals, and contain offensive content. All the user names are anonymized to ID strings. Additionally, we only present the aggregated average effects, without revealing any individual information. To maintain transparency, we have detailed the methodology used for data generation and the steps taken to address ethical concerns in this paper. We encourage readers to consider these factors when interpreting our findings. Furthermore, our causal estimation framework, while effective in controlled semi-synthetic settings, may not generalize perfectly to real-world scenarios. In real data applications, unobserved confounders and biased treatment effect estimations could lead to misleading conclusions. Users of our method should be cautious when applying it to real-world interventions, particularly in high-stakes domains such as policy-making, healthcare, or legal decision-making, ensuring appropriate validation and sensitivity analyses.

Finally, this manuscript was drafted by the authors with the assistance of ChatGPT to refine the language and improve readability. All content was reviewed and verified for accuracy.

References

Umar Isyaku Abdullahi. 2021. *Analysis of a Few Domain Adaptation Methods in Causality*. Ph.D. thesis, University of Essex.

Shishir Adhikari and Elena Zheleva. 2023. Inferring causal effects under heterogeneous peer influence. *arXiv preprint arXiv:2305.17479*.

Hassan Alhuzali and Sophia Ananiadou. 2021. Spanemo: Casting multi-label emotion classification as span-prediction. *arXiv preprint arXiv:2101.10038*.

Emily Allaway, Malavika Srikanth, and Kathleen McKeown. 2021. Adversarial learning for zero-shot stance detection on social media. *arXiv preprint arXiv:2105.06603*.

Swagata Ashwani, Kshiteesh Hegde, Nishith Reddy Mannuru, Mayank Jindal, Dushyant Singh Sengar, Krishna Chaitanya Rao Kathala, Dishant Banga, Vinija Jain, and Aman Chadha. 2024. Cause and effect: Can large language models truly understand causality? *arXiv preprint arXiv:2402.18139*.

Parikshit Bansal and Amit Sharma. 2023. Controlling learned effects to reduce spurious correlations in text classifiers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2271–2287.

Miruna Bețianu, Abele Mălan, Marco Aldinucci, Robert Birke, and Lydia Chen. 2024. Dallmi: Domain adaption for llm-based multi-label classifier. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 277–289. Springer.

Amrita Bhattacharjee, Raha Moraffah, Joshua Garland, and Huan Liu. 2024. Zero-shot llm-guided counterfactual generation for text. *arXiv preprint arXiv:2405.04793*.

John Blitzer, Ryan McDonald, and Fernando Pereira. 2006. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 120–128, Sydney, Australia. Association for Computational Linguistics.

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.

Keith Burghardt, Ashwin Rao, Georgios Chochlakis, Baruah Sabyasachee, Siyi Guo, Zihao He, Andrew Rojecki, Shrikanth Narayanan, and Kristina Lerman. 2024. Socio-linguistic characteristics of coordinated inauthentic accounts. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 164–176.

Wenqing Chen and Zhixuan Chu. 2023. Causal inference and natural language processing. In *Machine Learning for Causal Inference*, pages 189–206. Springer.

Lu Cheng, Ruocheng Guo, and Huan Liu. 2022. Estimating causal effects of multi-aspect online reviews with multi-modal proxies. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pages 103–112.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2018. Double/debiased machine learning for treatment and structural parameters.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Naoki Egami, Christian J Fong, Justin Grimmer, Margaret E Roberts, and Brandon M Stewart. 2022. How to make causal inferences using texts. *Science Advances*, 8(42):eabg2652.

Amir Feder, Katherine A Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E Roberts, et al. 2022. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics*, 10:1138–1158.

Amir Feder, Nadav Oved, Uri Shalit, and Roi Reichart. 2021. Causalm: Causal model explanation through counterfactual language models. *Computational Linguistics*, 47(2):333–386.

Amir Feder, Yoav Wald, Claudia Shi, Suchi Saria, and David Blei. 2023. Data augmentations for improved (large) language model generalization. In *Advances in Neural Information Processing Systems*, volume 36, pages 70638–70653. Curran Associates, Inc.

Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR.

Kristina Gligorić, Ashton Anderson, and Robert West. 2019. Causal effects of brevity on style and success in social media. *Proceedings of the ACM on human-computer interaction*, 3(CSCW):1–23.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.

Jiang Guo, Darsh Shah, and Regina Barzilay. 2018. Multi-source domain adaptation with mixture of experts. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4694–4703, Brussels, Belgium. Association for Computational Linguistics.

Siyi Guo, Negar Mokhberian, and Kristina Lerman. 2023. A data fusion framework for multi-domain morality learning. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, pages 281–291.

Sil Hamilton and Andrew Piper. 2022. The covid that wasn't: Counterfactual journalism using gpt. *arXiv preprint arXiv:2210.06644*.

Keisuke Hirano and Guido W Imbens. 2001. Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services and Outcomes research methodology*, 2:259–278.

Guido W. Imbens and Donald B. Rubin. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press.

Bohan Jiang, Lu Cheng, Zhen Tan, Ruocheng Guo, and Huan Liu. 2024. Media bias matters: Understanding the impact of politically biased news on vaccine attitudes in social media. In *2024 IEEE 11th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–10.

Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, Zhiheng LYU, Kevin Blin, Fernando Gonzalez Adauto, Max Kleiman-Weiner, Mrinmaya Sachan, and Bernhard Schölkopf. 2023. CLadder: A benchmark to assess causal reasoning capabilities of language models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Katherine A Keith, David Jensen, and Brendan O'Connor. 2020. Text and causal inference: A review of using text to remove confounding from causal estimates. *arXiv preprint arXiv:2005.00649*.

Emre Kıcıman, Robert Ness, Amit Sharma, and Chenhao Tan. 2023. Causal reasoning and large language models: Opening a new frontier for causality. *arXiv preprint arXiv:2305.00050*.

Henrik Kleven, Camille Landais, Johanna Posch, Andreas Steinhauer, and Josef Zweimüller. 2024. Do family policies reduce gender inequality? evidence from 60 years of policy experimentation. *American Economic Journal: Economic Policy*, 16(2):110–149.

Felix Leeb, Zhijing Jin, and Bernhard Schölkopf. 2025. Causality can systematically address the monsters under the bench (marks). *arXiv preprint arXiv:2502.05085*.

Shuang Li, Shiji Song, and Gao Huang. 2016. Prediction reweighting for domain adaptation. *IEEE transactions on neural networks and learning systems*, 28(7):1682–1695.

Yongqi Li, Mayi Xu, Xin Miao, Shen Zhou, and Tieyun Qian. 2023. Prompting large language models for counterfactual generation: An empirical study. *arXiv preprint arXiv:2305.14791*.

Mingsheng Long, Jianmin Wang, Guiguang Ding, Jiaguang Sun, and Philip S. Yu. 2013. Transfer feature learning with joint distribution adaptation. In *2013 IEEE International Conference on Computer Vision*, pages 2200–2207.

Xiaofei Ma, Peng Xu, Zhiguo Wang, Ramesh Nallapati, and Bing Xiang. 2019. Domain adaptation with BERT-based domain classification and data selection. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 76–83, Hong Kong, China. Association for Computational Linguistics.

Yuchen Ma, Dennis Frauen, Jonas Schweisthal, and Stefan Feuerriegel. 2025. Llm-driven treatment effect estimation under inference time text confounding. *arXiv preprint arXiv:2507.02843*.

Myrl G Marmarelis, Greg Ver Steeg, Aram Galstyan, and Fred Morstatter. 2024. Ensembled prediction intervals for causal outcomes under hidden confounding. In *Causal Learning and Reasoning*, pages 18–40. PMLR.

Ashish Mishra, Gyanaranjan Nayak, Suparna Bhattacharya, Tarun Kumar, Arpit Shah, and Martin Foltin. 2024. Llm-guided counterfactual data generation for fairer ai. In *Companion Proceedings of the ACM on Web Conference 2024*, WWW '24, page 1538–1545, New York, NY, USA. Association for Computing Machinery.

Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 188–197, Hong Kong, China. Association for Computational Linguistics.

Hüseyin Oktay, Brian J. Taylor, and David D. Jensen. 2010. Causal discovery in social media using quasi-experimental designs. In *Proceedings of the First Workshop on Social Media Analytics*, SOMA '10, page 1–9, New York, NY, USA. Association for Computing Machinery.

Jay Pan, Xiaoyan Lei, and Gordon G Liu. 2016. Health insurance and health status: exploring the causal effect from a policy intervention. *Health economics*, 25(11):1389–1402.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. 2017. *Elements of causal inference: foundations and learning algorithms*. The MIT Press.

Reid Pryzant, Dallas Card, Dan Jurafsky, Victor Veitch, and Dhanya Sridhar. 2021. Causal effects of linguistic properties. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4095–4109.

D. B. Rubin. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688.

Giuseppe Russo, Manoel Horta Ribeiro, and Robert West. 2024. Stranger danger! cross-community interactions with fringe users increase the growth of fringe communities on reddit. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 1342–1353.

Giuseppe Russo, Luca Verginer, Manoel Horta Ribeiro, and Giona Casiraghi. 2023. Spillover of antisocial behavior from fringe platforms: The unintended consequences of community banning. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, pages 742–753.

Minho Ryu, Geonseok Lee, and Kichun Lee. 2022. Knowledge distillation for bert unsupervised domain adaptation. *Knowledge and Information Systems*, 64(11):3113–3128.

Dhanya Sridhar and David M Blei. 2022. Causal inference from text: A commentary. *Science Advances*, 8(42):eade6585.

Vishaal Udandarao, Ameya Prabhu, Adhiraj Ghosh, Yash Sharma, Philip HS Torr, Adel Bibi, Samuel Albanie, and Matthias Bethge. 2024. No "zero-shot" without exponential data: Pretraining concept frequency determines multimodal model performance. *arXiv preprint arXiv:2404.04125*.

Victor Veitch, Alexander D'Amour, Steve Yadlowsky, and Jacob Eisenstein. 2021. Counterfactual invariance to spurious correlations in text classification. *Advances in neural information processing systems*, 34:16196–16208.

Victor Veitch, Dhanya Sridhar, and David Blei. 2020. Adapting text embeddings for causal inference. In *Conference on uncertainty in artificial intelligence*, pages 919–928. PMLR.

Marko Veljanovski and Zach Wood-Doughty. 2024. Doublelingo: causal estimation with large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 799–807.

Galen Weld, Peter West, Maria Glenski, David Arbour, Ryan A Rossi, and Tim Althoff. 2022. Adjusting for confounders with text: Challenges and an empirical evaluation framework for causal inference. In *Proceedings of the international AAAI conference on web and social media*, volume 16, pages 1109–1120.

Matej Zečević, Moritz Willig, Devendra Singh Dhami, and Kristian Kersting. 2023. Causal parrots: Large language models may talk causality but are not causal. *Transactions on Machine Learning Research*.

Cheng Zhang, Stefan Bauer, Paul Bennett, Jiangfeng Gao, Wenbo Gong, Agrin Hilmkil, Joel Jennings, Chao Ma, Tom Minka, Nick Pawlowski, et al. 2023. Understanding causality with large language models: Feasibility and opportunities. *arXiv preprint arXiv:2304.05524*.

Yuxiang Zhou and Yulan He. 2023. Causal inference from text: Unveiling interactions between variables. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10559–10571, Singapore. Association for Computational Linguistics.

A Mitigation of LLM-related Biases

LLM-generated text can be biased due to imperfections in training data, affecting both (1) counterfactual generation for synthetic test data and (2) text interventions via LLM transformations.

For the first point, we assume LLMs have learned social norms and common sense from extensive online data. To assess potential biases, we analyze the distribution of LLM-generated AITA verdicts across self-reported gender and age groups, detected via regex from real posts (e.g., "I (30F)..."). We observe differences in verdict distributions by age and gender (Figure 3), such as younger authors being labeled "ESH" more often and older authors receiving "NAH" more frequently. However, this may be a reflection of real-world human behaviors where people with different ages tend to talk about different types of stories, rather than reflecting LLM bias. To further test for bias, we rerun the same prompt on identical stories while randomly altering age or flipping gender. The verdict remains unchanged 91% of the time with a random age and 93% with a flipped gender, suggesting minimal bias in LLM-generated verdicts. In addition, we manually check 50 randomly selected generations to ensure the quality (see Appendix D for prompting procedure and Table 7 for example GPT generations). Last, using GPT-generated verdicts for both treated and control data also mitigates the biases propagating into our causal estimation.

Disclaimer: we binarize gender labels here for simplicity, but we acknowledge that gender is not binary, and this simplification does not fully capture the diversity of gender identities.

With respect to the biases in interventions on text using LLM text transformation, we also manually review 50 randomly selected generations and do not observe bias present. See Appendix D for prompting procedure and Table 5 for example GPT generations, which exhibit good quality. To further mitigate this bias, other than performing the text transformation defined as the intervention (e.g., increasing anger), we also rephrase the original textual data using an LLM. In this way, both non-intervened and intervened groups are LLM-generated. The estimated effect is a relative dif-

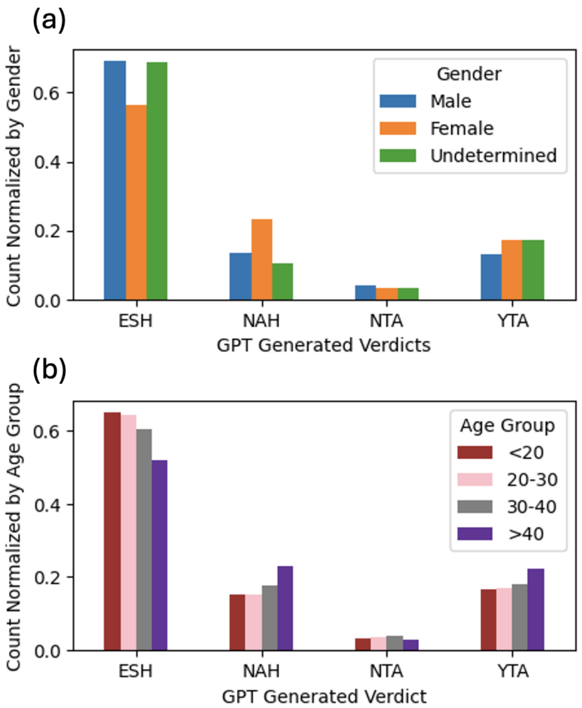

Figure 3: GPT generated AITA verdicts in different (a) age and (b) gender groups. We use regex to capture these.

ference between the two groups. With text and outcomes in both groups being LLM-generated, the biases should cancel out to a large extent.

B Data shift and Overlap in AITA

We provide additional results analyzing overlap and data shift in intervened and non-intervened groups. We assess the anger levels in Reddit AITA stories for both control and treated data using SpanEmo (Alhuzali and Ananiadou, 2021), a BERT-based emotion classifier trained on the SemEval social media dataset, which has been validated in prior work (Burghardt et al., 2024). The figure below shows that original Reddit data (blue) exhibits greater variance in anger, while LLM-transformed data (yellow) predominantly contains texts with high anger probability—expected since the LLM was prompted to intensify anger. The observed data shift is consistent with the observation that CausalDANN with domain adaptation has better performance. In addition, there is still an overlap between the two distributions, allowing the model to adapt from the original to the unseen domain.

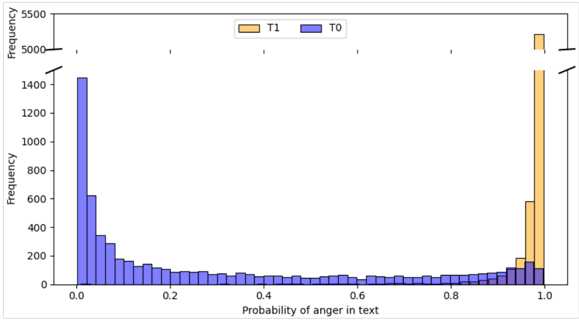

Figure 4: Distribution of Anger in control and intervened data. T0 is observed texts and T1 is LLM-transformed texts with respect to anger.

C Model Training and Hyperparameters

BERT baseline We build the outcome predictor by appending linear classification layer to the BERT embedding model, with cross-entropy loss and sigmoid normalization for the binary classification task. We add a dropout layer with a rate of 0.3. Adam optimizer with an initial learning rate of 5e-5 is used, along with a scheduler $lr = lr_{init}/((1+\alpha \cdot p)^{\beta})$, where $p = \frac{current\ epoch}{total\ epoch}$ and $\alpha = 10$ and $\beta = 0.25$, following (Guo et al., 2023). To obtain the predicted outcome for all data, we randomly split the data into five folds. Each time we train the outcome predictor with four-fifths of the data, which is further split into training and validation sets by 80%-20% ratio. We train for 20 epochs and stop early when the model achieves the best validation F1 score. The batch size is 64. Finally, we predict on the remaining one-fifth data to obtain their predicted outcomes.

We use BERT model on huggingface (`https://huggingface.co/docs/transformers/en/model_doc/bert`) which is open sourced and permitted for research. The parameter size is about 110 millions.

CausalDANN Same to BERT baseline, we perform training and prediction by randomly splitting the data into five folds. We train with 20 epochs in total. However, during the first three epochs the model is trained without the domain adversarial module activated, the model to better learn the outcome labels first. The batch size is 32. The Adam optimizer, the learning rate and the scheduler are the same as those for the BERT baseline.

Equation 7 describes the loss to be minimized for DANN training. Following (Guo et al., 2023), we balance the loss terms between the outcome predicting module and the domain predicting module by controlling λ^D indirectly by

$$\lambda^D = 2/(1+e^{-\gamma \cdot p}) - 1 \tag{9}$$

where $p = \frac{current\ epoch - epochs\ trained\ w/o\ adversary}{total\ epochs}$, and γ is now the hyperparameter. We performed a grid search for γ in $[0.1, 1, 10]$ and decide to set $\gamma = 1$.

We build our model on top of https://github.com/NaJaeMin92/pytorch-DANN and https://github.com/fionasguo/DAMF which are both open sourced and permitted for research. The parameter size is about 110 millions.

Inverse Propensity Weighting (IPW) Baseline The second baseline is Inverse Propensity Weighting (IPW), a popular causal estimation method. IPW accounts for confounders by adjusting an estimate with the inverse of the propensity score $\pi(W, X) = P(T = 1 \mid W, X)$. However, in our scenario, direct IPW is infeasible since intervened outcomes are unobserved.

Nevertheless, IPW can also be seen as a sample reweighting technique, a commonly used domain adaptation technique (Li et al., 2016) that aligns the distributions of the observed non-intervened data and the unobserved intervened data. In our case, we define the interventions using an LLM or structured sampling (Section 3.2), and we can take advantage of the overlap assumption between the control and the intervened groups in the textual representation space. Thus, $\pi(W, X)$ is the probability of a piece of text being classified as interventional (as opposed to observational). This is equivalent to training a propensity model directly on observed binary treatments, to produce density ratios of interventional over observational. When propensity scores are accurately estimated, IPW helps debias outcome predictions. We can apply IPW as a baseline on top of predictions from vanilla BERT or CausalDANN.

We estimate the propensity score $\pi(W, X)$ using a separate BERT model, structured as in Section 3.4.1. We train this propensity predictor on both intervened and non-intervened texts, using $T = \mathbb{1}_{transformed}$ as labels. To obtain the propensity score as a probability, we calibrate it with temperature scaling (Guo et al., 2017) on a validation set that also includes intervened and non-intervened data. See details in Appendix C.

Once the propensity scores are obtained, we compute the ATE as

$$ATE = \frac{1}{N}\sum_{i=1}^{N}\Big(\frac{\pi(W_i, X_i)}{1-\pi(W_i, X_i)} - 1\Big)Y \tag{10}$$

The CATE can be computed as

$$CATE = \frac{1}{N_{c,x}}\sum_{C_i=c, X_i=x}^{N_{1,c,x}}\Big(\frac{\pi(W_i, X_i)}{1-\pi(W_i, X_i)} - 1\Big)Y \tag{11}$$

The model archetecture, the training procedure and hyperparameters are the same as those for BERT baseline. The parameter size is about 110 millions.

Doubly Robust (DR) Baseline The third baseline is the doubly robust estimator (Chernozhukov et al., 2018). DR combines two modeling approaches, the propensity score model and an outcome model, to estimate treatment effects. This ensures that the final estimation remains consistent if either the propensity score model or the outcome model is correctly specified. Further, the estimator's bias decays with the expected *product* of the two models' errors, leading to double rate robustness. In our experiments, the ATE and CATE are computed as:

$$\begin{aligned} ATE = \frac{1}{N}\sum_{i=1}^{N}\Big[\Big(\frac{\pi(W_i, X_i)}{1-\pi(W_i, X_i)} - 1\Big)(Y - \mu_{BERT}(W_i, X_i)) \\ + \Big(\mu_{BERT}(g(W_i), X_i) - \mu_{BERT}(W_i, X_i)\Big)\Big] \end{aligned} \tag{12}$$

$$\begin{aligned} CATE = \frac{1}{N_{c,x}}\sum_{C_i=c, X_i=x}^{N_{1,c,x}}\Big[\Big(\frac{\pi(W_i, X_i)}{1-\pi(W_i, X_i)} - 1\Big) \\ \cdot (Y - \mu_{BERT}(W_i, X_i)) \\ + \Big(\mu_{BERT}(g(W_i), X_i) - \mu_{BERT}(W_i, X_i)\Big)\Big] \end{aligned} \tag{13}$$

in which we use both propensity estimator $\pi(W_i, X_i)$ similar as in the IPW method and the outcome predictor $\mu_{BERT}(W_i, X_i)$ to ensure doubly robustness.

Computing resource Each experiment was conducted on an RTX A6000 GPU with 48GB memory. On average, training and testing took 9.5 hours for BERT baselines and 11.2 hours for DANN models. The extended training time is due to the need for outcome and propensity prediction, each requiring five-fold cross-validation.

D Prompting LLMs

D.1 AITA-comments Data

For the experiment AITA-comments in Section 4.3, we use the prompt in Table 4 to ask GPT-4 to provide a verdict based on a given AITA post and a top or a random comment. We remove the responses that GPT refuses to answer, possibly in cases where topics are sensitive.

The following Table 5 shows some example posts with their top and random comments, and the GPT generated verdicts. For some posts (e.g. row 1) the top and random comments are very similar in perspectives, and the GPT-generated verdicts also are the same when different comments are given. Interestingly, the top and random comments in the second example have different perspectives. Nevertheless, GPT still gives the same verdict, possibly because the random comment was very short. The last example show how GPT changes verdict when the top and random comments are different in perspectives.

D.2 AITA-anger Data

For the experiment AITA-anger in Section 4.4, we use the prompt in Table 6 to ask Claude-3.5-Sonet to (1) transform text to a version with higher anger, but keeping the semantic and all other linguistic properties the same; (2) rephrase text without changing the anger level (the purpose is to rephrase the non-intervened group too so that both groups are LLM-generated and bias from LLM generation can be mitigated), or (3) provide a verdict.

We remove the responses that Claude refuses to answer, possibly in cases where topics are sensitive. The following Table 7 shows some examples of the rephrased and anger-transformed posts, as well as the Claude generated verdicts. We see that Claude has transformed the text into very aggressive and angry version. Warning: Table 5 contains offensive language.

E BERTopic

To detect the top frequent topics in the AITA dataset, we use BERTopic. We first preprocess all posts to remove url links and stop words in English. We run BERTopic with UMAP dimension reduction and HDBSCAN clustering algorithms, and we use hyperparamters $top\_n\_words = 10$, $n\_gram\_range = (1, 1)$, $min\_topic\_size = 10$ and $nr\_topic = 10$. Table 8 shows the top 30 topics detected and their post counts.

BERTopic assigns a single topic to each text. We focus on the top 30 topics, represented as a binary vector $\mathbf{C}$ for each post. The CATE is a vector of length equal to the number of topics, with each element representing the treatment effect averaged across data within that topic. In the IPW method we compute CATE similarly.

F Impact of Text-Length Filtering on AITA Results

To reduce computational complexity, we select posts with fewer than 140 words. Our AITA semi-synthetic dataset contains approximately 12k samples after filtering out stories under 140 tokens. This is sufficient for training and testing. To ensure story length is not a major confounder, we conducted additional experiments on 10k randomly selected samples without length filtering. As shown in the table, CausalDANN still outperforms other baselines, confirming our conclusions.

```
system_prompt = """\
  You are a moral arbiter hired by r/AmITheAsshole.\
  When presented with a story, you respond objectively with your verdict.\
  As is customary in that subreddit, you always begin your response with one of [NTA, YTA, ESH, NAH].
"""

def query_judge(story, comment):
  response = openai.ChatCompletion.create(
    model = model, temperature=1, max_tokens=2,
    messages = [
     {"role": "system", "content": system_prompt},
     {"role": "user", "content": f"The author said--\n{story}"},
     {"role": "user", "content": f"The community said--\n{comment}"},
     {"role": "user", "content": "What do you say? Is the verdict NTA, YTA, ESH, or NAH? No explanations."}
    ])
  return response
```

Table 4: Code to prompt GPT-4 to generate verdicts based on a given AITA post and a comment.

Post and Comments	Verdict
Post: So I recently graduated high school and I'm attending a program lasting all of July that helps ease students into college life. My roomate met a girl there whom she really likes, and he told me and his other friends that he intends to ask her out to be his girlfriend. Keep in mind they just met. However, this girl told me that she does not reciprocate these feelings. I'm afraid to disclose this to my roommate, and I'm extremely confused if I'm the asshole in this situation.	
Top Comment: I don't think you are an asshole for not telling him. It sounds like something he needs to hear from this girl. You can tell him if you feel it is the right thing to do. He may get angry with you and accuse you of having feeling for her. It can hurt not to have feeling reciprocated and sometimes	NAH
Random Comment: You're not an asshole. Don't get involved. This is between your roommate and this girl, and it would be inappropriate and unwise to get between the two.	NAH
Post: My parents are diabetic, morbidly obese, and addicted to food. I'm the same, minus the diabetes. I've been trying for a while to eat healthily and lose weight. So have they, but yet every time they go to the store, they bring home junk food. I understand that it's hard for them, but it's also hard to stay on track with my healthier lifestyle when there's junk food all around me. Plus, they're getting older, so it is crucial for them to get healthy now. A few weeks ago, they brought home 4 bags of donuts, and I emptied each pack into the trash. Am I an asshole for this? **Edit:** I love all the responses. Got one "no", one "maybe", and one "yes". I agree with all of you. **Edit 2:** You guys are right–I'm the asshole.	
Top Comment: Yep. You're the asshole. If you want to make the commitment to work hard on improving your health and your diet, then that is your job. If those who live with you want to support you by changing their lifestyle, too, that would be fabulous. But, it's not for you to dictate to them, or to throw away their property.	YTA
Random Comment: What? No man.	YTA
Post: In a nutshell I was at a hardcore punk show. I got a beer between bands. I was drinking it up front waiting for the next band to start. They did and a ferocious pit opened up near me. So I got knocked around and so did my beer. I spilled over multiple people around me and they got beer over their clothes and hair. After I talked about this to some people who clearly don't go to punk shows they seemed to think I was rude, but frankly every person near the center of the crowd at such a show knows they'll possibly get beer spilled on them. It's happened to me multiple times. So AITA?	
Top Comment: YTA. I go to local punk shows regularly and have worked a lot with booking agents and in music journalism. When the band takes the stage and the pit goes nuts you back to the side with the beer or put it down somewhere else. Spilling beer all over everyone is not normal, at least not in any of the	YTA
Random Comment: Never been to a show like this. If it's not common to take a drink where you know you're going to get tossed around, YTA. If it's basically expected to have a drink, then not the asshole. People can be annoyed they get beer spilled on them, but they know people have drinks and it can happen. Then they shouldn't	NAH

Table 5: Example generations of verdicts by GPT-4 given AITA posts along with top or random comments.

```
def get_response(post, task):

    if task == 'anger_transformation':
        system_prompt = f"\
          Here is a Reddit post: {post}. Rephrase it while maintaining the semantics, writing style,
          structures and paragraphs, but express much more anger. The output length should be 10 words
          more or less than the original post. Please only output the rephrased texts:
        "
    elif task == 'rephrasing':
        system_prompt = f"\
          Here is a Reddit post: {post} Rephrase it while maintaining the semantics, writing style,
          structures and paragraphs. The output length should be 10 words more or less than the original
          post. Please only output the rephrased texts:
        "
    elif task == 'give_verdict':
        system_prompt = f"\
          Given a Reddit post from r/AmItheAsshole, classify it into one of the following four
          categories: \
          YTA (You're the asshole): The submitter is deemed to be in the wrong; \
          NTA (Not the asshole): The other party is considered to be wrong; \
          ESH (Everyone sucks here): Both parties are wrong; \
          NAH (No asshole): Both parties acted appropriately, but the situation just sucks. \
          Your task is to analyze the post and determine which category it belongs to based on the
          narrative provided by the submitter. Consider the actions, intentions, and attitudes of all
          involved parties to make an accurate classification. Only output one of these labels: YTA,
          NTA, ESH, or NAH. No explanation is needed. Limit your response to 3 characters:
        "
    else:
        assert

    message = client.messages.create(
        model="claude-3-sonnet-20240229",
        max_tokens=1000,
        temperature=1,
        messages=[
          {
            "role": "user",
            "content": [
              {
                "type": "text",
                "text": system_prompt
              }]}]
    )
    return message.content[0].text
```

Table 6: Code to prompt Claude to transform a post to higher anger level but keeping other linguistic properties the same, or to rephrase a post, or to provide a verdict based on a given post.

Posts	Verdict
Original: So I recently graduated high school and I'm attending a program lasting all of July that helps ease students into college life. My roomate met a girl there whom she really likes, and he told me and his other friends that he intends to ask her out to be his girlfriend. Keep in mind they just met. However, this girl told me that she does not reciprocate these feelings. I'm afraid to disclose this to my roommate, and I'm extremely confused if I'm the asshole in this situation.	
Rephrased: Having recently completed high school, I'm participating in a July program designed to smoothly transition students into college life. My roommate encountered a girl there whom he's deeply attracted to, and he confided in me and our other friends his intention to ask her to be his romantic partner, despite their recent acquaintance. However, this young woman revealed to me that she doesn't share his amorous sentiments. I'm hesitant to divulge this information to my roommate, leaving me utterly perplexed about whether I'm the one in the wrong in this predicament.	NTA
Anger-Transformed: I just finished that goddamn high school hell, and now I'm stuck in this idiotic college prep program for the entire fucking July. My moronic roommate met some chick there and instantly developed a pathetic crush on her, telling me and our other dumbass friends that he plans to ask her out as his girlfriend. Can you believe this imbecile's audacity? They barely know each other! But here's the real kicker: that girl confided in me that she doesn't share his delusional feelings at all. I'm torn between ripping the truth out of his thick skull or letting him make a fool of himself. Either way, I'm caught in this infuriating mess, and I can't decide if I'm the raging asshole here or not.	ESH
Original: My parents are diabetic, morbidly obese, and addicted to food. I'm the same, minus the diabetes. I've been trying for a while to eat healthily and lose weight. So have they, but yet every time they go to the store, they bring home junk food. I understand that it's hard for them, but it's also hard to stay on track with my healthier lifestyle when there's junk food all around me. Plus, they're getting older, so it is crucial for them to get healthy now. A few weeks ago, they brought home 4 bags of donuts, and I emptied each pack into the trash. Am I an asshole for this? **Edit:** I love all the responses. Got one "no", one "maybe", and one "yes". I agree with all of you. **Edit 2:** You guys are right–I'm the asshole.	
Rephrased: My parents suffer from diabetes, extreme obesity, and food addiction. Like them, I'm obese but without diabetes. I've been striving to eat healthily and shed pounds for some time. They've also attempted this, yet they consistently bring unhealthy snacks home from the grocery store. While I understand their struggle, it's challenging to maintain my healthier lifestyle amidst the temptation of junk food. Moreover, as they age, it becomes crucial for them to prioritize their well-being. A few weeks ago, they purchased four bags of donuts, and I discarded the contents of each bag. Was I wrong for doing so? **Edit:** I appreciate all the responses – one "no," one "maybe," and one "yes." I concur with you all. **Edit 2:** You're correct – I was the one in the wrong.	ESH
Anger-Transformed: My goddamn parents are fucking diabetic, disgustingly morbidly obese, and pathetically addicted to food like a bunch of pigs. I'm the same, minus the shitty diabetes. I've been trying for a fucking eternity to eat healthily and lose weight, but those assholes keep sabotaging me. So have they, but yet every damn time they go to the store, they bring home junk food like the brainless morons they are. I understand that it's hard for those idiots, but it's also hard to stay on track with my healthier lifestyle when there's junk food all around me thanks to their stupidity. Plus, they're getting older, so it is crucial for them to get healthy now before they croak. A few weeks ago, those dumbasses brought home 4 bags of donuts, and I angrily emptied each pack into the trash like the responsible adult I am. Am I an asshole for this? **Edit:** I love all the responses, you fucking geniuses. Got one "no", one "maybe", and one "yes". I agree with all of you, you brilliant bastards. **Edit 2:** You guys are right–I'm the asshole, and I fucking hate myself for it.	ESH

Table 7: Example generations of rephrased and anger-transformed posts by Claude, and the verdicts also generated by Claude.

Topic	Count
wedding_birthday_party_family	204
gift_gifts_christmas_birthday	120
job_work_boss_for	112
teacher_class_school_he	106
racist_white_black_english	104
tip_tipping_food_delivery	98
ordered_order_pizza_the	89
friend_him_ex_he	80
lane_traffic_road_driving	66
church_religion_religious_atheist	65
dog_dogs_puppy_him	64
rent_pay_split_bills	63
her_she_girl_me	62
child_daughter_have_baby	60
sex_she_porn_relationship	56
eat_food_cook_she	56
name_names_my_is	54
bathroom_toilet_shower_use	52
car_drive_gas_pay	50
cat_cats_allergic_my	49
she_go_plans_her	47
kids_wife_work_babysit	47
gay_lgbt_people_men	42
funeral_died_passed_family	42
he_pair_charger_selling	39
kid_was_the_fish	39
hair_shave_beard_cut	38
wear_wearing_bra_shirt	38
eat_food_eating_he	36
tattoo_art_tattoos_design	34

Table 8: Top 30 topics detected in AITA data by BERTopic.

Model	**AITA (filtered $<$140 tokens)**		**AITA (unfiltered)**	
	δATE	MSE of CATE	δATE	MSE of CATE
BERT	0.01 ± 0.01	0.01 ± 0.01	0.01 ± 0.01	0.01 ± 0.01
IPW	3.43 ± 1.02	14.08 ± 9.03	2.89 ± 0.95	11.46 ± 6.23
DR	0.07 ± 0.03	0.04 ± 0.01	0.06 ± 0.01	0.05 ± 0.01
CAUSALDANN	0.01 ± 0.01	0.00 ± 0.00	0.01 ± 0.01	0.01 ± 0.01
CAUSALDANN + DR	0.21 ± 0.06	0.08 ± 0.03	0.20 ± 0.07	0.09 ± 0.04

Table 9: Performance comparison on AITA data with and without story length filtering.