Alternative Timelines: Generating Counterfactual Posts to Estimate Causal Effects in Social Media

Anonymous ACL submission

Abstract

We propose **CausalT5**, a generative framework for estimating causal effects in social media timelines. Given a user's posting history, the model estimates how an event at time i, for example, engaging with a low-credibility news outlet, influences subsequent posts. Building on neural architectures for causal inference, CausalT5 departs from outcome-only prediction by generating counterfactual post-treatment messages. A T5 language model is trained with three objectives: conditional generation of observed posts, treatment assignment classification, and outcome prediction via a differentiable count of attributes in generated posts. We evaluate CausalT5 on semi-synthetic data with known effects, finding that (a) generated posts are linguistically plausible and consistent with real post-intervention behavior, (b) CausalT5 estimates average treatment effects as accurately as strong outcome-prediction baselines, and (c) it captures heterogeneous effects and remains robust under topical shifts. These results suggest generative counterfactual modeling with CausalT5 is a promising tool for causal analysis of social media dynamics.

1 Introduction

Social media platforms provide an unprecedented lens into human behavior, offering fine-grained traces of how individuals and communities engage with informatione (Mendelsohn et al., 2023; Ganti et al., 2023; Pecile et al., 2025). Yet, while descriptive analyses of such data are abundant, identifying causal relationships underlying this data remains far more challenging. Most social media data are observational rather than interventional, meaning that user behaviors are shaped by many unobserved confounders such as political leaning, interests, or exposure to external events. Estimating cause-and-effect relationships in this setting requires methods that can account for such confounding while preserving the richness of language.

A growing body of work uses machine learning to estimate causal effects in text (Egami et al., 2018; Keith et al., 2020; Roberts et al., 2020; Grimmer et al., 2022), including with neural network approaches (Shi et al., 2019; Maiya, 2021; Koch et al., 2025; Li et al., 2025). While previous models achieve strong performance in semi-synthetic settings, they have focused on categorical or continuous scalar outputs rather than full text generation. This constrains interpretability and robustness. Without counterfactual text generation, it is difficult to judge whether estimated treatment effects align with realistic behavior. Outcome-only models also lack transparency: they output a scalar estimate without showing example posts that could plausibly occur under treatment or control, making validation and reasoning analysis difficult.

Advances in large language models (LLMs) offer a new opportunity. Recent work has shown that LLMs can faithfully simulate online interactions (Wang et al., 2024; Ng and Carley, 2025). In this paper, we investigate whether this emerging capability can be adapted to estimate causal effects. We propose **CausalT5**, a generative LLM framework for estimating treatment effects from observational social media data. CausalT5 builds on recent advances in causal representation learning but departs from outcome-only prediction by explicitly generating counterfactual text sequences for each user timeline. The model is based on a T5 LLM (Raffel et al., 2020; Chung et al., 2024) trained with three complementary objectives: (i) conditional generation of observed posts after treatment, (ii) treatment classification, and (iii) outcome prediction via a differentiable count of attributes (e.g., low-credibility news mentions) in the generated posts. This joint training setup ensures both linguistic fidelity and accurate causal estimation.

As a case study, we apply our approach to study how users engage with low-credibility news sources (those rated as "unreliable" by fact-

checkers). Specifically, we consider whether a user's first engagement with a low-credibility news source (i.e., the treatment) causally increases the number of low-credibility news sources mentioned in subsequent posts. Our approach is to fine-tune an LLM that, given pre-treatment posts from a user, simultaneously estimates the treatment propensity, generates counterfactual post-treatment posts, and provides outcome measures for treatment effect estimation. This formulation enables us to quantify the causal impact of first exposure on future user behavior, while also providing counterfactually generated text to further explore effects.

We evaluate CausalT5 on semi-synthetic Twitter data where true treatment effects are known, investigating three research questions: **RQ1**: Are the generated counterfactual posts linguistically plausible and consistent with observed post-treatment behavior? **RQ2**: How accurately does CausalT5 estimate average treatment effects compared to baseline methods? **RQ3**: Can CausalT5 capture heterogeneous treatment effects across user subgroups, and does it remain robust under topical shifts in low-credibility posts? We find that CausalT5 produces realistic post-treatment text, achieves treatment effect estimates comparable to strong baselines, and identifies heterogeneous effects that vary by user characteristics and topical domains. Furthermore, our case study results suggest that a user's first engagement with a low-credibility news source is followed by about half an additional low-credibility tweet per 100 subsequent posts. Given that the mean number of low-credibility tweets in a 100-post window in our data is 0.166, this suggests a meaningful positive shift in low-credibility prevalence resulting from first-time exposure.

2 Related Work

Recent work has investigated generative language models for social media (Gao et al., 2023; Törnberg et al., 2023; Jeon et al., 2025). RePALM aligns quote-tweet generation with popularity metrics using reinforcement learning (Wang et al., 2024). Other studies have proposed action-guided generation, where models predict an engagement type (retweet, quote, reply) before generating a response (Gao et al., 2025). Beyond individual posts, researchers have studied networks of synthetic accounts, highlighting both the realism of generated content and the challenges of detection (Li et al., 2024; Ng and Carley, 2025). Together, this line of work shows that LLMs can generate fluent, socially grounded text in online settings.

Fine-tuned models have also been shown to reproduce linguistic style and biases of particular communities (Hwang et al., 2023). Perception studies further indicate that humans often cannot reliably distinguish LLM-generated posts from authentic ones, suggesting that synthetic text can function as a plausible intervention in social media (Li et al., 2024).

Recent work has adapted causal inference methods to settings where confounding is embedded in text (Egami et al., 2018; Keith et al., 2020; Roberts et al., 2020; Feder et al., 2022; Grimmer et al., 2022; Sridhar and Getoor, 2019). A central approach is to use encoder-based neural networks to learn text representations that can be adjusted for in effect estimation. CausalBERT (Veitch et al., 2020), for example, fine-tunes BERT (Devlin et al., 2018) to produce embeddings sufficient for both treatment and outcome prediction, providing a neural baseline for estimating treatment effects with textual confounding. However, these approaches primarily target outcomes that are numeric or categorical (e.g., rate of paper acceptance, changes in sentiment), reducing the problem to scalar effect estimation. In contrast, our work models the outcome itself as text. We propose a generative framework based on an encoder–decoder LLM that models counterfactual post-intervention trajectories, allowing us to study how interventions causally shape language use rather than only aggregated numeric surrogates.

3 Methods

Our goal is to estimate how a specific event in a user's timeline—such as engaging with a low-credibility news source—affects subsequent posting behavior. We adopt the potential outcomes framework (Neyman, 1923), where each user has two potential post-intervention outcomes: one under treatment and one under control. Since only one outcome is observed, the task is to estimate the unobserved counterfactual outcome. Figure 1 shows a motivating example.

Formally, for each user i, we observe (X_i, T_i, Y_i). X_i denotes the user's pre-intervention posts, $T_i \in \{0, 1\}$ indicates whether the index post represents an intervention of interest (e.g., exposure to a particular content type or event), and Y_i denotes post-intervention posts. We assume

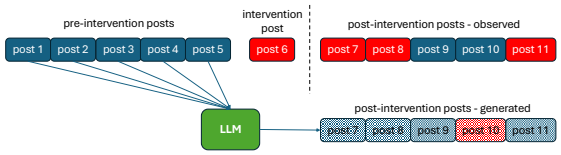

Figure 1: We design an LLM to generate posts that might have been written had the intervention post not occurred. By comparing properties of the observed and counterfactual posts (e.g., low-credibility posts, indicated in red), we can estimate effects of the intervention post. Analogously, we also generate posts for untreated users to simulate a world in which they had been treated.

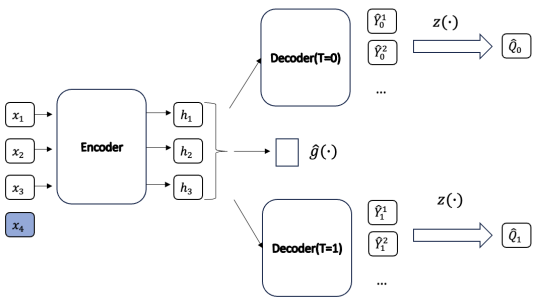

Figure 2: Model Pipeline. Pre-intervention posts x_1, x_2, x_3 are encoded to produce the propensity score $\hat{g}(\cdot)$, post-intervention posts under no treatment ($\hat{Y}_0^i$), and post-intervention posts under treatment ($\hat{Y}_1^i$). Here, x_4 is the intervention post and $z(\cdot)$ counts attributes of posts to produce outcome estimates $\hat{Q}_0$, $\hat{Q}_1$.

that the earlier $(k-1)$ posts $(x_{i,1}, \ldots, x_{i,k-1})$ are not exposed to treatment. Let $Y_i(0)$ and $Y_i(1)$ be post-intervention posts under control and treatment, respectively. We apply an outcome function $z(\cdot)$ (e.g., counting tweets mentioning a target property, such as a low-credibility news source). This gives scalar potential outcomes $Q_i(t) = z\big(Y_i(t)\big)$. The Average Treatment Effect (ATE) is then $\tau = \mathbb{E}[Q_i(1) - Q_i(0)]$. When studying heterogeneity, with subgroup indicator C_i, we consider the Conditional Average Treatment Effect (CATE), $\tau(c) = \mathbb{E}[Q_i(1) - Q_i(0) \mid C_i = c]$, and the aggregate effect $\tau = \sum_c \pi_c \tau(c)$, where $\pi_c = \Pr(C_i = c)$ is the proportion of units in subgroup c.

We assume standard causal identification assumptions: SUTVA, positivity, and conditional ignorability. We report ATE and CATE estimates using four standard estimators: $\hat{\tau}_{unadjust}$ (unadjusted difference-in-means), $\hat{\tau}_Q$ (outcome model), $\hat{\tau}_{IPW}$ (inverse probability weighting), and $\hat{\tau}_{AIPW}$ (augmented IPW (Glynn and Quinn, 2010)) (details in Appendix A.3).

3.1 CausalT5

We propose **CausalT5**, a transformer-based model designed for counterfactual text generation and causal effect estimation (Figure 2). The architecture is inspired by DragonNet (Shi et al., 2019), but extends it to the sequence-to-sequence setting by combining a shared encoder, two treatment-specific decoders, and a treatment prediction head. Specifically, we instantiate the encoder–decoder model with Flan-T5 (Chung et al., 2024), an LLM based on the T5 architecture (Raffel et al., 2020) that has been fine-tuned for instruction following. We prepend a fixed instruction prompt to each pre-intervention tweet sequence (Appendix A.1).

Given input sequence X (a user's pre-intervention tweets), the encoder produces contextual representations for each token $h = (h_1, h_2, \ldots, h_m) = E(X; \theta_e)$, where h_j is the hidden state for the j-th token. For treatment assignment prediction, we pool these token representations into a single vector $\bar{h}$ using attention-weighted averaging (see Appendix A.17 for details). This pooled vector is then passed to a sigmoid classifier to estimate the probability of treatment $\hat{g}$.

To generate post-intervention text, the two decoders (D_0, D_1) condition on encoder outputs, with $D_0(h; \theta_{d_0}) \to \hat{Y}_0$ for $T = 0$ and $D_1(h; \theta_{d_1}) \to \hat{Y}_1$ for $T = 1$, where Y_0 and Y_1 are sequences of tokens representing post-intervention texts under control and treatment.

Outcome model To model the outcome variables in the post-intervention generated text, we incorporate an auxiliary signal that encourages generated counterfactual texts to capture the prevalence of a target property. Let $\mathcal{V}_{target}$ denote the set of indicator tokens for this property (e.g., user handles for low-credibility news outlets). For each decoder time step t with vocabulary logits $z_t \in \mathbb{R}^{|\mathcal{V}|}$, we obtain a differentiable token distribution via the Gumbel–Softmax relaxation (Jang et al., 2016): $\pi_t = \text{GumbelSoftmax}(z_t, \tau)$, where $\pi_t \in \Delta^{|\mathcal{V}|-1}$ is a probability vector over the vocabulary and τ is the temperature parameter. The per-position probability of producing any target token is then

$$p_t^{target} = \sum_{v \in \mathcal{V}_{target}} \pi_t[v].$$

Because the outcome supervision is often not at the token level but at the tweet level (a tweet is considered positive if it contains at least one low-credibility news mention), we adopt a differentiable

Noisy–OR aggregator. Let S_j denote the index set of positions belonging to the j-th tweet (segmented by the special token [TWEET_SEP]). The probability that tweet j expresses the target property is

$$\hat{q}_j = 1 - \prod_{t \in S_j} (1 - p_t^{target}),$$

i.e., the probability of at least one low-credibility news outlet mention in tweet j is 1 minus the probability of zero low-credibility news outlet mentions in tweet j. The expected number of target-property tweets in a generated sequence of J tweets is then $\hat{c} = \sum_{j=1}^{J} \hat{q}_j$. This $\hat{c}$ provides a differentiable surrogate for the discrete count of property-bearing tweets. We then compare $\hat{c}$ against the ground-truth count c using mean squared error.

Loss function The overall training objective integrates three components: (i) conditional token generation loss for Y_0 or Y_1, (ii) treatment prediction loss via binary cross-entropy on propensity scores, and (iii) the target-property count regression loss:

$$\begin{aligned} L = \beta \cdot \frac{1}{n} \sum_{i=1}^{n} \Big[(1 - T_i) \sum_{t=1}^{m_i} -\log P(y_{i,t} \mid y_{i,<t}, h_i; \theta_{d_0}) \\ + T_i \sum_{t=1}^{m_i} -\log P(y_{i,t} \mid y_{i,<t}, h_i; \theta_{d_1}) \Big] \\ + \alpha \cdot \text{CrossEnt}(\hat{g}_i, T_i) \\ + (1 - \alpha - \beta) \cdot \text{MSE}(\hat{c}_i, c_i), \end{aligned}$$

where $y_{i,t}$ is the t-th target token, $\hat{g}_i$ the predicted propensity score, T_i the observed treatment, and $(\hat{c}_i, c_i)$ the predicted and ground-truth target-property counts. Hyperparameters α and β balance the three components.

At inference time, given a user's pre-intervention posts X, the model generates counterfactual futures $\hat{Y}_0$ and $\hat{Y}_1$. Outcome function $z(\cdot)$ is applied to obtain $\hat{Q}_0 = z(\hat{Y}_0)$ and $\hat{Q}_1 = z(\hat{Y}_1)$. Combined with the estimated propensity scores $\hat{g}$, these are plugged into one of four estimators: $\hat{\tau}_{unadjust}$ (unadjusted difference-in-means), $\hat{\tau}_Q$ (outcome model), $\hat{\tau}_{IPW}$ (inverse probability weighting), and $\hat{\tau}_{AIPW}$ (augmented IPW) (c.f. Appendix A.3).

4 Experiments

We conduct experiments on a case study of low-credibility news exposure on Twitter, evaluating CausalT5 on both real and semi-synthetic datasets. Our evaluation is guided by the following research questions: **RQ1:** To what extent does the text generated by CausalT5 resemble real-world post-intervention data? **RQ2:** How accurately does CausalT5 estimate ATE compared to baseline methods? **RQ3:** How accurately does CausalT5 estimate heterogeneous treatment effects, and is it robust to shifts in the topical composition of low-credibility posts?

Dataset We use a large-scale Twitter news engagement dataset (Shivaram et al., 2024). The raw corpus consists of 46.7 million tweets from 5,976 users posted over ten years. Following prior work, we define a *news engagement* event as either mentioning the official handle of a news source or sharing a URL to an article from that source. Each news source is assigned a partisan score on a seven-point scale $\{-3, -2, -1, 0, 1, 2, 3\}$, where ± 3 denote low-credibility outlets, as rated by fact checkers (Osmundsen et al., 2021).[1]

To construct our analytic population, we first identify 4,421 users who at some point engaged with low-credibility outlets. For each such user, the *first low-credibility engagement* defines the treatment pivot. Around this pivot, we extract a pre-intervention window of 5 tweets (X) and a post-intervention horizon of 100 tweets (Y), chosen to ensure sufficient signal given the sparsity of low-credibility engagements. Control pivots are selected by matching in time to each treatment pivot while enforcing two constraints: (i) the control user had no prior low-credibility engagement, and (ii) no control pivot is duplicated across matches, For each treatment pivot, we select 7 controls, and from each future post horizon we sample five tweets, identified to include up to five low-credibility outlets (more details in Appendix A.5). These choices yield a final analytic population of 35,368 instances. We further sample 3,000 balanced instances (1,500 treated, 1,500 control) for final effect estimation.

Semi-Synthetic Data Setup For evaluation, we construct semi-synthetic datasets with known treatment effects (following Weld et al. (2022)). We assign users to latent confounder classes C_i that influence both treatment probability and the distribution of outcomes Q_i (post-treatment low-credibility counts). We operationalize these classes by injecting topic-specific tweets (e.g., technology, sports) generated by an LLM into the pre-treatment history (prompt in Appendix A.2).

For **RQ2**, we design two classes to contrast spu-

[1] Classification of partisan scores follows conventions in prior work (Osmundsen et al., 2021).

rious correlations with genuine treatment effects. **Class 1** ($C = 1$) receives topic injections (Technology) but has no treatment effect, while **Class 2** ($C = 2$) receives no injections but exhibits a mild treatment effect. We induce confounding by correlating class membership with confounding strength p such that $P(T = 1 \mid C = 1) = 1 - p$ and $P(T = 1 \mid C = 2) = p$ ($p \in \{0.9, 0.7\}$). Outcomes are sampled from a Binomial distribution: $Q_i \mid (C_i = c, T_i = t) \sim \text{Binomial}(n, p_{ct})$; exact parameters are in Appendix Table 6.

For **RQ3**, we extend the framework to three classes to study heterogeneous effects: **Class 0** (no effect), **Class 1** (high effect), and **Class 2** (mild effect). To probe robustness, we apply a "swap" procedure where a proportion ($p_{\text{swap}} = 0.8$) of low-credibility posts are replaced with real posts examples from specific topics (health for Class 1, religion for Class 2). The full parameter specifications for priors, assignment probabilities, and outcome distributions are provided in Appendix Table 7.

5 Results

We conduct all evaluations under 5-fold cross-validation and report metrics averaged across folds.

5.1 RQ1: Fidelity of Generated Timelines

To assess the quality of generated timelines, we consider three types of measures: Factual Semantic Fidelity, Factual Behavioral Fidelity, and Counterfactual Semantic Fidelity.

Factual Semantic Fidelity We first evaluate whether CausalT5 generates realistic post-intervention tweets on the factual treatment arm. To do so, we measure the semantic similarity between generated post-intervention tweets and the corresponding observed tweets (e.g., tweets for a treated user are compared with generated tweets for the same user under T=1). For observed tweets $M = \{m_j\}_{j=1}^{J}$, generated tweets $N = \{n_j\}_{j=1}^{J}$, and embedding function $e(\cdot)$, we compute tweet-level *mean-of-max* similarity to match each generated tweet to its closest real counterpart:

$$S = \frac{1}{J} \sum_{j=1}^{J} \max_{m \in M} \cos\big(e(n_j), e(m)\big) \qquad (1)$$

We use the Sentence-BERT encoder (Reimers and Gurevych, 2019) as the embedding function $e(\cdot)$.

On average, the semantic similarity between observed and generated tweets for CasualT5 is 0.331 ($\pm$.002) (see Table 1). To contextualize this result, we compare with two baselines. First, the average semantic similarity between randomly selected pairs of users is 0.253 ($\pm$.071). Second, we consider a a **nearest-neighbor baseline** as follows: For each instance, we identify a different user whose pre-intervention tweets are most similar (highest cosine similarity). We compute the semantic similarity between the post-intervention tweets of the neighbor and the target user. This baseline estimates how much post-intervention similarity can be explained purely by pre-intervention resemblance. The semantic similarity of this nearest-neighbor baseline is 0.287 ($\pm$.003), which is 0.044 lower than CausalT5. The non-overlapping 95% confidence intervals provide strong evidence that CausalT5 produces more faithful generated timelines than the nearest-neighbor baseline.

	Factual Semantic Fidelity
CausalT5	$\mathbf{0.331 \pm 0.002}$
Nearest-Neighbor Baseline	0.287 ± 0.003
Random Pairs	0.253 ± 0.071

Table 1: Factual semantic fidelity results (mean $\pm$ std across 5 folds, N=3000) on real data.

Behavioral Fidelity To investigate fidelity beyond the words themselves, we additionally consider the attributes of the generated posts to assess behavioral fidelity. First, we find that the MAE of the predicted number of low-credibility tweets post-intervention is low ($0.392 \pm .015$), indicating that the model accurately captures signals of such behavior. Second, we assess the number of tweets that mention any news sources as a measure of user news engagement. We find an MAE of $0.835 \pm .017$ between the number of news sources mentioned in observed and generated posts, again indicating that the model can capture general news engagement levels based on prior tweets.

Finally, we assess whether generated low-credibility posts exhibit the expected hyper-partisan language patterns. We use GPT-4 as a zero-shot annotator (Ziems et al., 2024) with a standardized rubric prompt (see Appendix A.10) to quantify the political extremeness of tweets on a 1–5 scale. To validate this annotation procedure, we conducted a 100-example human validation study with two human annotators using the same scale. GPT–human agreement averaged linear weighted $\kappa = 0.51$ with 44% exact agreement, comparable to human–human agreement ($\kappa = 0.54$, 43%; Ap-

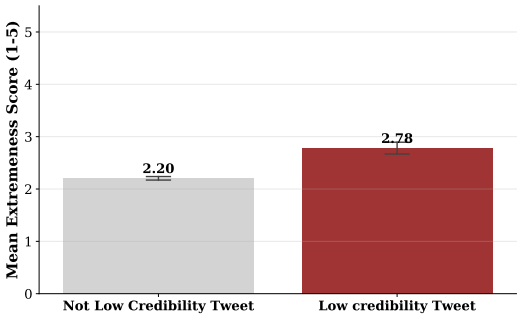

Figure 3: Mean extremeness scores of **generated text**, stratified by the presence of low-credibility content. Low-credibility tweets exhibit higher partisan intensity. Error bars represent 95% confidence intervals.

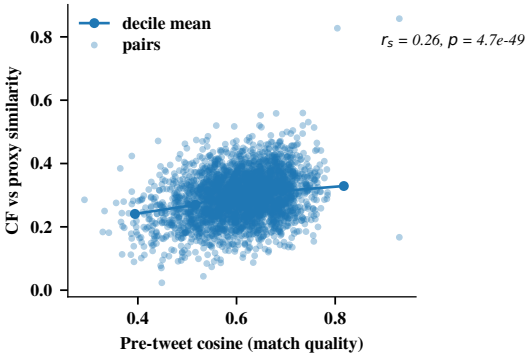

Figure 4: Relationship between pre-tweet similarity (matching quality) and similarity between generated counterfactuals (CF) and proxy counterfactuals. A positive Spearman correlation ($\rho = 0.264$) suggests that better matches yield more reliable proxies.

pendix A.18), suggesting that GPT-based annotations are broadly consistent with human judgments. Figure 3 shows mean extremeness scores (1–5) for tweets stratified by the presence of low-credibility news. We observe a significant divergence: tweets with low-credibility news exhibit higher partisan intensity ($\mu = 2.78$) compared to tweets without low-credibility content ($\mu = 2.20$). The proportion of non-political tweets (Extremeness Score 1) drops from 44% in the group without low-credibility content to 16% in the low-credibility group, confirming that the model captures the association between low-credibility and polarized rhetoric.

Counterfactual Semantic Fidelity Evaluating counterfactual text generation is challenging because true counterfactual post-intervention tweets are never observed. To address this, we use a matching-based proxy evaluation. For each user, we generate their counterfactual posts $\hat{Y}_{1...T}$. We then identify a user in the opposite treatment group with highly similar pre-intervention behavior and compare the generated counterfactual posts to this matched user's observed posts. The matched user's real post tweets $Y_{1...T}$ serve as a *proxy counterfactual*, and we assess the similarity between $\hat{Y}_{1...T}$ and this proxy using the same semantic similarity measure as before (Equation 1). This approach is grounded in the standard assumption of conditional ignorability in causal inference (Rosenbaum and Rubin, 1983): among users with comparable pre-treatment covariates, outcomes from opposite treatment groups can serve as unbiased stand-ins for the unobserved counterfactual outcomes.

Figure 4 plots the relationship between (i) the similarity between pre-intervention tweets of matched pairs and (ii) the similarity between the model's generated counterfactual tweets and their proxy counterfactuals. We observe a positive Spearman correlation ($\rho = 0.264$), indicating that better pre-treatment matching yields more reliable proxy counterfactuals—and that our generated counterfactuals align more closely with these reliable proxies. This supports the plausibility of the model's counterfactual generations.

For a qualitative assessment, Table 2 presents two representative examples of counterfactual generations produced by the model. In the first row, the user belongs to the control group ($T = 0$), having engaged exclusively with credible climate-related content before the intervention. The model-generated counterfactual imagines this same user under a hypothetical low-credibility treatment: while the content remains topically consistent with climate discourse, its tone shifts to partisan and conspiratorial. The second row shows the reverse setting for a treated user ($T = 1$), where the generated counterfactual replaces conspiratorial political content with a more neutral tone.

5.2 RQ2: Treatment Effect Estimation

Semi-synthetic data For semi-synthetic experiments where the ground-truth causal effects are known, we evaluate the model's ability to recover both ATE and CATE. We report absolute errors

$$\text{AE}_{\text{ATE}} = \left|\hat{\tau} - \tau\right|, \qquad \text{AE}_{\text{CATE}}(c) = \left|\hat{\tau}(c) - \tau(c)\right|,$$

where τ and $\tau(c)$ denote the true effects. We also monitor the treatment prediction head's classification accuracy as an auxiliary diagnostic of propensity estimation quality. To contextualize performance, we first compare CausalT5 against a non-generative encoder-based baseline model inspired by CausalBERT (Veitch et al., 2020) (Ap-

T	Pre-Intervention Tweet	Intervention Tweet	Post-Intervention - Obs	Post-Intervention - CF
0	RT @citizensclimate: Editorial in San Antonio! Carbon tax plan worthy of bipartisan support #climate @ImTedHalstead...	RT @CCLsaltlake: Arctic and Antarctic sea ice are both at record low levels right now @AndreaTWeather #climate	RT @CCLsaltlake: What next for #renewables in cities? – the expert view @tess_riley @guardian #climate	RT @CitizensClimate: I'm a big fan of @realDonaldTrump and @CCLsaltlake, but I don't think he's going to be able to stand up for what he does. He's a liar and a hypocrite. **#climatenews**
1	Ugliness in and out (*in reply to*: "WikiLeaks leaks Clinton election emails")	RT **@TPInsidr**: Voting Has Started, And Voter Fraud Is Happening... IN TEXAS! #tcot #WakeUpAmerica	BREAKING: Obama Just Caught Trying to Sabotage New Clinton Email Investigation... This is SICK **endingthefed**	RT @JosephMcDonaldTrump: "It's time to stop blaming the Trump administration for destroying our democracy." – @JamesHarris

Table 2: Example counterfactual generations. Each row shows one user, with (i) a representative pre-intervention tweet, (ii) the intervention (pivot) tweet, (iii) a representative observed (**Obs**) post-intervention tweet, and (iv) the model-generated counterfactual (**CF**) post-intervention tweet. Low-credibility sources are highlighted in **red**. For the untreated user in row 1, the model generates a low-credibility counterfactual tweet about the climate, following the topical focus of the pre-intervention tweets. For the treated user in row 2, the model generates a counterfactual about presidential politics, but avoids low-credibility news sources.

pendix A.4), which directly predicts numerical outcomes from pre-intervention representations but does not generate text. We additionally evaluate a no-training few-shot prompted LLM baseline, where a prompted model takes pre-intervention tweets together with example continuations and separately generates post-treatment text under treatment and control conditions, yielding a Q-only ATE estimate (details in Appendix A.20). Unlike CausalT5, this baseline uses no task-specific causal training or propensity estimation.

p	Estimator	CausalT5	Encoder Baseline
0.9	Q-only	0.21 ± 0.18	0.14 ± 0.04
	IPW	$\mathbf{0.09 \pm 0.07}$	$\mathbf{0.08 \pm 0.09}$
	AIPW	0.12 ± 0.08	0.11 ± 0.03
	Unadjusted	0.17 ± 0.00	0.17 ± 0.00
0.7	Q-only	0.33 ± 0.17	0.26 ± 0.05
	IPW	$\mathbf{0.05 \pm 0.04}$	0.19 ± 0.07
	AIPW	0.08 ± 0.06	$\mathbf{0.08 \pm 0.04}$
	Unadjusted	0.10 ± 0.00	0.10 ± 0.00

Table 3: ATE absolute error (mean $\pm$ std across 5 folds) under two-class confounding (**RQ2**). Best results are highlighted for CausalT5 and the Encoder Baseline.

Table 3 reports ATE estimation error under the two-class semi-synthetic setup, comparing CausalT5 with the encoder-only baseline (Appendix A.4). Across both confounding strengths ($p = 0.9$ and $p = 0.7$), CausalT5 achieves performance largely comparable to the encoder baseline across all four estimators. The best results for CausalT5 are statistically indistinguishable from those for the encoder baseline for both 0.9 and 0.7 confounding. This is notable because the encoder baseline is trained to directly regress on the outcome, whereas CausalT5 only indirectly regresses on the outcome via the attributes of the generated posts. These results suggest that CausalT5 can recover treatment effects as accurately as a purpose-built outcome regression model, while providing the additional benefit of interpretable counterfactual text. We also evaluate a no-training few-shot prompted LLM baseline using the Q-only estimator. The prompted baseline achieves ATE absolute error of 0.193 ± 0.047 at confounding strength $p = 0.7$ and 0.444 ± 0.662 at $p = 0.9$. Although this provides a useful modern LLM reference point, the prompted baseline becomes substantially less stable under stronger confounding, while the best causally adjusted CausalT5 estimators continue to achieve lower estimation error. Additionally, an ablation study (Appendix A.11) confirms that both the treatment prediction and target-property count regression objectives are necessary to minimize estimation error. For example, without the Count Loss, AIPW error more than doubles.

As a diagnostic, we also examine treatment prediction accuracy. As expected, higher confounding yielded higher accuracy: for p=0.9, treatment prediction accuracies were 0.88 ± 0.01 for CausalT5 and 0.89 ± 0.01 for the encoder baseline; for p=0.7, they were 0.66 ± 0.02 and 0.68 ± 0.01, respectively. These results confirm that CausalT5 can capture the relationship between treatment and confounder.

Real data Table 4 shows the results on the real Twitter dataset, estimating the causal effect of

Metric	CausalT5	Encoder Baseline
Treatment accuracy	0.70 ± 0.02	0.68 ± 0.04
ATE (Q-only)	0.16 ± 0.05	0.46 ± 0.07
ATE (IPW)	0.52 ± 0.08	0.56 ± 0.07
ATE (AIPW)	0.53 ± 0.08	0.54 ± 0.05

Table 4: ATE estimates (mean $\pm$ std across 5 folds) and treatment prediction accuracy on the real dataset.

first-time low-credibility news engagement on subsequent posts. CausalT5 with AIPW estimates an ATE of approximately $+0.53$, indicating that first-time engagement with low-credibility outlets is associated with about half an additional low-credibility tweet in the subsequent 100 tweets. The encoder baseline produces a similar estimate ($+0.54$), and both models achieve comparable treatment prediction accuracy (0.70 vs. 0.68). Given that the mean number of low-credibility tweets in a 100-post window is 0.166, both methods detect a meaningful positive shift in low-credibility prevalence following first-time exposure, with the caveat that these results rely on standard identification assumptions (e.g., ignorability).

5.3 RQ3: Heterogeneity and Robustness

Table 5 reports CATE estimation errors in the three-class semi-synthetic setting for confounding strength 0.9 (results for 0.7 are similar; see Appendix Table 11). Because CATE estimation requires recovering effects within each subgroup rather than over the full population, this setting poses a stricter test of model performance. The best results for CausalT5 are again statistically indistinguishable from those for the encoder baseline for both 0.9 and 0.7 confounding and for all three classes. However, the results show that CausalT5 struggles with the Q-only estimator—particularly in Classes 0 and 2—likely because the strong treatment effect signal in Class 1 dominates the learned regression function, overshadowing the weaker or null effects in the other classes. In contrast, the IPW and AIPW estimators remain comparatively stable and often achieve errors similar to or lower than the encoder baseline. Among them, AIPW gives the strongest results for CausalT5, bringing errors in the high-effect Class 1 and mild-effect Class 2 close to encoder baseline levels. (See additional results in Appendix A.15.)

Finally, we evaluate whether the model reproduces the intended topic composition introduced by the swap procedure in the generated post tweets

Confounding Strength $p = 0.9$			
Class	**Estimator**	**CausalT5**	**Encoder Baseline**
0	Q-only	0.95 ± 0.18	0.15 ± 0.10
	IPW	0.28 ± 0.23	0.18 ± 0.15
	AIPW	$\mathbf{0.23 \pm 0.22}$	$\mathbf{0.14 \pm 0.03}$
1	Q-only	0.50 ± 0.15	0.30 ± 0.08
	IPW	0.47 ± 0.26	0.55 ± 0.17
	AIPW	$\mathbf{0.12 \pm 0.14}$	$\mathbf{0.24 \pm 0.14}$
2	Q-only	0.77 ± 0.23	$\mathbf{0.11 \pm 0.11}$
	IPW	0.25 ± 0.20	0.30 ± 0.26
	AIPW	$\mathbf{0.19 \pm 0.16}$	0.18 ± 0.18

Table 5: CATE absolute error (mean $\pm$ std across 5 folds) in the three-class semi-synthetic setting for confounding strength 0.9. Lower is better. Best results are highlighted for CausalT5 and the Encoder Baseline.

on the factual arm (i.e., we use $\hat{Y}_1$ if $T{=}1$ and $\hat{Y}_0$ if $T{=}0$). By design, 80% of low-credibility tweets generated in Class 1 should be health-related, and 80% in Class 2 should be religion-related. The model captures this pattern far more accurately in Class 1 than in Class 2: 71.5% versus 3.4% and 58.2% versus 3.2%. This disparity likely reflects the higher prevalence of low-credibility tweets in Class 1 by design: for example, at $p{=}0.9$, Class 1 contains 684 instances with low-credibility tweets (1,324 such tweets in total), compared to 400 instances with 460 tweets in Class 2. This provides Class 1 instances more opportunities for topic-specific swaps and thus a stronger learning signal. Moreover, Class 1 coverage drops from 71.5% to 58.2% when the confounding strength decreases from $p{=}0.9$ to $p{=}0.7$, suggesting that stronger alignment between confounder class and treatment improves the model's ability to recover class-conditional topic patterns. Future work should investigate how sensitive this approach is to low-data regimes, where estimating relationships between pre- and post-treatment topics is more challenging.

6 Conclusions

We presented CausalT5, a generative framework for treatment effect estimation from social media timelines. Unlike outcome-only approaches, CausalT5 generates counterfactual text while jointly modeling treatment propensities and outcome counts. Experiments on semi-synthetic and real Twitter data show that CausalT5 produces realistic generations, achieves competitive ATE estimation, and captures heterogeneous effects across user subgroups.

7 Limitations

Our evaluation is limited to one domain (news engagement) and may not capture the full diversity of social media behavior. Moreover, because the ground-truth treatment effect is unknown in real data, we cannot directly evaluate the correctness of estimated effects in that setting. Furthermore, while a common assumption in related work, real-world pre-treatment text may not capture all confounding (e.g., latent political ideology, offline exposure). Finally, the results in RQ3 suggest that sufficient training data is required to discover nuanced relationships between counfounders in the pre-intervention period and topics discussed in the post-intervention period.

8 Ethical considerations

We analyze a large-scale dataset of publicly available Twitter posts about news engagement. While some low-credibility tweets contain offensive language, they are included only as text and not used to profile individuals or groups. Because the dataset contains only public, de-identified content, the study was deemed IRB exempt. To reduce potential harm, we report only aggregate treatment effects rather than individual-level predictions.

We also acknowledge potential dual-use risks: generative counterfactual models like CausalT5 could be misappropriated to produce realistic but fabricated content for disinformation or propaganda. To mitigate this, we report only aggregate treatment effects and focus on controlled, research-only code and data sharing. We encourage the community to pursue safeguards such as watermarking and detection mechanisms to reduce risks of harmful deployment.

References

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, and 1 others. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

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*.

Naoki Egami, Christian Fong, Justin Grimmer, Margaret Roberts, and Brandon M. Stewart. 2018. Causal inference with text as treatment. *Proceedings of the 35th International Conference on Machine Learning (ICML)*.

Amir Feder, Katherine A Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E Roberts, and 1 others. 2022. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics*, 10:1138–1158.

Achyutarama Ganti, Eslam Ali Hassan Hussein, Steven Wilson, Zexin Ma, and Xinyan Zhao. 2023. Narrative style and the spread of health misinformation on Twitter. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4266–4282, Singapore. Association for Computational Linguistics.

Chen Gao, Xiaochong Lan, Zhihong Lu, Jinzhu Mao, Jinghua Piao, Huandong Wang, Depeng Jin, and Yong Li. 2023. S3: Social-network simulation system with large language model-empowered agents. *arXiv preprint arXiv:2307.14984*.

Yuxuan Gao, Hong Lee, Wei Xu, and Lei Zhang. 2025. Can llms simulate social media engagement? a study on action-guided response generation. *arXiv preprint arXiv:2502.12073*.

Adam N Glynn and Kevin M Quinn. 2010. An introduction to the augmented inverse propensity weighted estimator. *Political analysis*, 18(1):36–56.

Justin Grimmer, Margaret Roberts, and Brandon M. Stewart. 2022. Text as data: A new framework for machine learning and the social sciences. *Princeton University Press*.

EunJeong Hwang, Rafal Kocielnik, Sarah Noble, and Mark Riedl. 2023. Exposing bias in online communities through large-scale language models. In *Proceedings of ICWSM*.

Eric Jang, Shixiang Gu, and Ben Poole. 2016. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*.

Min Soo Jeon, Marcelo Mendoza, Miguel Fernández, Eliana Providel, Felipe Rodríguez, Nicolás Espina, Andrés Carvallo, and Andrés Abeliuk. 2025. Simulating conversations on social media with generative agent-based models. *EPJ Data Science*, 14(1):79.

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*.

Bernard J Koch, Tim Sainburg, Pablo Geraldo Bastias, Song Jiang, Yizhou Sun, and Jacob G Foster. 2025. A primer on deep learning for causal inference. *Sociological Methods & Research*, 54(2):397–447.

Linsen Li, Aron Culotta, and Nicholas Mattei. 2025. Using text-based causal inference to disentangle factors influencing online review ratings. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11259–11277, Albuquerque, New Mexico. Association for Computational Linguistics.

Yifan Li, Rui Zhao, Ananya Kumar, and Elizabeth Clark. 2024. Human perception of llm-generated social media text. *arXiv preprint arXiv:2409.06653*.

Arun S Maiya. 2021. Causalnlp: A practical toolkit for causal inference with text. *arXiv preprint arXiv:2106.08043*.

Julia Mendelsohn, Sayan Ghosh, David Jurgens, and Ceren Budak. 2023. Bridging nations: quantifying the role of multilinguals in communication on social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, pages 626–637.

Jerzy Neyman. 1923. On the application of probability theory to agricultural experiments. essay on principles. *Ann. Agricultural Sciences*, pages 1–51.

Lynnette Hui Xian Ng and Kathleen M Carley. 2025. Are llm-powered social media bots realistic? *arXiv preprint arXiv:2508.00998*.

Mathias Osmundsen, Alexander Bor, Peter Bjerregaard Vahlstrup, Anja Bechmann, and Michael Bang Petersen. 2021. Partisan polarization is the primary psychological motivation behind political fake news sharing on twitter. *American political science review*, 115(3):999–1015.

Giulio Pecile, Niccolò Di Marco, Matteo Cinelli, and Walter Quattrociocchi. 2025. Mapping the global election landscape on social media in 2024. *PloS one*, 20(2):e0316271.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. 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)*, page 3982. Association for Computational Linguistics.

Margaret E Roberts, Brandon M Stewart, and Richard A Nielsen. 2020. Adjusting for confounding with text matching. *American Journal of Political Science*, 64(4):887–903.

Paul R Rosenbaum and Donald B Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Claudia Shi, David Blei, and Victor Veitch. 2019. Adapting neural networks for the estimation of treatment effects. *Advances in neural information processing systems*, 32.

Karthik Shivaram, Mustafa Bilgic, Matthew Shapiro, and Aron Culotta. 2024. Forecasting political news engagement on social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 1451–1462.

Dhanya Sridhar and Lise Getoor. 2019. Estimating causal effects of tone in online debates. *arXiv preprint arXiv:1906.04177*.

Petter Törnberg, Diliara Valeeva, Justus Uitermark, and Christopher Bail. 2023. Simulating social media using large language models to evaluate alternative news feed algorithms. *arXiv preprint arXiv:2310.05984*.

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.

Jian Wang, Shu Wang, Bo Li, Zhiqiang Zhang, and Xiangnan Chen. 2024. Repalm: Popular quote tweet generation via auto-regressive language models. In *Findings of the Association for Computational Linguistics: ACL 2024*.

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.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 40 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291.

A Appendix

A.1 Prompt Used for Generation

We prepend the following instruction to each pre-treatment sequence during training and generation:

```
"Assume you are a Twitter user who has posted these five messages (each tweet is separated by the token [TWEET_SEP]). Write five more messages in a similar style, using [TWEET_SEP] between tweets."
```

A.2 Confounder Tweet Generation Prompts

We used the following instruction to generate 25 technology-related tweets and 25 sports-related tweets for constructing topic-specific confounder pools.

Technology (Topic 1) prompt

```
"Write 25 short, informal tweets about recent technology news, gadgets, software updates, or innovations. Tweets should sound like typical user posts on social media (casual tone, personal reactions, no hashtags, no links)."
```

Sports (Topic 2) prompt

```
"Write 25 short, informal tweets about sports events, teams, or athletes. Tweets should sound like typical user posts on social media (casual tone, personal reactions, no hashtags, no links)."
```

A.3 Estimators

To estimate the ATE, we consider several core assumptions and estimators.

Assumptions. *Ignorability* requires that treatment assignment is independent of potential outcomes: $(Y_i(0), Y_i(1)) \perp\!\!\!\perp T_i$. *Conditional ignorability* strengthens this by conditioning on covariates: $(Y_i(0), Y_i(1)) \perp\!\!\!\perp T_i \mid X_i$. *Positivity* requires non-zero probability of receiving either treatment: $0 < g(X_i) < 1$, where $g(x) = \Pr(T = 1 \mid X = x)$. We also assume SUTVA.

Unadjusted difference-in-means.

$$\hat{\tau}_{unadjust} = \frac{1}{n_1} \sum_{i:T_i=1} Q_i - \frac{1}{n_0} \sum_{i:T_i=0} Q_i, \quad (2)$$

where $Q_i = z(Y_i)$, and $n_t = \sum_i \mathbb{I}\{T_i = t\}$.

Outcome model (Q-only).

$$\hat{\tau}_Q = \frac{1}{n} \sum_{i=1}^{n} \big(\hat{Q}(1, X_i) - \hat{Q}(0, X_i)\big). \quad (3)$$

Inverse probability weighting (IPW).

$$\hat{\tau}_{IPW} = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{T_i Q_i}{\hat{g}(X_i)} - \frac{(1 - T_i) Q_i}{1 - \hat{g}(X_i)} \right). \quad (4)$$

Augmented IPW (AIPW). The doubly robust estimator combines outcome regression and propensity weighting (Glynn and Quinn, 2010):

$$\hat{\tau}_{AIPW} = \frac{1}{n} \sum_{i=1}^{n} \Bigg[\left(\frac{T_i}{\hat{g}(X_i)} - \frac{1 - T_i}{1 - \hat{g}(X_i)} \right) \big(Q_i - \hat{Q}(T_i, X_i)\big) + \hat{Q}(1, X_i) - \hat{Q}(0, X_i) \Bigg]. \quad (5)$$

Conditional effects. For subgroup $C_i = c$, the CATE estimator $\hat{\tau}(c)$ is defined analogously by restricting sums to $\{i : C_i = c\}$. The overall ATE is recovered as

$$\hat{\tau} = \sum_c \frac{n_c}{n} \hat{\tau}(c), \quad (6)$$

where $n_c = |\{i : C_i = c\}|$.

A.4 Encoder Baseline Model

To disentangle the contribution of counterfactual text generation from standard causal prediction, we design an encoder-only baseline. This encoder baseline isolates the value of conditional outcome regression: rather than generating post-intervention texts, the model directly estimates expected outcomes under treatment and control conditions. By comparing against this simpler architecture, we can evaluate whether sequence generation provides additional benefits beyond standard outcome modeling.

The encoder baseline is inspired by CausalBERT (Veitch et al., 2020), employing a shared transformer encoder to obtain contextual representations.[2]. Given input sequence X, the encoder produces hidden states

$$h = (h_1, h_2, \ldots, h_k) = E(X; \theta_e).$$

We apply similar attention pooling to derive a compact sequence embedding

$$\alpha_i = \frac{\exp(w^\top h_i)}{\sum_{j=1}^{k} \exp(w^\top h_j)}, \quad \bar{h} = \sum_{i=1}^{k} \alpha_i h_i.$$

[2]Our formulation differs from CausalBERT primarily in using attention pooling rather than a [CLS] token for representation.

On top of $\bar{h}$, two regression heads predict the expected low-credibility tweet counts under control and treatment:

$$\hat{Q}_0 = W_0\bar{h} + b_0, \quad \hat{Q}_1 = W_1\bar{h} + b_1.$$

In parallel, a classification head estimates the propensity score:

$$\hat{g} = \sigma(W_g\bar{h} + b_g).$$

Training uses only the observed arm. For each instance with treatment assignment $T \in \{0, 1\}$, the predicted outcome is

$$\hat{Y} = (1 - T)\hat{Q}_0 + T\hat{Q}_1.$$

The objective combines outcome regression and treatment classification:

$$L = (1-\alpha) \cdot \text{MSE}(\hat{Y}, Y) \; + \; \alpha \cdot \text{BCE}(\hat{g}, T), \quad (7)$$

where MSE is the mean squared error, BCE is the binary cross-entropy loss, and $\alpha \in [0, 1]$ balances the two components.

At inference time, the encoder baseline outputs $\hat{Q}_0, \hat{Q}_1$, and $\hat{g}$. These quantities can be directly plugged into estimators such as IPW or AIPW to compute ATE.

A.5 Data Process Detail

To formalize our dataset, each dataset instance corresponds to a tuple (X, Pivot, Y, T), where $T = 1$ if the pivot is the user's first low-credibility engagement and $T = 0$ otherwise. The pre- and post-windows are concatenated into sequences with a special token `[TWEET_SEP]` marking tweet boundaries. To preserve information about source identity, each URL or mention matching a known news source is replaced with the source name and annotated with a partisan score token (e.g., `[SCORE=+3]`).

Our final population comprises 35,368 instances. For each instance, we track the number of low-credibility tweets and news tweets in both the pre- and post-windows, and additionally construct a standardized *sample window* of five tweets from the post horizon. If five or more low-credibility tweets appear, we randomly sample five; otherwise, we include all low-credibility tweets and supplement with non-low-credibility tweets until the window contains five in total. This procedure amplifies the sparse low-credibility signal while maintaining a fixed post-window size for comparability across users.

Formally, let K denote the number of low-credibility tweets in the full 100-post post-intervention window, and S denote the number of low-credibility tweets in the constructed 5-tweet sample window. By construction, $S = \min(K, 5)$. Thus, S and K are identical whenever $K \leq 5$ and only differ when $K \geq 5$. In our data, only 187 out of 35,368 instances ($\approx 0.53\%$) had $K \geq 5$. Accordingly, the sample-window counts S closely approximate the original 100-post counts K for the vast majority of users (overall means: $K = 0.166$, SD $= 1.26$; $S = 0.12$, SD $= 0.57$). Because $K \geq 5$ is so rare, treatment effect estimates obtained on S are numerically very close to those on K, and all reported effects in the main text are scaled back to the original 100-post outcome K for interpretability.

A.6 Semi-Synthetic Data Detail

Based on the definitions in Section 3, we formalize the semi-synthetic framework by extending each unit i with a confounder class C_i, yielding tuples (X_i, Y_i, T_i, C_i). The confounder determines the treatment assignment probabilities, specifies which topic-specific confounder tweets are injected into the observed pre-intervention text, and jointly with treatment status T_i governs the outcome distribution. To ensure natural variation, each topic-specific confounder set consists of 50 tweets generated with a large language model (ChatGPT) designed to mimic authentic online posts without artificial markers. We construct two such sets: Topic A (technology) and Topic B (sports).

RQ2. We define $C_i \in \{1, 2\}$ with equal class probabilities. For Class 1, we insert Topic A confounder tweets (technology), while for Class 2, no insertion occurs. Treatment assignment follows $P(T = 1 \mid C = 1) = 1 - p, \quad P(T = 1 \mid C = 2) = p$, with $p \in \{0.9, 0.7\}$. When $p = 0.9$, treatment is strongly confounded with C_i; when $p = 0.7$, the confounding is weaker. Outcomes are sampled as

$$Q_i \mid (C_i = c, T_i = t) \sim \text{Binomial}(n, p_{ct}),$$

where $Q_i = z(Y_i)$ is the number of low-credibility tweets in the post-intervention window of length n, and p_{ct} denotes the probability that an individual tweet is low-credibility given confounder class c and treatment status t. For RQ1&2 we set

Class	Pre-pivot topic injected	$\mathbf{P(T{=}1 \mid C)}$	$\mathbf{p_{c,0}}$ (T=0)	$\mathbf{p_{c,1}}$ (T=1)
1	tech	$1-p$	0.10	0.10
2	None	p	0.10	0.20

Table 6: RQ2 semi-synthetic design. Class 1 introduces Topic A confounders (tech) but has no treatment effect. Class 2 has no injected topic but exhibits a mild effect. Confounding strength set by $p \in \{0.9, 0.7\}$.

Class	Pre-pivot Topic	Post-pivot Topic Swap	$\mathbf{P(T{=}1 \mid C)}$	Outcome Prob. $\mathbf{p_{c,0}}$	Outcome Prob. $\mathbf{p_{c,1}}$
0	None	None	0.50	0.10	0.10
1	Tech	Health	p	0.10	0.40
2	Sports	Religion	$1-p$	0.10	0.20

Table 7: RQ3 semi-synthetic design. Three classes yield heterogeneous effects. A swap procedure ($p_{\text{swap}} = 0.8$) replaces topical content (e.g., health vs. religion) while keeping outcome probabilities ($p_{c,t}$) fixed.

$p_{1,0} = p_{1,1} = 0.10$ (Class 1, no treatment effect) and $p_{2,0} = 0.10$, $p_{2,1} = 0.20$ (Class 2, mild effect). Table 6 summarizes these parameters. This design creates a contrast where one group introduces lexical confounding without causal effect and the other exhibits a genuine treatment effect, thereby testing whether our model can recover causal signal while ignoring spurious correlations. To control variety of inserted text, we vary the confounder variety by sampling from the top-K items of the pool, with $K \in \{25, 50\}$.

RQ3. Building on the binary setup of RQ1&2, RQ3 generalizes the framework to three confounder classes and introduces a swap mechanism to capture heterogeneous effects and topic-specific shifts in low-credibility posts. We set $C_i \in \{0, 1, 2\}$ with prior probabilities $\{0.34, 0.33, 0.33\}$. Class 0 receives no confounder, Class 1 receives Topic A confounder tweets (technology), and Class 2 receives Topic B confounder tweets (sports). Treatment assignment is balanced for Class 0, while Classes 1 and 2 follow the p-dependent rule with $p \in \{0.9, 0.7\}$. Outcomes follow the same binomial model with parameters $(C, T) \mapsto p_{ct}$: $(0, 0) = (0, 1) = 0.10$ (Class 0, no effect), $(1, 0) = 0.10$, $(1, 1) = 0.40$ (Class 1, high effect), and $(2, 0) = 0.10$, $(2, 1) = 0.20$ (Class 2, mild effect). This design induces heterogeneous effects across classes—low, mild, and high—allowing us to test whether models can capture the variation in causal effect size.

In addition, we apply a *swap procedure* to modify the topical composition of low-credibility tweets in the post-intervention window. The swap pools are built from real low-credibility tweets: we cluster low-credibility posts using TF–IDF vectorization and MiniBatchKMeans, then manually inspect clusters to identify coherent themes. Two high-purity clusters emerge, one centered on health (*vaccines, GMO, monsanto, health ranger*) and another on anti-Muslim rhetoric (*muslims, islamic, jihad, migrant, invaders*). From these clusters we define two compact keyword sets and extract corresponding tweet pools. During generation, a proportion $p_{\text{swap}} = 0.8$ of low-credibility tweets in each window are replaced with tweets drawn from the class-specific real pool (health-related for Class 1, religion-related for Class 2). Table 7 details the full configuration. The total count of low-credibility posts remains fixed, but its topical distribution shifts in line with the assigned class. This procedure keeps the overall quantity of low-credibility posts fixed while altering its topical composition, thereby probing whether estimators are robust to shifts in *what kind* of low-credibility posts appears rather than just *how many*.

Population and Sample Sizes. The semi-synthetic data is drawn from the empirical population of 35,368 instances. For each configuration of (p, K, p_{swap}), we sample $n = 3000$ instances without replacement to form the final dataset used for evaluation. Thus, the only synthetic elements are the controlled sampling rules, the injection of LLM-generated topic confounder tweets, and the swap procedure introducing real low-credibility topical pools.

A.7 GPUs Usage

All experiments were conducted on a cluster using nodes equipped with two NVIDIA A100 80GB PCIe GPUs and 8 CPU cores per node. We ran 5-fold cross-validation, with each fold occupying one node (2 GPUs) concurrently, resulting in 10 GPUs active in parallel. Each fold took approximately 40 GPU-hours to complete, for a total of about 200 GPU-hours across all folds. Training was implemented in PyTorch 2.3.1 with CUDA enabled.

A.8 CausalT5 Parameter Detail

The CausalT5 model is built on top of the `google/flan-t5-base` checkpoint (Chung et al.,

2024) and consists of a shared encoder and two treatment-specific decoders. Each component adopts the standard T5-base configuration with a model dimension of 768 and a maximum input length of 512 tokens. The embeddings layer includes word embeddings of size $[32{,}117, 768]$ and positional encodings implicitly handled by the T5 architecture. The encoder comprises 12 transformer blocks, each containing a self-attention sublayer with projection matrices of size $[768, 768]$ for queries, keys, values, and outputs, and a feed-forward sublayer with projection sizes $[768, 2048]$ and $[2048, 768]$.

Each decoder also contains 12 transformer blocks, extending the encoder structure with an additional cross-attention sublayer that links to the encoder's output representations. Both decoders share the same embedding matrix $[32{,}117, 768]$ and output projection heads (`lm_head`) of size $[768, 32{,}117]$. Following the generation modules, a pooling attention layer (`pool_w`) aggregates the encoder representations, and a fully connected `treatment_head` predicts treatment assignment probabilities.

Overall, the model contains approximately 360 million parameters in total, comprising a shared encoder ($\sim$85 million), two treatment-specific decoder stacks ($\sim$ 113 million each, including self-attention, cross-attention, and feedforward layers), two token embedding matrices ($\sim$ 24.7 million each, with language modeling heads tied to their respective embeddings), and the downstream pooling and treatment prediction heads ($<$ 0.01 million).

A.9 CausalT5 Training Details

We train CausalT5 using 5-fold cross-validation on the semi-synthetic and real Twitter datasets, stratifying folds by treatment assignment to ensure balance. Each fold is fine-tuned for 17 epochs. Because conditional text generation is substantially more difficult than the other objectives, we first perform a warm-up phase of 5 *text-only* epochs optimizing the generation loss. This allows the model to learn to produce coherent sequences before introducing additional supervision. After this warm-up, we jointly optimize all three objectives—(i) conditional text generation, (ii) treatment prediction, and (iii) target-property count regression.

Hyperparameters were tuned through grid search over a held-out split, yielding the following configuration: learning rate 2×10^{-5} (AdamW), batch size 4, α=0.45 and β=0.45 for weighting treatment and outcome losses, token-level weight scalar 5.0, and Gumbel-softmax temperature 0.3.

A.10 Prompt for Extremeness Scoring

To evaluate the semantic consistency of the generated text, we employed GPT-4 with the following zero-shot prompt:

```
"You are analyzing political language on social media.
Question: Does the following tweet express extreme political or partisan sentiment?
Definition: Extreme political sentiment refers to highly polarized, strongly partisan, or ideologically rigid language...
Answer with a single integer from 1 to 5:
1 = Not extreme / non-political
2 = Mildly partisan
3 = Moderately partisan
4 = Strongly partisan / polarized
5 = Extremely partisan / polarized / ideologically extreme"
```

A.11 Ablation Study

To assess the individual contributions of the training objectives, we conduct an ablation study on the semi-synthetic dataset ($p = 0.9$, single fold). We evaluate the Full CausalT5 model against three variants:

- **w/o Target-Property Count Loss:** Removes the **target-property count regression loss** (MSE), relying only on generation and treatment prediction supervision.
- **w/o Treatment Prediction Loss:** Removes the **treatment prediction loss** ($CrossEnt$), training only on generation and target-property count regression. In this setting, propensity-based estimators (IPW, AIPW) are not applicable.
- **Generation Only:** Removes both auxiliary objectives, reducing the model to a standard fine-tuned T5.

Table 8 reports the Absolute Error for the ATE. We observe that the **Full CausalT5** model yields the lowest error across all valid estimators. Notably, removing the target-property count regression loss degrades the Q-only estimator (error increases from 0.24 to 0.37) and significantly harms the doubly robust AIPW estimator (error increases from 0.09 to 0.22). The Generation Only baseline performs poorly across the board, confirming that the auxiliary causal objectives are essential for recovering the true treatment effect signal.

Model Variant	ATE Absolute Error		
	Q-only	IPW	AIPW
Full CausalT5	**0.24**	**0.18**	**0.09**
w/o Count Loss	0.37	0.25	0.22
w/o Treatment Loss	0.37	–	–
Generation Only	0.66	–	–

Table 8: Ablation study on the semi-synthetic dataset ($p = 0.9$). **w/o Count Loss** removes the target-property count regression loss; **w/o Treatment Loss** removes the treatment prediction loss. The full multi-task objective achieves the lowest estimation error.

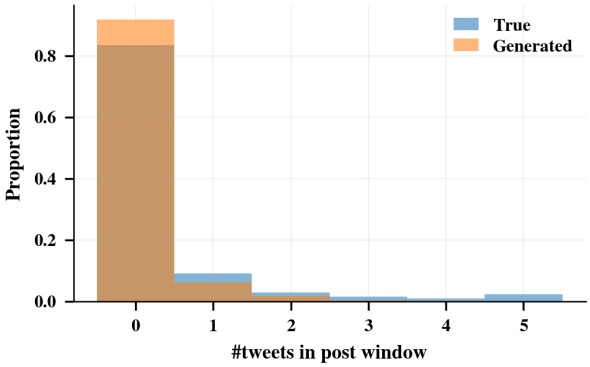

Figure 5: Distribution of low-credibility tweet counts in post windows (real vs. generated). The alignment suggests fidelity in outcome-relevant statistics.

A.12 Low-credibility Tweet Distribution

Figure 5 compares the distribution of low-credibility tweet counts between generated and real tweets in post-intervention window. The two distributions align closely, indicating that CausalT5 preserves the aggregate frequency of outcome events that are central to causal estimation.

A.13 Qualitative Examples of Topic Swaps

For additional qualitative assessment, Table 9 illustrates the effect of the swap mechanism on the tweets generated for a Class 1 user in the semi-synthetic dataset. By design, Class 1 post-intervention windows should contain low-credibility tweets that are predominantly focused on health-related content ($p_{\text{swap}} = 0.8$). Consistent with this design, the observed post-intervention tweet shown here is unrelated to health, whereas the generated counterfactual focuses on vaccines.

A.14 Effect of Propensity on Counterfactual Similarity

We examine how counterfactual–proxy similarity varies across units with low vs. high propensity scores. Figure 6 shows the similarity distributions for control-group instances (filtered to pre-match similarity ≥ 0.65 to ensure proxy quality). High-propensity controls ($p > 0.95$) exhibit substantially higher similarity scores than low-propensity controls ($p < 0.05$). This aligns with the theoretical expectation that controls with high propensity are closer to the treatment group and thus have more well-defined counterfactuals—making their generated counterfactuals more consistent with matched treated proxies.

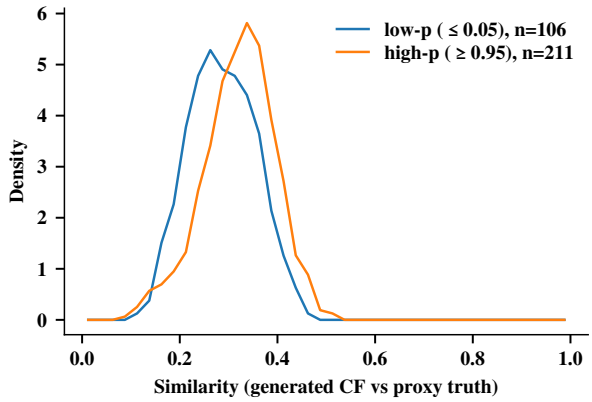

Figure 6: Distribution of counterfactual–proxy similarity for control units with high vs. low propensity scores ($p > 0.95$ vs. $p < 0.05$), filtered to pre-match similarity ≥ 0.65.

A.15 ATE result for three-class setting

Table 10 reports ATE estimation errors under the three-class semi-synthetic setup. The best results for CausalT5 are again statistically indistinguishable from those for the encoder baseline for both 0.9 and 0.7 confounding. While the Q-only estimator shows large errors for CausalT5 (especially at p=0.7), both IPW and AIPW remain stable and achieve performance comparable to the encoder baseline. Treatment prediction accuracy is slightly lower than in the two-class setting (0.79 vs. 0.77 at p=0.9 and 0.68 vs. 0.67 at p=0.7 for CausalT5 and the encoder baseline, respectively), likely because the introduction of Class 0—whose treatment is assigned at random—weakens the confounding signal and makes treatment prediction inherently harder.

A.16 CATE results for p=0.7

Table 11 shows the CATE results on semi-synthetic data with confounding strength 0.7.

Observed Post-Intervention Tweet (Y_1)	Counterfactual Post-Intervention Tweet ($\hat{Y}_1$)
RT @majorgauravarya: What the Chief is saying is logical. If you threaten India, it's army will fight back...through action on ground & statements of intent. And it's not just China or Pak. Even inside India there are people who ask "Why does the General speak so much?" **indiatimes**	RT **@OccupyDemocrats**: BREAKING NEWS: The University of Virginia disenrolls 238 students who failed to get their COVID vaccines, violating the mandate set by the school. RT IF YOU THINK THAT ALL UNIVERSITIES SHOULD FOLLOW THEIR LEAD!

Table 9: An example from the RQ3 semi-synthetic dataset (Class 1, treated). The observed post-intervention tweet (Y_1) is unrelated to health, while the generated counterfactual ($\hat{Y}_1$) produces a health-related, low-credibility tweet, as expected by the correlation introduced in the topic swap for Class 1. Tweets from low-credibility sources are highlighted in **red**.

p	Estimator	CausalT5	Encoder Baseline
0.9	Q-only	0.47 ± 0.18	$\mathbf{0.06 \pm 0.08}$
	IPW	0.13 ± 0.08	0.10 ± 0.09
	AIPW	$\mathbf{0.05 \pm 0.07}$	0.06 ± 0.06
	Unadjusted	0.32 ± 0.00	0.32 ± 0.00
0.7	Q-only	0.74 ± 0.11	$\mathbf{0.02 \pm 0.03}$
	IPW	0.14 ± 0.13	0.12 ± 0.10
	AIPW	$\mathbf{0.07 \pm 0.04}$	0.09 ± 0.05
	Unadjusted	0.18 ± 0.00	0.18 ± 0.00

Table 10: ATE absolute error (mean $\pm$ std across 5 folds) in the three-class semi-synthetic setting (**RQ3**). Best results are highlighted for CausalT5 and the Encoder Baseline.

Confounding Strength $p = 0.7$			
Class	Estimator	CausalT5	Encoder Baseline
0	Q-only	1.23 ± 0.10	0.18 ± 0.04
	IPW	0.12 ± 0.15	$\mathbf{0.05 \pm 0.02}$
	AIPW	$\mathbf{0.09 \pm 0.06}$	0.10 ± 0.08
1	Q-only	0.20 ± 0.08	0.20 ± 0.05
	IPW	0.50 ± 0.32	0.21 ± 0.16
	AIPW	$\mathbf{0.19 \pm 0.11}$	$\mathbf{0.16 \pm 0.12}$
2	Q-only	1.10 ± 0.08	$\mathbf{0.07 \pm 0.08}$
	IPW	0.38 ± 0.18	0.25 ± 0.04
	AIPW	$\mathbf{0.10 \pm 0.08}$	0.09 ± 0.04

Table 11: CATE absolute error (mean $\pm$ std across 5 folds) in the three-class semi-synthetic setting for confounding strength 0.7. Lower is better. Best results are highlighted for CausalT5 and the Encoder Baseline.

A.17 Attention Pooling Details

Given the encoder output $h = (h_1, h_2, \ldots, h_m)$, we compute a single pooled vector $\bar{h}$ for treatment prediction using attention weights α_i:

$$\alpha_i = \frac{\exp(w^\top h_i)}{\sum_{j=1}^{m} \exp(w^\top h_j)}, \qquad \bar{h} = \sum_{i=1}^{m} \alpha_i h_i. \quad (8)$$

Here, w is a learnable parameter vector. The pooled representation $\bar{h}$ is then fed into a binary classifier to estimate the propensity score:

$$\hat{g} = \sigma(W_g \bar{h} + b_g), \quad (9)$$

where $\sigma(\cdot)$ is the sigmoid activation function, and W_g, b_g are the classifier weights and bias.

A.18 Human Validation of GPT Extremeness Scores

To assess the reliability of the GPT-based political extremeness scores, we sampled 100 tweets from the behavioral fidelity evaluation and asked two human annotators to independently rate each tweet using the same 1–5 political extremeness rubric given to GPT (see the detail annotation instruction in Appendix A.19). Since the labels are ordinal, we report linear weighted Cohen's κ in addition to exact agreement. Table 12 suggest that GPT's ratings are broadly consistent with human judgments and fall within the range of agreement observed between human annotators.

Comparison	Linear weighted κ	Exact agreement
Human 1 vs. Human 2	0.537	43.0%
Human 1 vs. GPT	0.511	45.0%
Human 2 vs. GPT	0.508	43.0%

Table 12: Human validation of GPT-based political extremeness scores on 100 examples. GPT–human agreement is comparable to human–human agreement.

A.19 Annotation Instructions

Human annotators were provided the following instructions:

> You will read blocks of tweets and rate their political extremeness.
>
> Note that this dataset is highly political, so even lower scores may still contain political content. Please judge the intensity of the political language rather than whether the text is political.
>
> Each item may contain five tweets labeled Tweet 1, Tweet 2, ... Some tweets include "[Reference

text]", which provides additional context. Please rate the overall tone of the entire tweet block.

Before starting, please review the example dataset provided, which contains sample texts for each score to help calibrate your judgments.

Rate political extremeness on a 1–5 scale:

1 — Not extreme (neutral tone)

2 — Mildly partisan

3 — Moderately partisan

4 — Strongly partisan / polarized

5 — Extremely partisan or inflammatory

If unsure between two scores, choose the lower score.

A.20 Few-Shot Prompted LLM Baseline

To evaluate whether modern prompted LLMs can recover treatment effects without task-specific causal training, we implement a no-training few-shot prompted baseline using Qwen2.5-7B-Instruct (Yang et al., 2024). Unlike CausalT5, this baseline does not learn treatment propensities or perform supervised causal training.

For each validation fold, we construct a balanced few-shot prompt from the training split using examples from each treatment and latent confounder class combination. Each demonstration contains: (1) pre-intervention tweets X, (2) the treatment condition, and (3) observed post-intervention continuation tweets Y. We prioritize shorter examples that contain score-tagged tweets to ensure that the demonstrations expose the model to the target outcome format.

For each evaluation instance, the model receives the pre-intervention tweets X together with the few-shot demonstrations. We then use two separate prompts: one prompting the model to generate post-treatment tweets under the treatment condition ($\hat{Y}_1$), and another prompting the model to generate post-treatment tweets under the control condition ($\hat{Y}_0$). Generation is deterministic (`do_sample=False`) and produces exactly five continuation tweets separated by `[TWEET_SEP]`.

The generated texts are converted into scalar outcomes using the same outcome function $z(\cdot)$ as in CausalT5, yielding predicted potential outcomes

$$\hat{Q}_1 = z(\hat{Y}_1), \qquad \hat{Q}_0 = z(\hat{Y}_0).$$

Because the prompted baseline does not estimate treatment propensities, only the Q-only estimator is applicable. We therefore estimate ATE as

$$\hat{\tau}_Q = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{Q}_1(X_i) - \hat{Q}_0(X_i)\right).$$

Table 13 reports ATE absolute error for the prompted LLM baseline on the two-class semi-synthetic RQ2 setting.

p	Estimator	Prompted LLM Baseline
0.7	Q-only	0.193 ± 0.047
0.9	Q-only	0.444 ± 0.662

Table 13: ATE absolute error (mean $\pm$ std across 5 folds) for the few-shot prompted LLM baseline using Qwen2.5-7B-Instruct on the two-class semi-synthetic setting. The prompted baseline becomes substantially less stable under stronger confounding ($p = 0.9$).