Causal Representation Learning with Generative Artificial Intelligence: Application to Texts as Treatments

Kosuke Imai

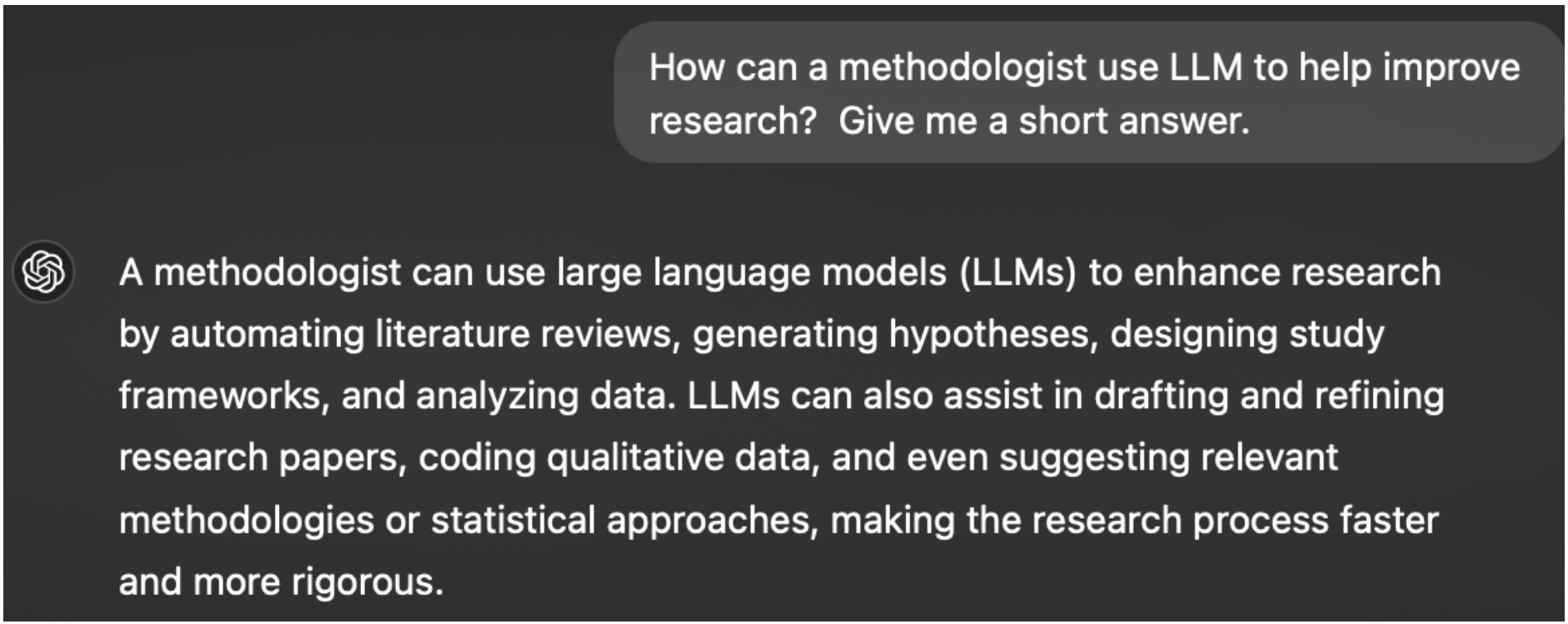
Harvard University

Oxford Spring School in Advanced Research Methods
University of Oxford
April 3, 2025

Joint work with Kentaro Nakamura (Harvard Kennedy School)

Motivation

- Generative AI is transforming medicine, education, marketing, etc.
- Can methodologists get some help from generative AI too?



Using Generative AI to Improve Causal Inference

- LLM-assisted causal inference with unstructured data
  1. use LLM to generate treatment texts
  2. randomly assign generated texts to survey respondents and measure their reactions
  3. extract the internal representation of generated texts from LLM
  4. machine learning using the extracted true vector representation
     - estimate a deconfounder that summarizes all confounding information
     - double machine learning for valid statistical inference

- Advantage: no need to estimate the representation
  - nonparametric modeling: no topic modeling
  - computational efficiency: no fine tuning
  - better empirical performance
- Extension: instrumental variable approach to perceived treatments

Motivating Application: Texts-as-Treatments

- Candidate Biography Experiment (Fong and Grimmer, 2016)
  - 1246 biographies of American politicians scraped from Wikipedia
  - 1,886 voters as respondents
  - randomly assign biographies to voters
  - feeling thermometer as the outcome

- Analysis
  - supervised topic model to discover 10 treatment features
  - estimate the average treatment effects of estimated topic proportions

- Existing approaches for texts-as-treatments:
  1. model-based approach (e.g., Egami *et al.* 2022; Fong and Grimmer, 2023)
  2. causal representation learning based on embedding
     (e.g., Veitch et al. 2020; Pryzant et al. 2021; Gui and Veitch, 2023)

Using LLM to Generate Biographies (Llama 3)

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.

- Requirements:
  1. open-source LLM: GPT, PaLM, Llama, OPT, etc.
  2. assume LLM is compliant with prompt or use human coder

Text Reuse: Using the Existing Texts as Treatments

Prompts	**Generated candidate biography**
(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.

Setup

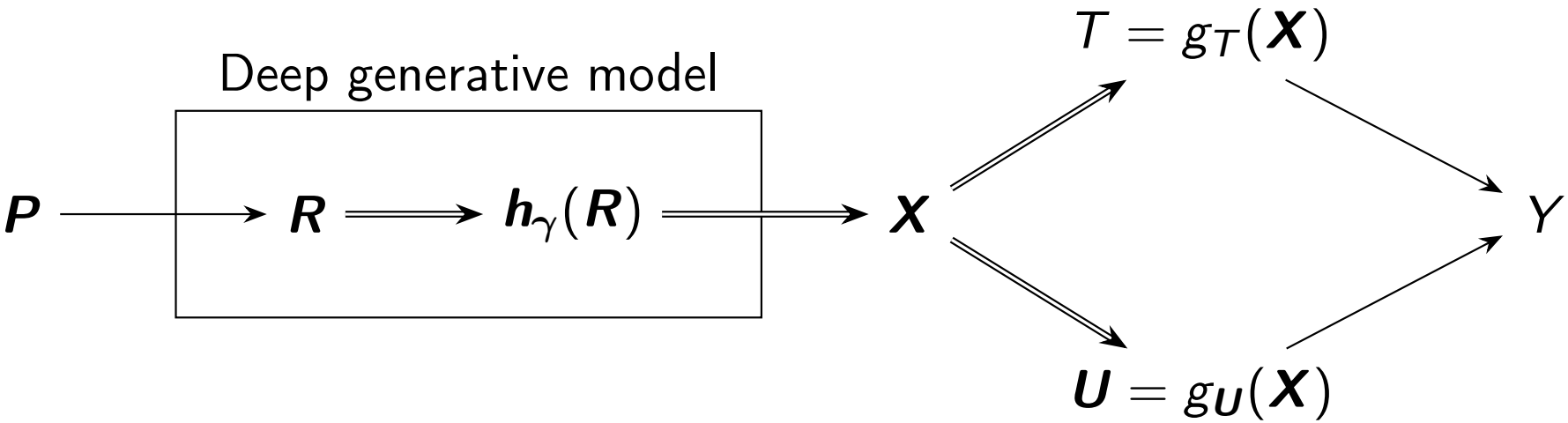
- Deep Generative Model

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

  - $\boldsymbol{P}_i$: prompt
  - $\boldsymbol{X}_i$: treatment object (generated texts)
  - $\boldsymbol{R}_i$: hidden states or internal representations
  - $\boldsymbol{h}_\gamma(\boldsymbol{R}_i)$: deterministic function from hidden states to last layer

- Other variables
  - $Y_i(\boldsymbol{x})$: Potential outcome when exposed to treatment object $\boldsymbol{x}$
  - Y_i: Outcome (collected from the survey respondents)
  - T_i: Binary treatment feature (e.g., military experiences)
  - $\boldsymbol{U}_i$: Confounding features (e.g., college education)

Assumptions



- Separability:

$$Y_i(\boldsymbol{X}_i) = Y_i(g_T(\boldsymbol{X}_i), g_{\boldsymbol{U}}(\boldsymbol{X}_i)) = Y_i(T_i, \boldsymbol{U}_i)$$

- Lemma: separability implies overlap

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

- Deterministic decording: $\mathbb{P}(\boldsymbol{X}_i \mid \boldsymbol{h}_\gamma(\boldsymbol{R}_i))$ is degenerate

Nonparametric Identification

- Average treatment effect (ATE):

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

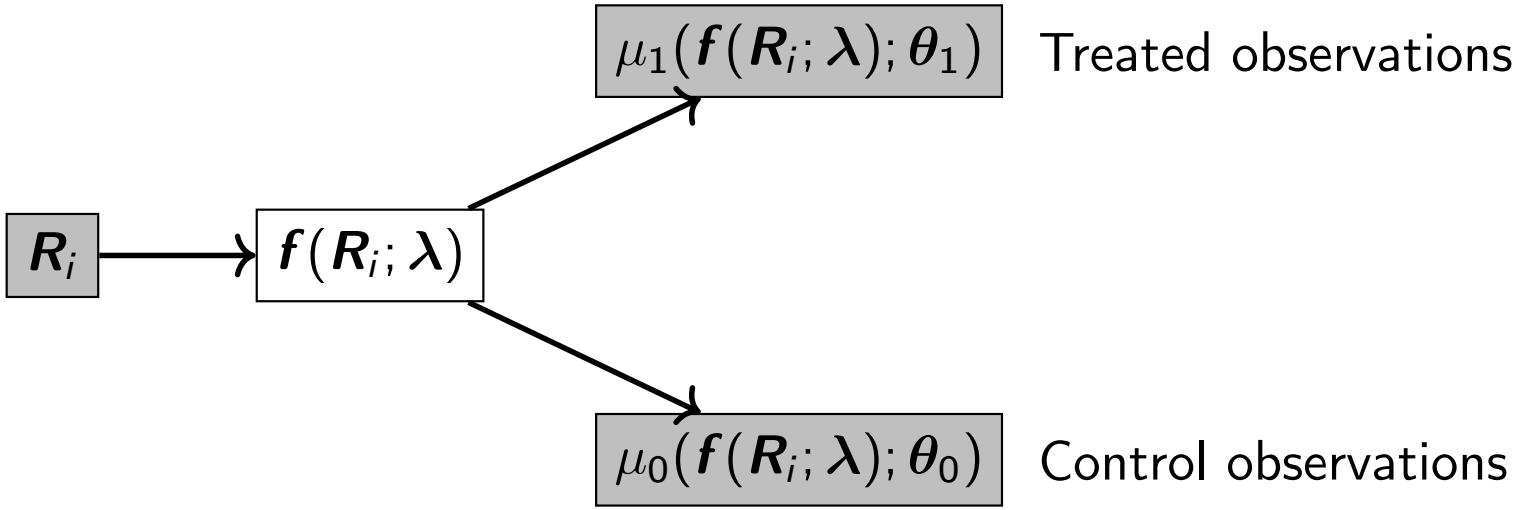
- Under these assumptions, there exists a **Deconfounder** $\boldsymbol{f} : \mathbb{R}^r \to \mathbb{R}^q$ with $q \leq r$ such that

$$Y_i \perp\!\!\!\perp \boldsymbol{R}_i \mid T_i = t, \boldsymbol{f}(\boldsymbol{R}_i), \quad t \in \{0, 1\}$$

  - Deconfounder does not have to be unique
  - Example: Confounding Features $\boldsymbol{U}_i$ (deterministic function of $\boldsymbol{R}_i$)

- By adjusting for this Deconfounder, we can identify the ATE
- Direct adjustment for $\boldsymbol{R}_i$ leads to the lack of overlap

Estimation and Inference



1. Estimate the outcome models and deconfounder via TarNet (Shalit et al. 2017):

$$\{\hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\theta}}_0, \hat{\boldsymbol{\theta}}_1\} = \underset{\boldsymbol{\lambda},\boldsymbol{\theta}_0,\boldsymbol{\theta}_1}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \{Y_i - \mu_{T_i}(\boldsymbol{f}(\boldsymbol{R}_i;\boldsymbol{\lambda});\boldsymbol{\theta}_{T_i})\}^2$$

2. Estimate the propensity score using the estimated Deconfounder

$$\pi(\boldsymbol{f}(\boldsymbol{R}_i, \hat{\boldsymbol{\lambda}})) = \mathbb{P}(T_i = 1 \mid \boldsymbol{f}(\boldsymbol{R}_i, \hat{\boldsymbol{\lambda}}))$$

Popular DragonNet (Shi et al. 2019) jointly estimates the outcome models, propensity score, and deconfounder, leading to the lack of overlap

Double Machine Learning (Chernozhukov et al. 2018)

- Cross-fitting:
  1. randomly divide the data into K folds
  2. for each $k = 1, \ldots, K$, use the kth fold as the test set and the remaining $k-1$ folds as the training set
     1. randomly split the training set further into two subsets
     2. use the first subset to estimate outcome models and deconfounder
     3. use the second subset to estimate propensity score given the estimated deconfounder
  3. Compute the ATE estimator as:

$$\begin{aligned}\hat{\tau} = \frac{1}{nK}\sum_{k=1}^{K}\sum_{i:I(i)=k} & \hat{\mu}_1^{(-k)}(\hat{\boldsymbol{f}}^{(-k)}(\boldsymbol{R}_i)) - \hat{\mu}_0^{(-k)}(\hat{\boldsymbol{f}}^{(-k)}(\boldsymbol{R}_i)) \\ & + \frac{T_i\{Y_i - \hat{\mu}_1^{(-k)}(\hat{\boldsymbol{f}}^{(-k)}(\boldsymbol{R}_i))\}}{\hat{\pi}^{(-k)}(\hat{\boldsymbol{f}}^{(-k)}(\boldsymbol{R}_i))} - \frac{(1-T_i)\{Y_i - \hat{\mu}_0^{(-k)}(\hat{\boldsymbol{f}}^{(-k)}(\boldsymbol{R}_i))\}}{1-\hat{\pi}^{(-k)}(\hat{\boldsymbol{f}}^{(-k)}(\boldsymbol{R}_i))}\end{aligned}$$

- Double robustness, asymptotic normality

Practical Implementation Details

- Internal representation extracted from LLM is still high-dimensional:

  $\dim(\boldsymbol{R}) = \text{number of tokens} \times 4096$ for Llama 3 (8 billion parameters)

- Pooling strategies depend on deep generative models
  - BERT: the first special classification token [CLS]
  - Llama 3: the hidden states of the last token

- TarNet requires hyperparameter tuning
  - size and depth of layers
  - learning rate
  - maximum epoch size
- Use of automatic hyperparameter optimization methods (e.g., Optuna)

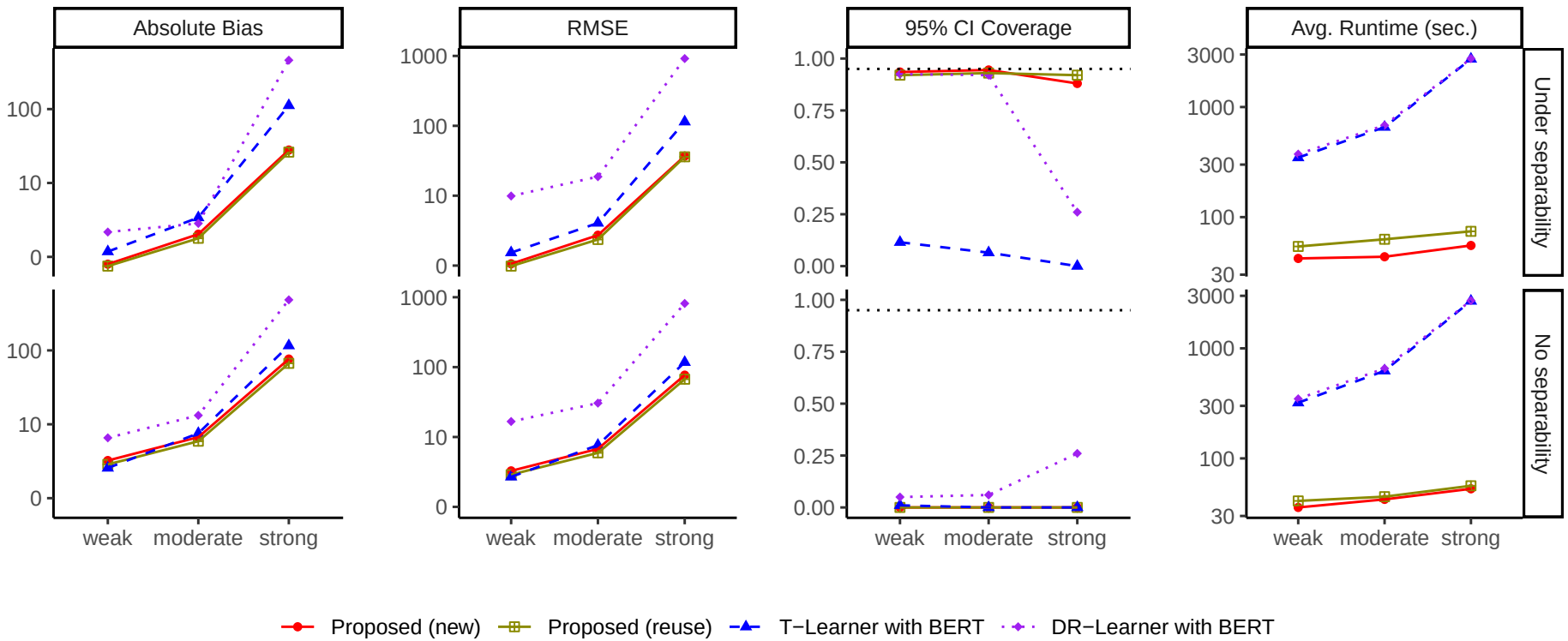
Simulation Study Setup

- A simulation based on the candidate biography experiment
  - Create 4,000 sets of the first, middle, and last names of political candidates via randomly sampling from the Fong and Grimmer data
  - Use Llama 3 to generate a biography for each US political candidate's
  - Instruct LLM to repeat the same texts for reuse
- The data generating process:

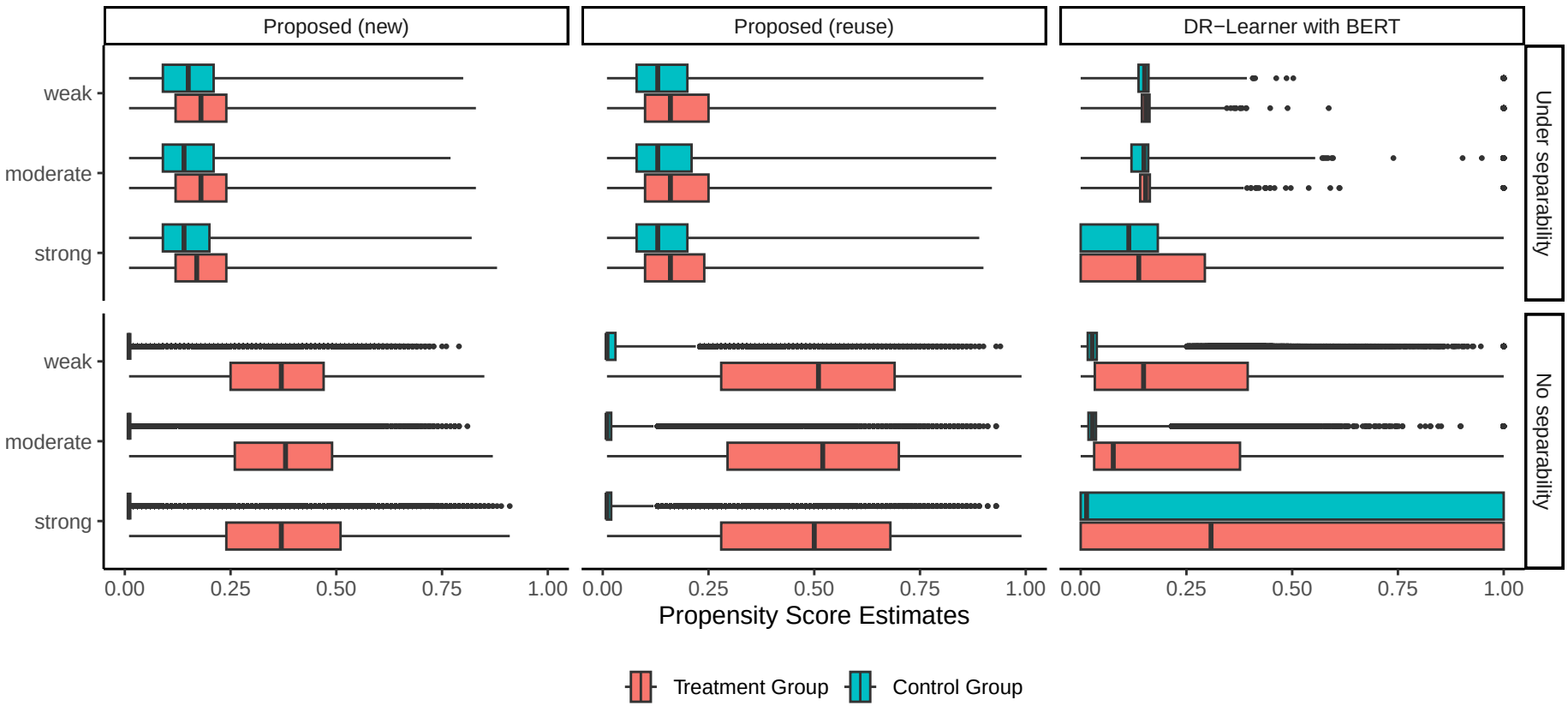
$$Y_i = \alpha_1 T_i + \alpha_2 T_i h_1(\boldsymbol{X}_i) - \alpha_3 h_1(\boldsymbol{X}_i) - \alpha_4 h_2(\boldsymbol{X}_i) + \epsilon_i$$
$$\epsilon_i \sim \mathcal{N}(\mu_i, 1)$$

  where
  - T_i: military background (binary)
  - $h_1(\boldsymbol{X}_i)$: topic-model based confounder
  - $h_2(\boldsymbol{X}_i)$: sentiment-analysis based confounder
- $2 \times 3 = 6$ scenarios:
  1. separability holds or does not hold (separate or overlapping topics)
  2. weak, medium, or strong confounding

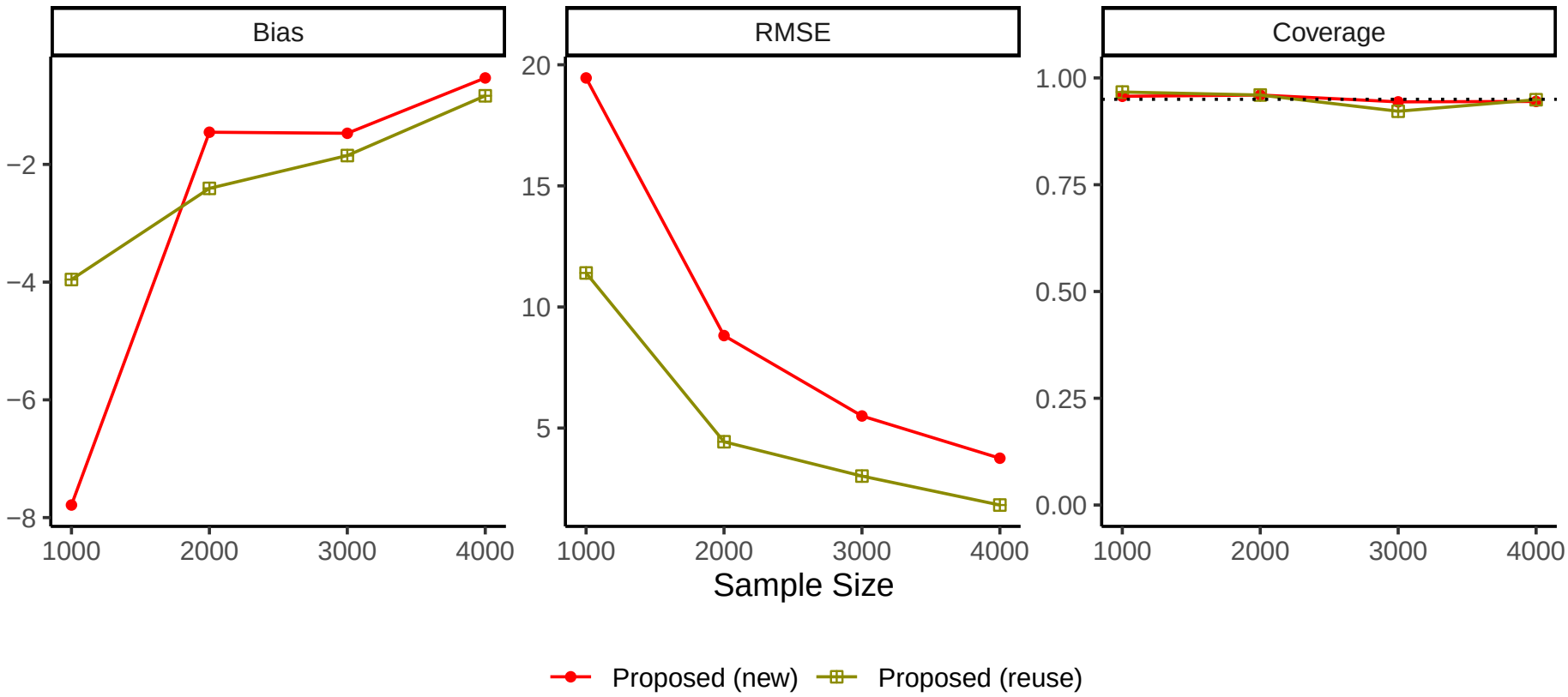
Simulation Results



Distribution of Estimated Propensity Score



Performance across Different Sample Sizes



Empirical Analysis I: Candidate Biography Experiment

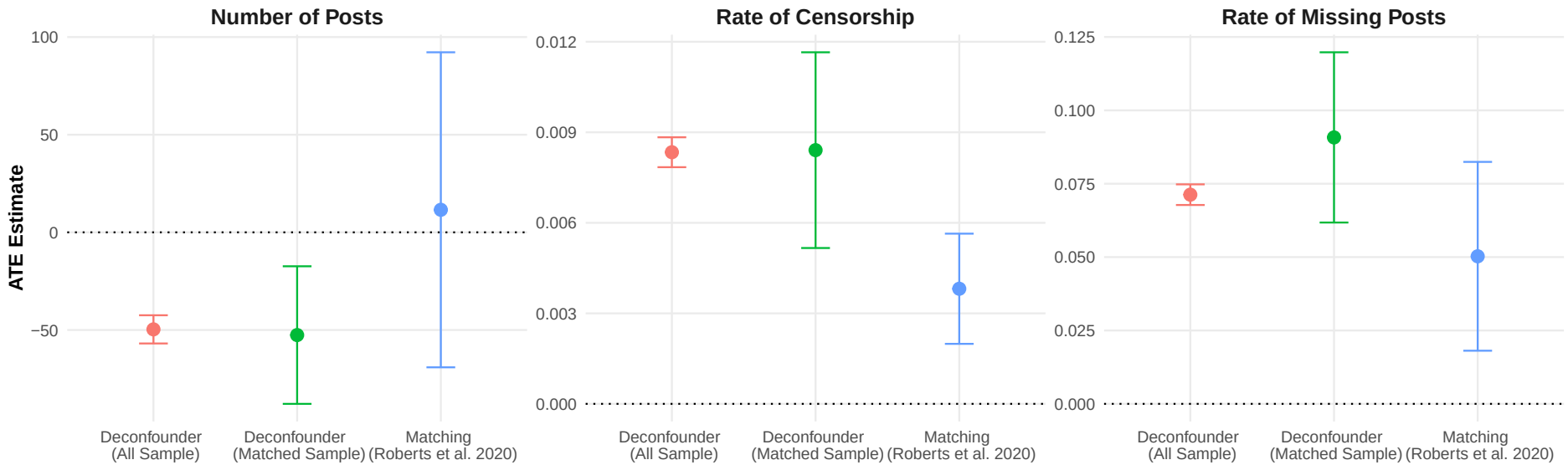
- Analyze the original survey by Fong and Grimmer (2016)
  - 1,246 Congressional candidate biographies from Wikipedia
  - 1,886 survey participants with a total of 5,291 observations
  - evaluate a biography using the feeling thermometer [0, 100]
  - Keyword-based treatment coding: "military", "war", "veteran", or "army"
  - use text-reuse approach with Llama 3

Methods	ATE	95% Conf. Int.	Runtime (sec.)
Proposed method (reuse)	5.462	[2.790, 8.135]	28.9
T-learner with BERT	−2.557	[−2.608, −2.505]	6139.7
DR-learner with BERT	−67.777	[−109.967, −25.587]	6210.3

Empirical Analysis II: Chinese Internet Sensorship

- Texts as confounders (Roberts *et al.* 2020)
- Question: Are Chinese social media users who had their post censored more likely to be censored for subsequent posts?
  - Treatment: whether or not a post was censored (at least once in the first half of 2012)
  - Outcomes: censorship during four weeks after a censored post
    1. number of posts
    2. proportion of censored posts
    3. proportion of removed accounts
  - text-as-confounder: contents of posts
- Original analysis: Matching (CEM) with topic proportions (STM) and propensity score (inverse regression)
- Our reanalysis:
  - Text reuse with Llama 3
  - Apply the deconfounder method:
    1. entire sample (4155 users; 75324 posts)
    2. matched samlpe (628 users; 879 posts)

ATE Estimates



Residual Correlations

- Proportion of top 30 censorship related keywords from Fu et al. (2013)

Outcome	**Deconfouder**	**TIRM (Roberts et al.)**
Number of Posts	-0.010	0.005
	[-0.076, 0.056]	[-0.072, 0.060]
Rate of censorship	-0.023	0.126
	[-0.089, 0.043]	[0.060, 0.190]
Rate of missing posts	-0.022	0.024
	[-0.089, 0.043]	[-0.042, 0.090]

Concluding Remarks

- Generative AI can be used to improve causal inference
  - generate treatments at scale
  - enables the extraction of true internal representation
  - better causal representation learning
- Open-source software GPI (GenAI Powered Inference) available at `https://gpi-pack.github.io/`

- Further extensions
  - images and videos
  - tabular data too
  - interpretation of estimated deconfounder
  - discovery of treatment concepts
  - policy learning with unstructured treatments