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

Kosuke Imai† Kentaro Nakamura‡

June 11, 2026

*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 at `https://github.com/k-nakam/gpi_replication`. We thank Christian Fong and Justin Grimmer for kindly sharing their dataset used in Candidate Profile Experiment and Naoki Egami for helpful comments. We also acknowledge useful comments from Naoki Egami, Max Kasy, and an anonymous reviewer of Harvard's Alexander and Diviya Magaro Peer Pre-Review Program.

†Professor, Department of Government and Department of Statistics, Harvard University, Cambridge, MA 02138. Phone: 617–384–6778, Email: Imai@Harvard.Edu, URL: https://imai.fas.harvard.edu

‡Ph.D. student, John F. Kennedy School of Government, Harvard University. Email: knakamura@g.harvard.edu

Supplementary Appendix

S1 Examples of Candidate Biography and Prompt

Candidate biography with military background Anthony Higgins was born in Red Lion Hundred in New Castle County, Delaware. He attended Newark Academy and Delaware College, and graduated from Yale College in 1861, where he was a member of Skull and Bones. After studying law at the Harvard Law School, he was admitted to the bar in 1864 and began practice in Wilmington, Delaware. He also served for a time in the United States Army in 1864.
Candidate biography without military background Benjamin Tappan was born in Northampton, Massachusetts, the second child and oldest son of Benjamin Tappan and Sarah (Homes) Tappan, who was a grandniece of Benjamin Franklin. Two of his younger brothers were abolitionists Arthur Tappan and Lewis Tappan. He attended the public schools in Northampton and traveled to the West Indies in his youth. He apprenticed as a printer and engraver, also studying painting with Gilbert Stuart. He read law to be admitted to the bar in Hartford, Connecticut, in 1799. Later that year, he moved to the Connecticut Western Reserve and founded what is now Ravenna, Ohio, laying out the original village in 1808. He married, March 20, 1801, Nancy Wright, sister of John C. Wright (congressman), afterwards a United States House of Representatives from Ohio. They had one son, Benjamin, born in 1812.

Table S1: Two examples of candidate biographies with and without military background from the experiment conducted by Fong and Grimmer (2016)

Prompts	**Generated candidate biography**
(System): You are a text generator who always produces a biography of the instructed person. **(User)**: Create a biography of an American politician named Nathaniel C. Gilchrist	Nathaniel C. Gilchrist was an American politician who served as a member of the United States House of Representatives from Maryland's 4th congressional district from 1895 to 1899. Born on November 24, 1847, in Baltimore, Maryland, Gilchrist was the son of a prominent family of merchants and shipbuilders. He was educated at the Baltimore City College and later attended the University of Maryland, where he studied law. After graduating, he was admitted to the bar and began practicing law in Baltimore.
(System): You are a text generator who just repeats the input text. **(User)**: Higgins was born in Red Lion Hundred in New Castle County, Delaware. He attended Newark Academy and Delaware College, and graduated from Yale College in 1861, where he was a member of Skull and Bones. After studying law at the Harvard Law School, he was admitted to the bar in 1864 and began practice in Wilmington, Delaware. He also served for a time in the United States Army in 1864.	Higgins was born in Red Lion Hundred in New Castle County, Delaware. He attended Newark Academy and Delaware College, and graduated from Yale College in 1861, where he was a member of Skull and Bones. After studying law at the Harvard Law School, he was admitted to the bar in 1864 and began practice in Wilmington, Delaware. He also served for a time in the United States Army in 1864.

Table S2: Two examples of generated candidate biographies with Llama 3. The system-level input (**System**) defines the type of tasks to be performed, whereas the user-level input (**User**) defines a specific task to be performed.

S2 Illustrative Examples of the Separability Assumption

In this Appendix section, we present illustrative example corpora of the separability assumption (Assumption 5). We present two examples, one in which the assumption is satisfied and the other where it is violated.

S2.1 Example corpus where the separability assumption is satisfied

Table S3 presents an illustrative example corpus, in which the separability assumption is satisfied. In this toy example, the treatment feature $T_i \in \{0,1\}$ equals 1 if a text contains a male pronoun (i.e., "he," "him," "his") and 0 otherwise. The confounding feature $U_i \in \{0,1\}$ equals 1 if the sentence contains a "lawyer" or "doctor" and 0 otherwise. In this small corpus, T_i and U_i are associated: $\Pr(T_i = 1 \mid U_i = 1) = \frac{2}{3}$ and $\Pr(T_i = 1 \mid U_i = 0) = \frac{1}{3}$. The separability assumption nevertheless holds: it is possible to change the value of T_i by swapping male and female pronouns while leaving U_i unchanged.

Original Text	T_i	U_i	**Text with** $T_i = 1$	**Text with** $T_i = 0$
He is a lawyer.	1	1	He is a lawyer.	<u>She</u> is a lawyer.
She is a nurse.	0	0	<u>He</u> is a nurse.	She is a nurse.
He writes a book.	1	0	He writes a book.	<u>She</u> writes a book.
He is a doctor.	1	1	He is a doctor.	<u>She</u> is a doctor.
She gets married.	0	0	<u>He</u> gets married.	She gets married.
She is a doctor.	0	1	<u>He</u> is a doctor.	She is a doctor.

Table S3: An example corpus where the separability assumption is satisfied. An underline represents a change made to the text when the treatment feature is altered.

S2.2 Example corpus where the separability assumption is violated

Table S4 presents an illustrative corpus, in which the separability assumption is violated. Here, $T_i = 1$ if the sentence contains the honorific "Mr." and $U_i = 1$ if it contains a male pronoun ("he," "him," "his"). In this setting, the editing operation that changes T_i from 0 to 1 (or vice versa) requires changing pronouns to maintain grammatical coherence. Consequently, U_i can change when T_i is altered, violating separability. For instance, in the first row, switching T_i from 0 to 1 changes U_i from 0 to 1.

Original Text	T_i	U_i	**Text with** $T_i = 1$	**Text with** $T_i = 0$
Mrs. Park loves her children.	0	0	<u>Mr.</u> Park loves <u>his</u> children.	Mrs. Park loves her children.
Mr. Lee met with the team.	1	0	Mr. Lee met with the team.	<u>Mrs.</u> Lee met with the team.
Mr. Zhou said he would pay.	1	1	Mr. Zhou said he would pay.	<u>Mrs.</u> Zhou said <u>she</u> would pay.
Mrs. Li met him.	0	1	<u>Mr.</u> Li met him.	Mrs. Li met him.

Table S4: A example corpus where the separability assumption is violated. An underline represents a change made to the text when the treatment feature is altered.

S3 Instrumental Variable Approach to the Perceived Treatment Feature

The methodology described in the previous section enables the estimation of the average causal effect of the treatment feature, which is assumed to be a deterministic function of the treatment object. In some cases, however, researchers may be interested in the causal effect of *perceived* treatment feature, which may not necessarily coincide with the treatment feature itself. In addition, the perception of the same treatment feature may vary across respondents. For example, in our application, respondents may disagree as to what constitutes a military background.

In this section, we extend the proposed methodology to the setting, in which the treatment feature is used as an instrumental variable for the perceived treatment feature. As before, we describe the required assumptions, establish nonparametric identification, and propose estimation and inference strategies.

S3.1 Assumptions and causal quantity of interest

We consider the same setting as in Section 3 except that we observe the perceived treatment feature $\widetilde{T}_i \in \{0, 1\}$, which may not equal the treatment feature itself, i.e., $\widetilde{T}_i \neq T_i$ for some i. We assume that a respondent's perceived treatment feature is a function of the treatment and confounding features of the assigned treatment object. We assume that both perceived treatment $\widetilde{T}_i$ and original treatment T_i are observed.

ASSUMPTION 1 (PERCEIVED TREATMENT FEATURE) *The perceived treatment feature $\widetilde{T}_i \in \{0, 1\}$ is a function of treatment and confounding features, i.e.,*

$$\widetilde{T}_i = \widetilde{T}_i(T_i, \boldsymbol{U}_i),$$

where $\widetilde{T}_i(t, \boldsymbol{u})$ is the potential value of the perceived treatment feature when the treatment variable T_i is equal to $t \in \{0, 1\}$ and the confounding variables $\boldsymbol{U}_i$ equal $\boldsymbol{u} \in \mathcal{U}$.

Importantly, under Assumption 1, different respondents may perceive the same treatment feature differently. In addition, it is also possible for confounding features to affect the perceived treatment feature. In practice, researchers may measure the perceived treatment feature by asking respondents directly. However, doing so may lead to the so-called *priming bias*, in which the act of asking this question itself draws a respondent's attention to the treatment feature and confounds the causal effect of interest. To avoid this bias, researchers may measure the perceived treatment feature after the outcome variable is realized. Addressing this methodological issue is beyond the scope of this paper, but interested readers should consult a recent literature on the topic (see e.g., Montgomery et al., 2018; Aronow et al., 2019; Klar et al., 2020; Blackwell et al., 2025).

To identify the causal effect of the perceived treatment feature, we use the treatment feature as an instrumental variable. To do this, we define the potential outcome as a function of the perceived treatment feature, and the treatment and confounding features. Formally, we replace Assumption 5 with the following assumption while maintaining the same separability between the treatment feature and the confounding features.

ASSUMPTION 2 (SEPARABILITY WITH THE PERCEIVED TREATMENT FEATURE) *The potential outcome is a function of the perceived treatment $\tilde{T}_i$, the treatment features of interest T_i, and the* *confounding features $\boldsymbol{U}_i$. That is, for any given $\boldsymbol{x} \in \mathcal{X}$ and all i, we have:*

$$Y_i(\boldsymbol{x}) = Y_i(\widetilde{T}_i(g_T(\boldsymbol{x}), \boldsymbol{g}_{\boldsymbol{U}}(\boldsymbol{x})), g_T(\boldsymbol{x}), \boldsymbol{g}_{\boldsymbol{U}}(\boldsymbol{x}))$$

where $\widetilde{T}_i(g_T(\boldsymbol{x}), \boldsymbol{g}_{\boldsymbol{U}}(\boldsymbol{x})) \in \{0,1\}$ is the perceived treatment feature, $g_T(\boldsymbol{x}) \in \{0,1\}$, and $\boldsymbol{g}_{\boldsymbol{U}}(\boldsymbol{x}) \in \mathcal{U}$. In addition, g_T and $\boldsymbol{g}_{\boldsymbol{U}}$ are separable in the same sense as Assumption 5.

Lastly, we adopt the standard instrumental variable assumptions in current settings (Imbens and Angrist, 1994). First, we assume monotonicity; the existence of treatment feature makes it no less likely for a respondent to perceive it as such. Second, we assume an exclusion restriction; the treatment feature only affects the outcome through the perceived treatment feature. Both assumptions are made while keeping the confounding features constant. We formally state these assumptions here.

ASSUMPTION 3 (VALIDITY OF THE INSTRUMENTAL VARIABLE) *We make the following instrumental variable assumptions:*

(a) (Monotonicity) For any $\boldsymbol{u} \in \mathcal{U}$, we have:

$$\widetilde{T}_i(1, \boldsymbol{u}) \geq \widetilde{T}_i(0, \boldsymbol{u}) \quad \textit{and} \quad \mathbb{P}(\widetilde{T}_i(1, \boldsymbol{u}) = 1) > \mathbb{P}(\widetilde{T}_i(0, \boldsymbol{u}) = 1).$$

(b) (Exclusion Restriction) For any $\tilde{t} \in \{0,1\}$, $\boldsymbol{u} \in \mathcal{U}$, and $i = 1, 2, \ldots, n$, we have:

$$Y_i(\tilde{t}, 1, \boldsymbol{u}) \;=\; Y_i(\tilde{t}, 0, \boldsymbol{u}) \;=\; Y_i(\tilde{t}, \boldsymbol{u}).$$

In many practical settings, the monotonicity assumption is reasonable. In our application, for example, if there is no military background in a candidate biography, a respondent should not notice the presence of this treatment feature. Exclusion restriction, however, may not be credible in some cases because it is possible for a respondent to be influenced by the treatment feature without noticing it.

Under this setup, we are interested in estimating the local average treatment effect (LATE) of the perceived treatment feature among the respondents who notice the presence of the treatment feature only when the treatment object actually contains such a feature. We define this LATE as follows:

$$\beta \;:=\; \mathbb{E}[Y_i(1, \boldsymbol{U}_i) - Y_i(0, \boldsymbol{U}_i) \mid \widetilde{T}_i(1, \boldsymbol{U}_i) = 1, \widetilde{T}_i(0, \boldsymbol{U}_i) = 0], \tag{S1}$$

where the first input of the potential outcome is the perceived treatment feature instead of the treatment feature, i.e., $Y_i(\widetilde{T}_i = \tilde{t}, \boldsymbol{U}_i = \boldsymbol{u})$.

S3.2 Nonparametric identification

We extend our nonparametric identification result obtained in Section 3.2 to the instrumental variable setting. Figure S1 summarizes the assumed data generation process with the perceived treatment feature. The absence of direct arrow from the treatment feature T into Y encodes exclusion restriction (Assumption 3(b)). In addition, we allow for the possible existence of unobserved confounders between the perceived treatment feature and the outcome, indicated by the dotted line in the DAG.

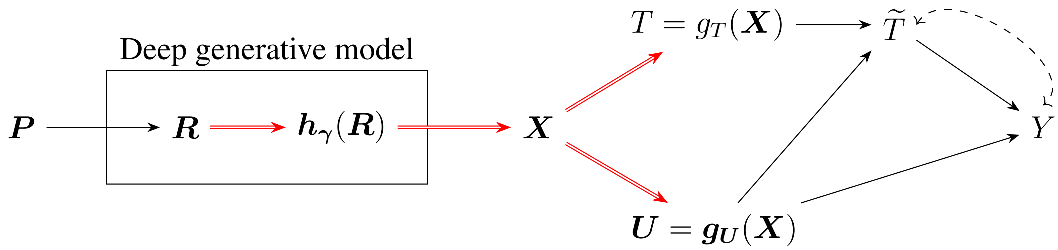

Figure S1: Directed Acyclic Graph (DAG) of the Assumed Data Generating Process with the Perceived Treatment Feature. This DAG is identical to that of Figure 1 except that the perceived treatment feature $\widetilde{T}$ is added. The perceived treatment feature may be affected by the treatment feature T and/or the confounding features $\boldsymbol{U}$. There may also be unobserved confounding variables that affect both the perceived treatment feature and the outcome Y. An arrow with red double lines represents a deterministic causal relation while an arrow with a single line indicates a possibly stochastic relationship.

The following theorem establishes the nonparametric identification of the LATE defined in Equation (S1). As in the case of ATE (see Theorem 1 in the main text), identification is achieved by adjusting for the deconfounder $\boldsymbol{f}(\boldsymbol{R}_i)$ and using the treatment feature T_i as an instrument for the perceived treatment feature. Similarly to the ATE case, the deconfounder satisfies the conditional independence relation $\{Y_i, \widetilde{T}_i\} \perp\!\!\!\perp \boldsymbol{R}_i \mid T_i = t, \boldsymbol{f}(\boldsymbol{R}_i)$. The difference is that the inner representation $\boldsymbol{R}_i$ is now independent of the perceived treatment feature as well as the outcome after conditioning on the treatment feature and the deconfounder. Finally, we emphasize that like Theorem 1 in the main text, this result does not require the deconfounder to be unique.

THEOREM 1 (NONPARAMETRIC IDENTIFICATION OF THE LATE) *Under Assumptions 1–4, 6 in the main text and Assumptions 1–3 in Appendix, there exists a deconfounder function* $\boldsymbol{f} : \mathcal{R} \to \mathcal{Q}$ *with* $d_Q = \dim(\mathcal{Q}) \leq d_R = \dim(\mathcal{R})$ *that satisfies the following conditional independence relation:*

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

for all $\boldsymbol{q} \in \mathcal{Q}$ *and* $t = 0, 1$. *In addition, the treatment feature and the deconfounder are separable. Then, by adjusting for such a deconfounder, we can uniquely and nonparametrically identify the local average treatment effect (LATE) defined in Equation* (S1) *as:*

$$\beta \;=\; \frac{\int_{\mathcal{R}} \mathbb{E}[Y_i \mid T_i = 1, \boldsymbol{f}(\boldsymbol{R}_i)] - \mathbb{E}[Y_i \mid T_i = 0, \boldsymbol{f}(\boldsymbol{R}_i)] dF(\boldsymbol{R}_i)}{\int_{\mathcal{R}} \mathbb{E}[\widetilde{T}_i \mid T_i = 1, \boldsymbol{f}(\boldsymbol{R}_i)] - \mathbb{E}[\widetilde{T}_i \mid T_i = 0, \boldsymbol{f}(\boldsymbol{R}_i)] dF(\boldsymbol{R}_i)}.$$

The proof is given in Appendix S4.4.

S3.3 Estimation and inference

Next, we extend the estimation and inference approaches developed in Section 3.3 to the instrumental variable setting. The main difference is that we additionally model the conditional expectation of the perceived treatment feature given the treatment feature and deconfounder,

$$m_t(\boldsymbol{f}(\boldsymbol{R}_i)) \;:=\; \mathbb{E}[\widetilde{T}_i \mid T_i = t, \boldsymbol{f}(\boldsymbol{R}_i)],$$

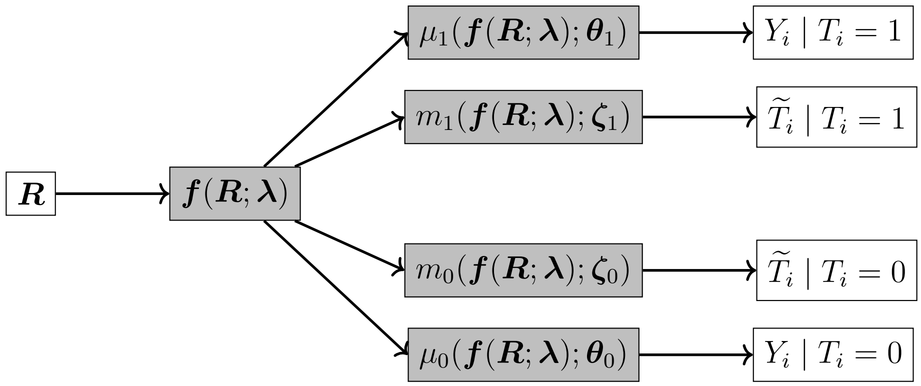

Figure S2: Diagram Illustrating the Proposed Model Architecture with the Instrumental Variable. The proposed model takes an internal representation of a treatment object $\boldsymbol{R}_i$ as an input, and finds a deconfounder $\boldsymbol{f}(\boldsymbol{R}_i)$, which is a lower-dimensional representation of $\boldsymbol{R}_i$, and then use it to predict the conditional expectations of the outcome $\mu_t(\boldsymbol{f}(\boldsymbol{R}_i)) := \mathbb{E}[Y_i \mid T_i = t, \boldsymbol{f}(\boldsymbol{R}_i)]$ and the perceived treatment feature $m_t(\boldsymbol{f}(\boldsymbol{R}_i)) := \mathbb{E}[\widetilde{T}_i \mid T_i = t, \boldsymbol{f}(\boldsymbol{R}_i)]$ under each treatment arm t.

for $t \in \{0, 1\}$. Figure S2 presents the proposed neural network architecture that extends the diagram shown in Figure 2 to the instrumental variable setting. The loss function is given by,

$$\{\hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\zeta}}\} \;=\; \underset{\boldsymbol{\lambda},\boldsymbol{\theta},\boldsymbol{\zeta}}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n}\left[(Y_i - \mu_{T_i}(\boldsymbol{f}(\boldsymbol{R}_i;\boldsymbol{\lambda});\boldsymbol{\theta}_{T_i}))^2 + \left(\widetilde{T}_i - m_{T_i}(\boldsymbol{f}(\boldsymbol{R}_i;\boldsymbol{\lambda});\boldsymbol{\zeta}_{T_i})\right)^2\right] \tag{S2}$$

Given this neural network architecture, we again use the DML framework for estimation and inference. The exact estimation procedure is described here for completeness.

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.

2. For each fold $k \in \{1, \cdots, K\}$, 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) obtain estimates of deconfounder and the conditional outcome function on the first fold, denoted by

   $$\begin{aligned}\hat{\boldsymbol{f}}^{(-k)}(\{\boldsymbol{R}_i\}_{i\in I_1^{(-k)}}) &:= \boldsymbol{f}(\{\boldsymbol{R}_i\}_{i\in I_1^{(-k)}}; \hat{\boldsymbol{\lambda}}^{(-k)}),\\ \mu_t^{(-k)} &:= \mu_t(\hat{\boldsymbol{f}}(\{\boldsymbol{R}_i\}_{i\in I_1^{(-k)}}; \boldsymbol{\lambda}^{(-k)}); \hat{\boldsymbol{\theta}}^{(-k)}),\end{aligned}$$

   and that of the perceived treatment feature, denoted by

   $$\hat{m}_t^{(-k)}(\{\boldsymbol{R}_i\}_{i\in I_1^{(-k)}}) := m_t(\boldsymbol{f}(\{\boldsymbol{R}_i\}_{i\in I_1^{(-k)}}; \hat{\boldsymbol{\lambda}}^{(-k)}); \hat{\boldsymbol{\zeta}}^{(-k)})$$

   by solving the optimization problem given in Equation (S2), and

   (c) obtain an estimate of the propensity score given the estimated deconfounder on the second fold, 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 an LATE estimate $\hat{\beta}$ as a solution to:

$$\frac{1}{nK}\sum_{k=1}^{K}\sum_{i:I(i)=k}\phi(\widetilde{\mathcal{D}}_i;\hat{\beta},\hat{\boldsymbol{f}}^{(-k)},\hat{\mu}_1^{(-k)},\hat{\mu}_0^{(-k)},\hat{m}_1^{(-k)},\hat{m}_0^{(-k)},\hat{\pi}^{(-k)}) \;=\; 0,$$

where $\widetilde{\mathcal{D}}_i := \{Y_i, \widetilde{T}_i, T_i, \boldsymbol{R}_i\}$ and

$$\begin{aligned}
&\phi(\widetilde{\mathcal{D}}_i;\beta,\boldsymbol{f},\mu_1,\mu_0,m_1,m_0,\pi)\\
=&\frac{T_i\{Y_i-\mu_1(\boldsymbol{f}(\boldsymbol{R}_i))\}}{\pi(\boldsymbol{f}(\boldsymbol{R}_i)}-\frac{(1-T_i)\{Y_i-\mu_0(\boldsymbol{f}(\boldsymbol{R}_i))\}}{1-\pi(\boldsymbol{f}(\boldsymbol{R}_i))}+\mu_1(\boldsymbol{f}(\boldsymbol{R}_i))-\mu_0(\boldsymbol{f}(\boldsymbol{R}_i))\\
&-\left[\frac{T_i\{\widetilde{T}_i-m_1(\boldsymbol{f}(\boldsymbol{R}_i))\}}{\pi(\boldsymbol{f}(\boldsymbol{R}_i))}-\frac{(1-T_i)\{\widetilde{T}_i-m_0(\boldsymbol{f}(\boldsymbol{R}_i))\}}{1-\pi(\boldsymbol{f}(\boldsymbol{R}_i))}+m_1(\boldsymbol{f}(\boldsymbol{R}_i))-m_0(\boldsymbol{f}(\boldsymbol{R}_i))\right]\cdot\beta.
\end{aligned}$$

Similar to the ATE case, we can establish the asymptotic property of this estimator. We first outline a set of additional regularity conditions required beyond Assumption 7.

ASSUMPTION 4 (ADDITIONAL 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 condition)

$$\mathbb{E}\left[\left|(Y_i-\mu_{T_i}(\boldsymbol{f}(\boldsymbol{R}_i)))-\beta\cdot\left(\widetilde{T}_i-m_{T_i}(\boldsymbol{f}(\boldsymbol{R}_i))\right)\right|^2\right]^{1/2}\geq c_2$$

(b) (Perceived treatment model estimation)

$$\begin{aligned}
&\mathbb{E}[m_1(\boldsymbol{f}(\boldsymbol{R}_i))-m_0(\boldsymbol{f}(\boldsymbol{R}_i))]\geq c_2,\quad \mathbb{E}[|\hat{m}_{T_i}(\hat{\boldsymbol{f}}(\boldsymbol{R}_i))-m_{T_i}(\boldsymbol{f}(\boldsymbol{R}_i))|^q]^{1/q}\leq c_1,\\
&\mathbb{E}[|\hat{m}_{T_i}(\hat{\boldsymbol{f}}(\boldsymbol{R}_i))-m_{T_i}(\boldsymbol{f}(\boldsymbol{R}_i))|^2]^{1/2}\leq\delta_n n^{-1/4}.
\end{aligned}$$

Together with Assumption 7, these regularity conditions are essentially equivalent to the assumptions required for DML inference on LATE (Chernozhukov et al., 2018).

Under the above assumptions, the asymptotic normality of the proposed estimator can be established.

THEOREM 2 (ASYMPTOTIC NORMALITY OF INSTRUMENT VARIABLE ESTIMATOR) *Under Assumptions 1–4, 6 in the main text and Assumptions 1–4 in Appendix, the estimator $\hat{\beta}$ obtained from the influence function ϕ satisfies asymptotic normality:*

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

where $\sigma^2=\mathbb{E}[\phi(\widetilde{\mathcal{D}}_i;\beta,\boldsymbol{f},\mu_1,\mu_0,m_1,m_0,\pi)^2]/\mathbb{E}[\gamma_1(\boldsymbol{f}(\boldsymbol{R}_i))-\gamma_0(\boldsymbol{f}(\boldsymbol{R}_i))]^2$.

Proof is omitted given that, like Theorem 2 in the main text, the result follows immediately from the application of DML theory (Chernozhukov et al., 2018).

S4 Proofs

S4.1 Proof of Lemma 1

We use proof by contradiction. Suppose that the overlap condition is not satisfied. That is, there exist $t \in \{0, 1\}$ and $\boldsymbol{u} \in \mathcal{U}$ such that $\mathbb{P}(T_i = t \mid \boldsymbol{U}_i = \boldsymbol{u}) = 0$. This implies that under Assumptions 3 and 4, there exist a deterministic function $\tilde{g}_T : \mathcal{U} \to \{0, 1\}$ and some $\boldsymbol{x} \in \mathcal{X}$ such that $t = \tilde{g}_T(\boldsymbol{u}) = \tilde{g}_T(\boldsymbol{g_U}(\boldsymbol{x}))$. This contradicts Assumption 5. □

S4.2 Proof of Theorem 1

Under a deep generative model of Definition 1, the distribution of $\boldsymbol{X}_i$ only depends on $\boldsymbol{P}_i$, and hence we have $Y_i(\boldsymbol{x}) \perp\!\!\!\perp \boldsymbol{X}_i \mid \boldsymbol{P}_i$. Together with Assumption 2, Lemma 4.3 of Dawid (1979) implies $Y_i(\boldsymbol{x}) \perp\!\!\!\perp \boldsymbol{X}_i$. Then, under Assumptions 3 and 5, we have:

$$Y_i(t, \boldsymbol{U}_i) \perp\!\!\!\perp T_i \mid \boldsymbol{U}_i. \tag{S3}$$

Next, under Assumptions 3, 4, and 5, $\boldsymbol{U}_i$ is a deterministic function of $\boldsymbol{R}_i$ such that we can write $\boldsymbol{U}_i = \boldsymbol{f}^*(\boldsymbol{R}_i)$ for some function $\boldsymbol{f}^* : \mathcal{R} \to \mathcal{U}$. Furthermore, Assumption 5 implies $Y_i \perp\!\!\!\perp \boldsymbol{R}_i \mid T_i = t, \boldsymbol{f}^*(\boldsymbol{R}_i) = \boldsymbol{u}$ and $0 < \Pr(T_i = t \mid \boldsymbol{f}^*(\boldsymbol{R}_i) = \boldsymbol{u}) < 1$ for all $t \in \{0, 1\}$ and $\boldsymbol{u} \in \mathcal{U}$. Thus, we have:

$$\begin{aligned}
\mathbb{P}(Y_i(t, \boldsymbol{U}_i) = y) &= \int_{\mathcal{U}} \mathbb{P}(Y_i(t, \boldsymbol{U}_i) = y \mid \boldsymbol{U}_i) dF(\boldsymbol{U}_i) \\
&= \int_{\mathcal{U}} \mathbb{P}(Y_i(t, \boldsymbol{U}_i) = y \mid T_i = t, \boldsymbol{U}_i) dF(\boldsymbol{U}_i) \\
&= \int_{\mathcal{U}} \mathbb{P}(Y_i = y \mid T_i = t, \boldsymbol{f}^*(\boldsymbol{R}_i)) dF(\boldsymbol{f}^*(\boldsymbol{R}_i)), \\
&= \int_{\mathcal{R}} \mathbb{P}(Y_i = y \mid T_i = t, \boldsymbol{f}^*(\boldsymbol{R}_i)) dF(\boldsymbol{R}_i),
\end{aligned}$$

where the second equality follows from Equation (S3) and Lemma 1, and the third equality is due to Assumption 1. Finally, suppose there is another function $\boldsymbol{f} : \mathcal{R} \to \mathcal{Q}$, which satisfies the conditional independence relation $Y_i \perp\!\!\!\perp \boldsymbol{R}_i \mid T_i = t, \boldsymbol{f}(\boldsymbol{R}_i) = q$ for all $q \in \mathcal{Q}$ and is separable from the treatment feature. Then,

$$\begin{aligned}
\int_{\mathcal{R}} \mathbb{P}(Y_i = y \mid T_i = t, \boldsymbol{f}(\boldsymbol{R}_i)) dF(\boldsymbol{R}_i) &= \int_{\mathcal{R}} \mathbb{P}(Y_i = y \mid T_i = t, \boldsymbol{f}(\boldsymbol{R}_i), \boldsymbol{R}_i) dF(\boldsymbol{R}_i) \\
&= \int_{\mathcal{R}} \mathbb{P}(Y_i = y \mid T_i = t, \boldsymbol{f}^*(\boldsymbol{R}_i), \boldsymbol{R}_i) dF(\boldsymbol{R}_i) \\
&= \int_{\mathcal{R}} \mathbb{P}(Y_i = y \mid T_i = t, \boldsymbol{f}^*(\boldsymbol{R}_i)) dF(\boldsymbol{R}_i)
\end{aligned}$$

Thus, any function of $\boldsymbol{R}_i$ that satisfies this conditional independence relation leads to the same identification formula for the marginal distribution of potential outcome. □

S4.3 Proof of Theorem 2

Assumptions 2(c)–(d) and the triangule inequality imply,

$$\mathbb{E}[|\hat{\pi}(\hat{\boldsymbol{f}}(\boldsymbol{R}_i)) - \pi(\boldsymbol{f}(\boldsymbol{R}_i))|^q]^{1/q} = \mathbb{E}[|\{\hat{\pi}(\hat{\boldsymbol{f}}(\boldsymbol{R}_i)) - \hat{\pi}(\boldsymbol{f}(\boldsymbol{R}_i))\} + \{\hat{\pi}(\boldsymbol{f}(\boldsymbol{R}_i)) - \pi(\boldsymbol{f}(\boldsymbol{R}_i))\}|^q]^{1/q}$$

$$
\begin{aligned}
&\leq \mathbb{E}[|\hat{\pi}(\hat{\boldsymbol{f}}(\boldsymbol{R}_i)) - \hat{\pi}(\boldsymbol{f}(\boldsymbol{R}_i))|^q]^{1/q} + c_1 \\
&\leq L \cdot \mathbb{E}\left[\left\|\hat{\boldsymbol{f}}(\boldsymbol{R}_i) - \boldsymbol{f}(\boldsymbol{R}_i)\right\|^q\right]^{1/q} + c_1 \\
&= (L+1)c_1 \qquad \text{(S4)}
\end{aligned}
$$

where L is a Lipschitz constant. Similarly, we can also show,

$$
\begin{aligned}
&\mathbb{E}\left[\left|\hat{\pi}(\hat{\boldsymbol{f}}(\boldsymbol{R}_i)) - \pi(\boldsymbol{f}(\boldsymbol{R}_i))\right|^2\right]^{1/2} \\
\leq\ & L \cdot \mathbb{E}\left[\left\|\hat{\boldsymbol{f}}(\boldsymbol{R}_i) - \boldsymbol{f}(\boldsymbol{R}_i)\right\|^2\right]^{1/2} + \mathbb{E}\left[|\hat{\pi}(\boldsymbol{f}(\boldsymbol{R}_i)) - \pi(\boldsymbol{f}(\boldsymbol{R}_i))|^2\right]^{1/2} \\
\leq\ & (L+1)\delta_n n^{-1/4}. \qquad \text{(S5)}
\end{aligned}
$$

Together with Assumption 2(b), Equation (S5) implies that there exists a sequence of positive constants δ_n' converging to zero as the sample size n increases such that the following inequality holds,

$$
\mathbb{E}[|\hat{\pi}(\hat{\boldsymbol{f}}(\boldsymbol{R}_i)) - \pi(\boldsymbol{f}(\boldsymbol{R}_i))|^2]^{1/2} \cdot \mathbb{E}[|\hat{\mu}_{T_i}(\hat{\boldsymbol{f}}(\boldsymbol{R}_i)) - \mu_{T_i}(\boldsymbol{f}(\boldsymbol{R}_i))|^2]^{1/2} \leq \delta_n' n^{-1/2}. \qquad \text{(S6)}
$$

Thus, the standard regularity conditions of the DML theory (Chernozhukov et al., 2018) are satisfied for the estimated propensity score with the estimated deconfounder, i.e., $\hat{\pi}(\hat{\boldsymbol{f}}(\boldsymbol{R}_i))$. Finally, Assumptions 2(a)–(d) and Equations (S4) and (S6) imply,

$$
\sqrt{n}\,(\hat{\tau} - \tau) \xrightarrow{d} \mathcal{N}(0, \sigma^2)
$$

where $\sigma^2 = \mathbb{E}[\psi(\mathcal{D}_i; \tau, \boldsymbol{f}, \eta_1, \eta_0, \pi)^2]$.

□

S4.4 Proof of Theorem 1

Under a deep generative model (Definition 1), the distribution of $\boldsymbol{X}_i$ only depends on $\boldsymbol{P}_i$, and hence we have $\{Y_i(\boldsymbol{x}), \widetilde{T}_i(\boldsymbol{x})\} \perp\!\!\!\perp \boldsymbol{X}_i \mid \boldsymbol{P}_i$. Together with Assumption 2 in the main text, Lemma 4.3 of Dawid (1979) implies $\{Y_i(\boldsymbol{x}), \widetilde{T}_i(\boldsymbol{x})\} \perp\!\!\!\perp \boldsymbol{X}_i$. Under Assumptions 3 and 4 in the main text and Assumptions 1, 2, and 3(b), we have, for any $t, \tilde{t} \in \{0,1\}$ and $\boldsymbol{u} \in \mathcal{U}$:

$$
\{Y_i(\tilde{t}, \boldsymbol{u}), \widetilde{T}_i(t, \boldsymbol{u})\} \perp\!\!\!\perp \{T_i, \boldsymbol{U}_i\}. \qquad \text{(S7)}
$$

Then, for any $\boldsymbol{u} \in \mathcal{U}$,

$$
\begin{aligned}
&\mathbb{E}[Y_i \mid T_i = 1, \boldsymbol{U}_i = \boldsymbol{u}] - \mathbb{E}[Y_i \mid T_i = 0, \boldsymbol{U}_i = \boldsymbol{u}] \\
=\ & \mathbb{E}[Y_i(\widetilde{T}_i(1, \boldsymbol{u}), \boldsymbol{u}) \mid T_i = 1, \boldsymbol{U}_i = \boldsymbol{u}] - \mathbb{E}[Y_i(\widetilde{T}_i(0, \boldsymbol{u}), \boldsymbol{u}) \mid T_i = 0, \boldsymbol{U}_i = \boldsymbol{u}] \\
=\ & \mathbb{E}[Y_i(\widetilde{T}_i(1, \boldsymbol{u}), \boldsymbol{u}) - Y_i(\widetilde{T}_i(0, \boldsymbol{u}), \boldsymbol{u})] \\
=\ & \mathbb{E}[Y_i(\widetilde{T}_i(1, \boldsymbol{u}), \boldsymbol{u}) - Y_i(\widetilde{T}_i(0, \boldsymbol{u}), \boldsymbol{u}) \mid \widetilde{T}_i(1, \boldsymbol{u}) = 1, \widetilde{T}_i(0, \boldsymbol{u}) = 0] \cdot \mathbb{P}(\widetilde{T}_i(1, \boldsymbol{u}) = 1, \widetilde{T}_i(0, \boldsymbol{u}) = 0) \\
=\ & \mathbb{E}[Y_i(1, \boldsymbol{u}) - Y_i(0, \boldsymbol{u}) \mid \widetilde{T}_i(1, \boldsymbol{u}) = 1, \widetilde{T}_i(0, \boldsymbol{u}) = 0] \cdot \mathbb{P}(\widetilde{T}_i(1, \boldsymbol{u}) = 1, \widetilde{T}_i(0, \boldsymbol{u}) = 0),
\end{aligned}
$$

where the second equality follows from Equation (S7), the third equality is due to Assumption 3. We also have:

$$\begin{aligned}
\mathbb{P}(\widetilde{T}_i(1,\boldsymbol{u})=1,\widetilde{T}_i(0,\boldsymbol{u})=0) &= \mathbb{E}[\widetilde{T}_i(1,\boldsymbol{u})-\widetilde{T}_i(0,\boldsymbol{u})] \\
&= \mathbb{E}[\widetilde{T}_i(1,\boldsymbol{u}) \mid T_i=1,\boldsymbol{U}_i=\boldsymbol{u}]-\mathbb{E}[\widetilde{T}_i(0,\boldsymbol{u}) \mid T_i=0,\boldsymbol{U}_i=\boldsymbol{u}] \\
&= \mathbb{E}[\widetilde{T}_i \mid T_i=1,\boldsymbol{U}_i=\boldsymbol{u}]-\mathbb{E}[\widetilde{T}_i \mid T_i=1,\boldsymbol{U}_i=\boldsymbol{u}]
\end{aligned}$$

where the second equality follows from Equation (S7). Together, under Assumption 3 (a), we have:

$$\begin{aligned}
&\mathbb{E}[Y_i(1,\boldsymbol{u})-Y_i(0,\boldsymbol{u}) \mid \widetilde{T}_i(1,\boldsymbol{u})=1,\widetilde{T}_i(0,\boldsymbol{u})=0] \\
=&\frac{\mathbb{E}[Y_i \mid T_i=1,\boldsymbol{U}_i=\boldsymbol{u}]-\mathbb{E}[Y_i \mid T_i=0,\boldsymbol{U}_i=\boldsymbol{u}]}{\mathbb{E}[\widetilde{T}_i \mid T_i=1,\boldsymbol{U}_i=\boldsymbol{u}]-\mathbb{E}[\widetilde{T}_i \mid T_i=1,\boldsymbol{U}_i=\boldsymbol{u}]}.
\end{aligned} \tag{S8}$$

Finally, the LATE is identified as:

$$\begin{aligned}
&\mathbb{E}[Y_i(1,\boldsymbol{U}_i)-Y_i(0,\boldsymbol{U}_i) \mid \widetilde{T}_i(1,\boldsymbol{U}_i)=1,\widetilde{T}_i(0,\boldsymbol{U}_i)=0] \\
=&\int_{\mathcal{U}} \mathbb{E}[Y_i(1,\boldsymbol{U}_i)-Y_i(0,\boldsymbol{U}_i) \mid \widetilde{T}_i(1,\boldsymbol{U}_i)=1,\widetilde{T}_i(0,\boldsymbol{U}_i)=0,\boldsymbol{U}_i]dF(\boldsymbol{U}_i \mid \widetilde{T}_i(1,\boldsymbol{U}_i)=1,\widetilde{T}_i(0,\boldsymbol{U}_i)=0) \\
=&\int_{\mathcal{U}} \frac{\mathbb{E}[Y_i \mid T_i=1,\boldsymbol{U}_i]-\mathbb{E}[Y_i \mid T_i=0,\boldsymbol{U}_i]}{\mathbb{E}[\widetilde{T}_i \mid T_i=1,\boldsymbol{U}_i]-\mathbb{E}[\widetilde{T}_i \mid T_i=1,\boldsymbol{U}_i]}dF(\boldsymbol{U}_i \mid \widetilde{T}_i(1,\boldsymbol{U}_i)>\widetilde{T}_i(0,\boldsymbol{U}_i)) \\
=&\int_{\mathcal{U}} \frac{\mathbb{E}[Y_i \mid T_i=1,\boldsymbol{U}_i]-\mathbb{E}[Y_i \mid T_i=0,\boldsymbol{U}_i]}{\mathbb{E}[\widetilde{T}_i \mid T_i=1,\boldsymbol{U}_i]-\mathbb{E}[\widetilde{T}_i \mid T_i=1,\boldsymbol{U}_i]} \cdot \frac{\mathbb{P}(\widetilde{T}_i(1,\boldsymbol{U}_i)=1,\widetilde{T}_i(0,\boldsymbol{U}_i)=0 \mid \boldsymbol{U}_i)}{\mathbb{P}(\widetilde{T}_i(1,\boldsymbol{U}_i)=1,\widetilde{T}_i(0,\boldsymbol{U}_i)=0)}dF(\boldsymbol{U}_i) \\
=&\int_{\mathcal{U}} \frac{\mathbb{E}[Y_i \mid T_i=1,\boldsymbol{U}_i]-\mathbb{E}[Y_i \mid T_i=0,\boldsymbol{U}_i]}{\mathbb{E}[\widetilde{T}_i \mid T_i=1,\boldsymbol{U}_i]-\mathbb{E}[\widetilde{T}_i \mid T_i=1,\boldsymbol{U}_i]} \cdot \frac{\mathbb{E}[\widetilde{T}_i \mid T_i=1,\boldsymbol{U}_i]-\mathbb{E}[\widetilde{T}_i \mid T_i=1,\boldsymbol{U}_i]}{\mathbb{P}(\widetilde{T}_i(1,\boldsymbol{U}_i)=1,\widetilde{T}_i(0,\boldsymbol{U}_i)=0)}dF(\boldsymbol{U}_i) \\
=&\frac{\int_{\mathcal{U}} \mathbb{E}[Y_i \mid T_i=1,\boldsymbol{U}_i]-\mathbb{E}[Y_i \mid T_i=0,\boldsymbol{U}_i]dF(\boldsymbol{U}_i)}{\int_{\mathcal{U}} \mathbb{E}[\widetilde{T}_i \mid T_i=1,\boldsymbol{U}_i]-\mathbb{E}[\widetilde{T}_i \mid T_i=0,\boldsymbol{U}_i]dF(\boldsymbol{U}_i)} \\
=&\frac{\int_{\mathcal{R}} \mathbb{E}[Y_i \mid T_i=1,\boldsymbol{f}(\boldsymbol{R}_i)]-\mathbb{E}[Y_i \mid T_i=0,\boldsymbol{f}(\boldsymbol{R}_i)]dF(\boldsymbol{R}_i)}{\int_{\mathcal{R}} \mathbb{E}[\widetilde{T}_i \mid T_i=1,\boldsymbol{f}(\boldsymbol{R}_i)]-\mathbb{E}[\widetilde{T}_i \mid T_i=0,\boldsymbol{f}(\boldsymbol{R}_i)]dF(\boldsymbol{R}_i)}
\end{aligned}$$

where the second equality follows from Equation (S8), the third equality is due to Bayes' theorem, and the last equality (as well as the existence of deconfounder) follows from the same argument used to establish Theorem 2 in the main text. □

S5 Results of Additional Simulation Studies

	Bias	RMSE	95% confidence interval coverage	95% confidence interval avg. length	Runtime (seconds)
Weak confounding w/ separability					
Proposed estimator (new)	-0.33	1.06	0.94	3.47	42.1
Proposed estimator (reuse)	-0.26	0.98	0.92	2.95	54.1
Difference-in-Means	3.61	3.61	0	4.72	0.0
Outcome model with BERT	1.17	1.00	0.12	0.53	296
DML with BERT	0.58	4.21	0.93	2.10	327
Moderate confounding w/ separability					
Proposed estimator (new)	-1.07	2.72	0.95	9.00	43.6
Proposed estimator (reuse)	-1.05	2.36	0.93	6.85	62.9
Difference-in-Means	7.95	7.95	0	9.50	0.0
Outcome model with BERT	3.44	2.27	0.05	0.92	673
DML with BERT	2.09	18.3	0.92	5.14	720
Strong confounding w/ separability					
Proposed estimator (new)	-14.6	36.9	0.88	113	55.3
Proposed estimator (reuse)	-15.1	36.0	0.92	96.3	74.3
Difference-in-Means	86.0	86.0	0	95.7	0.0
Outcome model with BERT	112	114	0	13.2	2731
DML with BERT	208	917	0.26	382	2756
Weak confounding w/o separability					
Proposed estimator (new)	3.23	3.27	0	1.45	35.9
Proposed estimator (reuse)	2.87	2.89	0	1.34	41.1
Difference-in-Means	2.20	2.20	0.03	4.03	0.0
Outcome model with BERT	2.55	2.70	0.01	1.02	320
DML with BERT	6.18	16.7	0.05	4.69	348
Moderate confounding w/o separability					
Proposed estimator (new)	6.70	6.76	0	2.89	42.5
Proposed estimator (reuse)	5.90	5.93	0	2.66	44.9
Difference-in-Means	4.39	4.40	0	8.04	0.0
Outcome model with BERT	7.56	7.66	0	1.85	624
DML with BERT	12.1	30.5	0.06	10.1	656
Strong confounding w/o separability					
Proposed estimator (new)	76.3	76.7	0	29.8	53.0
Proposed estimator (reuse)	66.8	67.1	0	27.7	56.4
Difference-in-Means	44.0	44.0	0	80.3	0.0
Outcome model with BERT	116	117	0	13.2	2689
DML with BERT	207	814	0.26	425	2716

Table S5: Simulation Results with 200 Monte Carlo trials

	Bias	RMSE	95% confidence interval coverage	avg. length	Average time (sec.)
Weak confounding w/ separability					
Proposed estimator (new)	-0.31	1.09	0.93	3.55	43.0
Proposed estimator (reuse)	-0.21	0.96	0.93	2.90	55.7
Moderate confounding w/ separability					
Proposed estimator (new)	-0.99	2.77	0.92	8.83	45.7
Proposed estimator (reuse)	-1.00	2.67	0.90	6.99	61.7
Strong confounding w/ separability					
Proposed estimator (new)	-14.3	36.1	0.89	108	57.7
Proposed estimator (reuse)	-16.0	38.1	0.90	98.8	76.3

Table S6: Simulation Results based on 1000 Monte Carlo trials

S6 Additional Empirical Application: Hong Kong Experiment

To further validate the proposed GPI methodology, we apply it to the Hong Kong experiment conducted by Fong and Grimmer (2023). This experiment examines the extent to which U.S. commitments to Hong Kong influence public perceptions of U.S. government support for Hong Kong protesters. To investigate this question, the authors carried out two experiments—one in December 2019 ($N = 1{,}983$) and another in October 2020 ($N = 2{,}072$). For each experiment, they first generated 555,660 unique candidate texts by randomly varying several text features: descriptions of commitments (commitments the United States made to Hong Kong), bravery (the bravery displayed by the protesters), mistreatment (China's mistreatment of its own citizens), flags (whether protesters were shown waving American flags), threat (the security threat China poses to the United States), economy (information about Hong Kong's political system and economy), and violations (how China's actions violate its treaty with the United Kingdom).

To mitigate confounding bias arising from these text features, the authors created roughly 15 variants of texts for each feature, randomly concatenated two or three variants to form a complete text, and then randomly assigned the resulting texts to participants. Participants read a text and then rated, on a 0–100 scale, how strongly they agreed with the view that the U.S. government should support Hong Kong protesters. Because textual features are randomized, the authors regressed participants' responses on the seven text features using ordinary least squares (OLS).

We estimate the average treatment effect separately for each experimental wave. Following the empirical application in Section 5, we use the GPI methodology based on the text-reuse approach with Llama 3-8B. After extracting the internal representation, we apply the proposed estimation procedure described in Section 3, using five-fold cross-fitting. We adopt a larger fold size than in Section 5 to ensure that the neural network is trained on a sufficiently large number of samples. For comparison—consistent with our simulation studies in Section 4 and the empirical application in Section 5—we also implement two existing BERT-based methods: the outcome-model approach with BERT (Pryzant et al., 2021) and DML with BERT (Gui and Veitch, 2023). Finally, we replicate the original OLS analysis. In this application, the experimental design should effectively mitigate confounding bias, so OLS serves as a reasonable approximation to the ground truth.

Methods	ATE Estimates	95% Confidence Interval	IOSS	Runtime (sec.)
Wave 1: December 2019				
OLS (original)	5.231	[1.814, 8.648]	-	0.0
GPI (reuse)	6.175	[2.784, 9.566]	0.04	35.9
Outcome model with BERT	26.591	[25.482, 27.701]	0.28	11890.6
DML with BERT	24.361	[20.163, 28.560]		11892.9
Wave 2: October 2020				
OLS (original)	2.680	[0.269, 5.091]	-	0.0
GPI (reuse)	2.043	[-0.790, 4.877]	0.04	27.9
Outcome model with BERT	1.676	[1.319, 2.033]	0.07	9940.0
DML with BERT	2.808	[−0.519, 6.136]		9942.7

Table S7: The Estimated Average Treatment Effect (ATE) for the Hong Kong Experiment.

Table S7 presents the estimated ATEs for both experimental waves. The GPI estimates are consistent with the original OLS estimates that controls the textual features directly. We find that IOSS for the GPI method is close to 0 (0.04 for both waves), indicating that the deconfounder extracted by GPI is disentangled from the treatment feature. In contrast, the BERT-based methods yield significantly larger ATE estimates for Wave 1 that significantly diverge from the OLS estimates. This unually large estiamte corresponds to a large value of IOSS for Wave 1, which is 0.28 (the IQSS score for Wave 2 is smaller, equaling 0.07). Moreover, the GPI method is computationally efficient, with runtimes significantly faster than those of the BERT-based methods.

References

Aronow, P. M., Baron, J., and Pinson, L. (2019). A Note on Dropping Experimental Subjects who Fail a Manipulation Check. *Political Analysis*, 27(4):572–589.

Blackwell, M., Brown, J. R., Hill, S., Imai, K., and Yamamoto, T. (2025). Priming bias versus post-treatment bias in experimental designs. *Political Analysis*, Forthcoming.

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(1):C1–C68.

Dawid, A. P. (1979). Conditional Independence in Statistical Theory. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(1):1–31.

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)*, pages 1600–1609.

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

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

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

Klar, S., Leeper, T., and Robison, J. (2020). Studying Identities with Experiments: Weighing the Risk of Posttreatment Bias Against Priming Effects. *Journal of Experimental Political Science*, 7(1):56–60.

Montgomery, J. M., Nyhan, B., and Torres, M. (2018). How Conditioning on Posttreatment Variables Can Ruin Your Experiment and What to Do about It. *American Journal of Political Science*, 62(3):760–775.

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