GenAI-Powered Inference

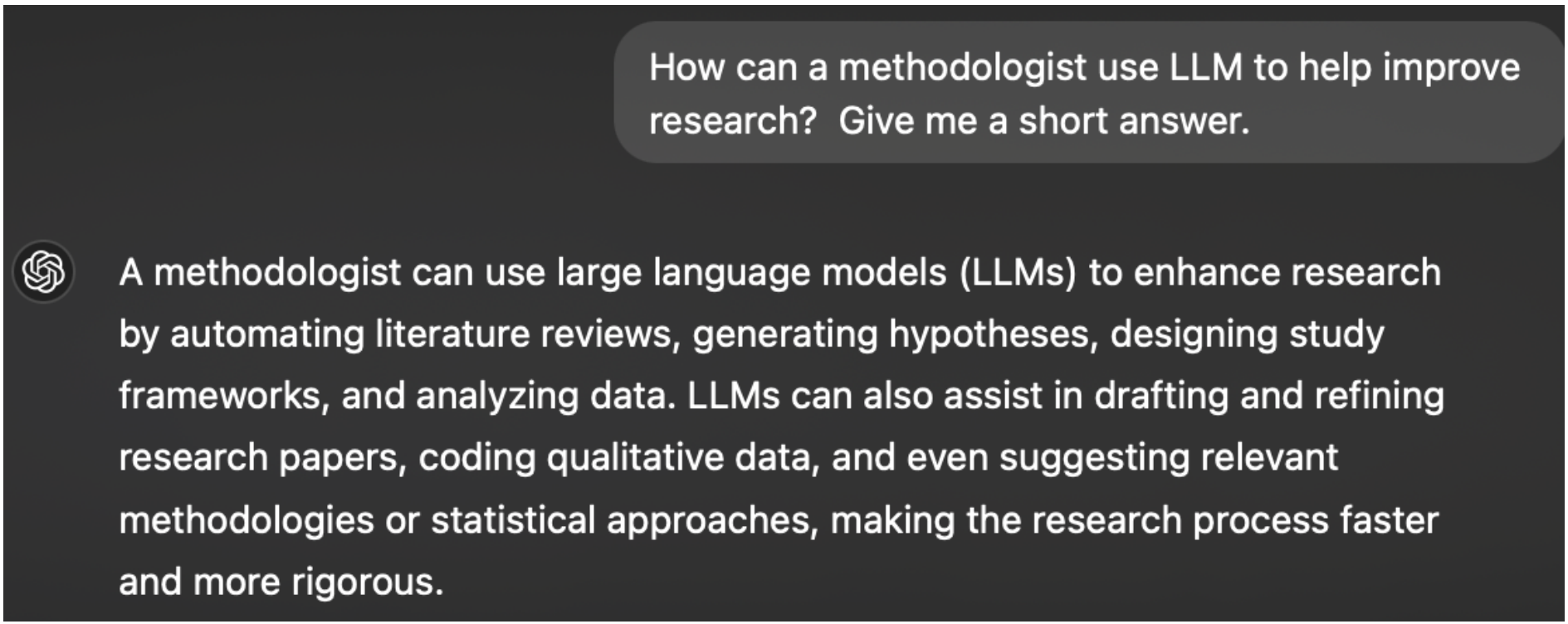
Kosuke Imai Kentaro Nakamura

Harvard University

Society for Political Methodology Annual Meeting
Emory University
June 17, 2025

Motivation

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



Using GenAI to Improve Statistical Inference

- GenAI-Powered Inference (GPI)
- GenAI-assisted statistical/causal inference with unstructured data
  1. (re)generate unstructured data at scale
  2. obtain true internal representation from GenAI
  3. use it directly for machine learning without fine tuning

- Advantages:
  - no need to estimate representation
  - avoid functional form assumptions
  - better empirical performance

GenAI: Definition and Assumption

- Deep generative model:

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

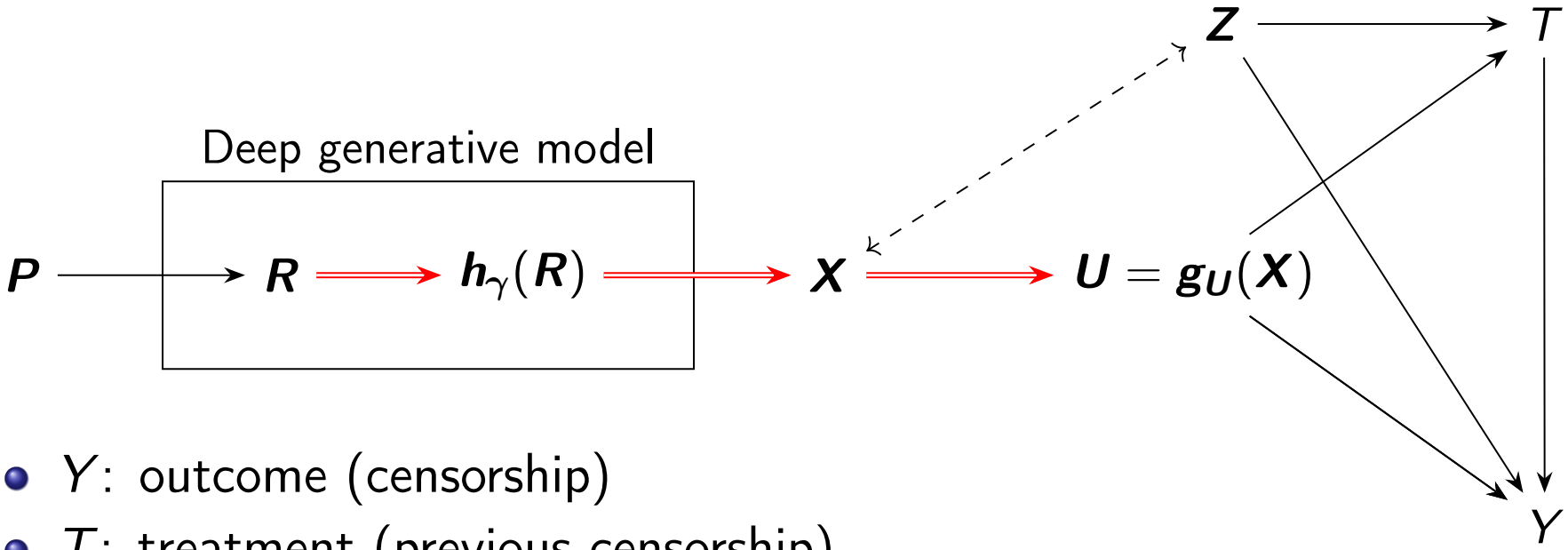
  - $\boldsymbol{P}_i$: prompt
  - $\boldsymbol{X}_i$: texts or images
  - $\boldsymbol{R}_i$: hidden states or internal representations
  - $\boldsymbol{h}_\gamma(\boldsymbol{R}_i)$: deterministic function from hidden states to the last layer

- Deterministic decoding: $\mathbb{P}(\boldsymbol{X}_i \mid \boldsymbol{h}_\gamma(\boldsymbol{R}_i))$ is degenerate
- Use of open-source GenAI

Text as Confounder: Chinese Censorship (Roberts *et al.* 2020)

- Do Chinese social media users who had their post censored become more likely to be censored for later posts or self-censor themselves?
  - Treatment: whether or not a post was censored
  - Outcomes: censorship during four weeks after a censored post
    1. number of posts
    2. proportion of censored posts
    3. proportion of missing posts
  - text-as-confounder: contents of posts
- Original analysis: Matching (CEM) with topic proportions (STM) and treatment projection
- Our reanalysis:
  - Regenerate texts with Llama 3.1 (8 billion) and Gemma3 (1 billion)
  - Apply the proposed method:
    1. entire sample (4155 users; 75324 Weibo posts)
    2. matched sample (628 users; 879 posts)

Assumptions



- Y: outcome (censorship)
- T: treatment (previous censorship)
- $\boldsymbol{Z}$: observed structured confounding variables
- $\boldsymbol{X}$: unstructured objects
- $\boldsymbol{U} = \boldsymbol{g_U}(\boldsymbol{X})$: unknown confounding features

- Strong latent ignorability:

$$\{Y_i(t)\}_{t\in\mathcal{T}} \perp\!\!\!\perp T_i \mid \boldsymbol{Z}_i = \boldsymbol{z}, \boldsymbol{U}_i = \boldsymbol{u}, \quad \text{for all } \boldsymbol{z} \in \mathcal{Z}, \boldsymbol{u} \in \mathcal{U}$$

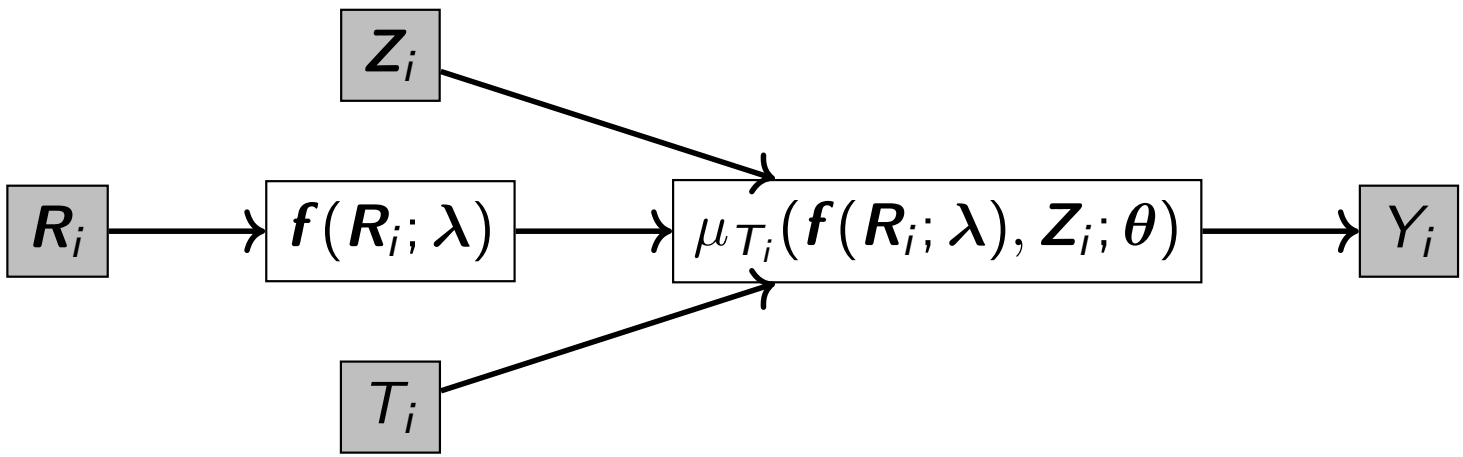
$$\mathbb{P}(T_i = t \mid \boldsymbol{Z}_i = \boldsymbol{z}, \boldsymbol{U}_i = \boldsymbol{u}) > 0 \quad \text{for all } t \in \mathcal{T}, \boldsymbol{z} \in \mathcal{Z}, \boldsymbol{u} \in \mathcal{U}$$

Identification, Estimation, and Inference

- There exists a deconfounder $\boldsymbol{f} : \mathbb{R}^r \mapsto \mathbb{R}^q$ with $q \leq r$ that satisfies:

$$Y_i \perp\!\!\!\perp \boldsymbol{R}_i \mid T_i, \boldsymbol{Z}_i, \boldsymbol{f}(\boldsymbol{R}_i)$$

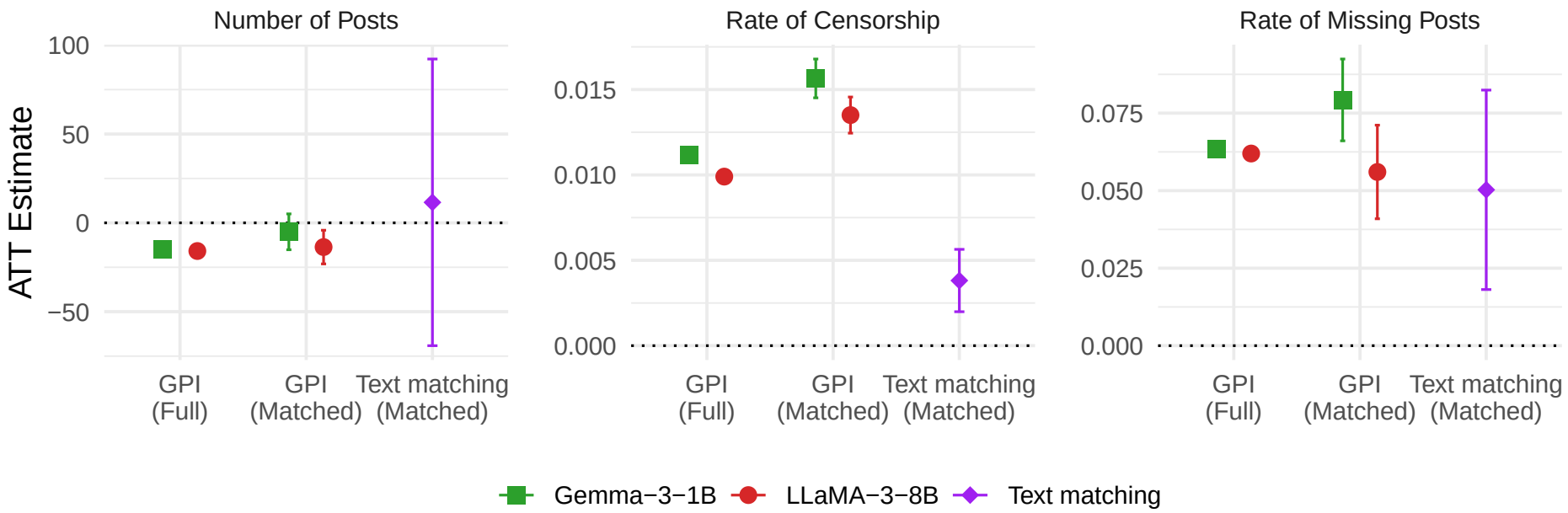
- Adjusting for the deconfounder $\boldsymbol{f}(\boldsymbol{R}_i)$ and the observed confounder $\boldsymbol{Z}$ identifies the ATE $\mathbb{E}[Y_i(1, \boldsymbol{U}_i) - Y_i(0, \boldsymbol{U}_i)]$
- Estimation via neural network



1. Estimate the deconfounder and outcome model
2. Estimate the propensity score given the estimated deconfounder

- Inference via double machine learning (DML)

Empirical Results



- Our analysis shows higher rates of censorship and self-censorship
- Full sample analysis is much more efficient
- Similar estimates across LLMs

Efficient Score Correlations with Candidate Confounder

- Confounder: proportion of 60 censorship related keywords (Fu et al. 2013)
- Calculate the efficient score
- Pearson's correlation with the confounder
  - If confounders are properly controlled, the efficient score should be uncorrelated with the keywords

	GPI (LLaMA3-8B)		**Text matching**
Outcome	Full	Matched	Matched
Number of posts	0.004	0.073	0.032
	(0.273)	(0.094)	(0.398)
Rate of censorship	−0.001	−0.016	0.089
	(0.783)	(0.587)	(0.001)
Rate of missing posts	0.001	0.007	0.062
	(0.735)	(0.826)	(0.063)

- Similar results for Gemma3-1B

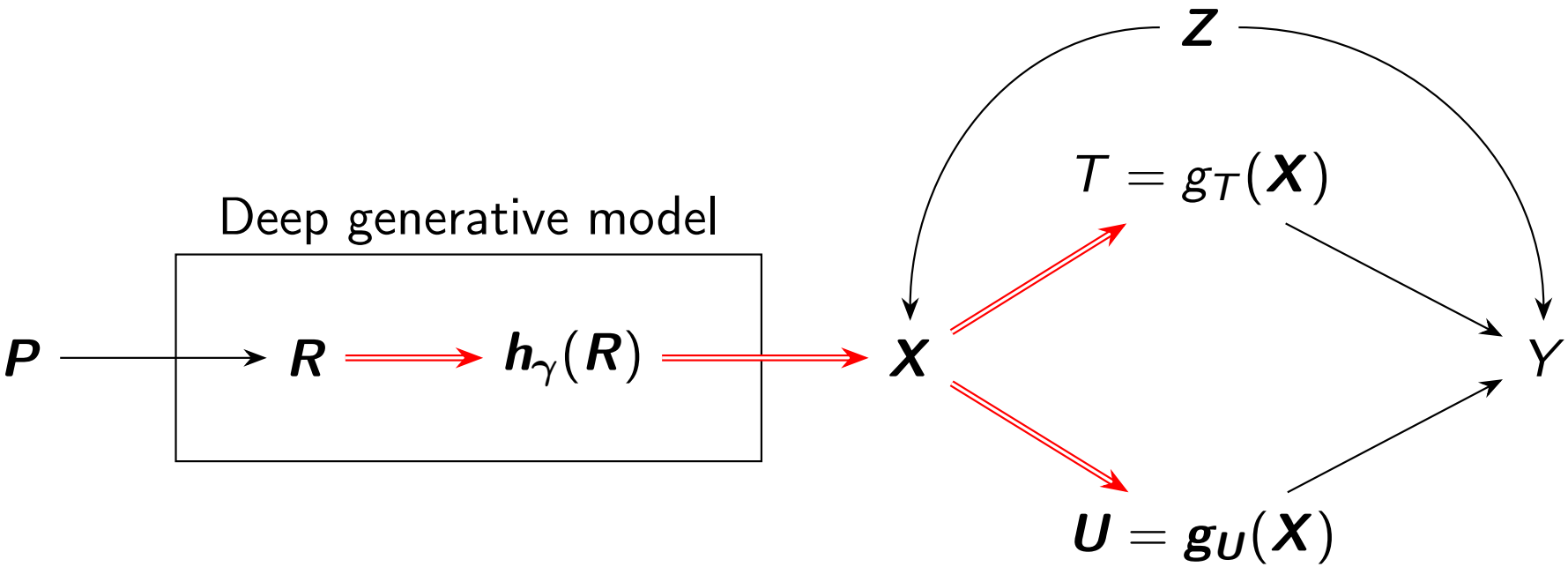
Image as Treatment: Facial Features and Election Results

(Lindholm et al. 2024)

- How does the visual appearance of political candidate predict their electoral success?
- Data: 7,080 Danish politicians with candidate photos

- Treatment variables: facial features (discretized into 10 quantile bins)
  1. attractiveness
  2. trustworthiness
  3. dominance
- Outcome: Election results (number of votes standardized via z-score)
- Structured confounding variables: age, gender, education
- We wish to adjust other facial confounding features

Assumptions



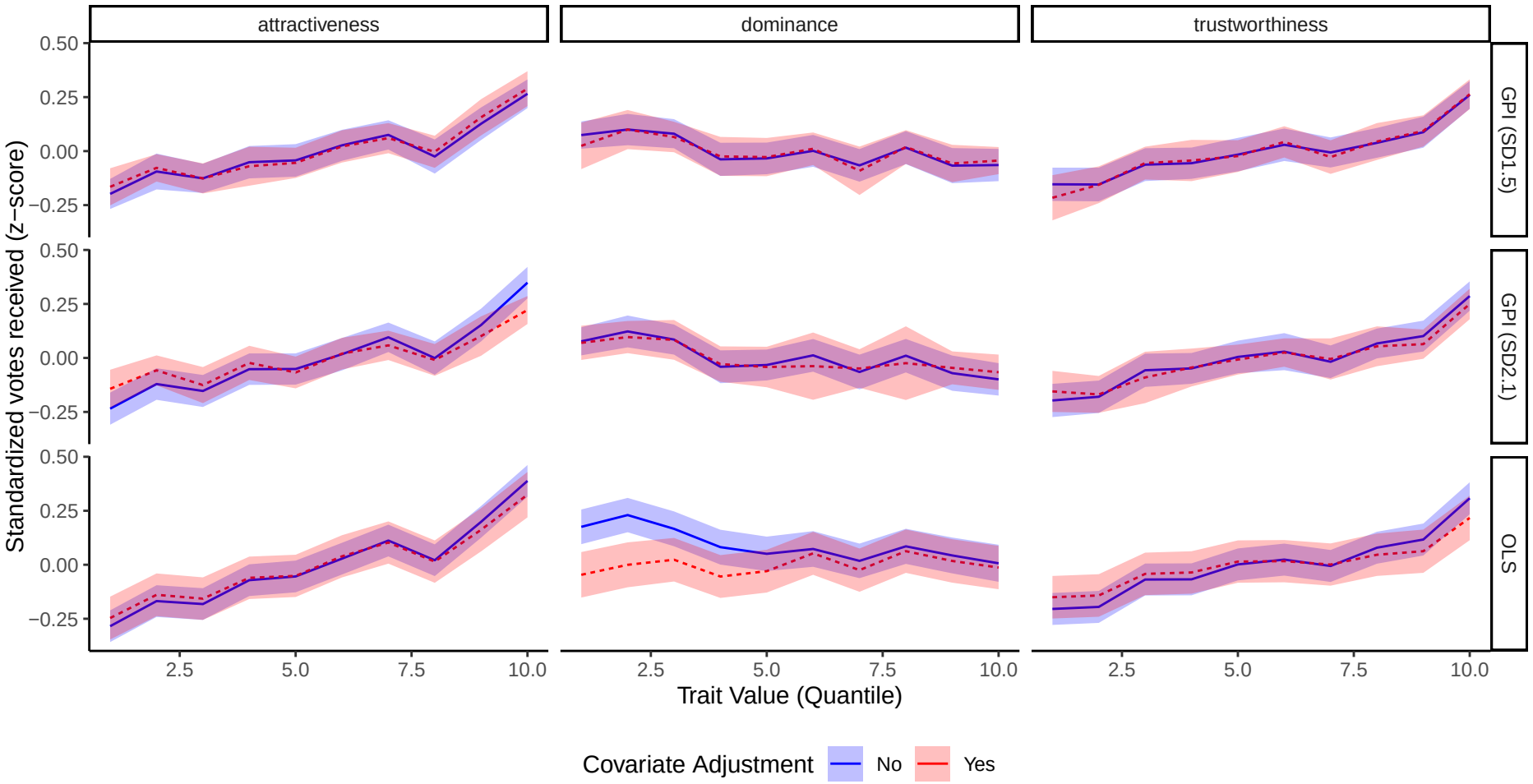
- Separability:

$$Y_i(X_i) = Y_i(g_T(X_i), g_U(X_i)) = Y_i(T_i, U_i)$$

- Lemma: separability implies overlap

$$\mathbb{P}(T_i = t \mid U_i = u, Z_i = z) > 0 \quad \text{for all } u \in \mathcal{U}, z \in \mathcal{Z}$$

Empirical Results



- Unlike OLS, the proposed method is not sensitive to the inclusion of structured confounding variables

Text as Treatment: Persuasion and Rhetoric

(Blumenau and Lauderdale, 2022)

- Which types of political rhetorics are most persuasive?

- Forced choice conjoint experiment with texts
- Total of 336 political arguments
  - 12 policy issues: tuition fees, fracking, etc.
  - 14 rhetorical elements: cost and benefit, morality, etc.
  - 2 sides: for or against

- Outcome: Persuasiveness of arguments
  - one argument is more persuasive than the other
  - equally persuasive

Example Text Pair

- Policy topic: building a third runaway at Heathrow:

Appeal to authority / For The Airports Commission, an independent body established to study the issue, have argued that expanding Heathrow is the most effective option to address the UK's aviation capacity challenge
Appeal to history / Against History show us that most large infrastructure projects do not lead to significant economic growth, which suggests that the expansion of Heathrow will fail to pay for itself

- Can we adjust for the unstructured confounding features of texts?

The Structural Model

- The original Bradley-Terry model:

$$\log\left[\frac{\mathbb{P}(Y_{jj'(i)} \leq k)}{\mathbb{P}(Y_{jj'(i)} > k)}\right] = \delta_k + \left(\alpha_{P_j S_j} + \beta_{T_j} + \gamma_j\right) - \left(\alpha_{P_{j'} S_{j'}} + \beta_{T_{j'}} + \gamma_{j'}\right)$$

  where i indexes respondents, j indexes arguments, P_j denotes policy area, S_j denotes for/against, and T_j denotes rhetoric

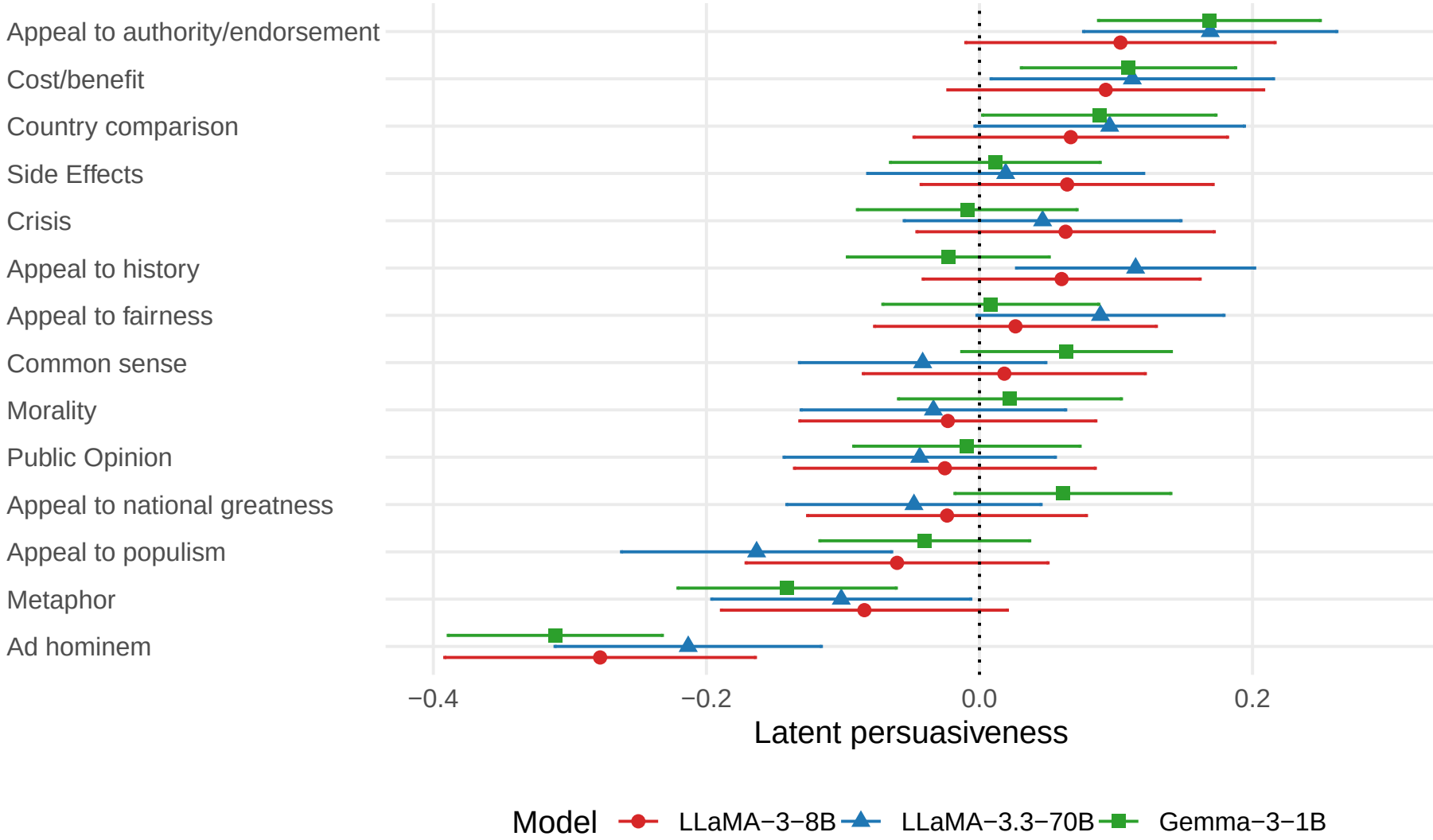
- Our semiparametric model:

$$\log\left[\frac{\mathbb{P}(Y_{jj'(i)} \leq k)}{\mathbb{P}(Y_{jj'(i)} > k)}\right] = \delta_k + \mu(T_j, \boldsymbol{U}_j) - \mu(T_{j'}, \boldsymbol{U}_{j'})$$

- Persuasiveness of rhetoric $T_j = t$

$$\beta(t) := \mathbb{E}[\mu(t, \boldsymbol{U}_j)]$$

- Estimate $\beta(t)$ using the deconfounder $\boldsymbol{f}(\boldsymbol{R}_j)$



- Stronger effects for ad hominem, appeal to authority, and cost/benefit
- Similar effects across models

Concluding Remarks

- Generative AI can be used to improve causal/statistical inference
  - can generate unstructured objects at scale
  - enables the extraction of true internal representation
  - more robust and efficient causal/statistical inference

- Paper:

  https://imai.fas.harvard.edu/research/GPI.html

- Open-source software GPI (GenAI Powered Inference):

  https://gpi-pack.github.io/

- Further extensions
  - causal inference with multimodal data (e.g., videos)
  - interpretation of estimated deconfounder
  - discovery of treatment concepts
  - policy learning with unstructured treatments