DoubleLingo: Causal Estimation with Large Language Models

Marko Veljanovski, Zach Wood-Doughty
Northwestern University
Department of Computer Science
marko@u.northwestern.edu, zach@northwestern.edu

Abstract

Estimating causal effects from non-randomized data requires assumptions about the underlying data-generating process. To achieve unbiased estimates of the causal effect of a treatment on an outcome, we typically adjust for any confounding variables that influence both treatment and outcome. When such confounders include text data, existing causal inference methods struggle due to the high dimensionality of the text. The simple statistical models which have sufficient convergence criteria for causal estimation are not well-equipped to handle noisy unstructured text, but flexible large language models that excel at predictive tasks with text data do not meet the statistical assumptions necessary for causal estimation. Our method enables theoretically consistent estimation of causal effects using LLM-based nuisance models by incorporating them within the framework of Double Machine Learning. On the best available dataset for evaluating such methods, we obtain a 10.4% reduction in the relative absolute error for the estimated causal effect over existing methods.

1 Introduction

A common goal of scientific research is the analysis of causal relationships (Triantafillou et al., 2017; Sanna et al., 2019; Chang et al., 2022). Consider the following motivating example, where a pharmaceutical company wants to estimate the causal effect of the prescription of antibiotics (treatment) on the patient's disease progression (outcome). The causal effect is defined as the expected change in disease progression across two *counterfactual* worlds which only differ in whether the patient is given antibiotics (Hernán, 2004). When randomization is impossible or unethical, we estimate causal effects from observational data using assumptions about the underlying data distribution. Confounders – variables affecting both the treatment and outcome – introduce potential bias that must be addressed.

When data is low-dimensional, confounding can be controlled for using various methods from the literature (Pearl, 2009). However, several challenges arise in the case of high-dimensional confounders. Suppose the pharmaceutical company has free-text clinical notes that may include information about patients' histories, diagnoses, or relationships with their doctors (Rajkomar et al., 2018). If these variables appear nowhere else in the patients' records, then account for potential confounding should use text-based causal methods (Rosenbloom et al., 2011; Wu et al., 2013). Since text is high-dimensional, it requires sophisticated modeling that captures semantic meaning.

Existing models often utilize overly simplified representations of the text (Wood-Doughty et al., 2018; Keith et al., 2020), such as a *bag-of-words* (BoW) representation. While such representations combined with simple estimation models allow for consistent[1] estimation, they may fail to capture the true complexity of the text's underlying relationships. The use of large language models (LLMs) in causal estimation has only recently been studied (Veitch et al., 2020), and many researchers suggest the need for more sophisticated natural language processing (NLP) techniques (Wood-Doughty et al., 2021; Feder et al., 2022; Keith et al., 2023). However, while LLMs excel at predictive tasks, they do not meet the necessary statistical assumptions for a consistent causal estimation.

We present **DoubleLingo**, combining Double Machine Learning with LLM-based nuisance models to enable a theoretically consistent estimation of causal effects with text-based confounding. We test our model on a novel dataset (Keith et al., 2023), obtaining the best causal effect estimates reported thus far. In particular, our relative absolute error is over 10% lower than the best current models.

[1] Defined in more detail in §3.

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers), pages 799–807
June 16-21, 2024 ©2024 Association for Computational Linguistics

2 Causal Inference Background

While causal inference is a broad and diverse field (Robins et al., 2000; Pearl, 2009), we provide a brief introduction here. For recent surveys of causal inference and natural language processing, see Keith et al. (2020) or Feder et al. (2022).

2.1 DAGs & Counterfactuals

The motivating example described above is illustrated by the directed acyclic graph (DAG) in Figure 1, where we use a binary random variable A to indicate whether the patient receives $(A = 1)$ antibiotics or not $(A = 0)$. We similarly use a binary Y to denote whether the disease progresses $(Y = 1)$ or not $(Y = 0)$. An arrow in the DAG such as $A \rightarrow Y$ indicates that A has a potential causal effect on Y. Finally, we denote T as the patient medical records, and C as the set of all confounding variables contained in the records. For example, such variables could include income status or family disease history (Acharya et al., 2021). Most importantly, C is unobserved — we don't know the exact confounding variables, but we have access to the text T containing them. In particular, T is related to A and Y through C. The *counterfactual* outcome $Y^{a=1}$ represents the hypothetical disease progression had we intervened to assign $A = 1$ (prescribe antibiotics), and $Y^{a=0}$ is defined analogously. In causal inference, the most common estimand is the average treatment effect (ATE) of A on Y, computed as:

$$ATE = \mathbb{E}[Y^{a=1} - Y^{a=0}] \tag{1}$$

A fundamental problem is that we can never simultaneously observe both *counterfactuals* $Y^{a=1}, Y^{a=0}$ (Holland, 1986), thus we need a way to compute the ATE only utilizing observed data.

2.2 Identification Assumptions

We proceed by assuming *consistency*, requiring that the outcome we observe for any possible treatment a is equal to the *counterfactual* outcome we would have observed had we intervened to assign $A = a$. Formally:

$$A = a \Rightarrow Y^a = Y \tag{2}$$

We then assume *conditional exchangeability*, requiring the independence between our counterfactual Y^a and the observed treatment A conditioned on all confounders C, formalized as:

$$Y^a \perp A \mid C \quad \forall a \in \{0, 1\} \tag{3}$$

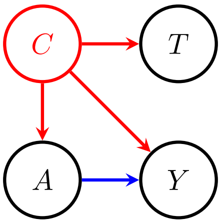

Figure 1: Textual Confounding DAG with Treatment A, Outcome Y, Confounders C, and Text T. We assume the $C \rightarrow T$ edge is such that adjusting for T can control C's confounding of the $A \rightarrow Y$ relationship.

Using these assumptions, we may compute the counterfactual $\mathbb{E}[Y^a]$ as follows:

$$\begin{aligned} \mathbb{E}[Y^a] &= \sum_C \mathbb{E}[Y^a \mid C]\mathbb{P}(C) && (4) \\ &\overset{(3)}{=} \sum_C \mathbb{E}[Y^a \mid A = a, C]\mathbb{P}(C) && (5) \\ &\overset{(2)}{=} \sum_C \mathbb{E}[Y \mid A = a, C]\mathbb{P}(C) && (6) \end{aligned}$$

Equation (6) expresses our counterfactual as a function of observed data. However, we are interested in the case where the low-dimensional C is unobserved but encoded inside the high-dimensional T. Thus, if we could adequately model T, we would be able to adjust for C's confounding effect.

2.3 Causal Effect Estimation

To estimate (1) using (6), we thus require (a) a representation of the text and (b) an appropriate causal estimation method. As mentioned in §1, a BoW text representation is commonly used by existing text-based causal estimators. For (b), there are countless estimation methods, and we refer the reader to a much more exhaustive guide by Peters et al. (2017). One such commonly used method is the *Inverse Propensity of Treatment Weighting* (IPTW), where $\mathbb{E}[Y^a]$ is calculated as follows for a dataset of size N:

$$\mathbb{E}[Y^a] = \frac{1}{N} \sum_{i \in [N]} Y_i \frac{\mathbb{1}(A_i = a)}{\mathbb{P}(A_i = a \mid T)} \tag{7}$$

A simple way to combine (a) and (b) is to use IPTW and train a *Logistic Regression* model $\mathbb{P}(A \mid T)$ for the propensity of the treatment A given a BoW text representation T. However, BoW will fail to model the complexities of real-world text, introducing bias into our estimates.

3 Model

We now introduce notation to formalize our proposed method to use LLMs to replace a simplistic BoW text representation. Consider this partially linear model corresponding to Figure 1:

$$Y = A\theta_0 + g_0(T) + U, \quad \mathbb{E}[U \mid T, A] = 0 \quad (8)$$

$$A = m_0(T) + V, \quad \mathbb{E}[V \mid T] = 0 \quad (9)$$

Here θ_0 is the true ATE we hope to estimate, $\eta_0 = (m_0, g_0)$ are nuisance parameters, and U, V are our error terms. Following Keith et al. (2023),[2] we similarly assume the causal effect $A \rightarrow Y$ is linear. Any estimator $\widehat{\theta}_0$ of θ_0 must be both unbiased and consistent such that:[3]

$$\mathbb{E}[\widehat{\theta}_0] = \theta_0 \quad \text{and} \quad \widehat{\theta}_0 \xrightarrow{p} \theta_0 \quad (10)$$

While LLMs have drastically changed the field of NLP (Vaswani et al., 2017; Min et al., 2023), they are not consistent estimators of causal parameters due to both explicit and implicit regularization (Neyshabur, 2017; Chernozhukov et al., 2018). Thus, a naive approach of using an LLM such as BERT (Devlin et al., 2019) to learn the propensity $\mathbb{P}(A \mid T)$ in Equation (7) would be biased.

3.1 Double Machine Learning

To overcome this challenge, we turn to Double Machine Learning (DML), which has never previously been used in the context of LLMs. As introduced by Chernozhukov et al. (2018), DML is an estimation procedure which removes regularization bias and overfitting on estimation by combining (a) Neyman-orthogonal moments with (b) sample-splitting. Let $\widehat{m}_0$ and $\widehat{g}_0$ be ML estimators of η_0. For (a), we partial out the effect of T from A to obtain the orthogonalized regressor $\widehat{V} = A - \widehat{m}_0(T)$. For (b), we randomly split our dataset of size N into a main and auxiliary sample with their indices denoted respectively by I and I^C, both of size $n = N/2$. We first train $\widehat{m}_0$ and $\widehat{g}_0$ on I^C, and then subsequently estimate θ_0 from I as follows:

$$\widehat{\theta}_0 = \left(\frac{1}{n} \sum_{i \in I} \widehat{V}_i A_i \right)^{-1} \frac{1}{n} \sum_{i \in I} \widehat{V}_i (Y_i - \widehat{g}_0(T_i)) \quad (11)$$

Now, as shown by Chernozhukov et al. (2018), the scaled estimation error can be decomposed as:

$$\sqrt{n}(\widehat{\theta}_0 - \theta_0) = A + B + C \quad (12)$$

The A term from (12) converges in distribution to a mean-zero Gaussian with variance Σ:

$$\frac{1}{\mathbb{E}[V^2]\sqrt{n}} \sum_{i \in I} V_i U_i \xrightarrow{d} \mathcal{N}(0, \Sigma) \quad (13)$$

Sample-splitting guarantees that the C term is $O_p(1)$, as it contains terms of form:

$$\frac{1}{\sqrt{n}} \sum_{i \in I} V_i(\widehat{g}_0(T_i) - g_0(T_i)) \quad (14)$$

Finally, the regularization bias from training our two ML estimators $\widehat{m}_0$ and $\widehat{g}_0$ is captured by the B term, which equals:

$$\frac{1}{\mathbb{E}[V^2]\sqrt{n}} \sum_{i \in I} (\widehat{m}_0(T_i) - m_0(T_i))(\widehat{g}_0(T_i) - g_0(T_i)) \quad (15)$$

Observe that due to orthogonalization via (a), term B contains the product of the estimation errors, which Chernozhukov et al. (2018) show to be upper-bounded by $\sqrt{n}n^{-(\varphi_m + \varphi_g)}$, where $n^{-\varphi_m}$ and $n^{-\varphi_g}$ denote the respective convergence rates of $\widehat{m}_0$ and $\widehat{g}_0$. Hence, this term vanishes even in cases where $\widehat{m}_0$ and $\widehat{g}_0$ converge at relatively slower rates. In particular, if these two ML estimators converge at $n^{-1/4}$, the estimation of the ATE is $\sqrt{n}$-consistent, where:

$$\widehat{\theta}_0 - \theta_0 = O_p(n^{-1/2}) \quad (16)$$

For proofs of the above claims, and more general cases covering unequal split-sizes, please see Chernozhukov et al. (2018). Finally, as we train both $\widehat{m}_0$ and $\widehat{g}_0$, the estimation is doubly robust such that only one of the two need to be correctly specified to obtain an unbiased ATE (Funk et al., 2011).

3.2 Faster Converging Model Variations

A potential concern is that our two ML estimators must converge at $n^{-1/4}$ to obtain the desired $\sqrt{n}$-consistent estimation of θ_0. While there is research on the rate of convergence of misclassification probability (Gurevych et al., 2022) for encoder-based transformer classifiers such as BERT, its convergence rate for semiparametric inference is unknown. Since fully fine-tuning BERT classifiers within the DML framework may not be appropriate, we present **DoubleLingo**, utilizing two faster converging model variations.

[2] See §4.1 for more detail.

[3] $\xrightarrow{p}$ denotes convergence in probability.

BERT+Adapter. Our first configuration utilizes parameter efficient transfer learning in the form of adapters (Houlsby et al., 2019). Thus, instead of fine-tuning all of BERT, we only fine-tune the adapter layers. Here, it's crucial to note that there are no theoretical bounds for the convergence of adapters. While a proof that **BERT+Adapter** converges at $n^{-1/4}$ would be desirable, it is outside the scope of this paper. However, see §5.2 for an empirical justification.

Embedding+FFN. Fully-connected feedforward neural networks (FFNs) with the ReLU activation function have been proven to converge at $n^{-1/4}$ rates for their use in semiparametric inference (Farrell et al., 2021). Thus, instead of fine-tuning BERT at all, a potential approach is to fine-tune a feed-forward layer on top of BERT's pre-trained $[CLS]$ encoding. However, this encoding is pre-trained on next sentence prediction which may not necessarily result in a semantically meaningful representation of the sentence. Consequently, we utilize embeddings from pre-trained sentence transformers (Reimers and Gurevych, 2019), which are much more semantically meaningful. While transformer embeddings have been widely influential in many NLP tasks (Ethayarajh, 2019), to our knowledge we are the first to compare their potential for causal estimation against simpler text representations.

4 Causal Dataset & Experiment

Unlike supervised learning models, which can be evaluated on held-out test sets with ground-truth labels, causal estimation methods require evaluations with *counterfactual* ground-truth, which is impossible to measure from observed data (Holland, 1986). Researchers often turn to (semi-)synthetic data, for which there is a tension between generating realistic text and maintaining full knowledge of the underlying data-generating process (DGP) (Wood-Doughty et al., 2021). Most current datasets fail to accomplish both, either fully specifying the DGP but with unrealistic text (Johansson et al., 2016; Yao et al., 2019), or using real-world text inside a semi-synthetic DGP (Veitch et al., 2020).

4.1 Dataset and Baselines

A recent novel dataset employs a randomized controlled trial (RCT) rejection sampling algorithm to create datasets with real text that build on a real-world DGP (Keith et al., 2023). In particular, the authors fix C to be a single binary confounding variable contained in the text and choose RCT's with an existing $C \rightarrow Y$ relationship. They then sample the dataset to artificially create a $C \rightarrow A$ relationship and evaluate 8 different models over 100 sampled dataset subsets. They train *Logistic Regression* and *CatBoost* nuisance models based on a BoW representation for the text, combining both with 4 different causal estimation techniques, including IPTW, Augmented-IPTW (AIPTW), Outcome Regression, and DML. They finally evaluate an *Oracle* with full access to the (otherwise unobserved) C variable. We include their empirical results in our Table 1.

4.2 DoubleLingo Experiments

We now describe our methods that use LLMs inside the DML framework. Our **BERT+Adapter** method fine-tunes adapters within BERT classifiers for both $\widehat{m}_0$ and $\widehat{g}_0$ (Houlsby et al., 2019). Our **Embedding+FFN** configuration uses embeddings from two transformers. First, *all-mpnet-base-v2*,[4] based on *MPNet* (Song et al., 2020) and fine-tuned on over 1 billion sentence pairs including paper abstracts from S2ORC (Lo et al., 2020). Second, *SPECTER* (Cohan et al., 2020), pre-trained on a dataset of scientific paper titles and abstracts which matches the exact format of Keith et al. (2023). For both **Embedding+FFN** methods, we use a single hidden layer, ReLU activation functions, and the AdamW optimizer (Loshchilov and Hutter, 2018). Finally, we implement a *TF-IDF+FFN* baseline, following Manzoor et al. (2023), which uses DML with FFNs with batch normalization (Ioffe and Szegedy, 2015) and a TF-IDF text representation. A more detailed implementation, including specific hyper-parameters and RCT parameterization choices are provided in Appendix A.

5 Results and Conclusions

5.1 Main Findings

Table 1 shows that our three **DoubleLingo** estimators obtain the lowest ATE relative absolute error (0.103), a 10.4% decrease from the prior best (0.115). These results provide strong empirical evidence that the DML framework successfully enables the use of LLMs in causal estimation. Notably, the prior best was achieved by both a BoW model (CB$_{\text{AIPTW}}$) and the *Oracle* estimator which calculates the estimates using the unobserved C values. If C contained all causes of A and Y, it would

[4] https://hf.co/sentence-transformers/all-mpnet-base-v2

	LR	CB
Outcome	1.408 (1.00)	0.237 (0.10)
IPTW	0.470 (0.16)	0.141 (0.11)
AIPTW	1.579 (0.66)	0.115 (0.10)
DML	1.899 (0.91)	0.128 (0.10)
BERT+Adapter	0.104 (0.08)	
MPNetV2+FFN	0.103 (0.08)	
SPECTER+FFN	0.104 (0.08)	
TF-IDF+FFN	0.118 (0.09)	
Unadjusted	0.214 (0.08)	
Oracle (C)	0.115 (0.09)	

Table 1: Relative Absolute Error mean (variance) for all methods over 100 subsets. Oracle, Unadjusted, Logistic Regression (LR), and CatBoost (CB) baselines are from Keith et al. (2023). Oracle and Unadjusted models use Outcome regressions. Our **DoubleLingo** methods and TF-IDF baseline use DML, as described in §4.2. Our methods achieve the best (lowest) error and variance.

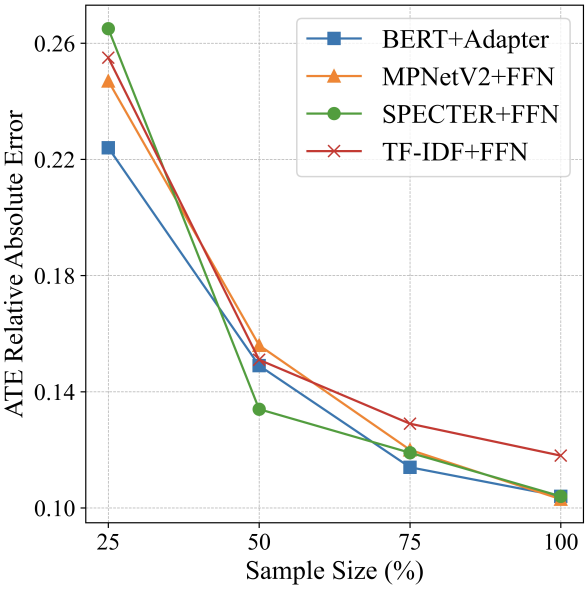

Figure 2: Empirical convergence comparison of **BERT+Adapter** with FFN configurations. We plot the ATE relative absolute error at 4 sample sizes.

be the theoretically-optimal efficient adjustment set (Rotnitzky and Smucler, 2020) and the Oracle should – asymptotically – be impossible to outperform. However, while the $C \to A$ relationship is artificially induced by the sampling procedure of Keith et al. (2023), the authors verified that $C \not\perp Y$ using an odds-ratio test. We hypothesize that the underlying complexity of the $T \to Y$ relationship is not fully captured by the binary topic C, and there exists some $T \dashrightarrow Y$ relationship. If true, then modeling T allows for more efficient estimation reflected in **DoubleLingo**'s outperformance of the Oracle.

Our results specifically support the hypothesis that the text representation itself matters to causal estimation. Among all DML methods with feed-forward classifiers, our **Embedding+FFN** methods' outperformance of our *TF-IDF+FFN* baseline shows that better representations can enable lower estimation error. Appendix B also shows our models' slightly better classification accuracy than the *TF-IDF+FFN* baseline during estimation.

Between our three proposed methods, we see no large differences in performance. This suggests that while the incorporation of LLMs into the estimators is essential, the specific architecture and training setup matters less. However, **BERT+Adapter** trains two to three times slower than **Embedding+FFN**. We also see little difference between the two pre-trained embeddings, despite the similarity of the *SPECTER* embedding's dataset to that of our evaluation data.

5.2 Convergence Experiment

The assumptions of DML require that $\widehat{m}_0$ and $\widehat{g}_0$ must converge at $n^{-1/4}$ to enable a $\sqrt{n}$-consistent estimation of θ_0. Analysis and proof of **BERT+Adapter** convergence is left for future work. However, we empirically compare its convergence rate to that of the three FFN configurations which are proven to converge at $n^{-1/4}$ (Farrell et al., 2021). Figure 5.2 plots the ATE relative absolute error mean as we increase the available data. Regressing the logarithms of the means against the sample sizes, we obtain rough estimates that **BERT+Adapter**, **MPNetV2+FFN**, **SPECTER+FFN**, and *TF-IDF+FFN* converge respectively at $(n^{-0.57}, n^{-0.64}, n^{-0.67}, n^{-0.56})$, all faster than our desired $n^{-0.25}$ rate.

5.3 Conclusion

This work proposes **DoubleLingo**, a theoretically consistent causal estimator that uses LLM nuisance models inside the DML framework. We show that both adapters and sentence transformers can achieve the lowest estimation error on the best available dataset for evaluating methods that account for text confounding. We release our code which reproduces our results to enable future research.[5]

[5] https://github.com/markov24/DoubleLingo

Limitations

The main limitation of our estimation procedure is compute time – training the **BERT+Adapter** configuration on 100 sampled dataset subsets takes 10 hours parallelized across 2 RTX 8000's, significantly longer than the baseline *Linear Regression* or *CatBoost* models. In particular, our model's reliance on sample-splitting and double robustness to obtain a consistent final estimate requires training 4 times as many models per each dataset subset. However, it's important to note that the **Embedding+FFN** configurations only take a third of the time, yet achieve identical results.

While DML provides solid theoretical grounding for our methods, we have necessarily focused on a specific DAG and dataset. We have assumed that the relationship between C and T is such that DML nuisance models fit to T can control for the confounding effect of C. In the dataset released by Keith et al. (2023), this is plausible given the underlying connections between text and topic. In other datasets (e.g., if T were only loosely predictive of C), additional methods might be necessary to account for measurement error, for example following Kuroki and Pearl (2014).

Additionally, our work only focuses on causal estimation with text-based confounding. Dealing with textual treatments or outcomes is still an open problem in the field (Feder et al., 2022). Finally, we only train on a single English-language dataset; we encourage future work to expand on this by testing other types of text-based RCTs.

Acknowledgments

We would like to express our gratitude to the Northwestern WNS Undergraduate Research Fellowship for funding our work through the 2023 McCormick Summer Undergraduate Research Award. We further thank the Northwestern Undergraduate Innovation Fellowship for providing additional financial support. The resources and support from these fellowships have been instrumental in the completion of this research.

References

Mahip Acharya, Thomas Kim, and Chenghui Li. 2021. Broad-spectrum antibiotic use and disease progression in early-stage melanoma patients: A retrospective cohort study. *Cancers*, 13(17).

Chun-Wei Chang, Stephan B Munch, and Chih-hao Hsieh. 2022. Comments on identifying causal relationships in nonlinear dynamical systems via empirical mode decomposition. *Nature communications*, 13(1):2860.

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

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. 2020. SPECTER: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2270–2282, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65.

Max H Farrell, Tengyuan Liang, and Sanjog Misra. 2021. Deep neural networks for estimation and inference. *Econometrica*, 89(1):181–213.

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

Michele Jonsson Funk, Daniel Westreich, Chris Wiesen, Til Stürmer, M. Alan Brookhart, and Marie Davidian. 2011. Doubly Robust Estimation of Causal Effects. *American Journal of Epidemiology*, 173(7):761–767.

Iryna Gurevych, Michael Kohler, and Gözde Gül Şahin. 2022. On the rate of convergence of a classifier based on a transformer encoder. *IEEE Transactions on Information Theory*, 68(12):8139–8155.

Miguel Angel Hernán. 2004. A definition of causal effect for epidemiological research. *Journal of Epidemiology & Community Health*, 58(4):265–271.

Paul W. Holland. 1986. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR.

Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 448–456, Lille, France. PMLR.

Fredrik Johansson, Uri Shalit, and David Sontag. 2016. Learning representations for counterfactual inference. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 3020–3029, New York, New York, USA. PMLR.

Katherine 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*, pages 5332–5344, Online. Association for Computational Linguistics.

Katherine A Keith, Sergey Feldman, David Jurgens, Jonathan Bragg, and Rohit Bhattacharya. 2023. Rct rejection sampling for causal estimation evaluation. *Transactions on Machine Learning Research*.

Manabu Kuroki and Judea Pearl. 2014. Measurement bias and effect restoration in causal inference. *Biometrika*, 101(2):423–437.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Emaad Manzoor, George H Chen, Dokyun Lee, and Michael D Smith. 2023. Influence via ethos: On the persuasive power of reputation in deliberation online. *Management Science*.

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. 2023. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, 56(2):1–40.

Behnam Neyshabur. 2017. Implicit regularization in deep learning. *arXiv preprint arXiv:1709.01953*.

Judea Pearl. 2009. *Causality*. Cambridge university press.

Jonas Peters, Dominik Janzing, and Bernhard Schlkopf. 2017. *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press.

Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M Dai, Nissan Hajaj, Michaela Hardt, Peter J Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, et al. 2018. Scalable and accurate deep learning with electronic health records. *NPJ digital medicine*, 1(1):18.

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)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

James M Robins, Miguel Angel Hernan, and Babette Brumback. 2000. Marginal structural models and causal inference in epidemiology. *Epidemiology*, pages 550–560.

S Trent Rosenbloom, Joshua C Denny, Hua Xu, Nancy Lorenzi, William W Stead, and Kevin B Johnson. 2011. Data from clinical notes: a perspective on the tension between structure and flexible documentation. *Journal of the American Medical Informatics Association*, 18(2):181–186.

Andrea Rotnitzky and Ezequiel Smucler. 2020. Efficient adjustment sets for population average causal treatment effect estimation in graphical models. *The Journal of Machine Learning Research*, 21(1):7642–7727.

Serena Sanna, Natalie R van Zuydam, Anubha Mahajan, Alexander Kurilshikov, Arnau Vich Vila, Urmo Võsa, Zlatan Mujagic, Ad AM Masclee, Daisy MAE Jonkers, Marije Oosting, et al. 2019. Causal relationships among the gut microbiome, short-chain fatty acids and metabolic diseases. *Nature genetics*, 51(4):600–605.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *Advances in Neural Information Processing Systems*, 33:16857–16867.

Sofia Triantafillou, Vincenzo Lagani, Christina Heinze-Deml, Angelika Schmidt, Jesper Tegner, and Ioannis Tsamardinos. 2017. Predicting causal relationships from biological data: Applying automated causal discovery on mass cytometry data of human immune cells. *Scientific reports*, 7(1):12724.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

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.

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.

Zach Wood-Doughty, Ilya Shpitser, and Mark Dredze. 2021. Generating synthetic text data to evaluate causal inference methods. *arXiv preprint arXiv:2102.05638*.

Chia-Yi Wu, Chin-Kuo Chang, Debbie Robson, Richard Jackson, Shaw-Ji Chen, Richard D Hayes, and Robert Stewart. 2013. Evaluation of smoking status identification using electronic health records and open-text information in a large mental health case register. *PloS one*, 8(9):e74262.

Liuyi Yao, Sheng Li, Yaliang Li, Hongfei Xue, Jing Gao, and Aidong Zhang. 2019. On the estimation of treatment effect with text covariates. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 4106–4113. International Joint Conferences on Artificial Intelligence Organization.

A Implementation

This section gives a more detailed overview of our implementation, including specific hyper-parameter values for both model configurations and parameterization choices of $\mathbb{P}(A \mid C)$ required by the RCT rejection sampling algorithm.

BERT+Adapters. For our BERT adapter configuration, we use a batch size of 128, the maximum that can fit parallelized across two RTX 8000's. We use default values for beta and weight decay, setting $B_1 = 0.9$, $B_2 = 0.999$, $\lambda = 0$. We manually optimize for the learning rate and number of epochs based on validation accuracy on a small subset of the 100 datasets, resulting in a learning rate of $3e$-4 over 5 epochs. Our estimation takes around 10 hours to complete. For the estimation of a single dataset, we suggest practitioners perform a larger search over hyper-parameters, however the use of sample-splitting and doubly-robust estimation requires training 4 times the number of models. Thus, a simple grid-search over just 10 hyper-parameter combinations with 4-fold cross-validation over 100 dataset seeds would require the training of $16,000$ models. Finally, we use BERT$_{\text{BASE}}$ which has $109,482,240$ parameters, however the use of adapters allows us to only fine-tune $894,528$ parameters.

Embedding+FFN. For all of our FFN configurations, we use the same batch size of 128 and the same default beta and weight decay values. We use a single hidden layer with the same number of nodes as the input layer, equal to 768 for both sentence transformers. Since these FFNs are much quicker to train, we perform a search over the learning rates, $\{1e\text{-}5, 1e\text{-}4, 1e\text{-}3, 1e\text{-}2\}$, combined with early-stopping for each one of the 100 dataset subsets.

TF-IDF Tokenization For the *TF-IDF+FFN* baseline, we follow the same tokenization and vocabulary selection procedure as used for BoW by Keith et al. (2023) to allow for a fair comparison. In particular, the text is first preprocessed to remove numbers. We then utilize the following parameters:

- `max_features=2000`: The maximum number of features to consider based on term frequency across the corpus.
- `lowercase=True`: Convert all characters to lowercase before tokenizing.
- `strip_accents="unicode"`: Remove accents and perform other character normalization during the preprocessing step.
- `stop_words="english"`: Exclude common English stop words from the vocabulary.
- `max_df=0.9`: Ignore terms that appear in more than 90% of the documents.
- `min_df=5`: Ignore terms that appear in fewer than 5 documents.
- `binary=True`: All non-zero term counts are set to 1.

For the remaining parameters unique to TF-IDF (not present for BoW), we use the default sklearn parameters:

- `norm='l2'`: Sum of squares of vector elements is 1. The cosine similarity between two vectors is their dot product when l2 norm has been applied.
- `use_idf=True`: Enable inverse-document-frequency reweighting.

- `smooth_idf=True`: Smooth idf weights by adding one to document frequencies, as if an extra document was seen containing every term in the collection exactly once. Prevents zero divisions.

- `sublinear_tf=False`: Apply sublinear tf scaling, i.e. replace tf with $1 + \log(\mathrm{tf})$.

RCT parameterization. The RCT rejection sampling algorithm requires practitioners to specify $\mathbb{P}(A \mid C)$. In particular, the authors choose C to be a binary random variable representing the specific text topic. We accordingly utilize the default provided RCT using medicine $(C = 0)$ and physics $(C = 1)$ articles. Authors then define $\mathbb{P}(A \mid C)$ as follows:

$$\mathbb{P}(A = 1 \mid C) = \begin{cases} \zeta_0 & \text{if } C = 0 \\ \zeta_1 & \text{if } C = 1 \end{cases} \tag{17}$$

which is used in sampling the RCT to create an artificial $C \rightarrow A$ effect. We utilize the default choices of $\zeta_0 = 0.85$ and $\zeta_1 = 0.15$ which induce the highest amount of confounding. For a much more thorough explanation, we direct readers to Keith et al. (2023).

B Nuisance Model Predictive Accuracy

Model	Accuracy	
	$\widehat{m}_0$	$\widehat{g}_0$
Logistic Regression	75.5	82.8
CatBoost	80.3	95.5
TF-IDF+FFN	80.6	95.3
SPECTER+FFN	82.8	95.7
MPNetV2+FFN	83.2	95.7
BERT+Adapter	83.2	95.7

Table 2: Average Predictive Accuracy over 100 dataset subsets

Specific values for the average predictive accuracy during estimation of all tested nuisance models are provided in Table 2. A similar trend appears compared to causal estimation results in Table 1, where the largest improvement occurs from simply switching to non-linear nuisance models (*CatBoost* vs. *LogisticRegression*).

While our three **DoubleLingo** model configurations achieve the best predictive accuracies $(83.2\%, 95.7\%)$, the values are only slightly higher than those for the *TF-IDF+FFN* implementation. Here, it's important to note that predictive accuracy alone does not directly contribute to a more accurate estimation (Wood-Doughty et al., 2018).

C Use of Scientific Artifacts & Licensing

Our work uses the RCT rejection sampling dataset by Keith et al. (2023). In particular, the dataset is fully in English, containing publicly available paper titles and abstracts. The authors remove any potentially personally identifiable information from the dataset (author names, user ids, user IP addresses, or session ids). The dataset is made publically available for research purposes (apache-2.0).

Finally, **DoubleLingo** uses the Hugging Face implementations for *bert-base-uncased*, *allenai/specter*, and *all-mpnet-base-v2*, all made publically available for research purposes (apache-2.0).