JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2026, VOL. 00, NO. 0, 1–14: Special Issue on Statistical Science in Artificial Intelligence
https://doi.org/10.1080/01621459.2026.2689629

Causal Inference with Generative Artificial Intelligence: Application to Texts as Treatments

Kosuke Imai[a] and Kentaro Nakamura[b]

[a]Department of Government and Department of Statistics, Harvard University, Cambridge, MA; [b]John F. Kennedy School of Government, Harvard University, Cambridge, MA

ABSTRACT

In this article, we demonstrate how to enhance the validity of causal inference with unstructured high-dimensional treatments like texts, by leveraging the power of generative Artificial Intelligence (GenAI). Specifically, we propose to use a deep generative model such as large language models (LLMs) to efficiently generate treatments and use their internal representation for subsequent causal effect estimation. We show that the knowledge of this true internal representation helps disentangle the treatment features of interest, such as specific sentiments and certain topics, from other possibly unknown confounding features. Unlike existing methods, the proposed GenAI-Powered Inference (GPI) methodology eliminates the need to learn representation directly from the data, and hence produces more accurate and efficient estimates. We formally establish the conditions required for the nonparametric identification of the average treatment effect, propose an estimation strategy that avoids the violation of the overlap assumption, and derive the asymptotic properties of the proposed estimator through the application of double machine learning. Finally, using an instrumental variables approach, we extend the proposed GPI methodology to the settings in which the treatment feature is based on human perception. The GPI is also applicable to text reuse where an LLM is used to regenerate existing texts. We conduct simulation and empirical studies, using the generated text data from an open-source LLM, Llama 3, to illustrate the advantages of our estimator over state-of-the-art representation learning algorithms. An open-source software package is available for implementing the proposed GPI methodology. Supplementary materials for this article are available online, including a standardized description of the materials available for reproducing the work.

ARTICLE HISTORY
Received January 2025
Accepted June 2026

KEYWORDS
Deep generative models;
Double machine learning;
Large language models;
Representation learning;
Unstructured treatments

1. Introduction

The emergence of generative artificial intelligence (GenAI) technology such as large language models (LLMs) has had an enormous impact on many fields, including medicine (Thirunavukarasu et al. 2023), education (Kasneci et al. 2023), marketing (Kshetri et al. 2024), and social sciences (Bisbee et al. 2024). These tools come with advanced capabilities to generate realistic texts, images, and videos at scale, based on the user-provided prompts.

In this article, we demonstrate that GenAI can enhance the performance of causal inference based on unstructured data such as text and images. We focus on the problem of estimating the causal effect of a specific treatment feature embedded in text—such as topic or sentiment—while adjusting for other confounding features. Although we assume the treatment feature is pre-specified and measurable, the central challenge lies in learning a low-dimensional representation of the unknown confounders and properly adjusting for them. We show that internal representations extracted from GenAI models can be leveraged within a causal machine learning framework to estimate the causal effects of interest. Specifically, we introduce an experimental design (Section 2) and propose the GenAI-Powered Inference (GPI) methodology (Section 3), which enables efficient estimation of causal effects associated with specific features embedded in unstructured data.

In the proposed experiment, we first generate texts by providing treatment and control prompts to an LLM. If we wish to use existing texts rather than generate new ones, we ask an LLM to reproduce the same texts exactly. We then present each generated text to a randomly selected survey respondent and measure their reactions. Lastly, we directly extract the true internal representation of the generated text from the LLM and analyze it with a machine learning algorithm to disentangle the treatment features from other confounding features contained within the same texts. The GPI leverages the true vectorized representation of treatment text that is available from an open-source deep generative model. This enables us to efficiently learn the representation of relevant features in unstructured data even when texts contain strong confounding features.

We establish the nonparametric identification based on the key assumption of *separability* between treatment and confounding features. This assumption states that the treatment feature is not a deterministic function of confounding features and the latter are also not a function of the former. The assumption is closely related to the concept of disentanglement in the literature (Wang and Jordan 2024) and implies that

CONTACT Kosuke Imai *imai@harvard.edu* Department of Government and Department of Statistics, Harvard University, Cambridge, MA 02138.
Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/JASA*.

one can intervene the treatment feature without changing confounding features. We discuss diagnostic tools to detect the potential violation of this assumption.

As part of the proposed estimation approach, we develop a neural network architecture based on TarNet (Shalit, Johansson, and Sontag 2017) to separately learn treatment and confounding features. We show that it is possible to nonparametrically identify a *deconfounder*, which summarizes all confounding features as a lower-dimensional function of the internal representation obtained from a deep generative model. Once a deconfounder is estimated, we use it to model the treatment features and estimate the propensity score. We then apply the double machine learning (DML) methodology to obtain the asymptotically valid confidence intervals (Chernozhukov et al. 2018). An open-source Python package, "GPI: Generative-AI Powered Inference," that implements the proposed methodology is available at *https://gpi-pack.github.io/*.

To investigate the empirical performance of the proposed methodology, we design simulation studies based on the candidate profile experiment of Fong and Grimmer (2016) (Section 4). In this experiment, survey respondents are asked to evaluate biographies of different political candidates. Our goal is to infer the causal effects of certain features of these biographies, such as military background and education levels. We use an open-source LLM, Llama 3, to generate a set of new candidate biographies and design simulation studies to compare the performance of the proposed estimator with that of state-of-the-art methods. We also have Llama 3 regenerate the existing biographies to examine the performance of the proposed methodology in the case of text reuse.

We find that the GPI outperforms the state-of-the-art methods, which estimate the representation of confounding features using the existing embedding (Pryzant et al. 2021; Gui and Veitch 2023). Specifically, the proposed estimator has a smaller bias and root mean squared error, while its confidence interval retains a proper nominal coverage level. These findings hold even when the sample size is relatively small, and in the case of text reuse. Furthermore, we apply the GPI to the candidate profile experiment (Section 5). Our analysis shows that the previous military experience of the candidates significantly affects the voter evaluation on average. Appendix S6 presents an additional empirical application based on experiments about the public perceptions of US government support for Hong Kong protest (Fong and Grimmer 2023). We show that the GPI yields much more reasonable estimates than the existing state-of-the-art methods. Imai and Nakamura (2025) presents additional applications of the GPI including one based on images.

Finally, we extend the GPI to the estimation of the causal effects of *perceived* treatment features, motivated by the fact that the respondents may perceive the same treatment features differently (Appendix S3). A key challenge is that the perceived treatment features may be confounded by possibly unobserved respondent characteristics as well as the confounding features of texts. To address this, we adopt the instrumental variable approach (Imbens and Angrist 1994) by using the actual treatment features as instruments for their perceived counterparts.

Related literature. Several existing works estimate the causal effects of textual features (see Feder et al. 2022, for a review). The key difference between the GPI and these previous approaches is the use of GenAI to produce treatment objects. We exploit the fact that the *true* vectorized representation of generated texts can be obtained directly from open-source LLMs. In contrast, existing methods must learn representation from the treatment texts. For example, Fong and Grimmer (2016) and Ahrens et al. (2021) impose topic models, whereas Pryzant et al. (2021) and Gui and Veitch (2023) estimate the vector representation using the BERT (Bidirectional Encoder Representations from Transformers) embeddings. We show that the use of true representation not only improves the estimation performance but also significantly increases computational efficiency.

We advance the literature on causal inference with text by formalizing the assumptions necessary for causal identification. Most prior studies implicitly assume that confounding features are not functions of the treatment feature (e.g., Pryzant et al. 2021; Gui and Veitch 2023), with the exception of Daoud, Jerzak, and Johansson (2022), who states this assumption explicitly. Consequently, no existing estimation method directly incorporates this assumption. For instance, Fong and Grimmer (2023) recommends using topic models and adjusting only for those estimated topics that are not functions of the treatment feature. In practice, however, reliably identifying such topics is challenging, since any topic may be entangled with the treatment feature. Relatedly, the representation learning literature discusses this condition under the label of disentanglement (Wang and Jordan 2024), but does not consider interventions on the treatment feature while holding confounding features fixed. To address this gap, we introduce the separability assumption, which implies that it be possible to intervene on the treatment feature without altering the confounding features.

Our work has broader implications for the literature on causal inference and unstructured objects in general. For example, scholars have developed methods that adjust for texts as confounders, but all of these methods estimate a low-dimensional representation of confounding information from texts (Veitch, Sridhar, and Blei 2020; Roberts, Stewart, and Nielsen 2020; Mozer et al. 2020, 2024; Klaassen et al. 2024). Although our paper focuses on the use of texts as treatments rather than confounders, the proposed use of GenAI should be beneficial for these other cases where existing estimation methods are likely to suffer from estimation error (Keith, Jensen, and O'Connor 2020). Similarly, the GPI can also be applied to causal inference problems with images and even videos. For example, Jerzak, Johansson, and Daoud (2023a,b) have considered the use of causal inference with images in an observational setting. Given the availability of deep generative models for images, it would be of interest to use GenAI for improved causal inference with images (e.g., Ramesh et al. 2021; Rombach et al. 2022).

Our work also contributes to the literature on representation learning. Since unstructured objects such as texts are high-dimensional, learning a low-dimensional representation is crucial (e.g., Shi, Blei, and Veitch 2019; Wang and Jordan 2024). Some propose learning a low-dimensional representation that predicts both treatment and outcome (e.g., Veitch, Sridhar, and Blei 2020; Gui and Veitch 2023). Although we also

use a low-dimensional representation to adjust for confounding features, the GPI disentangles confounding features from treatment features without violating the overlap assumption. In addition, the GPI focuses on estimating causal effects rather than discovering of causal structure, which is an important goal of representation learning (Schölkopf et al. 2021).

Finally, the GPI differs from that of Wang and Blei (2019) though both use the estimated "deconfounder" for confounding adjustment. Wang and Blei adjust the *unobserved* confounding by estimating the deconfounder as a function of treatments (Imai and Jiang 2019). In contrast, the GPI learns a representation of *observed* confounding by estimating the deconfounder as a function of generative model's internal representation of treatment objects. In this setting, we formally establish the nonparametric identification of the average treatment effect.

2. The Experimental Design for Texts as Treatments

While the GPI is a general methodological approach, it is helpful to consider a concrete application. We use the candidate profile experiment of Fong and Grimmer (2016) which investigates how various features of a political candidate affect voter evaluation. In the context of this experiment, we describe our alternative approach that uses LLMs to generate treatment texts. In Sections 4 and 5, we return to this application to empirically evaluate the performance of the GPI.

2.1. Candidate Profile Experiment

Fong and Grimmer (2016) collected 1246 candidate biographies (written in texts) from Wikipedia, and then asked a total of 1,886 voters to answer an online survey evaluating up to four randomly assigned candidate profiles. Specifically, the survey asked these voters to rate each candidate biography using a feeling thermometer that ranges from 0 (cold) to 100 (warm). The authors first used a supervised Bayesian model based on the Indian buffet process to discover a total of 10 treatment features and then estimated the marginal association between each treatment feature and the observed feeling thermometer value.

Unlike the original analysis, we consider a setting in which researchers have a specific treatment feature of interest. Suppose that we are interested in estimating the causal effect of having a military background on voter evaluation. The relevant political science literature suggests that the experience and occupation of candidates play an important role (e.g., Campbell and Cowley 2014; Kirkland and Coppock 2018; Pedersen, Dahlgaard, and Citi 2019). Indeed, the original analysis shows that candidate biographies with military background tend to receive a high feeling thermometer score.

The challenge, however, is that military background may be correlated with other features present in candidate biographies. Table S1 of Appendix S1 displays two example biographies used in the original experiment. The first describes a candidate with military background, whereas the second shows another without military background. Yet, these two biographies also differ in terms of other features, including educational background, marital status, and family structure. The length of each biography and their levels of detail are also different. If these differences are correlated with military background and influence voter evaluation, a simple comparison between biographies with military background and those without it would lead to biased causal estimates.

2.2. Using Large Language Models to (Re)generate Treatment Texts

We use an LLM to generate treatment texts. In the current context, we can achieve this in two ways. First, we can provide a prompt to an LLM, asking it to generate a candidate biography from scratch. Alternatively, we can use existing biographies (e.g., those collected by Fong and Grimmer 2016) and then ask an LLM to reproduce the same biographies without any modification.

Both approaches require (1) the coding of the treatment variable (e.g., the existence of military background) and (2) the extraction of the internal representation of LLM used to generate treatment texts. The first requirement may mean that human coders have to read treatment texts unless one is willing to assume that LLM has a good compliance with the instruction in the prompts. The second requirement implies that we should use open-source LLMs including GPT (Generative Pre-trained Transformer), Llama (Large Language Model Meta AI), and OPT (Open Pre-trained Transformer).

Table S2 of Appendix S1 shows an example from each approach. Here, we use Meta's Llama 3 instruction-tuned model with eight billion parameters. This model takes two types of inputs: system-level inputs (**System**), which define the type of task to be performed, and user-level inputs (**User**), which define a specific task to be performed. The first example in the table shows how the model generates a new candidate biography with military background from scratch, whereas the second shows how the model reproduces a given biography. The former suggests that an LLM can create realistic treatments, while the latter indicates that it can follow the instruction accurately.

We emphasize that the use of LLM in itself does not automatically solve the confounding bias. This is because the LLM learns the associations of words in real-world texts, and even if one manipulates the concepts with instructions, other correlated concepts might also be influenced, causing the confounding bias (Hu and Li 2021).

3. The Proposed Methodology

We turn to the proposed GPI methodology that adjusts for confounding features in unstructured treatment objects such as texts to estimate the causal effects of the specific treatment feature. We begin by defining the causal quantity of interest, establish its nonparametric identification, and then develop an estimation strategy.

3.1. Assumptions and Causal Quantity of Interest

Consider a simple random sample of N respondents drawn from a population of interest. For each respondent $i = 1, 2, \ldots, N$, we assign a prompt $\boldsymbol{P}_i$ that is randomly and independently sam-

pled from a set of potential prompts $\mathcal{P}$. In our application, the prompts are based on natural language (e.g., "Create a biography of an American politician who has some military experience"). Given each prompt, we use a deep generative model to generate a *treatment object* such as text, denoted by $\boldsymbol{X}_i \in \mathcal{X}$ where $\mathcal{X}$ is the support of $\boldsymbol{X}_i$.

We use a broad definition of a deep generative model to encompass LLMs and other foundation models. Indeed, our definition includes many models for texts (e.g., Zhang et al. 2022; Touvron et al. 2023; Jiang et al. 2023a) and images (e.g., Ramesh et al. 2021; Rombach et al. 2022).

Definition 1 (*Deep Generative Model*). A deep generative model is the following probabilistic model that takes prompt $\boldsymbol{P}_i$ as an input and generates the treatment object $\boldsymbol{X}_i$ as an output:

$$\mathbb{P}(\boldsymbol{X}_i \mid \boldsymbol{h}_{\boldsymbol{\gamma}}(\boldsymbol{R}_i))$$
$$\mathbb{P}(\boldsymbol{R}_i \mid \boldsymbol{P}_i)$$

where $\boldsymbol{R}_i \in \mathcal{R} \subset \mathbb{R}^{d_R}$ denotes an observable internal representation of $\boldsymbol{X}_i$ contained in the model and $\boldsymbol{h}_{\boldsymbol{\gamma}}(\boldsymbol{R}_i)$ is a deterministic function parameterized by $\boldsymbol{\gamma}$ that completely characterizes the conditional distribution of $\boldsymbol{X}_i$ given $\boldsymbol{R}_i$.

Under this definition, $\boldsymbol{R}$ represents a lower-dimensional representation of the treatment object $\boldsymbol{X}$ and is a hidden representation of neural networks. In addition, $\boldsymbol{h}_{\boldsymbol{\gamma}}(\boldsymbol{R})$ is the propensity function (Imai and van Dyk 2004) and is both known and observable in an open-source deep generative model. Thus, the treatment object $\boldsymbol{X}$ depends on $\boldsymbol{P}$ only through $\boldsymbol{h}_{\boldsymbol{\gamma}}(\boldsymbol{R})$. Note that any given text $\boldsymbol{X}_i$ can have multiple internal representations. For example, one can first instruct an LLM to generate a new text and then ask it to repeat the generated text exactly. These two prompts will produce the same text but yield different internal representations. In Section 4.3, our simulation study shows that the internal representation of regenerated text leads to more efficient causal estimates.

In the proposed experiment, each respondent i is exposed to the generated treatment object $\boldsymbol{X}_i$, and subsequently generates the outcome variable $Y_i \in \mathcal{Y} \subset \mathbb{R}$ where $\mathcal{Y}$ is its support. In our application, the treatment object is a candidate biography and the outcome is a respondent's evaluation of the biography. Let $Y_i(\boldsymbol{x})$ denote the potential outcome of respondent i when exposed to a treatment object $\boldsymbol{x} \in \mathcal{X}$. Then, the observed outcome is solely determined by the assigned treatment object, that is, $Y_i = Y_i(\boldsymbol{X}_i)$. This experimental design implies the following two assumptions.

Assumption 1 (*Consistency*). The potential outcome under the treatment object $\boldsymbol{x} \in \mathcal{X}$, denoted by $Y_i(\boldsymbol{x})$, equals the observed outcome Y_i under the realized treatment object $\boldsymbol{X}_i$:

$$Y_i = Y_i(\boldsymbol{X}_i).$$

Assumption 2 (*Random Assignment of Prompt*). Prompt is randomly assigned such that the following independence holds for all i and $\boldsymbol{x} \in \mathcal{X}$:

$$Y_i(\boldsymbol{x}) \perp\!\!\!\perp \boldsymbol{P}_i.$$

We estimate the causal effect of a particular feature that can be part of a treatment object. For simplicity, we consider a binary treatment feature denoted by T_i for each i. In our application, $T_i = 1$ if candidate biography i contains military background and $T_i = 0$ otherwise. We assume that the treatment feature is determined solely by the treatment object (Egami et al. 2022).

Assumption 3 (*Treatment Feature*). There exists a deterministic function $g_T : \mathcal{X} \to \{0, 1\}$ that maps a treatment object $\boldsymbol{X}_i$ to a binary treatment feature of interest T_i, that is,

$$T_i = g_T(\boldsymbol{X}_i).$$

This assumption is violated, for example, if respondents infer different values of the treatment feature from the same treatment object. We address this issue in Appendix S3 by considering perceived treatment, which reflects heterogeneity between respondents.

Next, we define confounding features, which represent all features of $\boldsymbol{X}$ other than the treatment feature T that influence the outcome Y. These confounding features, denoted by $\boldsymbol{U} \in \mathcal{U}$, are based on a vector-valued deterministic function of $\boldsymbol{X}$ where $\mathcal{U}$ denotes their support. The confounding features may be multidimensional, though we assume that its dimensionality is much smaller than that of the treatment object. In our application, confounding features may include other candidate characteristics and textual features of biographies, such as their length. These confounding features are possibly correlated with the treatment feature. Unlike the treatment feature, however, we do not directly observe the confounding features. Formally, the confounding features are defined as follows (Pryzant et al. 2021; Fong and Grimmer 2023).

Assumption 4 (*Confounding Features*). There exists an unknown vector-valued deterministic function $\boldsymbol{g}_{\boldsymbol{U}} : \mathcal{X} \to \mathcal{U}$ that maps an unstructured object $\boldsymbol{X}_i \in \mathcal{X}$ to the confounding features $\boldsymbol{U}_i \in \mathcal{U}$, that is,

$$\boldsymbol{U}_i = \boldsymbol{g}_{\boldsymbol{U}}(\boldsymbol{X}_i),$$

where $\dim(\boldsymbol{U}_i) \ll \dim(\boldsymbol{X}_i)$.

Finally, we introduce our key assumption that the potential outcome is a function of treatment and confounding features and that we can intervene the treatment feature without changing the confounding features. In other words, the treatment feature cannot be represented as a deterministic function of the confounding features. In addition, confounding features should not vary as a deterministic function of treatment feature in order to avoid "posttreatment" bias (Daoud, Jerzak, and Johansson 2022). In the candidate biography application, the assumption essentially implies that researchers need to be able to imagine hypothetical candidate biographies with and without military background while keeping the confounding features constant. We formalize this assumption as follows.

Assumption 5 (*Separability of Treatment and Confounding Features*). The potential outcome is a function of the treatment feature of interest t and another separate function of the confounding features $\boldsymbol{u}$. That is, for any given $\boldsymbol{x} \in \mathcal{X}$ and all i, we

have:

$$Y_i(\boldsymbol{x}) = Y_i(t, \boldsymbol{u}) = Y_i(g_T(\boldsymbol{x}), \boldsymbol{g_U}(\boldsymbol{x})),$$

where $t = g_T(\boldsymbol{x}) \in \{0, 1\}$ and $\boldsymbol{u} = \boldsymbol{g_U}(\boldsymbol{x}) \in \mathcal{U}$. In addition, g_T and $\boldsymbol{g_U}$ are separable. That is, there exists no deterministic function $\tilde{g}_T : \mathcal{U} \to \{0, 1\}$, which satisfies $g_T(\boldsymbol{x}) = \tilde{g}_T(\boldsymbol{g_U}(\boldsymbol{x}))$ for some $\boldsymbol{x} \in \mathcal{X}$. Similarly, there exist no deterministic functions $\boldsymbol{g}' : \mathcal{X} \to \mathcal{X}'$ and $\tilde{\boldsymbol{g}}_{\boldsymbol{U}} : \{0, 1\} \times \mathcal{X}' \to \mathcal{U}$, which satisfy $\boldsymbol{g_U}(\boldsymbol{x}) = \tilde{\boldsymbol{g}}_{\boldsymbol{U}}(g_T(\boldsymbol{x}), \boldsymbol{g}'(\boldsymbol{x}))$ for all $\boldsymbol{x} \in \mathcal{X}$ and $\tilde{\boldsymbol{g}}_{\boldsymbol{U}}(1, \boldsymbol{g}'(\boldsymbol{x}')) \neq \tilde{\boldsymbol{g}}_{\boldsymbol{U}}(0, \boldsymbol{g}'(\boldsymbol{x}'))$ for some $\boldsymbol{x}' \in \mathcal{X}$.

The first requirement of separability (i.e., no existence of $\tilde{g}_T$) implies that the treatment should not be a deterministic function of confounding features, whereas the second requirement (i.e., no existence of $\boldsymbol{g}'$ and $\tilde{\boldsymbol{g}}_{\boldsymbol{U}}$) means that the confounding features cannot vary as a deterministic function of treatment. Thus, the separability assumption holds when the treatment feature does not completely overlap with the other features that influence the outcome (i.e., confounding features). In Appendix S2, we provides two simple examples: one, in which the separability assumption holds, and the other one, where it is violated.

As shown in Section 3.2, Assumption 5 plays an essential role in the identification of causal effects. Importantly, Assumptions 3–5 imply the standard overlap condition in causal inference, which enables to identify causal effects without extrapolation. In practice, as demonstrated in Section 4.3, the violation of Assumption 5 can be diagnosed by examining if the estimated propensity scores take extreme values. The following lemma formally establishes that the separability assumption implies the overlap condition.

Lemma 1 (Overlap). Under Assumptions 3–5, for any $t \in \{0, 1\}$ and $\boldsymbol{u} \in \mathcal{U}$,

$$\mathbb{P}(T_i = t \mid \boldsymbol{U}_i = \boldsymbol{u}) > 0.$$

The proof is given in Appendix S4.1.

Under this setup, we estimate the average causal effect of the treatment feature while controlling for the confounding features. This average treatment effect (ATE) is defined as follows:

$$\tau := \mathbb{E}[Y_i(1, \boldsymbol{U}_i) - Y_i(0, \boldsymbol{U}_i)]. \tag{1}$$

3.2. Nonparametric Identification

Figure 1 presents a directed acyclic graph (DAG) that summarizes the data generating process described above. In the DAG, an arrow with double lines represents a deterministic causal relationship, while an arrow with a single line indicates a possibly stochastic causal relationship. We consider a deep generative model whose decoding is deterministic, meaning that the output object is a deterministic function of the input prompt. Formally, we make the following assumption.

Assumption 6 (Deterministic Decoding). The output layer of a deep generative model is deterministic. That is, $\mathbb{P}(\boldsymbol{X}_i \mid \boldsymbol{h_\gamma}(\boldsymbol{R}_i))$ in Definition 1 is a degenerate distribution.

If $\mathbb{P}(\boldsymbol{X}_i \mid \boldsymbol{h_\gamma}(\boldsymbol{R}_i))$ is stochastic, the noise introduced by the deep generative model can confound the treatment in an unknown way. Fortunately, almost all LLMs have the option of deterministic decoding (e.g., greedy, beam, and contrastive searches), thereby easily satisfying the assumption (e.g., Su et al. 2022). For example, for greedy decoding, we typically only need to set the temperature parameter to zero, instructing a model to always produce the same texts. Indeed, many LLMs also have deterministic *encoding* architectures, making the entire text generation process deterministic. Similarly, for images, we can make the final decoding step deterministic for stochastic diffusion models, including Stable Diffusion (Rombach et al. 2022).

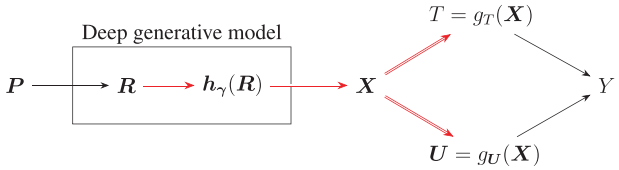

Figure 1. Directed acyclic graph of the assumed data generating process. A treatment object $\boldsymbol{X}$ is generated using a deep generative model (rectangle), in which a prompt $\boldsymbol{P}$ produces an internal representation $\boldsymbol{R}$ that generates $\boldsymbol{X}$ through a propensity function $\boldsymbol{h_\gamma}(\boldsymbol{R})$. The treatment object affects the outcome Y through its treatment feature of interest T and other confounding features $\boldsymbol{U}$. An arrow with red double lines represents a deterministic causal relation while an arrow with a single line indicates a possibly stochastic relationship.

Given this setup, we establish the nonparametric identification of the marginal distribution of potential outcome. We prove the existence of a deconfounder function $\boldsymbol{f}(\boldsymbol{R}_i)$ that satisfies the conditional independence relation $Y_i \perp\!\!\!\perp \boldsymbol{R}_i \mid T_i, \boldsymbol{f}(\boldsymbol{R}_i)$. One trivial example of the deconfounder is the confounding features $\boldsymbol{U}_i$, which are a deterministic function of $\boldsymbol{R}_i$ under Assumption 6. But, the deconfounder need not be unique. In fact, we show that it is possible to identify the marginal distribution of potential outcome by adjusting for any deconfounder.

Theorem 1 (Nonparametric Identification of the Marginal Distribution of Potential Outcome). Under Assumptions 1–6, there exists a deconfounder function $\boldsymbol{f} : \mathcal{R} \to \mathcal{Q} \subset \mathbb{R}^{d_Q}$ with $d_Q = \dim(\mathcal{Q}) \leq d_R = \dim(\mathcal{R})$ that satisfies the following conditional independence relation:

$$Y_i \perp\!\!\!\perp \boldsymbol{R}_i \mid T_i = t, \boldsymbol{f}(\boldsymbol{R}_i) = \boldsymbol{q}, \tag{2}$$

where $0 < \mathbb{P}(T_i = t \mid \boldsymbol{f}(\boldsymbol{R}_i) = \boldsymbol{q}) < 1$ for all $t = 0, 1$ and $\boldsymbol{q} \in \mathcal{Q}$. In addition, the treatment feature and a deconfounder are separable. By adjusting for such a deconfounder, we can uniquely and nonparametrically identify the marginal distribution of the potential outcome under the treatment condition $T_i = t$ for $t \in \{0, 1\}$ as:

$$\mathbb{P}(Y_i(t, \boldsymbol{U}_i) = y) = \int_{\mathcal{R}} \mathbb{P}(Y_i = y \mid T_i = t, \boldsymbol{f}(\boldsymbol{R}_i)) dF(\boldsymbol{R}_i),$$

for all $t \in \{0, 1\}$ and $y \in \mathcal{Y}$.

The proof is given in Appendix S4.2. The definition of separability is given in Assumption 5 for the treatment and confounding features given the treatment object. In Theorem 1, we apply the same definition to the treatment feature and a deconfounder given the internal representation. Specifically, let $f_T : \mathcal{R} \to \{0, 1\}$ be the function that maps the internal representation to the treatment feature. Such a function exists because the treatment

feature is a deterministic function of treatment object, which is in turn a deterministic function of the internal representation by Assumption 6. Then, we assume that f_T and $\boldsymbol{f}$ are separable. That is, there exists no deterministic function $\tilde{f}_T : \mathcal{Q} \to \{0,1\}$, which satisfies $f_T(\boldsymbol{r}) = \tilde{f}_T(\boldsymbol{f}(\boldsymbol{r}))$ for all $\boldsymbol{r} \in \mathcal{R}$. Similarly, there exist no deterministic functions $\boldsymbol{f}' : \mathcal{Q} \to \mathcal{Q}'$ and $\tilde{\boldsymbol{f}} : \{0,1\} \times \mathcal{Q}' \to \mathcal{Q}$, which satisfy $\boldsymbol{f}(\boldsymbol{r}) = \tilde{\boldsymbol{f}}(f_T(\boldsymbol{r}),\boldsymbol{f}'(\boldsymbol{r}))$ for all $\boldsymbol{r} \in \mathcal{R}$ and $\tilde{\boldsymbol{f}}(1,\boldsymbol{f}'(\boldsymbol{r}')) \neq \tilde{\boldsymbol{f}}(0,\boldsymbol{f}'(\boldsymbol{r}'))$ for some $\boldsymbol{r}' \in \mathcal{R}$. The identification of the ATE defined in (1) can be established in the same way, except that the deconfounder assumption is relaxed to the mean independence condition

$$\mathbb{E}[Y_i \mid T_i = t, \boldsymbol{f}(\boldsymbol{R}_i)] = \mathbb{E}[Y_i \mid T_i = t, \boldsymbol{R}_i], \tag{3}$$

which we will use for estimation in the next section.

We emphasize that one cannot directly adjust for the internal representation of the treatment object. Such direct adjustment leads to the lack of overlap because, under Assumptions 3 and 6, the treatment feature T_i is a deterministic function of $\boldsymbol{R}_i$. In addition, since the deconfounder $\boldsymbol{f}(\boldsymbol{R}_i)$ is typically of much lower dimension than the internal representation of the treatment object $\boldsymbol{R}_i$, making adjustments for the former leads to a more effective estimation strategy.

3.3. Estimation and Inference

Given the identification result, we next consider estimation and inference. Our estimation strategy is based on the following two observations. First, Assumption 5 implies that the deconfounder $\boldsymbol{f}(\boldsymbol{R}_i)$ should not be a function of the treatment feature T_i. Second, the deconfounder should satisfy the conditional mean independence relation given in (3). For simplicity, we assume independence across observations. Technically, if a deep generative model has a stochastic component, we can only assume the conditional mean independence of observations given the internal representation. However, as discussed earlier, many language models have the option of making the whole text generation process deterministic, guaranteeing this conditional mean independence.

We use a neural network architecture based on TarNet (Shalit, Johansson, and Sontag 2017) to estimate the conditional potential outcome function given the deconfounder, that is,

$$\mu_t(\boldsymbol{f}(\boldsymbol{R}_i)) \;:=\; \mathbb{E}[Y_i(t, \boldsymbol{U}_i) \mid \boldsymbol{f}(\boldsymbol{R}_i)], \qquad \text{for } t = 0, 1.$$

Our architecture, which is summarized in Figure 2, simultaneously estimates the deconfounder and the outcome model. Specifically, we minimize the following loss function:

$$\{\hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\theta}}_0, \hat{\boldsymbol{\theta}}_1\} \;=\; \underset{\boldsymbol{\lambda},\boldsymbol{\theta}_0,\boldsymbol{\theta}_1}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \left\{ Y_i - \mu_{T_i}(\boldsymbol{f}(\boldsymbol{R}_i;\boldsymbol{\lambda});\boldsymbol{\theta}_{T_i}) \right\}^2, \tag{4}$$

where n is the sample size. We make the parameters of neural network explicit by letting $\boldsymbol{\lambda}$ represent the parameters of deconfounder $\boldsymbol{f}$ to be estimated, and using $\boldsymbol{\theta}_t$ to denote the parameters of the nuisance function μ_t.

Given the above architecture, we estimate the ATE using the double machine learning (DML) framework of Chernozhukov et al. (2018), in which both the outcome and the propensity score models are estimated. Here, we estimate the propensity score model as a function of the estimated deconfounder, that is, $\pi(\boldsymbol{f}(\boldsymbol{R}_i;\hat{\boldsymbol{\lambda}})) = \Pr(T_i = 1 \mid \boldsymbol{f}(\boldsymbol{R}_i;\hat{\boldsymbol{\lambda}}))$ after solving the minimization problem in (4). Crutially, we do not model the propensity score as a function of internal representation $\boldsymbol{R}_i$ to avoid violating the assumption of overlap. Thus, the deconfounder only captures the features of treatment object that are predictive of the outcome beyond the treatment, including the features that are completely unrelated to the treatment. The deconfounder, however, will not capture the features of treatment object that affect the outcome only through the treatment feature. So long as such features exist (Assumption 5), the overlap assumption will hold (Lemma 1).

In the representation learning literature, DragonNet (Shi, Blei, and Veitch 2019) is a popular estimation method used by many researchers (see e.g., Veitch, Sridhar, and Blei 2020; Gui and Veitch 2023, as well as Pryzant et al. 2021 who also use it in their implementation code). Unlike our approach, DragonNet includes the cross-entropy loss between the propensity score model and the treatment variable when estimating the deconfounder $\boldsymbol{f}(\boldsymbol{R}_i)$. However, this joint estimation leads to $\mathbb{P}(T_i = 1 \mid \boldsymbol{f}(\boldsymbol{R}_i)) = \mathbb{P}(T_i = 1 \mid \boldsymbol{R}_i)$ due to the fact that $\boldsymbol{f}(\boldsymbol{R}_i)$ is a balancing score satisfying $T_i \perp\!\!\!\perp \boldsymbol{R}_i \mid \boldsymbol{f}(\boldsymbol{R}_i)$. This is problematic because $\mathbb{P}(T_i = 1 \mid \boldsymbol{R}_i)$ is degenerate under Assumptions 3 and 6. Thus, we first estimate the deconfounder using (4) and then model the propensity score given the estimated deconfounder.

In sum, the entire estimation procedure can be described as follows. Denote the observed data by $\mathcal{D} := \{\mathcal{D}_i\}_{i=1}^N$ where $\mathcal{D}_i := \{Y_i, T_i, \boldsymbol{R}_i\}$. We use the following K-fold cross-fitting procedure, assuming that N is divisible by K.

1. Randomly partition the data into K folds of equal size where the size of each fold is $n = N/K$. The observation index is denoted by $I(i) \in \{1, \ldots, K\}$ where $I(i) = k$ implies that the ith observation belongs to the kth fold.

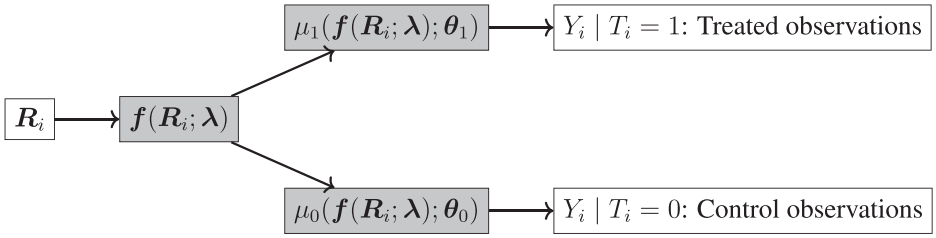

Figure 2. Diagram illustrating the proposed model architecture. The proposed model takes an internal representation of a treatment object $\boldsymbol{R}_i$ as an input, and finds a deconfounder $\boldsymbol{f}(\boldsymbol{R}_i)$, which is a lower-dimensional representation of $\boldsymbol{R}_i$, and then use it to predict the conditional expectation of the outcome $\mu_t(\boldsymbol{f}(\boldsymbol{R}_i)) := \mathbb{E}[Y_i \mid T_i = t, \boldsymbol{f}(\boldsymbol{R}_i)]$ under each treatment arm $t = 0, 1$.

2. For each fold $k \in \{1, \ldots, K\}$, use observations with $I(i) \neq k$ as training data:
   (a) split the training data into two folds, $I_1^{(-k)}$ and $I_2^{(-k)}$
   (b) simultaneously obtain an estimated deconfounder and an estimated conditional outcome function on the first fold, which are denoted by $\hat{\boldsymbol{f}}^{(-k)}(\{\boldsymbol{R}_i\}_{i\in I_1^{(-k)}}) := \boldsymbol{f}(\{\boldsymbol{R}_i\}_{i\in I_1^{(-k)}}; \hat{\boldsymbol{\lambda}}^{(-k)})$ and $\hat{\mu}_t^{(-k)}(\{\boldsymbol{R}_i\}_{i\in I_1^{(-k)}}) := \mu_t(\boldsymbol{f}(\{\boldsymbol{R}_i\}_{i\in I_1^{(-k)}}; \hat{\boldsymbol{\lambda}}^{(-k)}); \hat{\boldsymbol{\theta}}^{(-k)})$, respectively, by solving the optimization problem given in (4), and
   (c) obtain an estimated propensity score given the estimated deconfounder on the second fold, which is denoted by $\hat{\pi}^{(-k)}(\hat{\boldsymbol{f}}^{(-k)}(\{\boldsymbol{R}_i\}_{i\in I_2^{(-k)}})) := \hat{\pi}^{(-k)}(\boldsymbol{f}(\{\boldsymbol{R}_i\}_{i\in I_2^{(-k)}}; \hat{\boldsymbol{\lambda}}^{(-k)}))$.
3. Compute the GPI estimator $\hat{\tau}$ as a solution to:

$$\frac{1}{nK}\sum_{k=1}^{K}\sum_{i:I(i)=k}\psi(\mathcal{D}_i; \hat{\tau}, \hat{\boldsymbol{f}}^{(-k)}, \mu_1^{(-k)}, \mu_0^{(-k)}, \hat{\pi}^{(-k)}) = 0, \tag{5}$$

where

$$\begin{aligned}\psi(\mathcal{D}_i; \tau, \boldsymbol{f}, \mu_1, \mu_0, \pi) &= \frac{T_i\{Y_i - \mu_1(\boldsymbol{f}(\boldsymbol{R}_i))\}}{\pi(\boldsymbol{f}(\boldsymbol{R}_i))} - \frac{(1-T_i)\{Y_i - \mu_0(\boldsymbol{f}(\boldsymbol{R}_i))\}}{1-\pi(\boldsymbol{f}(\boldsymbol{R}_i))} \\ &\quad + \mu_1(\boldsymbol{f}(\boldsymbol{R}_i)) - \mu_0(\boldsymbol{f}(\boldsymbol{R}_i)) - \tau.\end{aligned} \tag{6}$$

To derive the asymptotic properties of the proposed GPI estimator, we assume the following regularity conditions.

Assumption 7 (Regularity Conditions). Let c_1, c_2, and $q > 2$ be positive constants and δ_N be a sequence of positive constants approaching zero as the sample size N increases. Then, the following conditions hold.

(a) (Primitive conditions)

$$\mathbb{E}[|Y_i|^q]^{1/q} \leq c_1, \quad \sup_{\boldsymbol{r}\in\mathcal{R}} \mathbb{E}[|Y_i - \mu_{T_i}(\boldsymbol{f}(\boldsymbol{r}))|^2 \mid \boldsymbol{R}_i = \boldsymbol{r}] \leq c_1,$$
$$\mathbb{E}[|Y_i - \mu_{T_i}(\boldsymbol{f}(\boldsymbol{R}_i))|^2]^{1/2} \geq c_2.$$

(b) (Outcome model estimation)

$$\mathbb{E}[|\hat{\mu}_{T_i}(\hat{\boldsymbol{f}}(\boldsymbol{R}_i)) - \mu_{T_i}(\boldsymbol{f}(\boldsymbol{R}_i))|^q]^{1/q} \leq c_1,$$
$$\mathbb{E}[|\hat{\mu}_{T_i}(\hat{\boldsymbol{f}}(\boldsymbol{R}_i)) - \mu_{T_i}(\boldsymbol{f}(\boldsymbol{R}_i))|^2]^{1/2} \leq \delta_N N^{-1/4}.$$

(c) (Deconfounder estimation)

$$\mathbb{E}\left[\left\|\hat{\boldsymbol{f}}(\boldsymbol{R}_i) - \boldsymbol{f}(\boldsymbol{R}_i)\right\|^q\right]^{1/q} \leq c_1,$$
$$\mathbb{E}\left[\left\|\hat{\boldsymbol{f}}(\boldsymbol{R}_i) - \boldsymbol{f}(\boldsymbol{R}_i)\right\|^2\right]^{1/2} \leq \delta_N N^{-1/4}$$

(d) (Propensity score estimation) $\pi(\cdot)$ is Lipschitz continuous at the every point of its support, and satisfies:

$$\mathbb{E}[|\hat{\pi}(\boldsymbol{f}(\boldsymbol{R}_i)) - \pi(\boldsymbol{f}(\boldsymbol{R}_i))|^q]^{1/q} \leq c_1,$$
$$\mathbb{E}[|\hat{\pi}(\boldsymbol{f}(\boldsymbol{R}_i)) - \pi(\boldsymbol{f}(\boldsymbol{R}_i))|^2]^{1/2} \leq \delta_N N^{-1/4}.$$

Like the standard application of DML, the required rate for nonparametric estimation of nuisance models is slower than the usual parametric estimation rate of $n^{-1/2}$. Recall that we use neural networks for the joint estimation of outcome model and deconfounder. The required convergence rate of $n^{-1/4}$ is achievable with the standard neural network architecture (Farrell, Liang, and Misra 2021). The propensity score function can be estimated using various nonparametric methods, including the feedforward neural networks with regularization to ensure Lipschitz continuity (Gouk et al. 2021) and the kernel-based methods with nonexpansive kernels (van Waarde and Sepulchre 2022).

Given the above assumptions, the asymptotic normality of the proposed GPI estimator follows immediately from the DML theory.

Theorem 2 (Asymptotic Normality of the GPI Estimator). Under Assumptions 1–7, the estimator $\hat{\tau}$ obtained from the influence function ψ satisfies asymptotic normality:

$$\frac{\sqrt{N}(\hat{\tau} - \tau)}{\sigma} \xrightarrow{d} \mathcal{N}(0, 1)$$

where $\sigma^2 = \mathbb{E}[\psi(\mathcal{D}_i; \tau, \boldsymbol{f}, \mu_1, \mu_0, \pi)^2]$.

The proof is given in Appendix S4.3. In addition, we can consistently estimate the asymptotic variance using the plugin estimator based on (6).

3.4. Practical Implementation Details

We discuss some important practical implementation details. First, to satisfy Assumption 6, researchers must choose a deep generative model that has the option of deterministic decoding. Aside from appropriately setting a hyper-parameter, the assumption also implies that we should not use batches with LLMs, which may induce unknown correlations across observations. The use of LLMs that have memory should also be avoided.

In addition, the effective implementation of the proposed estimation method requires a careful choice of dimension reduction strategy for the internal representation, as well as hyperparameter tuning for TarNet. First, the internal representation $\boldsymbol{R}_i$, which typically corresponds to the last hidden states of a deep generative model, is of high dimension. Specifically, it is a matrix of size equal to the length of texts $\times$ the size of representation for each token, which is equal to 768 for BERT-base, 1024 for BERT-large and T5-3B, and 4096 for Llama3-8B. In theory, we can directly incorporate this matrix in TarNet. In practice, however, given the limited computational resources available to researchers, it is advisable to apply a pooling operation to reduce the dimensionality.

The choice of pooling strategy depends on the architecture of a deep generative model. For example, in BERT, the first special classification token [CLS] contains all semantic information (Devlin et al. 2019). Thus, we could use the hidden states that correspond to this [CLS] token alone. In Bidirectional and Auto-Regressive Transformers (BART), the special token is added at the end, so researchers can extract the hidden states of the last token (Lewis et al. 2020). In contrast, encoder-decoder models

like T5 (Text-to-Text Transfer Transformer, Raffel et al. 2023) do not have such special tokens, and mean pooling is often applied (e.g., Ni et al. 2021).

For decoder-only models, such as Llama, the autoregressive structure forces all tokens to pay attention only to the past tokens. Hence, we can use the hidden states of the last token. This pooling strategy is frequently used as an approximation of the input text representation in the literature (e.g.,Neelakantan et al. 2022; Jiang et al. 2023b; Ma et al. 2024). We show the validity of this approximation in a simulation study (Section 4). Our experience shows that this approximation generally works well. If the confounding features are deemed too complex to be adequately captured by the last token alone, researchers may consider a sensitivity analysis (Lin, Morency, and Ben-Michael 2024).

Second, for TarNet, we must carefully choose hyperparameters such as the size and depth of layers, learning rate, and maximum epoch size. Together, they determine the success of optimization and the quality of the deconfounder estimate. The dimension of the deconfounder should be sufficiently large to capture the confounding information, but not too large to violate the overlap assumption. The learning rate and the epoch size are also crucial, as the performance is highly dependent on the success of the optimization process.

A practical strategy is to try different hyperparameter values and select the one that minimizes the loss. If the loss does not decrease within the first few epochs, the optimization has likely failed, and researchers should try different hyperparameter values. The process can be automated with advanced hyperparameter optimization methods, such as Optuna (Akiba et al. 2019), that search the optimal hyperparameters efficiently by dynamically constructing the search space.

3.5. Diagnostic Tools

The key assumption of the GPI methodology is the separability between the treatment and confounding features (Assumption 5). This assumption concerns the functional relationship between the treatment and confounding features: (i) the treatment feature T_i is not a deterministic function of $\boldsymbol{U}_i$, and (ii) the confounding feature $\boldsymbol{U}_i$ is not a function of T_i. As we show in Lemma 1, condition (i) implies the overlap assumption. Condition (ii) implies that the confounding feature $\boldsymbol{U}_i$ is disentangled from the treatment feature T_i, so that $\mathbb{P}\big(\boldsymbol{U}_i \mid \text{do}(T_i = t)\big) = \mathbb{P}\big(\boldsymbol{U}_i\big)$ for all $t \in \mathcal{T}$. Wang and Jordan (2024) show that disentanglement implies the independence of support, that is, $\text{supp}(T_i) \times \text{supp}(\boldsymbol{U}_i) = \text{supp}(T_i, \boldsymbol{U}_i)$, which is a necessary condition for positivity. Therefore, checking positivity is crucial for diagnosing the potential violation of separability assumption.

We propose two ways to diagnose the positivity assumption. The first way is to plot the distribution of the estimated propensity scores $\hat{\pi}(\hat{\boldsymbol{f}}(\boldsymbol{R}_i)) = \widehat{\mathbb{P}(T_i = 1 \mid \hat{\boldsymbol{f}}(\boldsymbol{R}_i))}$ and assess whether they are bounded away from 0 and 1. If some estimated scores are too close to 0 or 1, one may trim them (Crump et al. 2009), clip them (Dorn 2025), or use overlap weights (Li, Thomas, and Li 2019).

Another way to diagnose the potential violation of positivity is to compute the independence-of-support score (IOSS) proposed by Wang and Jordan (2024). We use IOSS to measure the dependence of support between the deconfounder and the treatment variable. After standardization, IOSS lies in [0, 1] and can be interpreted as the fraction of the standardized range by which the support of the two variables would need to be shifted in order to achieve perfect overlap.

The separability assumption also requires that the treatment feature can be manipulated independently of the confounding features. This implication is challenging to verify statistically. One practical approach is to instruct either an LLM or human to alter only the treatment feature while keeping the remaining aspects of the original text unchanged. The GPI methodology can then be applied to estimate treatment effects. If the manipulation is successful, the resulting treatment effect estimates should closely match those based on the original data. A limitation of this approach is that one needs to collect outcome data for the altered texts. Nevertheless, this provides a direct test of a core component of the separability assumption.

4. Simulation Studies

We conduct simulation studies to evaluate the empirical performance of the GPI estimator and compare it to existing estimators.

4.1. Simulation Setup

We use the candidate profile experiment introduced in Section 2 to make the simulation setup realistic. We first generate candidate biographies using an open source LLM, Llama3–8B, that is fine-tuned for instruction-based prompt generations with 8 billion parameters. We ask the model to create a biography of politicians under a hypothetical name using the system and user level prompts shown in Table S2 of Appendix S1. We create hypothetical names by randomly drawing a surname and a first/middle name with replacement from the original corpus of Fong and Grimmer (2016). The total number of biographies in our sample is 4,000.

We also examine the performance of the GPI methodology with the text reuse approach. To do this, we instruct the same Llama3 model to exactly repeat each generated biography. This allows us to compare our two approaches using the same set of texts.

After generating candidate biographies, we label them based on treatment and confounding features of interest. We consider two scenarios: the first is designed to adhere to the separability assumption (Assumption 5), which is our key assumption, while the second is likely to violate it. For the first scenario, we use the candidate's military background as the treatment variable. A biography is assigned to the treatment group if it contains at least one of the following keywords: "military," "veteran," or "army."

For confounding features, we first consider a combined topic of politics and education, denoted as $h_1(\boldsymbol{X}_i)$ where h_1 is a complex function of the treatment object. We employ a widely-used embedding-based topic model, `BERTopic` (Grootendorst 2022), and then assign $h_1(\boldsymbol{X}_i) = 1$ if a generated biography is classified to a topic whose representative words include "politics," "student," "college," "elected," "university," "political,"

"advocate," and "education". As the second confounding feature, denoted by $h_2(\boldsymbol{X}_i)$, we use the sentiment analysis module available in the Python `TextBlob` package, which yields a continuous sentiment score ranging from -1 to 1. Since the treatment feature is based on a set of specific keywords that are quite different from the confounding features, the separability assumption is likely to be satisfied under this scenario.

For the second scenario, we use two overlapping topics to define the treatment and confounding features so that the separability assumption is likely to be violated. We again use topics obtained from `BERTopic`, and assign $T_i = 1$ if a generated biography is classified to a topic whose representative words include "college," "political," "elected," "politics," "student," "senator," "education," and "legislative". For the confounding concept, we set $h_1(\boldsymbol{X}_i) = 1$ if a biography is assigned to a topic whose representative words include "college," "political," "senator," "politics," "elected," "student," and "career". Thus, although the treatment and confounding concepts are coded based on two different topics, they share many representative words, making it likely for the separability assumption to be violated.

Finally, we use the following linear model to generate the outcome variable,

$$Y_i = \alpha_1 T_i + \alpha_2 T_i h_1(\boldsymbol{X}_i) - \alpha_3 h_1(\boldsymbol{X}_i) - \alpha_4 h_2(\boldsymbol{X}_i) + \epsilon_i,$$

where ϵ_i is the standard normal random variable. Although the model may appear to be a relatively simple function of the treatment and confounding features, these variables themselves are complex functions of text. Thus, in this simulation setting, inferring the ATE is not straightforward because it requires learning an accurate representation of confounding features.

To evaluate the performance of each estimator, we assume that researchers do not have access to the confounding features, $h_1(\boldsymbol{X}_i)$ and $h_2(\boldsymbol{X}_i)$. We wish to infer the ATE, which is given by $\tau = \alpha_1 + \alpha_2 \mathbb{E}[h_1(\boldsymbol{X}_i)]$ where we set $\alpha_1 = \alpha_2 = 10$ throughout the simulations. We consider the three scenarios: (a) weak confounding $\alpha_3 = \alpha_4 = 50$, (b) moderate confounding $\alpha_3 = \alpha_4 = 100$, (c) strong confounding $\alpha_3 = \alpha_4 = 1000$. Given the computational cost, our evaluation is conditional on a single set of generated biographies and its associated treatment and confounding variables. Therefore, $\mathbb{E}[h_1(\boldsymbol{X}_i)]$ is set to its sample mean, and the randomness comes only from the error term of the outcome model ϵ_i.

4.2. Estimators to be Compared

Once Llama3 generates all biographies, we extract an internal representation of each biography from the last hidden layer whose dimension is 4096 $\times$ the number of tokens contained in the biography. For text reuse, we ask the LLM to repeat each generated text. To compute the GPI estimator, we follow the discussion presented in Section 3.4 and use the representation of the last token, yielding a 4096 dimensional vector of internal representation for each candidate biography.

Our neural network architecture uses one linear layer for the deconfounder $\boldsymbol{f}(\boldsymbol{R}_i)$ whose output dimension is 2048. Similarly, we utilize two consecutive linear layers for the potential outcome models whose output dimensions are 500 and 1. We apply ReLU as an activation function between each layer, and also use a dropout rate of 0.15 to prevent overfitting. We select the neural network architecture, including dropout rates, layer depth, and dimension of each layer, based on additional hyperparameter tuning under the weak confounding setting. While we do not conduct hyperparameter tuning for each setting separately, we find that minor changes in these hyperparameters do not significantly affect the value of the loss function.

We use 40% of our data as a training set, 10% as a validation set, and the remaining 50% is used to estimate downstream causal effects. For optimization, we set the batch size to 32 and train the model for 500 epochs, and use early stopping to prevent overfitting. Specifically, we stop training if the estimated loss does not improve for more than 15 epochs. Since the performance of the methodology can depend on the choice of hyperparameters, especially learning rates, we select learning rates using an automated hyperparameter tuning package `Optuna`. We do not include the size of the deconfounder as a hyperparameter to be tuned because tuning it for the BERT-based methods is computationally too expensive. We also do not observe any substantial difference in performance when varying the dimensions of the deconfounder for the GPI methodology. Once the outcome model is fitted, we estimate the propensity score using a random forest classifier with the estimated deconfounder, using the `scikit-learn` library.

We evaluate the performance of the GPI estimator in comparison to the following three estimators. First, as a baseline, we use the difference-in-means estimator, which makes no adjustment for confounding. Second, we implement the two existing approaches that estimate the vectorized representation of texts using the BERT embedding $\boldsymbol{b}(\cdot)$; Pryzant et al. (2021) estimates nuisance functions with TarNet and calculates the causal effect using outcome models, while Gui and Veitch (2023) uses the same outcome models as Pryzant et al. (2021) but applies DML with the estimated propensity score. Importantly, both of these approaches are based on the following loss function,

$$\begin{aligned} &\frac{\lambda}{n}\sum_{i=1}^{n}\left\{Y_i - Q_{T_i}(\boldsymbol{b}(\boldsymbol{X}_i))\right\}^2 \\ &\quad - \frac{\alpha}{n}\sum_{i=1}^{n}\left\{T_i \log g(\boldsymbol{b}(\boldsymbol{X}_i)) + (1-T_i)\log[1-g(\boldsymbol{b}(\boldsymbol{X}_i))]\right\} \\ &\quad + \frac{1}{n}\sum_{i=1}^{n} B(\boldsymbol{b}_{\text{full}}(\boldsymbol{X}_i)) \end{aligned} \tag{7}$$

where $Q_t(\cdot)$ is the outcome model under $T_i = t$, $g(\cdot)$ is the treatment prediction, $B(\cdot)$ is the original BERT masked language loss for estimating vector representations of texts from the embedding $\boldsymbol{b}(\cdot)$, and $\alpha, \lambda \in \mathbb{R}$ are the hyperparameters (Veitch, Sridhar, and Blei 2020). Only the vector representation of the first token $\boldsymbol{b}(\cdot)$ is used for prediction of outcome and treatment because it is the special token called the [CLS] token that contains the semantic information. On the other hand, for the masked language loss, the entire embedding $\boldsymbol{b}_{\text{full}}(\cdot)$ is used. The loss function given in (7) differs from our loss function (4) in that in addition to the outcome model, it optimizes the representation learning from texts and the prediction of treatment.

After estimating nuisance functions, Gui and Veitch (2023) proposes to estimate the propensity score model by treating $Q_1(\boldsymbol{b}(\boldsymbol{X}_i))$ and $Q_0(\boldsymbol{b}(\boldsymbol{X}_i))$ as covariates, that is, $\Pr(T_i = 1 \mid$

$\boldsymbol{X}_i) = g(Q_1(\boldsymbol{b}(\boldsymbol{X}_i)), Q_0(\boldsymbol{b}(\boldsymbol{X}_i)))$. For propensity score estimation, we use a Gaussian Process, as recommended by Gui and Veitch (2023). We truncate the extreme values of the estimated propensity scores at 0.01 and 0.99 for DML with BERT, although such a truncation is not necessary for the GPI methodology. For DML with BERT, this truncation occurs even when the separability assumption is satisfied (5% of time for weak and moderate confounding, and 45% for strong confounding). Without truncation, the bias and RMSE are not computable for these methods. Finally, we follow the default hyperparameter used in the authors' original code and set $\alpha = \lambda = 1$ while tuning the other hyperparameters regarding the learning rate in the same way as done for the GPI estimator.

4.3. Simulation Results

Figure 3 graphically displays the results of our simulation studies while Appendix S5 reports the corresponding numerical results. The results are based on 200 Monte Carlo trials. We choose this relatively small number of trials because, unlike the GPI estimator, the BERT-based estimators are computationally intensive. Appendix S5 also presents the simulation results based on 1000 Monte Carlo trials for the proposed estimator alone. As expected, the results are qualitatively similar to those presented in this section.

We find that when the separability assumption holds (top row), the GPI estimators (red line for new texts and dark yellow for text reuse) exhibit a smaller bias and RMSE compared to all other estimators, with the 95% confidence interval coverage closely matching the nominal rate. The performance differences are particularly striking in the strong confounding setting, where DML with BERT (purple) performs poorly, unlike the weak and moderate confounding scenarios. Additionally, outcome model with BERT (blue) has severe undercoverage across all simulation scenarios, whereas the confidence interval of DML with BERT breaks down only under the strong confounding scenario. Lastly, the GPI estimators are more than ten times as computationally efficient as BERT-based estimators, when measured in terms of the average runtime.

In contrast, when the separability assumption is violated (bottom row), all estimators, including ours, perform poorly. In this setting, both bias and RMSE grow as the strength of confounding increases. As a result, the coverage of confidence intervals no longer approximates the nominal coverage rate even for the GPI estimator.

We further examine the difference in performance between the GPI estimator and the DML with BERT. Figure 4 shows the distribution of the estimated propensity scores without truncation (recall that truncation is not needed for the proposed methodology). For the GPI methodology, when the separability assumption is met (top panel), the distribution of estimated propensity scores is relatively symmetric regardless of the confounding strength and is similar between the treatment and control groups. Indeed, most observations have estimated propensity scores far from zero. This explains why the GPI estimator performs well in this scenario.

On the other hand, when the separability assumption is violated (bottom left and middle panels), the estimated propensity scores for the control group are heavily skewed to the right and close to zero, indicating that the overlap assumption is violated. This implies that the estimated propensity scores can help diagnose the potential violation of the separability assumption for our proposed method, which leads to complete entanglement between treatment and confounding features and causes the deconfounder to perfectly predict the treatment assignment. In contrast, DML with BERT yields a wide range of estimated propensity scores under the strong confounding settings (top right panel), in which the estimator performs poorly. This pattern is observed even when the separability assumption is violated (bottom right panel). The finding implies that, unlike the proposed estimator, extreme values of estimated propensity scores may not serve as a reliable diagnostic for DML with BERT.

Finally, we examine the performance of the GPI estimator for new texts and text reuse as we vary the sample size from

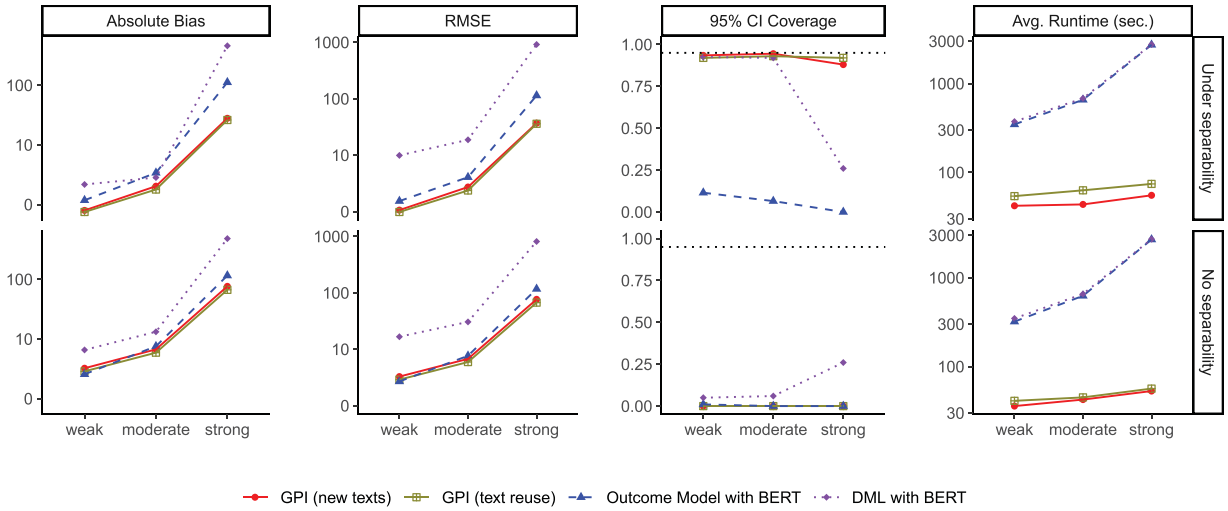

Figure 3. Performance of five estimators under different confounding scenarios (weak, moderate, and strong) under separability (top row) and no separability (bottom row). A red solid line represents the proposed GPI methodology for the new texts, while a dark yellow solid line represents that for the regenerated texts (text reuse). For comparison, we also include outcome model with BERT (blue) and DML with BERT (purple). The black horizontal dotted line in the 95% Confidence Interval (CI) Coverage panel represents the nominal coverage of 95%.

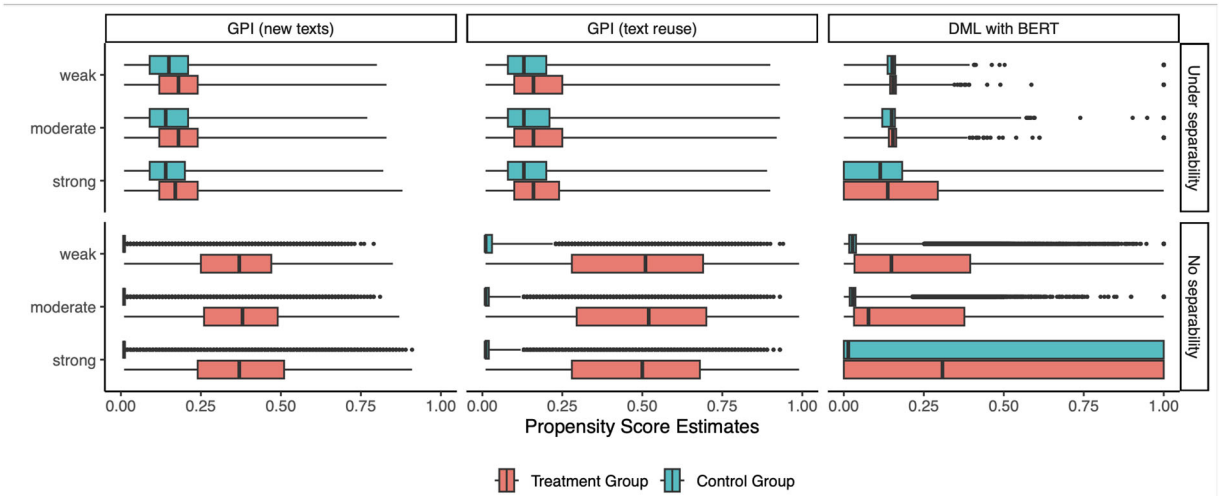

Figure 4. Distribution of the estimated propensity scores for the treatment (red) and control (blue) groups. We present the results for the proposed GPI estimator (new texts in the left panels, and text reuse in the middle panels), and DML with BERT (right) under the separability assumption (top) and no separability (bottom). Under each scenario, we present the results based on three different strengths of confounding; weak, moderate, and strong. For the proposed GPI methodology, the estimated propensity scores are distributed similarly across different confounding scenarios under the separability assumption. In contrast, for DML with BERT, the distribution is heavily skewed right under the strong confounding scenario. When the separability assumption is violated, both methods have extremely small estimated propensity scores for the control group.

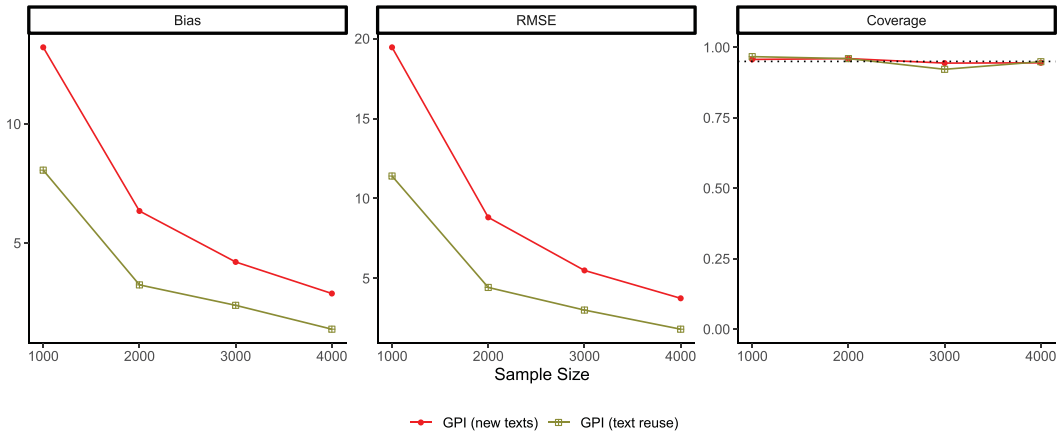

Figure 5. Performance of the GPI estimator on the created texts (red) and the reused texts (blue) for different sample size based on 1000 Monte Carlo trials. The data generating process is no interaction setting ($\alpha_1 = 10, \alpha_2 = 0, \alpha_3 = \alpha_4 = 100$). The black dotted line in the coverage panel represents the nominal coverage of 95%.

1000 to 4000 under the assumption of separability. For this simulation, we use the moderate confounding setting without the interaction term ($\alpha_1 = 10, \alpha_2 = 0, \alpha_3 = \alpha_4 = 100$) so that the true ATE stays identical across sample sizes. We conduct 1000 Monte Carlo trials as we focus only on the GPI estimator, which is computationally efficient. Figure 5 presents bias, RMSE, and coverage for each sample size. As expected, both bias and RMSE become smaller as the sample size increases. We also find that the empirical coverage of the 95% confidence intervals remains close to the nominal coverage even when the sample size is as small as 1000. Interestingly, while bias is similar between new texts and text reuse, RMSE (hence. variance) is substantially smaller for text reuse. In sum, the proposed GPI methodology performs well in different sample sizes, provided that the separability assumption is satisfied.

5. Empirical Analysis

Finally, we apply the proposed GPI methodology to human responses (rather than simulated data) from the candidate profile experiment of Fong and Grimmer (2016) using pre-defined treatment coding. In the original experiment, the authors scraped 1246 biographies of congressional candidates from Wikipedia, randomly assigned up to four biographies to 1886 survey participants, and asked respondents how they felt about each candidate using the standard feeling thermometer ranging from 0 to 100, where a higher value indicates a more favorable evaluation. The total number of observations we analyze is 5291 after dropping some observations with empty texts.

While Fong and Grimmer (2016) use a topic model to discover treatments from the data, we use the pre-defined treatment

Table 1. The estimated average treatment effect (ATE) for the candidate profile experiment.

Methods	ATE estimates	95% Confidence Interval	IOSS	Runtime (sec.)
GPI (reuse)	4.852	[1.902, 8.580]	0.10	62.3
Outcome model with BERT	−4.277	[−4.312, −4.241]	0.41	5914.0
DML with BERT	45.708	[33.730, 57.686]		5986.2

coding. As in simulation studies, we use a candidate's military background as the treatment variable by assigning a biography to the treatment group if it contains at least one of the following keywords: "military," "veteran," and "army." According to this coding rule, only seven percent of the biographies (362 out of 5291) are in the treatment group.

For causal effect estimation, we take the text reuse approach by regenerating each biography using Llama3-8B. We then apply the proposed GPI methodology following the procedure described in Section 3 with two-fold cross fitting. For comparison, as in our simulation studies, we also apply the two existing BERT-based methods—Pryzant et al. (2021) (Outcome model with BERT) and Gui and Veitch (2023) (DML with BERT).

Table 1 presents the results. The analysis based on the proposed GPI methodology implies that military experience has a positive effect and is statistically significant. This echoes with the finding of Fong and Grimmer (2016) that the topic corresponding to military experience has a statistically significant positive association with a higher feeling thermometer score. In contrast, outcome model with BERT yields a negative and statistically significant estimate, while DML with BERT produces a positive estimate that is unreasonably large given the scale is between 0 and 100.

To diagnose the potential violation of positivity, we also compute IOSS for both the GPI and BERT-based methods. While GPI yields IOSS of 0.10, the BERT-based methods yield a much larger IOSS of 0.41. This suggests that the separability assumption is much more likely to be voilated for the BERT-based methods than GPI. Indeed, for the BERT-based methods, all 5291 observations have estimated propensity scores outside of the range of $[0.01, 0.99]$.

Lastly, as observed in our simulation studies, the runtime for GPI is around 100 times shorter than that of the two BERT-based estimators.

6. Concluding Remarks

In this article, we demonstrate that the use of GenAI can significantly enhance the validity of causal inference with unstructured treatments, such as texts and images. We leverage GenAI to both efficiently produce a variety of treatments and precisely control confounding bias. By using the true vector representation of generated texts, we avoid estimating such representation as done in the previous methods, leading to more efficient and robust causal effect estimation.

We formalize the conditions required for nonparametric identification, showing that the separability of treatment and confounding features plays an essential role. We also develop an estimation method based on a neural network architecture that mitigates the risk of positivity violation, a common problem of existing methods. Lastly, we extend the proposed GPI methods to the settings of perceived treatments, using an instrumental variables approach. Our simulation study shows that the GPI estimator outperforms existing methods.

Although we have focused on texts as treatments, our GPI approach can be extended to other types of unstructured data, such as images and videos. For images, we expect the proposed GPI methodology to be directly applicable so long as the dimensionality of internal representation is relatively low. For videos, both treatment and confounding features are likely to exhibit complex relationships due to the combination of audio and images with the additional temporal dimension. Thus, the GPI methodology described here is not readily applicable. Future work should consider how to leverage the internal representation of videos obtained from GenAI.

Supplementary Materials

Supplementary materials contain the details of assumptions, the extension of our methodology to perceived treatments, additional simulation results, and mathematical proofs.

Acknowledgments

An open-source Python package, "GPI: Generative-AI Powered Inference," that implements the proposed methodology is available at *https://gpi-pack.github.io/*. The replication code and data are available as Nakamura and Imai (2026). We thank Christian Fong and Justin Grimmer for kindly sharing their dataset used in Candidate Profile Experiment. We also acknowledge useful comments from Naoki Egami, Max Kasy, Sooahn Shin, and an anonymous reviewer of Harvard's Alexander and Diviya Magaro Peer Pre-Review Program.

Disclosure Statement

The authors report there are no competing interests to declare.

Code and Data Availability

The code and data necessary to replicate all results reported in this article are publicly available through the Harvard Dataverse at *https://doi.org/10.7910/DVN/2DZQ6T* (Nakamura and Imai 2026). The Python package implementing the proposed GenAI-Powered Inference (GPI) methodology is available at *https://gpi-pack.github.io/*.

ORCID

Kosuke Imai https://orcid.org/0000-0002-2748-1022
Kentaro Nakamura https://orcid.org/0009-0002-1573-9150

References

Ahrens, M., Ashwin, J., Calliess, J.-P., and Nguyen, V. (2021), "Bayesian Topic Regression for Causal Inference," arXiv:2109.05317 [cs, stat]. [2]

Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. (2019), "Optuna: A Next-Generation Hyperparameter Optimization Framework," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, pp. 2623–2631, New York, NY, USA. Association for Computing Machinery. [8]

Bisbee, J., Clinton, J. D., Dorff, C., Kenkel, B., and Larson, J. M. (2024), "Synthetic Replacements for Human Survey Data? The Perils of Large Language Models," *Political Analysis*, 32, 401–416. [1]

Campbell, R., and Cowley, P. (2014), "What Voters Want: Reactions to Candidate Characteristics in a Survey Experiment," *Political Studies*, 62, 745–765. [3]

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018), "Double/Debiased Machine Learning for Treatment and Structural Parameters," *The Econometrics Journal*, 21, C1–C68. [2,6]

Crump, R. K., Hotz, V. J., Imbens, G. W., and Mitnik, O. A. (2009), "Dealing with Limited Overlap in Estimation of Average Treatment Effects," *Biometrika*, 96, 187–199. [8]

Daoud, A., Jerzak, C. T., and Johansson, R. (2022), "Conceptualizing Treatment Leakage in Text-based Causal Inference," arXiv:2205.00465 [cs]. [2,4]

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019), "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv:1810.04805 [cs]. [7]

Dorn, J. (2025), "How Much Weak Overlap Can Doubly Robust t-statistics Handle?" arXiv preprint arXiv:2504.13273. [8]

Egami, N., Fong, C. J., Grimmer, J., Roberts, M. E., and Stewart, B. M. (2022), "How to Make Causal Inferences Using Texts," *Science Advances*, 8, eabg2652. [4]

Farrell, M. H., Liang, T., and Misra, S. (2021), "Deep Neural Networks for Estimation and Inference," *Econometrica*, 89, 181–213. [7]

Feder, A., Keith, K. A., Manzoor, E., Pryzant, R., Sridhar, D., Wood-Doughty, Z., Eisenstein, J., Grimmer, J., Reichart, R., Roberts, M. E., Stewart, B. M., Veitch, V., and Yang, D. (2022), "Causal Inference in Natural Language Processing: Estimation, Prediction, Interpretation and Beyond," arXiv:2109.00725 [cs]. [2]

Fong, C., and Grimmer, J. (2016), "Discovery of Treatments from Text Corpora," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1600–1609. [2,3,8,11,12]

Fong, C., and Grimmer, J. (2023), "Causal Inference with Latent Treatments," *American Journal of Political Science*, 67, 374–389. [2,4]

Gouk, H., Frank, E., Pfahringer, B., and Cree, M. J. (2021), "Regularisation of Neural Networks by Enforcing Lipschitz Continuity," *Machine Learning*, 110, 393–416. [7]

Grootendorst, M. (2022), "BERTopic: Neural Topic Modeling with a Class-based TF-IDF Procedure," arXiv:2203.05794 [cs]. [8]

Gui, L., and Veitch, V. (2023), "Causal Estimation for Text Data with (Apparent) Overlap Violations," arXiv:2210.00079 [cs, stat]. [2,6,9,10,12]

Hu, Z., and Li, L. E. (2021), "A Causal Lens for Controllable Text Generation," in *Advances in Neural Information Processing Systems* (Vol. 34), pp. 24941–24955. [3]

Imai, K., and Jiang, Z. (2019), "Comment: The Challenges of Multiple Causes," *Journal of the American Statistical Association*, 114, 1605–1610. [3]

Imai, K., and Nakamura, K. (2025). GenAI-Powered inference. *arXiv preprint*, 2507.03897. [2]

Imai, K., and van Dyk, D. A. (2004), "Causal Inference With General Treatment Regimes: Generalizing the Propensity Score," *Journal of the American Statistical Association*, 99, 854–866. [4]

Imbens, G. W., and Angrist, J. D. (1994), "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467–475. [2]

Jerzak, C. T., Johansson, F., and Daoud, A. (2023a), "Integrating Earth Observation Data into Causal Inference: Challenges and Opportunities," arXiv:2301.12985 [cs, stat]. [2]

Jerzak, C. T., Johansson, F. D., and Daoud, A. (2023b), "Image-based Treatment Effect Heterogeneity," in *Proceedings of the Second Conference on Causal Learning and Reasoning*, pp. 531–552, PMLR. [2]

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. d. l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. (2023a), "Mistral 7B," arXiv:2310.06825. [4]

Jiang, T., Huang, S., Luan, Z., Wang, D., and Zhuang, F. (2023b), "Scaling Sentence Embeddings with Large Language Models," arXiv preprint arXiv:2307.16645. [8]

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., Stadler, M., Weller, J., Kuhn, J., and Kasneci, G. (2023), "Chatgpt For Good? On Opportunities and Challenges of Large Language Models for Education," *Learning and Individual Differences*, 103, 102274. [1]

Keith, K. A., Jensen, D., and O'Connor, B. (2020), "Text and Causal Inference: A Review of Using Text to Remove Confounding from Causal Estimates," arXiv:2005.00649 [cs]. [2]

Kirkland, P. A., and Coppock, A. (2018), "Candidate Choice Without Party Labels," *Political Behavior*, 40, 571–591. [3]

Klaassen, S., Teichert-Kluge, J., Bach, P., Chernozhukov, V., Spindler, M., and Vijaykumar, S. (2024), "DoubleMLDeep: Estimation of Causal Effects with Multimodal Data," arXiv:2402.01785 [cs, econ, stat]. [2]

Kshetri, N., Dwivedi, Y. K., Davenport, T. H., and Panteli, N. (2024), "Generative Artificial Intelligence in Marketing: Applications, Opportunities, Challenges, and Research Agenda," *International Journal of Information Management*, 75, 102716. [1]

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2020), "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, eds. D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, pp. 7871–7880. [7]

Li, F., Thomas, L. E., and Li, F. (2019), "Addressing Extreme Propensity Scores via the Overlap Weights," *American Journal of Epidemiology*, 188, 250–257. [8]

Lin, V., Morency, L.-P., and Ben-Michael, E. (2024), "Isolated Causal Effects of Natural Language." [8]

Ma, X., Wang, L., Yang, N., Wei, F., and Lin, J. (2024), "Fine-Tuning Ilama for Multi-Stage Text Retrieval," in *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2421–2425. [8]

Mozer, R., Kaufman, A. R., Celi, L. A., and Miratrix, L. (2024), "Leveraging Text Data for Causal Inference Using Electronic Health Records," arXiv:2307.03687 [cs, stat]. [2]

Mozer, R., Miratrix, L., Kaufman, A. R., and Anastasopoulos, L. J. (2020), "Matching with Text Data: An Experimental Evaluation of Methods for Matching Documents and of Measuring Match Quality," *Political Analysis*, 28, 445–468. [2]

Nakamura, K., and Imai, K. (2026), "Replication Data for: Causal Inference with Generative Artificial Intelligence: Application to Texts as Treatments," DOI:10.7910/DVN/2DZQ6T [12]

Neelakantan, A., Xu, T., Puri, R., Radford, A., Han, J. M., Tworek, J., Yuan, Q., Tezak, N., Kim, J. W., Hallacy, C., et al. (2022), "Text and Code Embeddings by Contrastive Pre-training," arXiv preprint arXiv:2201.10005. [8]

Ni, J., Ábrego, G. H., Constant, N., Ma, J., Hall, K. B., Cer, D., and Yang, Y. (2021), "Sentence-T5: Scalable Sentence Encoders from Pre-trained Text-to-Text Models," arXiv:2108.08877 [cs]. [8]

Pedersen, R. T., Dahlgaard, J. O., and Citi, M. (2019), "Voter Reactions to Candidate Background Characteristics Depend on Candidate Policy Positions," *Electoral Studies*, 61, 102066. [3]

Pryzant, R., Card, D., Jurafsky, D., Veitch, V., and Sridhar, D. (2021), "Causal Effects of Linguistic Properties," arXiv:2010.12919 [cs]. [2,4,6,9,12]

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2023), "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," arXiv:1910.10683 [cs, stat]. [8]

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., and Sutskever, I. (2021), "Zero-Shot Text-to-Image Generation," in *Proceedings of the 38th International Conference on Machine Learning*, pp. 8821–8831, PMLR. [2,4]

Roberts, M. E., Stewart, B. M., and Nielsen, R. A. (2020), "Adjusting for Confounding with Text Matching," *American Journal of Political Science*, 64, 887–903. [2]

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022), "High-Resolution Image Synthesis With Latent Diffusion Models," pp. 10684–10695. [2,4,5]

Schölkopf, B., Locatello, F., Bauer, S., Ke, N., Kalchbrenner, N., Goyal, A., and Bengio, Y. (2021), "Toward Causal Representation Learning," *Proceedings of the IEEE*, 109, 612–634. [3]

Shalit, U., Johansson, F. D., and Sontag, D. (2017), "Estimating Individual Treatment Effect: Generalization Bounds and Algorithms," arXiv:1606.03976 [cs, stat]. [2,6]

Shi, C., Blei, D., and Veitch, V. (2019), "Adapting Neural Networks for the Estimation of Treatment Effects," in *Advances in Neural Information Processing Systems* (Vol. 32). [2,6]

Su, Y., Lan, T., Wang, Y., Yogatama, D., Kong, L., and Collier, N. (2022), "A Contrastive Framework for Neural Text Generation," arXiv:2202.06417 [cs]. [5]

Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., and Ting, D. S. W. (2023), "Large Language Models in Medicine," *Nature Medicine*, 29, 1930–1940. [1]

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. (2023), "LLaMA: Open and Efficient Foundation Language Models," arXiv:2302.13971 [cs]. [4]

van Waarde, H., and Sepulchre, R. (2022), "Training Lipschitz Continuous Operators Using Reproducing Kernels," in *Learning for Dynamics and Control Conference*, pp. 221–233, PMLR. [7]

Veitch, V., Sridhar, D., and Blei, D. M. (2020), "Adapting Text Embeddings for Causal Inference," arXiv:1905.12741 [cs, stat]. [2,6,9]

Wang, Y., and Blei, D. M. (2019), "The Blessings of Multiple Causes," *Journal of the American Statistical Association*, 114, 1574–1596. [3]

Wang, Y., and Jordan, M. I. (2024), "Desiderata for Representation Learning: A Causal Perspective," *Journal of Machine Learning Research*, 25, 1–65. [1,2,8]

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V., Mihaylov, T., Ott, M., Shleifer, S., Shuster, K., Simig, D., Koura, P. S., Sridhar, A., Wang, T., and Zettlemoyer, L. (2022), "OPT: Open Pre-trained Transformer Language Models," arXiv:2205.01068 [cs]. [4]