Challenges and an Empirical Evaluation Framework for Text-based Confounding Adjustment

Galen Weld[*=], Peter West[*=], Maria Glenski[†], David Arbour[‡], Ryan A. Rossi[‡] and Tim Althoff[*]

[*]Paul G. Allen School of Computer Science and Engineering
[†]Pacific Northwest National Laboratory
[‡]Adobe Research
[=]Contributed equally

Abstract

Leveraging text, such as social media posts, for *causal inferences* requires the use of NLP models to 'learn' and adjust for confounders, which could otherwise impart bias. However, evaluating such models is challenging, as ground truth is almost never available. We demonstrate the need for empirical evaluation frameworks for causal inference in natural language by showing that existing, commonly used models regularly disagree with one another on real world tasks. We contribute the first such framework, generalizing several challenges across these real world tasks. Using this framework, we evaluate a large set of commonly used causal inference models based on propensity scores and identify their strengths and weaknesses to inform future improvements. We make all tasks, data, and models public to inform applications and encourage additional research.

1 Introduction

A frequent goal for computational social science practitioners is to understand the casual effect of intervening on a treatment of interest. Researchers often operationalize this by estimating the *average treatment effect* (ATE) of a specific treatment variable (*e.g.* therapy) on a specific outcome (*e.g.* suicide) (Pearl, 1995, 2009; Rosenbaum, 2010; Keith et al., 2020). A major challenge is adjusting for *confounders* (*e.g.* comments mentioning depression) that affect both the treatment and outcome (depression affects both an individual's propensity to receive therapy and their risk of suicide) (Keith et al., 2020). Without adjusting for depression as a confounder, we might look at suicide rates among therapy patients and those not receiving therapy, and wrongly conclude that therapy causes suicide.

The gold standard for avoiding confounders is to assign treatment via a *randomized controlled trial* (RCT). Unfortunately, in many domains, assigning treatments in this manner is not feasible (*e.g.* due to ethical or practical concerns). Instead, researchers conduct *observational studies* (Rosenbaum, 2010), using alternate methods to adjust for confounders.

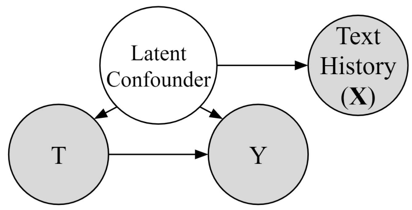

Figure 1: Causal graph representing the the context of our evaluation framework. All edges have known probabilities. While our framework naturally generalizes to more complex scenarios, we chose binary treatments and outcomes, and a binary latent confounder, as even in this simple scenario, current models struggle (§7).

Text (*e.g.* users' social media histories) can be used to adjust for confounding by training a model to recognize confounders (or proxies for confounders) in the text, so that similar treated and untreated observations can be compared. However, a recent review (Keith et al., 2020) finds that evaluating the performance of such models is "a difficult and open research question" as *true* ATEs are almost never known, and so, unlike in other NLP tasks, we cannot know the correct answer. In this work, we find that this challenge is amplified, as models disagree with one another on real world tasks (§3) – how do we know which is correct?

As ground truth is almost never available, the only[1] practical method to evaluate causal inference models is with semi-synthetic data, where synthetic treatments and outcomes are assigned to real observations, as in Fig. 1 (Dorie et al., 2019; Jensen, 2019; Gentzel et al., 2019). While widely-used semi-synthetic benchmarks have produced positive results in the medical domain (Dorie et al., 2019),

[1]With the extremely rare exception of *constructed* observational studies, conducted with a parallel RCT.

no such benchmark exists for causal inference models using text (Gentzel et al., 2019; Dorie et al., 2019; Jensen, 2019; Keith et al., 2020).

In this work, we contribute the first evaluation framework for causal inference with text, consisting of five tasks inspired by challenges from a wide range of studies (Keith et al., 2020) (§4): Linguistic Complexity, Signal Intensity, Strength of Selection Effect, Sample Size, and Placebo Test. Each semi-synthetic task is generated (§5) from real Reddit users' profiles, perturbed with synthetic posts to create increasing levels of difficulty (§5.2). This principled approach lets us evaluate the specific strengths and weakness of widely-used models (§6.1) and estimators (§6.2).

Concerningly, we find that almost every model predicts a false significant treatment effect when none is present, which could be greatly misleading to unwary practitioners (§7). While we find that each model struggles with at least one challenge, models leveraging recent, hierarchical, transformer-based architectures perform best, although such models are not yet widely used (Keith et al., 2020). **For NLP researchers:** We make our tasks, data and models publicly available[2] to encourage the development of stronger models for causal inference with text and identify areas for improvement (§8). **For CSS practitioners:** We identify strengths and weaknesses of commonly used models, identifying those best suited for specific applications, and make these publicly available[2] (§ 8).

2 Background and Related Work

Causal Inference Primer. We formalize causal inference using notation from Pearl (1995). Given a series of n observations (in our context, a social media user), each observation is a tuple $O_i = (Y_i, T_i, \mathbf{X}_i)$, where Y_i is the outcome (*e.g.* did user i develop a suicidal ideation?), T_i is the treatment (*e.g.* did user i receive therapy?), and $\mathbf{X}_i$ is the vector of observed covariates (*e.g.* user i's textual social media history).

The *Fundamental Problem of Causal Inference* is that each user is either treated *or* untreated, and so we can never observe *both* outcomes. Thus, we cannot compute the $ATE = \frac{1}{n}\sum_{i=1}^{n} Y_i\left[T_i = 1\right] - Y_i\left[T_i = 0\right]$ directly, and must estimate it by finding comparable treated and untreated observations. To do so, it is common practice to use a model to estimate the **propensity score**, $\hat{p}(\mathbf{X}_i) \approx p(T_i = 1|\mathbf{X}_i)$, for each observation i. As treatments are typically known, propensity score models are effectively supervised classifiers, predicting T_i, given $\mathbf{X}_i$. Matching, stratifying, or weighting using these propensity scores will produce an unbiased ATE estimate (§6.2) if three assumptions hold: all confounders must be observed, propensity scores must be accurate, and there must be overlap in the distribution of covariates in the treated and untreated groups (*common support assumption*) (Rosenbaum, 2010; Hill and Su, 2013). In practice, verifying these assumptions is difficult, hence the need for empirical evaluation.

Causal Inference and NLP. Until recently, there has been little interaction between causal inference researchers and the NLP research community (Keith et al., 2020). There are many ways to consider text in a causal context, such as text as a mediator (Veitch et al., 2019), text as treatment (Wood-Doughty et al., 2018; Egami et al., 2018; Fong and Grimmer, 2016; Tan et al., 2014), text as outcome (Egami et al., 2018; Zhang et al., 2018), and causal discovery from text (Mani and Cooper, 2000; Mirza and Tonelli, 2016). However, we narrow our focus to text *as a confounder*, as in Keith et al. (2020). This is an important area of research because the challenge of adjusting for confounding underlies most causal contexts, such as text as treatment or outcome (Keith et al., 2020). Effective adjusting for confounding with text enables causal inference in any situation where observations can be represented with text – *e.g.* social media, news articles, and dialogue.

Adjusting for Confounding with Text. A recent ACL review (Keith et al., 2020, Table 1) summarizes common practices across a diverse range of studies. Models and text representations used in these applications do not yet leverage recent breakthroughs in NLP, and generally fall into three groups: those using uni- and bi-gram representations (De Choudhury et al., 2016; Johansson et al., 2016; Olteanu et al., 2017), those using LDA or topic modeling (Falavarjani et al., 2017; Roberts et al., 2020; Sridhar et al., 2018), and those using neural word embeddings such as GLoVe (Pham and Shen, 2017) and BERT (Veitch et al., 2019). Three classes of estimators are commonly used to compute the ATE from text data: **inverse probability of treatment weighting** (IPTW), propensity score **stratification**, and **matching**, either using propensity scores or some other distance metric.

[2] Dataset Website

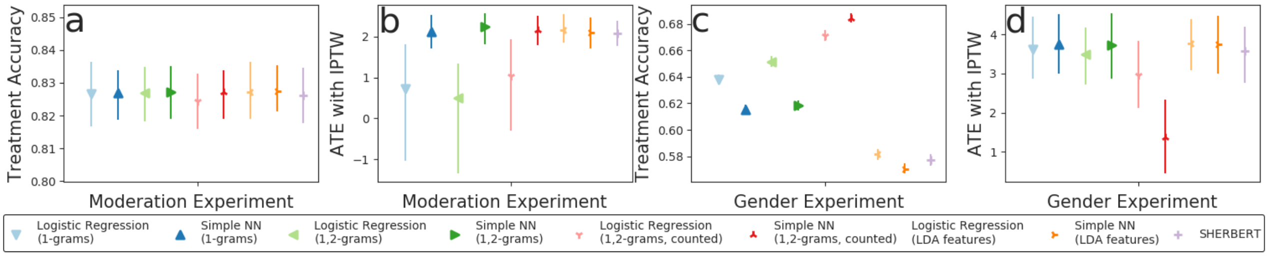

Figure 2: Treatment accuracy and ATE for both real world experiments, with bootstrapped 95% confidence intervals. Note that for the Gender Experiment, the models with the highest accuracy have the lowest ATE.

In this work, we evaluate at least one variant of every commonly used model (§6.1) and estimator (§6.2). While we focus on the propensity score methods, which are the most popular (Keith et al., 2020), our framework's structure enables evaluation of *any* ATE estimation method, including those computed using non-propensity score-based matching, such as TIRM (Roberts et al., 2020) and exact matches (Mozer et al., 2020).

Evaluation of Causal Inference. In rare specialized cases, researchers can use the unbiased outcomes of a parallel RCT to evaluate those of an observational study, as in Eckles and Bakshy (2017). This practice is known as a *constructed* observational study, and, while useful, is only possible where parallel RCTs can be conducted. Outside these limited cases, proposed models are typically evaluated on synthetic data generated by their authors. These synthetic datasets often favor the proposed model, and do not reflect the challenges faced by real applications (Keith et al., 2020).

Outside of the text domain, *widely used* evaluation datasets have been successful, most notably the 2016 Atlantic Causal Inference Competition (Dorie et al., 2019), and a strong case has been made for the empirical evaluation of causal inference models (Gentzel et al., 2019; Jensen, 2019). In the text domain, matching approaches have been evaluated empirically (Mozer et al., 2020), but this approach evaluates only the quality of *matches*, not the causal effect estimates. In contrast, our work applies to all estimators, not just matching, and evaluates the entire causal inference pipeline.

3 Current Models Disagree

Recent causal inference papers (Veitch et al., 2019; Roberts et al., 2020; De Choudhury et al., 2016; Chandrasekharan et al., 2017; Bhattacharya and Mehrotra, 2016) have used social media histories to adjust for confounding. Each of these papers uses a different model: BERT in Veitch et al. (2019), topic modeling in Roberts et al. (2020), and logistic regression in De Choudhury et al. (2016). For all of these studies, ground truth causal effects are unavailable, and so we cannot tell if the chosen model was correct. However, we *can* compute their prediction accuracy on propensity scores, and see if their ATE estimates agree—if they don't, then at most one disagreeing model can be correct.

Methods. We conducted two experiments using real world data from Reddit, inspired by these recent papers. In the **Moderation Experiment**, we test if having a post removed by a moderator impacts the amount a user later posts to the same community again. In the **Gender Experiment**, we use data from Veitch et al. (2019) to study the impact of the author's gender on the score of their posts. For details on data collection, see §A.

Results. Comparing the performance of nine different models (Fig. 2), we find that all models have similar treatment accuracy in the Moderation Experiment. However, the models using 1,2-gram features perform better in the Gender Experiment than the LDA and SHERBERT models. Most importantly, we see that all models have mediocre treatment accuracy (Fig. 2a,c) and the models with the highest treatment accuracy produce the lowest ATE estimates (Fig. 2b,d), which in many cases disagree entirely with estimates from other models.

Implications. This should come as a great concern to the research community. We do not know which model may be correct, and we do not know whether there may be a more accurate model that would even further decrease the estimated treatment effect. We derive theoretical bounds and compute them, finding that in 99+% of cases, these bounds are looser than those computed empirically using our framework (§C), making them less useful for model selection. This concern motivates our research questions (§1) and underlines the importance and urgency of empirical evaluation for causal inference in natural language. Next, we describe key challenges in adjusting for confounding with text and present a principled evaluation framework that highlights these challenges and generates actionable insights for future research.

4 Challenges for Causal Inference with Natural Language

Using the common setting of real social media histories (De Choudhury et al., 2016; Olteanu et al., 2017; Veitch et al., 2019; Choudhury and Kiciman, 2017; Falavarjani et al., 2017; Kiciman et al., 2018; Saha et al., 2019; Roberts et al., 2020), we identify five challenges consistently present when representing natural language for causal inference:

1. **Linguistic Complexity:** Different expressions can be indicative of important underlying commonalities and signals. Someone who struggles with mental health might write "I feel depressed" or "I am isolated from my peers," which have distinct meanings but both may be indicative of depression. *Can models recognize that both are relevant?*
2. **Signal Intensity:** Some users only have a few posts that contain a specific signal (such as poor mental health) whereas others may have many posts with this signal. Signals are especially weak when posts containing the signal constitute only a small fraction of a user's posts. *Can models detect weak signals?*
3. **Strength of Selection Effect:** Many studies have few comparable treated and untreated users (§2) (Li et al., 2018; Crump et al., 2009). *Can models adjust for strong selection effects?*
4. **Sample Size:** Observational studies often face data collection limitations.[3] *Can models perform well with limited data samples?*
5. **Placebo Test:** Oftentimes, no causal effect is present between a given treatment and an outcome. *Do models falsely predict causality when none is present?*

While natural language is far more complex than any finite set of challenges can capture, the five we have chosen to highlight are challenges that regularly need to be addressed in causal inference tasks that use natural language (Keith et al., 2020). They also cover three key concepts of model performance: generalizability (*linguistic complexity*), sensitivity (*signal intensity*, *strength of selection effect*), and usability (*sample size*, *placebo test*) that are critical for comprehensive evaluation. To produce our evaluation framework, we derive a concrete task from each challenge.

[3]In Keith et al. (2020, Table 1), 8/12 studies had fewer than 5,000 observations, and 4/12 had fewer than 1,000.

5 Framework for Evaluation

We generate five tasks, each with discrete levels of difficulty, and corresponding semi-synthetic task datasets based on real social media histories. Without the semi-synthetic component, it would not be possible to empirically evaluate a model, as we would not know the true ATE. By basing our user histories on real data, we are able to include much of the realism of unstructured text found 'in the wild.' This semi-synthetic approach to evaluation preserves the best of both worlds: the empiricism of synthetic data with the realism of natural data (Jensen, 2019; Gentzel et al., 2019; Jensen, 2019).

5.1 Semi-Synthetic Dataset Generation

The method for generating a semi-synthetic dataset can be arbitrarily complex, however, for simplicity and clarity, we generate our datasets according to a simplified model of the universe; where all confounding is present in the text, and where there are only two types of people, `class 1` and `class 2` (Fig. 1). In the context of mental health, for example, these two classes could simply be people who struggle with depression (`class 1`), and those who don't (`class 2`). If models struggle on even this simple two-class universe, as we find, then it is highly unlikely they will perform better in the more complex real world. In this universe, the user's (latent) class determines the probability of treatment and outcome conditioned on treatment. Dependent on class, but independent of treatment and outcome is the user's comment history, which contains both synthetic and real posts that are input to the model to produce propensity scores.

We produce each dataset using a generative process (§B). For each task, we start with the same collection of real world user histories from public Reddit profiles. We randomly assign (with .5/.5 probability) each user to `class 1` or `class 2`. Into each profile, we insert synthetic posts using a function f_n for `class` n specific to each task, described in §5.2. We assign binary treatments (conditioned on class) and binary outcomes (conditioned on class and treatment) according to a known probability distribution (§B). These outcomes and treatments could represent anything of interest, and they need not be binary.

To estimate the ATE, there must be overlap between the treated and untreated groups (§2), so we cannot make all users in `class 1` treated and all users in `class 2` untreated; instead, we assign treatment with a biased coin-flip: Treated with $P = .9$ and untreated with $P = .1$ for `class 1`, and the opposite for `class 2`. These true propensities are not more extreme than those commonly accepted in practice (Crump et al., 2009; Lee et al., 2011; Yang and Ding, 2018).

Once a treatment has been assigned according to the class' probabilities, a positive outcome is assigned with probability .9 (treated) and .1 (untreated) for `class 1`, and .9 for both treatments for `class 2`. These probabilities are the 'default' and are used in all Tasks except Tasks 5.2.3 and 5.2.5, where we vary them to explore those specifically. The objective for models (§6.1) is to recover these probabilities in the form of a propensity score.

5.1.1 Real World User Histories

We use Reddit user histories as the real world component of our semi-synthetic datasets. Reddit was selected as our natural data source due to its use in De Choudhury et al. (2016), its public nature, and its widespread use in the research community.

We downloaded all Reddit comments for the 2014 and 2015 calendar years from the Pushshift archives (Baumgartner et al., 2020) and grouped comments by user. After filtering out users with fewer than 10 comments, we randomly sampled 8,000 users and truncated users' histories to a maximum length of 60 posts for computational practicality.[4] These users were randomly partitioned into three sets: a 3,200 user training set, an 800 user validation set, and a 4,000 user test set used to compute Treatment Accuracy and ATE Bias.

5.1.2 Synthetic Posts

When generating semi-synthetic tasks, we insert three types of synthetic posts (§D), representative of major life events that could impact mental health, into real users' histories:

- *Sickness Posts* describe being ill (*e.g.* 'The doctor told me I have AIDS'). We vary both the illness, as well as way the it is expressed.
- *Social Isolation Posts* indicate a sense of isolation or exclusion. ('I feel so alone, my last friend said they needed to stop seeing me.')
- *Death Posts* describe the death of companion (*e.g.* 'I just found out my Mom died'). We vary the phrasing as well as the companion.

A complete list of all posts of each type is in §D.

5.2 Tasks

We consider five tasks focused around the common challenges for text-based causal inference methods previously highlighted in §4.

5.2.1 Linguistic Complexity

This task tests a model's resilience to the linguistic complexity of text inputs, *i.e.* the ability to recognize synonyms and the shared importance of dissimilar phrases. We increase the difficulty in four steps by increasing the diversity of synthetic sentences inserted into user histories assigned to `class 1` (*i.e.* the linguistic complexity of the dataset): f_1 initially appends the same Sickness Post to the end of each `class 1` user's history; At the second level of difficulty, f_1 selects a Sickness Post uniformly at random; At the third level, f_1 selects either a Sickness or Social Isolation Post; and at the fourth level, f_1 selects a Sickness, Social Isolation, or Death Post. For each level of difficulty, f_2 is the identity function, *i.e.* user histories assigned to `class 2` are unchanged.

5.2.2 Signal Intensity

This task tests a model's ability to distinguish between the number of similar posts in a history. There are two levels of difficulty. At the easier level, f_1 appends 10 randomly sampled (with replacement) Sickness Posts, while f_2 is the identity function. At the harder level, f_1 appends only three Sickness Posts, while f_2 appends one.

5.2.3 Strength of Selection Effect

In this and the following tasks, we do not vary f_1 or f_2. For *Strength of Selection Effect*, we make causal inference more challenging by increasing the strength of the selection effect, decreasing the overlap between treated and untreated users (§2). We test two levels of difficulty: a weaker selection effect (easier) with the same .9/.1 split to assign the majority of `class 1` to the treated group and `class 2` to the control group. For the stronger selection effect (harder), we increase this split for `class 1` to .95/.05. For both the weak and strong selection effects, we use f_1 to append a single random Sickness Post and f_2 as the identity function. Outcome probabilities, conditioned on treatment, are unchanged from §5.1.

[4] The resulting set of users had a mean of 41 posts/user, mean of 37.37 tokens/post, and a mean of 1523.28 tokens/user.

5.2.4 Sample Size

In this task, we test how the models' performance drops off as the amount of available training data is reduced.[5] As before, we use f_1 to append a single random Sickness Post and f_2 as the identity function. For the easiest case, we train on all 3,200 users' histories in the training set. We then create smaller training sets by randomly sampling subsets with 1,600 and 800 users.

5.2.5 Placebo Test

The final task assesses a model's tendency to predict a treatment effect when none is present. To do so, we must have asymmetric treatment probabilities between `class 1` and `class 2`. Without this asymmetry, the unadjusted estimate would be equal to the true ATE of zero. We use the same asymmetric `class 1` treatment split as in §5.2.3.

We set $P(Y = 1|T = 0, \texttt{class=1}) = .05$, $P(Y = 1|T = 1, \texttt{class=2}) = .95$, and the opposite for $Y = 0$. This gives a treatment effect of +.9 to `class 1` and a treatment effect of -.9 to `class 2`, making the true ATE for the entire task equal 0. As in previous tasks, f_1 appends one random Sickness Post and f_2 is the identity function.

6 Causal Inference Pipeline

We evaluate commonly used text representations, propensity score models, and ATE estimators (§2).

6.1 Propensity Score Models

The **Oracle** uses the true propensity scores, which are known in our semi-synthetic evaluation framework (§5). The Oracle provides an upper-bound on model performance, only differing from the theoretical optimum due to finite sample effects.

We include an **Unadjusted Estimator**, which uses the naive method of not adjusting for selection effects, producing an estimated treatment effect of $\bar{Y}_{T=1} - \bar{Y}_{T=0}$, and as such is a lower-bound for models that attempt to correct for selection effects.

We train a **Simple Neural Net** (with one fully connected hidden layer) in four variants with different text representations: 1-grams with a binary encoding, 1,2-grams with a binary encoding, 1,2-grams with counts, and Latent Dirichlet Allocation (LDA) features (Blei et al., 2003) based on 1,2-grams, counted. We also train **Logistic Regression** models on the same four text representations.

Finally, we propose and evaluate a novel cauSal HiERarchical variant of BERT, which we call **SHERBERT**. SHERBERT expands upon Causal BERT proposed by Veitch et al. (2019), which is too computationally intensive to scale to user histories containing more than 250 tokens, let alone ones orders of magnitude longer, such as in our tasks. In SHERBERT, we use one pretrained BERT model per post to produce a post-embedding (Appendix Fig. 5), followed by two hierarchical attention layers to produce a single embedding for the entire history, with a final linear layer to estimate the propensity score. This architecture is similar to HIBERT (Zhang et al., 2019), but is faster to train on long textual histories, as SHERBERT fixes the pretrained BERT components.

6.2 Average Treatment Effect Estimators

We consider three commonly used ATE estimators – IPTW, stratification, and matching. All three estimators use propensity scores (§2) but differ in how they weight or group relevant samples.

Inverse Propensity of Treatment Weighting estimates the ATE by weighting each user by their relevance to selection effects:

$$\widehat{ATE}_{\texttt{IPTW}} = \sum_{i=1}^{n} \frac{(2 * T_i - 1) * Y_i}{\hat{p}_{T_i}(\mathbf{X}_i) * \left[\sum_{j=1}^{n} \frac{1}{\hat{p}_{T_j}(\mathbf{X}_j)}\right]}$$

where T_i, Y_i, and X_i are treatment, outcome, and features for sample i, and $\hat{p}_T(\mathbf{X})$ is the estimated propensity for treatment T on features $\mathbf{X}$. Use of the Hajek estimator (1970) adjustment improves stability compared to simple inverse propensity.

Stratification divides users into strata based on their propensity score, and the ATE for each is averaged: $\widehat{ATE}_{\texttt{strat}} = \frac{1}{n} \sum_k n_k * \widehat{ATE}_k$ where n is the total number of users, n_k is the number of users in the k-th stratum, and $\widehat{ATE}_k$ is the unadjusted ATE within the k-th stratum. We report results on 10 strata divided evenly by percentile, but results are qualitatively similar for other k.

Matching can be considered as a special case of stratification, where each strata contains only one treated user. As matching produces extremely similar results to stratification, we include details of our approach and plots of the results in §F.1.

[5] In Keith et al. (2020, Table 1), 8/12 studies had fewer than 5,000 observations, and 4/12 had fewer than 1,000.

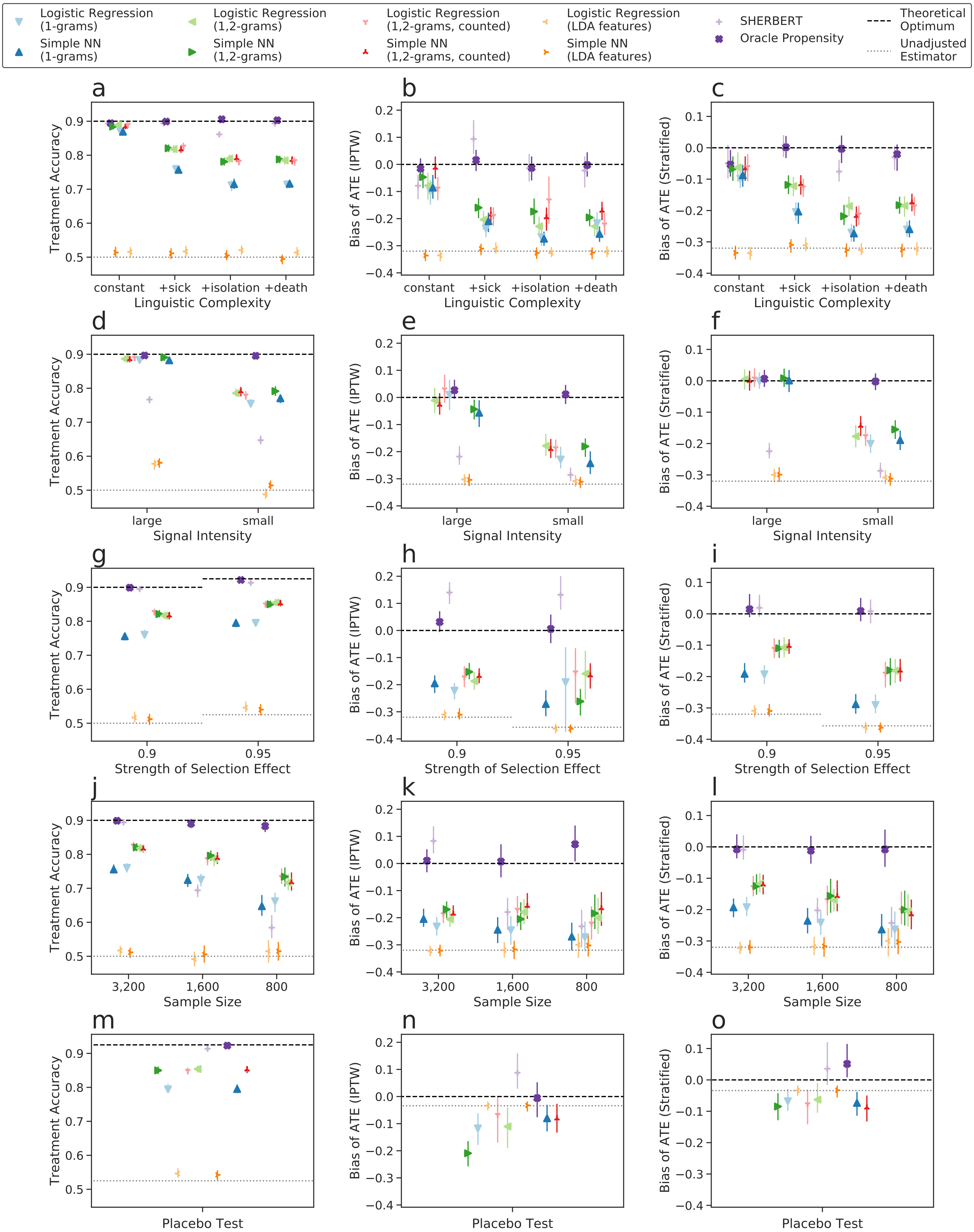

Figure 3: Results for tasks, with bootstrapped 95% confidence intervals, perturbed along the x-axis for readability. Within each plot, difficulty increases from left to right. SHERBERT generally does well, especially on *Strength of Selection Effect* and *Absence of Non-Zero Treatment Effect*, but struggles on *Signal Intensity*.

6.3 Metrics for Evaluation

Our semi-synthetic tasks are generated such that we know the true ATE and thus can compute the **Bias of $\widehat{\mathbf{ATE}}$**. A bias of zero is optimal, indicating a correct estimated ATE. The greater the bias, positive or negative, the worse the model performance. This is the primary metric we use in evaluation, and we compute it for both $\widehat{ATE}_{\texttt{strat}}$ and $\widehat{ATE}_{\texttt{IPTW}}$. We also consider **Treatment Accuracy**, the accuracy of the model's predictions of binary treatment assignment. While higher accuracy is often better, high accuracy does not guarantee low bias. We include additional metrics (Spearman Correlation of Estimated Propensity Scores and Mean Squared Error of IPTW for each task) in §F.2.

7 Evaluation of Common Models

Transformers better model relevant linguistic variation. Many trends in the results manifest in the Linguistic Complexity task (§5.2.1), including treatment accuracy clustering by text representation (Fig. 3a). SHERBERT performs well, with uni- and bi-gram models somewhere in between. Accuracy correlates fairly well with bias (Fig. 3b,c). As in nearly all tasks, LDA models perform worst, not even outperforming the unadjusted estimator.

Transformer models struggle with counting and ordering. The Signal Intensity task (§5.2.2) requires models to effectively 'count' the number of posts to distinguish between classes.This is the only task where n-gram models outperform SHERBERT (Fig. 3e,f) and LDA models perform slightly better than not adjusting at all, due to the stronger presence of tokens correlated with treatment.

High accuracy often reflects strong selection effects, not low *ATE* bias. In the Strength of Selection Effect task (§5.2.3), we decrease the overlap in propensity scores between treated and untreated users which makes it *easier* to distinguish between the two groups. We see corresponding *increases* in Treatment Accuracy (Fig. 3g), however, bias worsens (Fig. 3h,i). In context of observational studies, models with high treatment accuracy should be used with extreme caution — high accuracy likely reflects that the common support assumption is violated, preventing causal inference. This highlights the importance of empirical evaluation of the *complete* causal inference pipeline.

Transformer models fail with limited data. The Sample Size task (§5.2.4) explores models' performance on small datasets, a common occurrence in real world applications. SHERBERT outperforms other models except when there is very limited data available, with accuracy and bias dropping below n-gram features when data is reduced (Fig. 3j,k,l).

Models predict causality when none is present. Alarmingly, in the Placebo Test (§5.2.5), every model except SHERBERT failed to include the (correct) null hypothesis ($ATE = 0$) in their 95% confidence intervals (Fig. 3n,o), including high accuracy models using bigram features (Fig. 3m). This result is of greatest concern, as eight out of nine methods falsely claim a non-zero effect.

Models have greater impact than estimators. Each estimator evaluated produced overall similar results (Fig. 6), with the quality of the propensity scores being far more impactful. However, IPTW is more sensitive to extreme propensity scores (Fig. 3h). See §F.2 for more details.

8 Implications & Conclusions

Causal inferences are difficult to evaluate in the absence of ground truth causal effects – a limitation of virtually all real world observational studies. Despite this absence, we *can* compare different models' estimates and demonstrate that different models regularly disagree with one another.

Empirical evaluation requires knowledge of the true treatment effects. Our proposed evaluation framework is reflective of five key challenges for causal inference in natural language.

We evaluate every commonly used propensity score model to produce key insights:

For **NLP Researchers**, we find that continued development of transformer-based models offers a promising path towards rectifying deficiencies of existing models. Models are needed that can effectively represent the order of text, variability in expression, and the counts of key tokens. Given the limited availability of training data in many causal inference applications, more research is needed in adapting pretrained transformers to small data settings (Gururangan et al., 2020). We hope our public framework[6] will provide a principled method for evaluating future NLP models for causal inference.

For **CSS Practitioners**, we find that transformer-based models such as SHERBERT, which we make publicly available,[5] perform the best in all cases except those with very limited data. Models with high accuracy should be applied with great care, as

[6]Dataset Website

this is likely indicative of a strong and unadjustable selection effect. Many models failed our placebo test by making false causal discoveries, a major problem (Aarts et al., 2015; Freedman et al., 2015).

References

Alexander Aarts, Joanna Anderson, Christopher Anderson, Peter Attridge, Angela Attwood, Jordan Axt, Molly Babel, Štěpán Bahník, Erica Baranski, Michael Barnett-Cowan, Elizabeth Bartmess, Jennifer Beer, Raoul Bell, Heather Bentley, Leah Beyan, Grace Binion, Denny Borsboom, Annick Bosch, Frank Bosco, and Mike Penuliar. 2015. Estimating the reproducibility of psychological science. *Science*, 349.

Alberto Abadie and Guido W. Imbens. 2016. Matching on the estimated propensity score. *Econometrica*, 84(2):781–807.

David Arbour and Drew Dimmery. 2019. Permutation weighting.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset.

P. Bhattacharya and R. Mehrotra. 2016. The information network: Exploiting causal dependencies in online information seeking. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*, CHIIR '16.

D.M. Blei, A.Y. Ng, and M. I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.

Glenn W. Brier. 1950. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3.

Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. 2017. You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech. *Proc. ACM Hum.-Comput. Interact.*, 1(CSCW).

Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. 2018. The internet's hidden rules: An empirical study of reddit norm violations at micro, meso, and macro scales. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW).

Munmun De Choudhury and Emre Kiciman. 2017. The language of social support in social media and its effect on suicidal ideation risk. *Proceedings of the ... International AAAI Conference on Weblogs and Social Media. International AAAI Conference on Weblogs and Social Media*, 2017:32–41.

Richard K. Crump, V. Joseph Hotz, Guido W. Imbens, and Oscar A. Mitnik. 2009. Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96(1):187–199.

Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. 2016. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

Vincent Dorie, Jennifer Hill, Uri Shalit, Marc Scott, and Dan Cervone. 2019. Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statist. Sci.*, 34(1):43–68.

Dean Eckles and Eytan Bakshy. 2017. Bias and high-dimensional adjustment in observational studies of peer effects.

Naoki Egami, Christian J. Fong, Justin Grimmer, Margaret E. Roberts, and Brandon M. Stewart. 2018. How to make causal inferences using texts.

Seyed Amin Mirlohi Falavarjani, Hawre Hosseini, Zeinab Noorian, and Ebrahim Bagheri. 2017. Estimating the effect of exercising on users' online behavior. In *AAAI 2017*.

Christian Fong and Justin Grimmer. 2016. Discovery of treatments from text corpora. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1600–1609, Berlin, Germany. Association for Computational Linguistics.

Leonard P. Freedman, Iain M. Cockburn, and Timothy S. Simcoe. 2015. The economics of reproducibility in preclinical research. *PLOS Biology*, 13(6):1–9.

Amanda Gentzel, Dan Garant, and David Jensen. 2019. The case for evaluating causal models using interventional measures and empirical data. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 11722–11732. Curran Associates, Inc.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks.

J. Hájek. 1970. A characterization of limiting distributions of regular estimates. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*.

Jennifer Hill and Yu-Sung Su. 2013. Assessing lack of common support in causal inference using bayesian nonparametrics: Implications for evaluating the effect of breastfeeding on children's cognitive outcomes. *Ann. Appl. Stat.*, 7(3):1386–1420.

David Jensen. 2019. Comment: Strengthening empirical evaluation of causal inference methods. *Statist. Sci.*, 34(1):77–81.

Fredrik Johansson, Uri Shalit, and David Sontag. 2016. Learning representations for counterfactual inference. In *ICML*.

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. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Seattle, WA. Association for Computational Linguistics.

Emre Kiciman, Scott Counts, and Melissa Gasser. 2018. Using longitudinal social media analysis to understand the effects of early college alcohol use. In *ICWSM*.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization.

Brian K. Lee, Justin Lessler, and Elizabeth A. Stuart. 2011. Weight trimming and propensity score weighting. *PLOS ONE*, 6(3):1–6.

Fan Li, Laine E Thomas, and Fan Li. 2018. Addressing Extreme Propensity Scores via the Overlap Weights. *American Journal of Epidemiology*, 188(1):250–257.

Subramani Mani and Gregory F. Cooper. 2000. Causal discovery from medical textual data. *Proceedings. AMIA Symposium*, pages 542–6.

Paramita Mirza and Sara Tonelli. 2016. CATENA: CAusal and TEmporal relation extraction from NAtural language texts. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 64–75, Osaka, Japan. The COLING 2016 Organizing Committee.

Reagan Mozer, Luke Miratrix, Aaron Russell Kaufman, and L. Jason Anastasopoulos. 2020. Matching with text data: An experimental evaluation of methods for matching documents and of measuring match quality. *Political Analysis*, page 1–24.

Alexandra Olteanu, Onur Varol, and Emre Kiciman. 2017. Distilling the outcomes of personal experiences: A propensity-scored analysis of social media. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17.

Judea Pearl. 1995. Causal diagrams for empirical research. *Biometrika*, 82(4).

Judea Pearl. 2009. *Causality*. Cambridge University Press.

Thai T. Pham and Yuanyuan Shen. 2017. A deep causal inference approach to measuring the effects of forming group loans in online non-profit microfinance platform.

Margaret E Roberts, Brandon M Stewart, and Richard A Nielsen. 2020. Adjusting for confounding with text matching.

Paul R. Rosenbaum. 2010. *Design of Observational Studies*. Springer.

Koustuv Saha, Benjamin Sugar, John B Torous, Bruno D. Abrahao, Emre Kiciman, and Munmun De Choudhury. 2019. A social media study on the effects of psychiatric medication use. *Proceedings of the ... International AAAI Conference on Weblogs and Social Media. International AAAI Conference on Weblogs and Social Media*, 13:440–451.

Dhanya Sridhar, Aaron Springer, Victoria Hollis, Steve Whittaker, and Lise Getoor. 2018. Estimating causal effects of exercise from mood logging data.

Chenhao Tan, Lillian Lee, and Bo Pang. 2014. The effect of wording on message propagation: Topic- and author-controlled natural experiments on twitter. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 175–185, Baltimore, Maryland. Association for Computational Linguistics.

Victor Veitch, Dhanya Sridhar, and David M. Blei. 2019. Using text embeddings for causal inference. *CoRR*, abs/1905.12741.

Yongji Wang, Hongwei Cai, Chanjuan Li, Zhiwei Jiang, Ling Wang, Jiugang Song, and Jielai Xia. 2013. Optimal caliper width for propensity score matching of three treatment groups: A monte carlo study. *PLOS ONE*, 8(12):1–7.

Zach Wood-Doughty, Ilya Shpitser, and Mark Dredze. 2018. Challenges of using text classifiers for causal inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4586–4598, Brussels, Belgium. Association for Computational Linguistics.

S Yang and P Ding. 2018. Asymptotic inference of causal effects with observational studies trimmed by the estimated propensity scores. *Biometrika*, 105(2):487–493.

Justine Zhang, Jonathan Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Dario Taraborelli, and Nithum Thain. 2018. Conversations gone awry: Detecting early signs of conversational failure. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1350–1361, Melbourne, Australia. Association for Computational Linguistics.

Xingxing Zhang, Furu Wei, and Ming Zhou. 2019. HIBERT: Document level pre-training of hierarchical bidirectional transformers for document summarization. In *ACL*.

A Moderation and Gender Experiments – Data Collection Details

A.1 Moderation Experiment

In the Moderation Experiment, we test if having a post removed by a moderator impacts the amount a user later posts to the same community. For this experiment, we use 13,786 public Reddit histories (all of which contain more than 500 tokens) from users in `/r/science` from 2015-2017 who had a not had a post removed prior to 2018. Our treated users are those who *have* had a post removed in 2018. Out untreated users are those who have *not* had a post removed in 2018 (nor before). The outcome of interest is the number of posts they made in 2019.

To determine which users have had posts removed, we utilize the Pushshift Reddit API (Baumgartner et al., 2020). The data acessible via this API, in combination with publicly available Pushshift dump archives, allow us to compare two snapshots of each Reddit post: one snapshot made within a few seconds of posting, and one made approximately 2 months later. By comparing these two versions, we can tell a) which user made the post, and b) if it was removed. This approach is similar to that of Chandrasekharan et al. (2018).

This experiment mimics the setup in De Choudhury et al. (2016), where each user is represented by their entire Reddit comment history within specific subreddits. While (De Choudhury et al., 2016) has been influential in our work, their dataset is not public, and publicly available comparable data contains only a relatively small set of Reddit users, leading to underpowered experiments with large, uninformative confidence intervals that fail to reproduce the findings in the original paper.

A.2 Gender Experiment

In the Gender Experiment, we use the dataset made public by Veitch et al. (2019), which consists of single posts from three subreddits: `/r/okcupid`, `/r/childfree`, and `/r/keto`. Each post is annotated with the gender (male or female) of the poster, which is considered the treatment. The outcome is the score of the post (number of 'upvotes' minus number of 'downvotes').

B Model of Conditional Probabilities used for Assignment of Treatment and Outcome

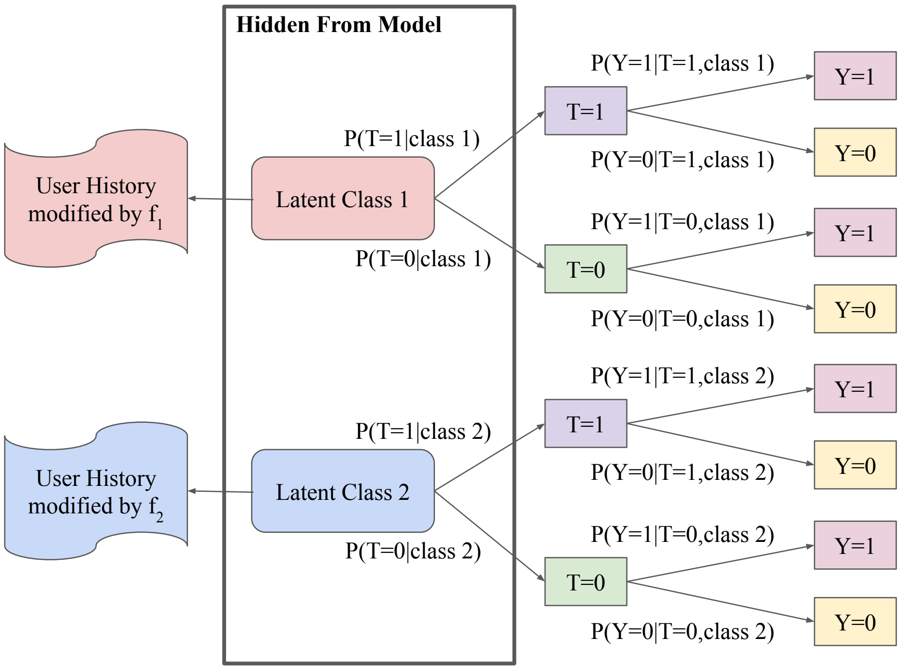

Figure 4: The latent class is used to assign treatments and outcomes to users, and to modify their histories (§5).

Below are the 'default' outcome probabilities used in the synthetic data generation process, conditioned on user class:

$P(Y=1|T=0,\texttt{class}=1)=.1$, $P(Y=0|T=0,\texttt{class}=1)=.9$
$P(Y=1|T=1,\texttt{class}=1)=.9$, $P(Y=0|T=1,\texttt{class}=1)=.1$
$P(Y=1|T=0,\texttt{class}=2)=.9$, $P(Y=0|T=0,\texttt{class}=2)=.1$
$P(Y=1|T=1,\texttt{class}=2)=.9$, $P(Y=0|T=1,\texttt{class}=2)=.1$

These probabilities are used unless otherwise indicated in §5.2.

C Theoretical Bounds

We leverage recent results of Arbour and Dimmery (2019) to bound the expected bias of the ATE, $\bar{Y}[T=1] - \bar{Y}[T=0]$, by considering the weighted risk of the propensity score:

$$\left|\mathbb{E}\left[\hat{Y}(T)\right] - \mathbb{E}\left[Y(T)\right]\right| \leq \left|\mathbb{E}\left[\frac{Y}{p(T|X)} \frac{S(\hat{p}(T|X), p(T|X))}{\hat{p}(T|X)^2}\right]\right|$$

where $\hat{p}$ and p are the estimated and true propensity score, and S is the Brier score (1950). Conceptually, this bound suggests that the bias grows as a function of the Brier score between estimated and true propensity score (numerator), and the inverse of the squared estimate of the propensity score, significantly penalizing very small scores.

Findings. We compute these bounds using the estimated propensity score and find that they are largely uninformative in practice. In 250/252 cases, the empirical confidence interval (Fig. 6) provides a tighter bound than the theoretical bound, and in 230/252 cases the Unadjusted Estimator (§6.2) also provides a tighter bound than the theoretical bound. These results again highlight the importance of the principled empirical evaluation framework presented here.

Details of Derivation. The central challenge is estimating the error of the counterfactual quantities, $Y(1)$, and $Y(0)$. Recall that in the case of weighting estimators, when the true propensity score ($p(\cdot)$) is available, these are estimated as $\mathbb{E}\left[y(T)\right] = \mathbb{E}\left[\frac{Y}{p(T)}\right]$, where y is the observed outcome. For the problem addressed in this paper, the propensity must be estimated. Estimating the error for each potential outcome under an estimated propensity score results in a bias of

$$\left|\mathbb{E}\left[\hat{Y}(T)\right] - \mathbb{E}\left[Y(T)\right]\right| = \left|\mathbb{E}\left[\frac{Y}{p(T|X)}\right] - \mathbb{E}\left[\frac{Y}{\hat{p}(T|X)}\right]\right|$$

following Proposition 1 of Arbour and Dimmery (2019).

More concretely, an empirical upper bound can be obtained for Equation 1 given a lower bound on the true propensity score. Specifically, replacing the p with the lower bound and using the weighted cross-validated Brier score will provide a conservative bound on the bias of the counterfactual. This bound can be tightened with further assumptions, for example by assuming instance level bounds on p instead of a global bound. Balancing weights may also be used to estimate the bias directly using only empirical quantities (Arbour and Dimmery, 2019).

Note that due to the evaluation framework in this paper, the true propensity score p is known, and therefore we do not need to apply loose bounds.

$$\begin{aligned}
\left|\mathbb{E}\left[\hat{Y}(T)\right] - \mathbb{E}\left[Y(T)\right]\right| &= \left|\mathbb{E}\left[\frac{y}{p(T)} - \frac{y}{p(T) + (\hat{p}(T) - p(T))}\right]\right| \\
&= \left|\mathbb{E}\left[\frac{y}{p(T)} - \frac{y}{p(T) + (\hat{p}(T) - p(T))}\right]\right| \\
&= \left|\mathbb{E}\left[\frac{y(1 + \frac{1}{p}(\hat{p}(T) - p(T)))}{p(T)(1 + \frac{1}{p}(\hat{p}(T) - p(T)))} - \frac{y}{p(T) + (\hat{p}(T) - p(T))}\right]\right| \\
&= \left|\mathbb{E}\left[\frac{y + \frac{y}{p}(\hat{p}(T) - p(T))}{p(T) + (\hat{p}(T) - p(T))} - \frac{y}{p(T) + (\hat{p}(T) - p(T))}\right]\right| \\
&= \left|\mathbb{E}\left[\frac{y}{p}\frac{(\hat{p}(T) - p(T))}{\hat{p}(T)}\right]\right| \\
&\leq \left|\mathbb{E}\left[\frac{y}{p}\frac{(\hat{p}(T) - p(T))^2}{\hat{p}(T)^2}\right]\right| \\
&\leq \left|\mathbb{E}\left[\frac{y}{p(T)}\frac{S(\hat{p}(T), p(T))}{\hat{p}(T)^2}\right]\right| \qquad (1)
\end{aligned}$$

After obtaining the bounds on the individual counterfactual quantities, the corresponding lower and upper bias bounds on the average treatment effect can be constructed by considering

$$\hat{Y}(0) + \left| \mathbb{E}\left[\frac{y}{p(T=0|X)} \frac{S(\hat{p}(0|X), p(0|X))}{\hat{p}(T=0|X)^2} \right] \right| \tag{2}$$

$$\hat{Y}(1) - \left| \mathbb{E}\left[\frac{y}{p(T=1 \mid X)} \frac{S(\hat{p}(T=1 \mid X), p(T=1 \mid X))}{\hat{p}(T=1 \mid X)^2} \right] \right| \tag{3}$$

and

$$\hat{Y}(0) - \left| \mathbb{E}\left[\frac{y}{p(T=0|X)} \frac{S(\hat{p}(T=0|X), p(0))}{\hat{p}(T=0|X)^2} \right] \right| \tag{4}$$

$$\hat{Y}(1) + \left| \mathbb{E}\left[\frac{y}{p(T=1|X)} \frac{S(\hat{p}(T=1|X), p(T=1|X))}{\hat{p}(T=1|X)^2} \right] \right| \tag{5}$$

respectively.

D Templates for Synthetic Posts

As described in §5.1.2, synthetic *sickness*, *social isolation*, and *death* posts are used to generate our evaluation tasks. These synthetic posts are selected and inserted into social media histories of real world users by randomly sampling a template and word pair, or, in the case of Social Isolation Posts, by randomly sampling a complete post.

D.1 Sickness Posts

Sickness Posts are created by randomly sampling a Sickness Word and inserting it into a randomly sampled Sickness Template.
Sickness Templates are sampled from:

{`The doctor told me I have` **x**,
`I was at the hospital earlier and I have` **x**.,
`I got diagnosed with` **x** `last week.`,
`Have anyone here dealt with` **x**? `I just got diagnosed.`,
`How should I handle a` **x** `diagnosis?`,
`How do I tell my parents I have` **x**? }

Sickness Words are sampled from {`cancer, leukemia, HIV, AIDS, Diabetes, lung cancer, stomach cancer, skin cancer, parkinson's`}

D.2 Social Isolation Posts

Social Isolation Posts are randomly sampled from the following set of complete synthetic posts:

```
{My friends stopped talking to me.,
 My wife just left me.,
 My parents kicked me out of the house today.,
 I feel so alone, my last friend said they needed to stop seeing me.,
 My partner decided that we shouldn't talk anymore last night.,
 My folks just cut me off, they won't talk to me anymore.,
 I just got a message from my brother that said he can't talk to me anymore.  He was my
      last contact in my family.,
 My last friend at work quit, now there's no one I talk to regularly.,
 I tried calling my Mom but she didn't pick up the phone.  I think my parents may be done
      with me.,
 I got home today and my partner was packing up to leave.  Our apartment feels so empty
      now.                                                                                  }
```

D.3 Death Posts

Death Posts are created by randomly sampling a Death Word and inserting it into a Death Template.
Death Templates are sampled from:

{`My` **x** `just died`,
`I just found out my` **x** `died`,
`My` **x** `died last weekend`,
`What do you do when your` **x** `dies? This happened to me.`,
`Has anyone else had a` **x** `die recently?`,
`I lost my` **x** `yesterday.`,
`My` **x** `passed away recently.`,
`I am in shock. My` **x** `is gone.` }

Death Words are sampled from {`Mom, Mother, Mama, Father, Dad, Papa, Brother, Wife, girlfriend, partner, spouse, husband, son, daughter, best friend`}

E Model Implementation, Tuning, and Parameters

E.1 SHERBERT Architecture

Fig. 5 depicts the architecture of SHERBERT as part of the broader ATE estimation pipeline.

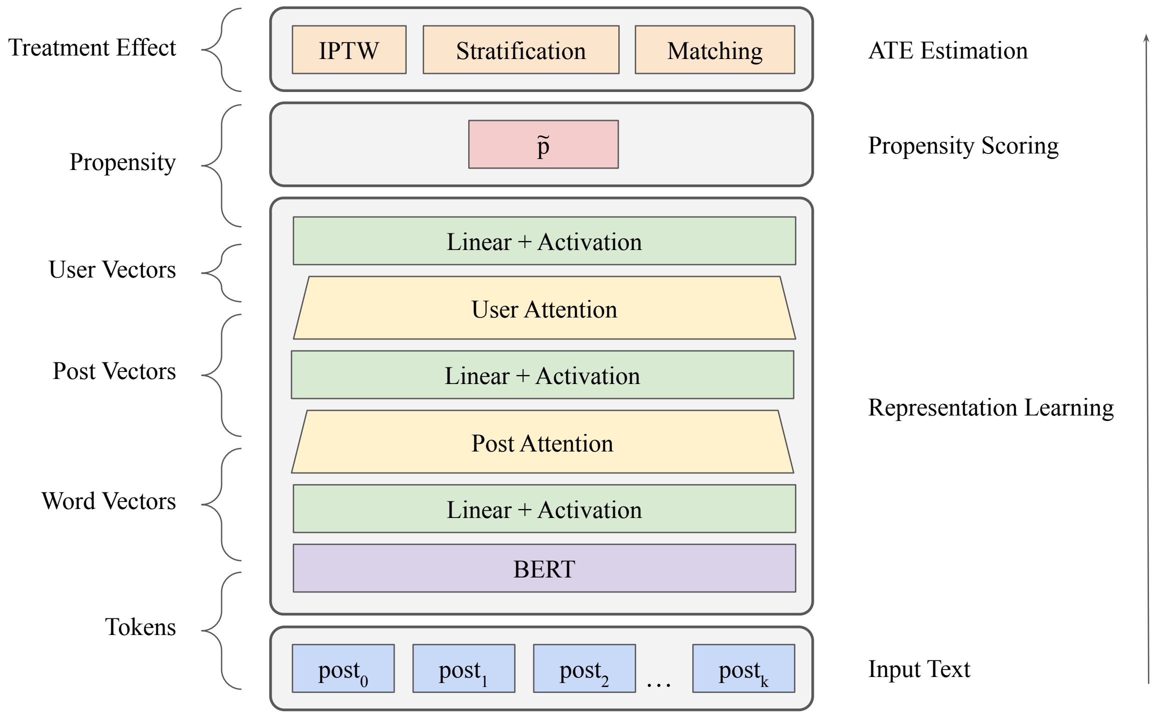

Figure 5: The complete ATE estimation pipeline, with tokens input at the bottom, and an estimate propensity at the top. ATE estimates are computed with IPTW, Stratification, and Matching (§6.2) based on models' propensity scores. This example is instantiated using SHERBERT and detailing its hierarchical architecture. In this pipeline, other propensity score models could replace the 'Representation Learning' box (*e.g.*, Bag-of-n-grams with Logistic Regression).

Our work attempts to expand the success of large pretrained transformers to long history length using a hierarchical attention, which is a problem also explored by the HIBERT model in Zhang et al. (2019). Essentially, SHERBERT differs from HIBERT in that SHERBERT trains a light-weight hierarchical attention on top of the pretrained BERT model (Devlin et al., 2019) whereas HIBERT is trained from scratch. This results in a relatively simple training procedure for SHERBERT, and lighter limitations on history length, both at the local (50 words for HIBERT v. 512 wordpiece tokens for SHERBERT) and global (30 sentences for HIBERT v. 60 for SHERBERT) scales. This reflects differing tradeoffs; where HIBERT has a more sophisticated attention mechanism for combining local and global information, SHERBERT sacrifices some complexity for fast and simple training and longer text histories.

E.2 Practicality of Models

SHERBERT trades-off practicality for performance in comparison to simpler models. For instance, in most experiments we found SHERBERT takes 10 - 12 hours to train, sometimes requiring multiple starts to converge to a reasonable model. In contrast, training all other models collectively requires less than 1 hour. Further, the performance of SHERBERT sharply suffered as the number of users was reduced (Fig. 3j). While effectively training SHERBERT on 1 GPU (Tesla V100) in under 24 hours is quite practical compared to contemporary text pretraining regimes (Devlin et al., 2019), these issues should be considered when deciding on a causal text model.

E.3 Hyperparameters

A complete description of parameters and hyperparameters is included in the code repository at Dataset Website. Basic details are included here.

In producing n-gram features, a count threshold of 10 is used to filter out low frequency words, and word tokenization is done using the NLTK word tokenizer. In producing LDA features, we use the Scikit Learn implementation, with 20 topics. To produce BERT word embedding features, we use the uncased model of the 'base' size.

All models use the Adam optimizer (Kingma and Ba, 2014), with various learning rates decided empirically depending on model and task to maximize treatment accuracy on the validation set.

For the simple neural network model, we use a hidden size of 10. For SHERBERT, we use hidden sizes of 1000 and dot-product attention.

F Additional Estimators and Metrics

In order to further detail our findings (§7), we include several additional ATE estimators (§6.2) and metrics for evaluation (§6.3).

F.1 Matching Estimator

Matching can be considered as a special case of stratification, where each strata contains only one treated user (§6.2). As our treated and untreated groups are approximately balanced, we implement 1:1 matching, where each treated user is matched to exactly one untreated user.

While there are many implementations of matching, we implement matching *with* replacement, as in Abadie and Imbens (2016, pg. 784):

$$\widehat{ATE}_{\texttt{match}} = \frac{1}{n}\sum_{i=1}^{n}(2T_i - 1)(Y_i - Y_j)$$

where j is the matched observation, *i.e.* $j = \min_{j\in\{1\ldots n\}} |\hat{p}(\mathbf{X}_i) - \hat{p}(\mathbf{X}_i)|$ where $T_i \neq T_j$.
A recent evaluation of matching techniques for text found no significant difference in match quality between matches produced with and without replacement (Mozer et al., 2020). We use a caliper value of $.2\times$ the standard deviation of propensity scores in the population, as was found to perform the best by Wang et al. (2013) and recommended by Rosenbaum (2010, pg. 251).

For each of the five tasks, the matching estimator produces results extremely similar to those of the stratified estimator (Fig. 6).

F.2 Mean Squared Error of IPTW and Spearman Correlation

In addition to Treatment Accuracy and Bias (§6.3), we computed the Mean Squared Error (MSE) of the Inverse Probability of Treatment Weights, and the Spearman Correlation of propensity scores.

Mean Squared Error of IPTW shows the absolute error in the calibration of a models' causal weights:

$$MSE_{IPTW} = \sum_{i=1}^{n}\left(\frac{\left[\sum_{j=1}^{n}\frac{1}{\hat{p}_{T_j}(\mathbf{X}_j)}\right]^{-1}}{\hat{p}_{T_i}(\mathbf{X}_i)} - \frac{\left[\sum_{j=1}^{n}\frac{1}{p_{T_j}(\mathbf{X}_j)}\right]^{-1}}{p_{T_i}(\mathbf{X}_i)}\right)^2$$

Notation is the same as in §6.2, with the addition of p as true propensity, which is known in our semi-synthetic tasks. The MSE is fairly correlated with the Treatment Accuracy, with MSE increasing as accuracy decreases as the tasks become more difficult. This is especially evident in Fig. F.2b,k.

Spearman Correlation instead shows the relative calibration of a models' propensity scores. Propensity scores may have poor absolutely calibration, but still have meaningful relative ordering, in which case the Spearman Rank Correlation is close to its maximum value of 1. The Spearman Correlation coefficient is simply the Pearson correlation coefficient between the rank variables for the estimated and actual propensity scores. We find that Spearman Correlation is also quite correlated with the Treatment Accuracy (Fig. 7c,i)

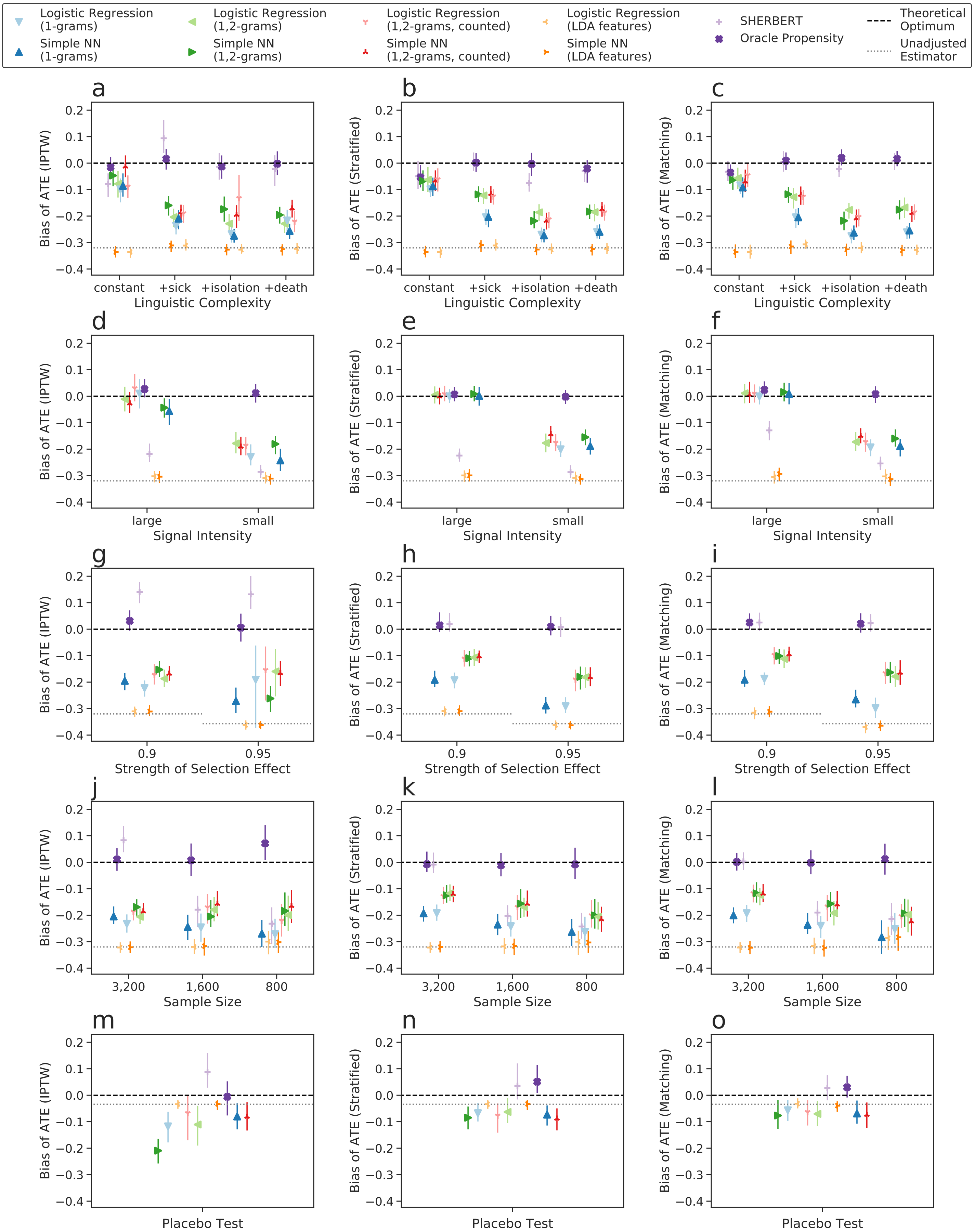

Figure 6: Comparison of bias computed using IPTW, Stratification, and Propensity Score Matching, for each task. Note that matching produces extremely similar results to stratification.

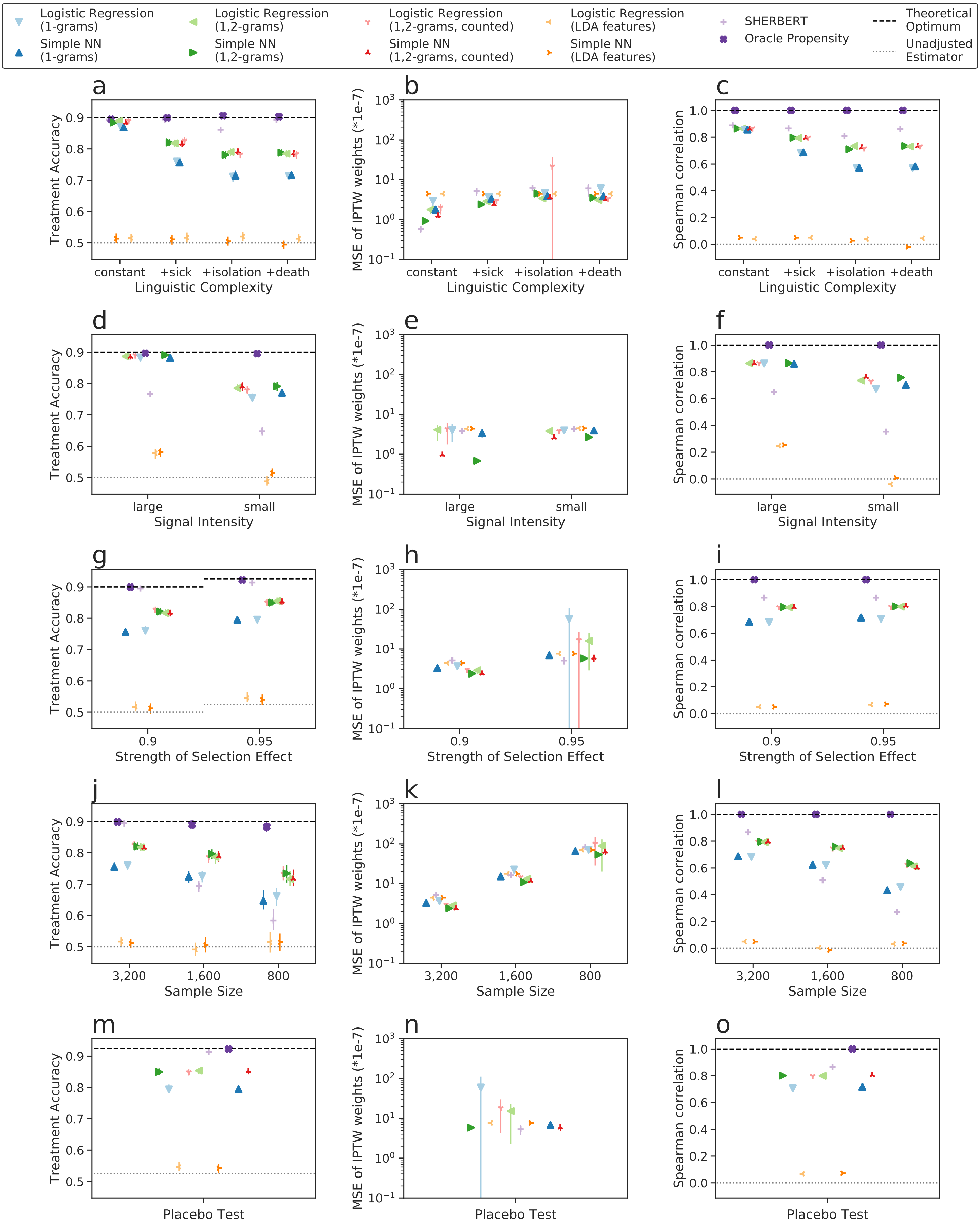

Figure 7: Treatment Accuracy, Mean Squared Error, and Spearman Correlation for each task. Spearman Correlation varies directly with Treatment Accuracy, whereas Mean Squared Error increases as accuracy falls.