GenAI-Powered Causal Inference

Kosuke Imai

Harvard University

Seminar Talk at Department of Government
University of Vienna
June 16, 2025

Joint work with Kentaro Nakamura (Harvard Kennedy School)

Motivation

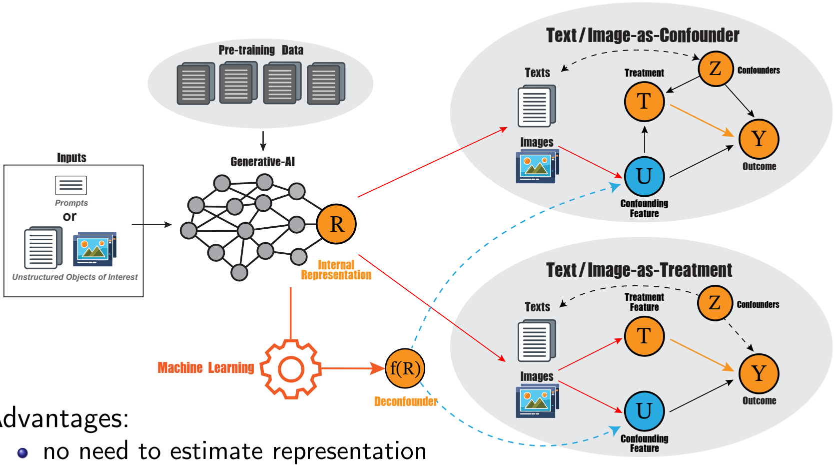
- Generative AI is transforming medicine, education, marketing, etc.
- Can methodologists get some help from GenAI too?

How can a methodologist use LLM to help improve research? Give me a short answer.

A methodologist can use large language models (LLMs) to enhance research by automating literature reviews, generating hypotheses, designing study frameworks, and analyzing data. LLMs can also assist in drafting and refining research papers, coding qualitative data, and even suggesting relevant methodologies or statistical approaches, making the research process faster and more rigorous.

Using Generative AI to Improve Causal Inference

- GenAI-Powered Causal Inference (GPI)
- GenAI-assisted causal inference with unstructured data



- 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
  - $\boldsymbol{X}_i$: unstructured generated
  - $\boldsymbol{R}_i$: 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 for replicability

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
  - structural confounders: lagged outcomes, date of the post (dummies)
  - 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 proposed method:
    1. entire sample (4155 users; 75324 Weibo posts)
    2. matched sample (628 users; 879 posts)

Assumptions

Deep generative model

$$\begin{tikzcd}
 & & & & \boldsymbol{Z} \arrow[r] \arrow[ddr] & T \arrow[dd] \\
\boldsymbol{P} \arrow[r] & \boldsymbol{R} \arrow[r, Rightarrow] & \boldsymbol{h}_\gamma(\boldsymbol{R}) \arrow[r, Rightarrow] & \boldsymbol{X} \arrow[r, Rightarrow] \arrow[ur, dashed, leftrightarrow] & \boldsymbol{U} = \boldsymbol{g}_{\boldsymbol{U}}(\boldsymbol{X}) \arrow[ur] \arrow[dr] & \\
 & & & & & Y
\end{tikzcd}$$

- Y: outcome (censorship)
- T: treatment (previous censorship)
- $\boldsymbol{Z}$: observed structured confounding variables
- $\boldsymbol{X}$: unstructured confounding object
- $\boldsymbol{U} = \boldsymbol{g}_{\boldsymbol{U}}(\boldsymbol{X})$: unknown and unstructured confounding variables

- 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

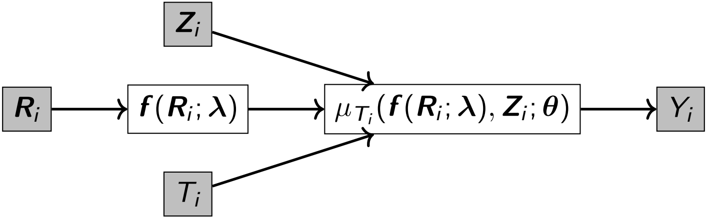
- There exists a deconfounder $\boldsymbol{f} : \mathbb{R}^r \mapsto \mathbb{R}^q$ with $q \leq r$ that satisfies:

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

- Adjusting for the deconfounder $\boldsymbol{f}(\boldsymbol{R}_i)$ and the observed confounder $\boldsymbol{Z}$ identifies the marginal distribution of any potential outcome $Y(t)$:

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

Estimation via Neural Network



- Conditional expectation function:

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

- Loss function for the outcome model and deconfounder:

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

- Estimate the propensity score using the estimated deconfounder

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

Double Machine Learning (Chernozhukov et al. 2018)

- Cross-fitting for the binary treatment case:
  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 model and deconfounder
     3. use the second subset to estimate propensity score given deconfounder
  3. Compute the ATE estimator as:

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

- Double robustness, asymptotic normality

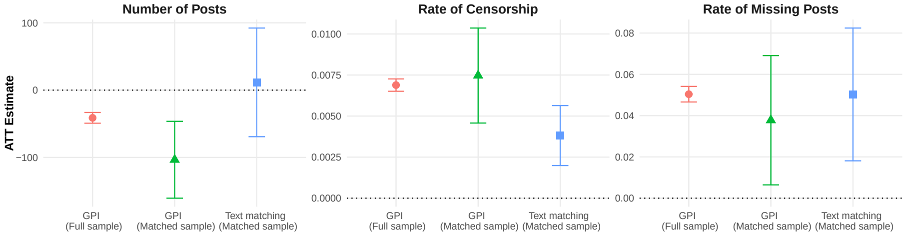
Empirical Analysis

- Reproduced all the texts using open-source LLaMa3–8B
- Internal representation: last token of the final layer, $\dim(\boldsymbol{R}) = 4080$

- Automated hyperparameter tuning via Optuna (Akiba et al. 2019)
  - $\dim(\boldsymbol{f}(\boldsymbol{R})) = 2048$
  - depth of hidden layers $= 2$
  - size of hidden layers after deconfounder $= [50, 1]$

- 2-fold cross-fitting:
  - clustered standard errors at the user level
  - truncation of extreme propensity scores (Dorn, 2025)

Empirical Results



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

Residual Correlations with Candidate Confounder

- Confounder: proportion of 30 censorship related keywords (Fu et al. 2013)
- Extract the residuals from each method
- Compute Spearman's rank correlation with the confounder and *p*-value

	GenAI-Powered Inference		**Matching**
Outcome	Full sample	Matched sample	Matched sample
Number of posts	0.005	−0.027	0.022
	(0.330)	(0.476)	(0.353)
Rate of censorship	−0.003	0.017	0.071
	(0.421)	(0.647)	(0.005)
Rate of missing posts	−0.002	−0.025	0.043
	(0.653)	(0.504)	(0.062)

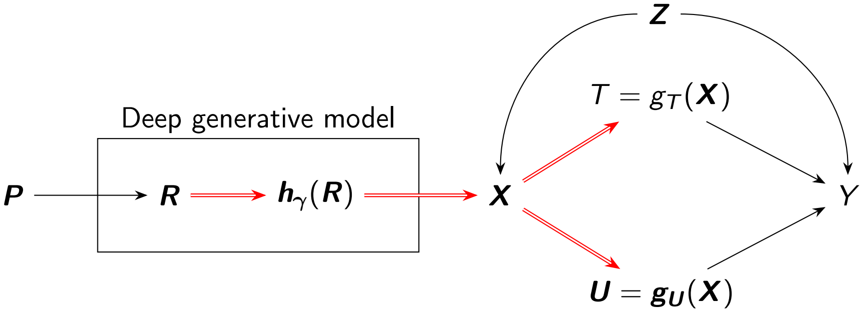
Image as Treatment: Facial Features and Election Results

(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 (continuous scores)
  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:

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

- Lemma: separability implies overlap

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

Identification, Estimation, and Inference

- Identification
  - Existence of (possibly non-unique) deconfounder
  - Adjusting for the deconfounder yields nonparametric identification

- Estimation and inference (same as before)
  1. estimate the outcome models and deconfounder via Neural Network
  2. estimate the propensity score using the estimated Deconfounder
  3. inference via Double Machine Learning

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

Empirical Analysis

- Reproduce all images using Stable diffusion (ver. 1.5)
- Original image:

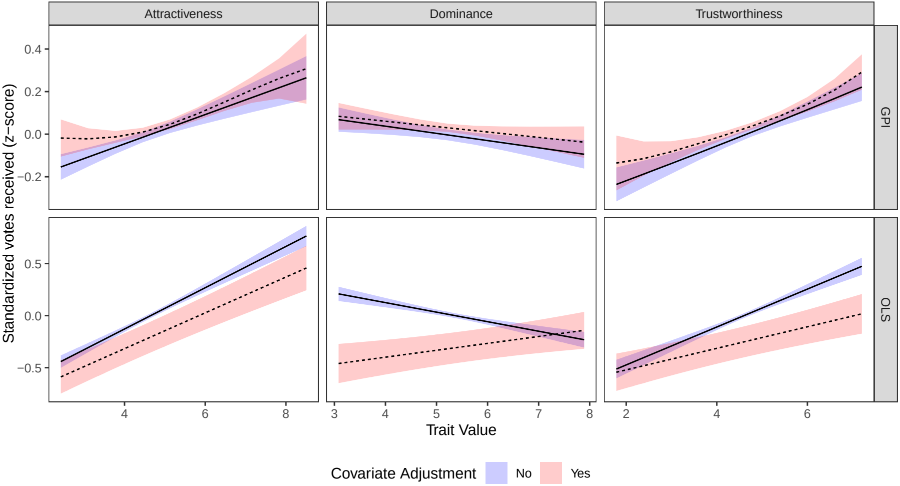
$$\dim(\boldsymbol{X}) = 304(\text{width}) \times 304(\text{height}) \times 3(\text{RGB}) = 277248$$

- Internal representation: $\dim(\boldsymbol{R}) = 16384$
- Neural network architecture:
  - $\dim(\boldsymbol{f}(\boldsymbol{R})) = 1024$
  - depth of hidden layers $= 2$
  - size of hidden layers after deconfounder $= [200, 1]$
- Nonparametrically estimate the average effect curve

$$\xi_t := \mathbb{E}[Y_i(t, \boldsymbol{U}_i)]$$

- Doubly-robust pseudo-outcome approach (Kennedy et al. 2017)

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.
  - 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:

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

  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 $T_j = t$

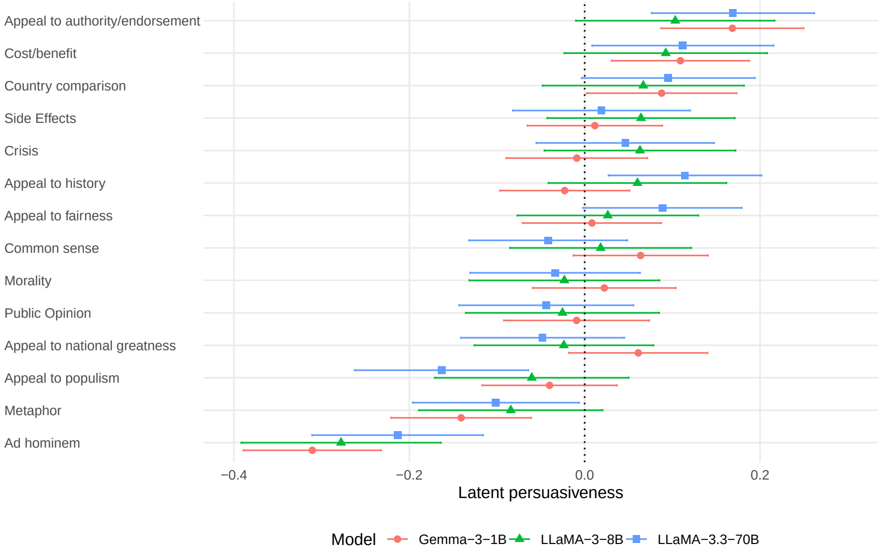
$$\beta(t) := \mathbb{E}[\mu(t, \boldsymbol{U}_j)]$$

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

Empirical Analysis

- Reproduce all texts using Llama3–8B
- Internal representation: last token of the final layer, $\dim(\boldsymbol{R}) = 4096$
- Neural network architecture:
  - $\dim(\boldsymbol{f}(\boldsymbol{R})) = 1024$
  - depth of hidden layers $= 2$
  - size of hidden layers after deconfounder $= [200, 1]$
- Quantify uncertainty via Monte Carlo dropout (Gal and Ghahramani 2016)

- We also use newer models: Llama3.3–70B and Gemma3–1B



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

Concluding Remarks

- Generative AI can be used to improve causal inference
  - can generate treatments at scale
  - enables the extraction of true internal representation
  - better causal representation learning
- Open-source software GPI (GenAI Powered Inference) is available at `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