Generative AI Powered Inference

Kosuke Imai

Harvard University

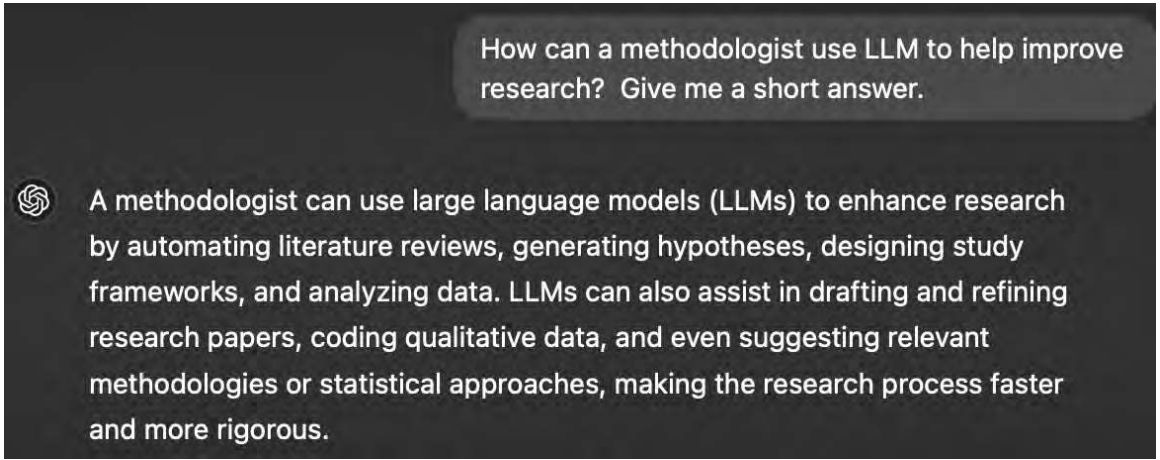
Keynote Talk
CLASS Advanced Methods School (CAMS)
City University of Hong Kong

August 10, 2026

Joint work with Kentaro Nakamura, Adam Breuer, Bryce Dietrich, and Mike Crespin

Motivation

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



GenAI-Powered Inference (GPI)

- Causal inference framework for unstructured data (text, image, video, etc.)
  1. leverage a GenAI model to (re)generate the unstructured data
  2. extract its internal representations
  3. estimate feature-specific causal effects using extracted representations

- Advantages
  1. representations contain *all* the information of generated objects in a low-dimensional format
  2. use neural network to avoid functional-form assumptions
  3. asymptotic statistical guarantees with good performance under moderate sample sizes

- Applications for today
  1. text as treatment (Imai and Nakamura, *JASA*, forthcoming)
  2. video as treatment (Nakamura et al. Working paper)
- Other applications (Imai and Nakamura, *PNAS*, Forthcoming)

Motivating Application: Text-as-Treatment

- 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 $[0, 100]$ 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 fine-tuned BERT embedding (e.g., Veitch et al. 2020; Pryzant et al. 2021; Gui and Veitch, 2023)

Example Biographies

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.

Setup

- Notation
  - $Y_i(\boldsymbol{x})$: Potential outcome when exposed to treatment object $\boldsymbol{x}$
  - $Y_i = Y_i(\boldsymbol{X}_i)$: 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
  1. Treatment Feature:
  $$T_i \;=\; g_T(\boldsymbol{X}_i)$$
  2. Confounding Features:
  $$\boldsymbol{U}_i \;=\; \boldsymbol{g_U}(\boldsymbol{X}_i) \quad \text{where } \dim(\boldsymbol{U}_i) \ll \dim(\boldsymbol{X}_i)$$
  3. Separability:
  $$Y_i(\boldsymbol{x}) = Y_i(g_T(\boldsymbol{x}), \boldsymbol{g_U}(\boldsymbol{x})),$$
     1. T is not a function of $\boldsymbol{U}$
     2. $\boldsymbol{U}$ is not a function of T

Generative AI: 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 object
  - $\boldsymbol{R}_i$: hidden states or internal representations
  - $\boldsymbol{h}_\gamma(\boldsymbol{R}_i)$: deterministic function from hidden states to the last layer

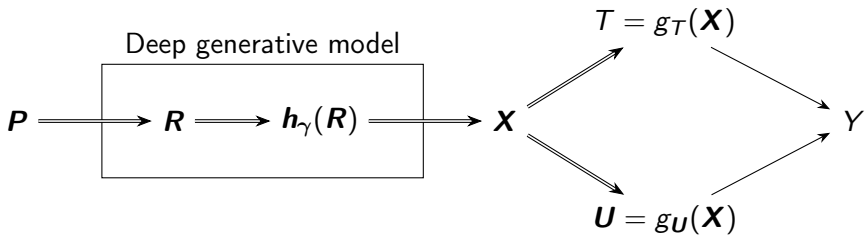
- Deterministic decoding:

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

  - can be achieved by setting a hyperparameter
  - use of open-source GenAI and deterministic encoding for replicability

Summary of Assumptions

- Assumptions:



- Overlap: The above assumptions imply that for any $t \in \{0, 1\}$ and $\boldsymbol{u} \in \mathcal{U}$, we have

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

Nonparametric Identification

- Average treatment effect (ATE):

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

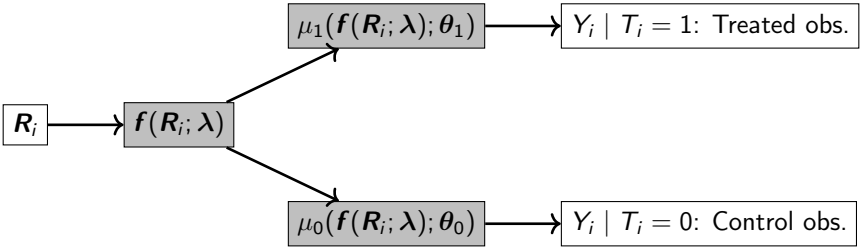
- Under these assumptions, there exists a **Deconfounder** $\boldsymbol{f} : \mathbb{R}^r \to \mathbb{R}^q$ with $q \leq r$ such that

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

  - 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

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

3. Estimate ATE by Double Machine Learning (Chernozhukov et al. 2018)

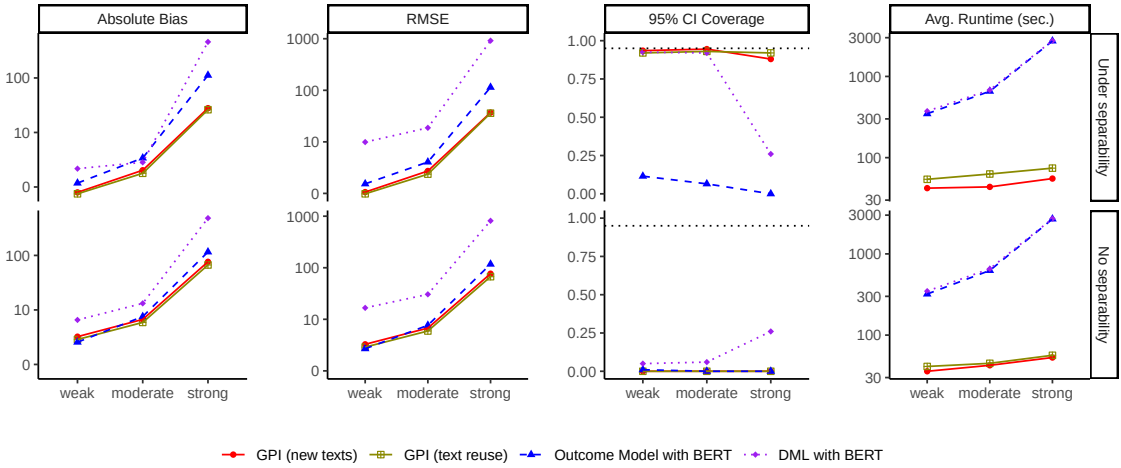
Simulation Study Setup

- 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:

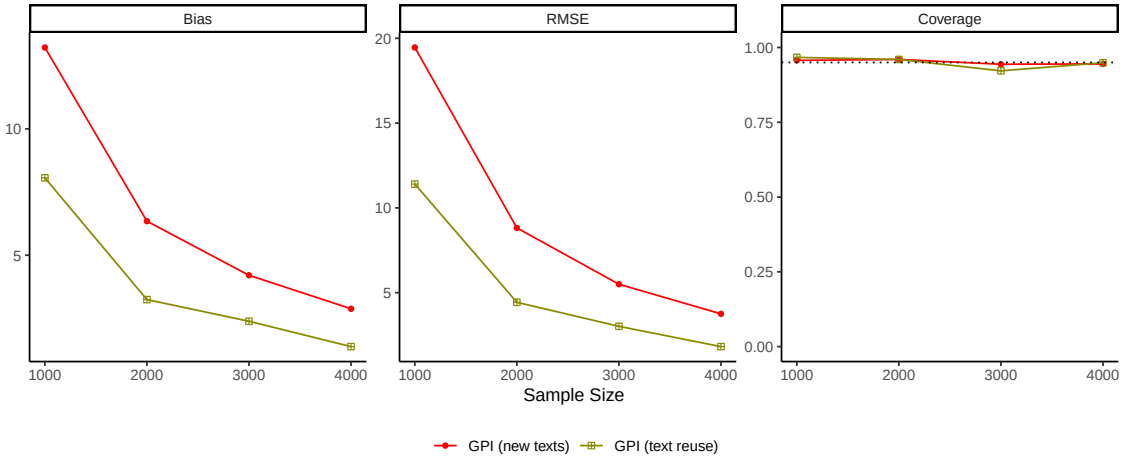
$$Y_i = \alpha_1 T_i + \alpha_2 T_i h_1(\boldsymbol{X}_i) - \alpha_3 h_1(\boldsymbol{X}_i) - \alpha_4 h_2(\boldsymbol{X}_i) + \epsilon_i$$

  where $\epsilon_i \sim \mathcal{N}(\mu_i, 1)$ and
  - 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



Performance across Different Sample Sizes



Empirical Analysis

- 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
Proposed method (reuse)	5.462	[2.790, 8.135]	28.9 sec.
Outcome model with BERT	−2.557	[−2.608, −2.505]	6139.7
DML with BERT	−67.777	[−109.967, −25.587]	6210.3

Motivation: Video-as-treatment

- Video is a dominant medium of modern communication

- How do we conduct causal inference with video?
- Standard approach:
  1. classify videos into treatment and control groups based on features of interest
  2. randomly assign viewers to videos
  3. compare viewers' reactions to treated and control videos
- Challenge: treatment features are confounded with many other visual and auditory features

- How can we identify the causal effects of specific video features?
- How can we estimate their dynamic effects on viewers' reactions over time?
- No causal inference method is available for video data

Empirical Application: Presidential Campaign Advertisement

- How does the candidate's appearance affect a voter's real-time evaluation?

- Data: 2020 campaign videos from the Wesleyan Media Project
  - 849 videos (10–120 seconds; median: 30 seconds)
  - Joe Biden (389 videos) / Donald Trump (128 videos)

- Treatment feature: candidate appearance
  - survey conducted on Prolific
  - asked respondents to press the "R" key when candidates appeared
  - each video coded by up to five coders
  - videos divided into 5-second segments
  - segment is "treated" if a majority of coders identify the candidate

Real-time Evaluation of Campaign Advertisement

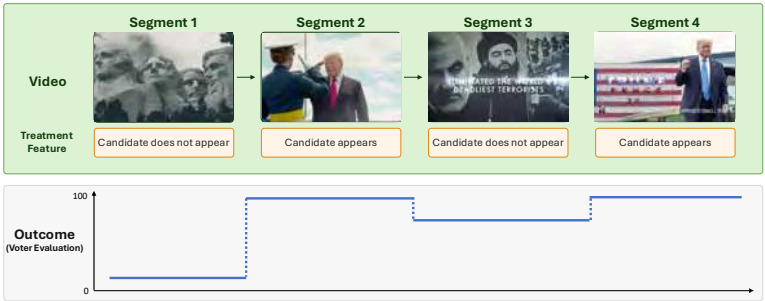
- separate survey conducted on Prolific ($N = 1,108$)
- randomly assign videos to respondents
- respondents continuously evaluated the candidate using a feeling thermometer (0–100)
- dial positions were recorded throughout the video
- each respondent rated on average 9.81 advertisements

Methodological Challenges

1. High-dimensionality of videos
   - even a short video contains many audiovisual features
   - pre-trained embedding is less developed for video data and its quality is unknown
2. Confounding by other video features
   - candidates do not appear in ads randomly
   - e.g., candidates are more likely to appear in ads with positive emotions
   - random assignment of videos is not enough
3. Dynamic real-time outcome
   - viewers' current evaluation might depend on any of the previous frames
   - unknown carryover effects

Can we estimate the causal effects of video features on real-time responses?

Setup



- $\boldsymbol{X}_i$: videos assigned to individual i
- $\boldsymbol{X}_{is}$: s-th segment of video assigned to individual i
- S_i: length of video segments assigned to individual i
- Y_{is}: outcome of respondent i at segment s (e.g., voter evaluation)
- $Y_{is}(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_s)$: potential outcome for individual i at segment s
- W_{is}: Treatment feature of videos assigned to individual i at segment s

Assumptions

1. Consistency: $Y_{is} = Y_{is}(\boldsymbol{X}_{i1}, \cdots, \boldsymbol{X}_{is})$
2. Randomization: $Y_{is}(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_s) \perp\!\!\!\perp \boldsymbol{X}_i$
3. Treatment Feature: $W_{is} = g_W(\boldsymbol{X}_{is})$ (i.e., objective feature rather than perceived one)
4. Confounding Features: Confounding features in $\boldsymbol{X}_{is}$ for $Y_{is'}$ are unknown

$$\boldsymbol{U}_{is}^{(s')} = \boldsymbol{g}_{\boldsymbol{U}_s}^{(s')}(\boldsymbol{X}_{is})$$

5. Sequential Separability:

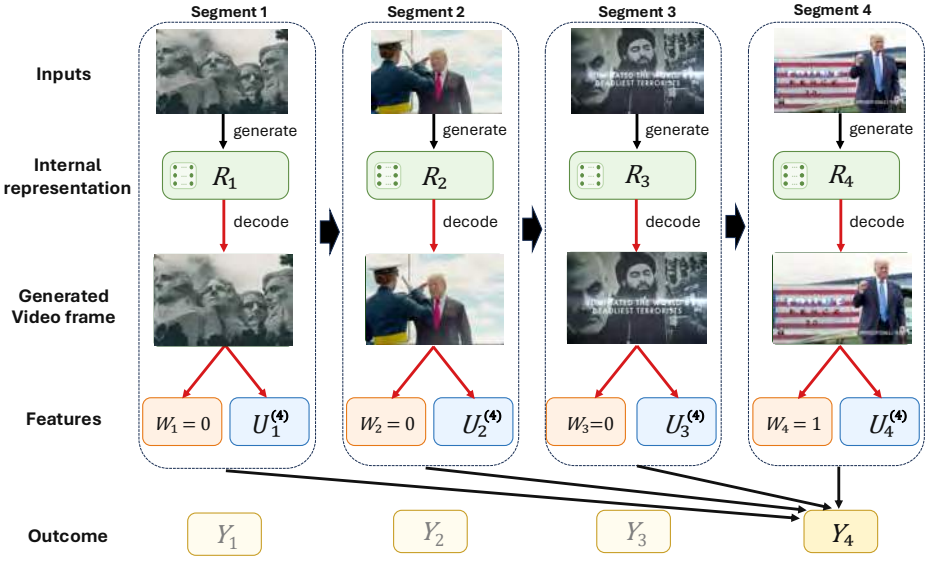
$$Y_{is}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_s) = Y_{is}(g_W(\boldsymbol{x}_1), \boldsymbol{g}_{\boldsymbol{U}_1}^{(s)}(\boldsymbol{x}_1), \ldots, g_W(\boldsymbol{x}_s), \boldsymbol{g}_{\boldsymbol{U}_s}^{(s)}(\boldsymbol{x}_s)).$$

   - we can intervene on the treatment feature while holding the confounding features constant
   - implies positivity given confounding features

6. Factorized deterministic decoding for deep GenAI

Reproducing Video Data from Generative AI

Summary of Assumptions



Stochastic Intervention

- High-dimensional data $\rightsquigarrow$ violation of overlap assumption
- Want our counterfactuals to be close to the observed data
- Stochastic intervention: for each segment s,

$$q_s(\delta_s; \overline{\boldsymbol{w}}_{s-1}) := \frac{\delta_s p_s(\overline{\boldsymbol{w}}_{s-1})}{\delta_s p_s(\overline{\boldsymbol{w}}_{s-1}) + 1 - p_s(\overline{\boldsymbol{w}}_{s-1})}, \quad \delta_s \in (0, \infty)$$

  where $p_s(\overline{\boldsymbol{w}}_{s-1}) = \mathbb{P}(W_{is} = 1 \mid \overline{\boldsymbol{W}}_{i,s-1} = \overline{\boldsymbol{w}}_{s-1}, S_i \geq s)$

- Greater value of $\delta_s \rightarrow$ higher assignment probability
- q_s intervenes only when possible, given the treatment history
  - When $p_s(\overline{\boldsymbol{w}}_{s-1}) = 1$, then intervention q_s always assigns 1
  - When $p_s(\overline{\boldsymbol{w}}_{s-1}) = 0$, then intervention q_s always assigns 0

$\rightarrow$ We only require the positivity given confounding features

Causal Identification

- Average potential outcome for each segment under stochastic intervention:

$$\mathbb{E}\left[\int_{\mathcal{W}^s} Y_{is}(W_{i1} = w_1, \ldots, W_{is} = w_s) \prod_{s'=1}^{s} dQ_{s'}(w_{s'}; \overline{\boldsymbol{w}}_{s'-1}, \delta_{s'})\right]$$

- We can identify the target parameter by adjusting a low-dimensional deconfounder

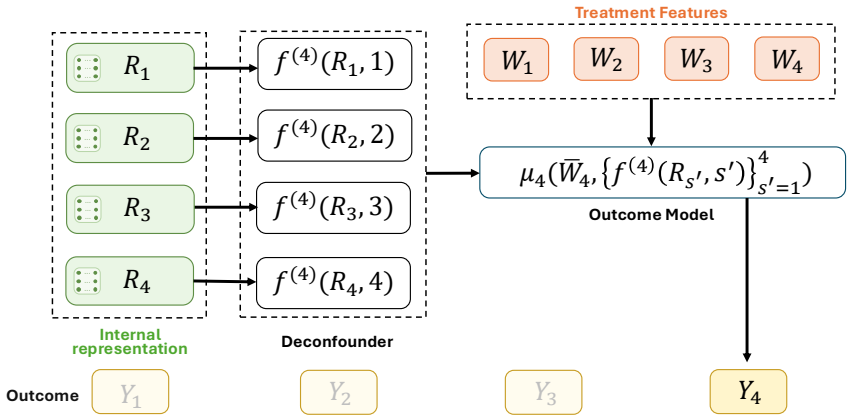
$$\begin{aligned}&\mathbb{E}[Y_{is} \mid \overline{\boldsymbol{W}}_{is}, \boldsymbol{f}_1^{(s)}(\boldsymbol{R}_{i1}), \ldots, \boldsymbol{f}_s^{(s)}(\boldsymbol{R}_{is}), S_i \geq s] \\ &\quad = \mathbb{E}[Y_{is} \mid \overline{\boldsymbol{W}}_{is}, \boldsymbol{f}_1^{(s)}(\boldsymbol{R}_{i1}), \ldots, \boldsymbol{f}_s^{(s)}(\boldsymbol{R}_{is}), \overline{\boldsymbol{R}}_{is}, S_i \geq s]\end{aligned}$$

  - a deconfounder always exists (e.g., confounding features)
  - identification works with any deconfounder

- Direct adjustment of $\boldsymbol{R}_{is}$ leads to overlap violation

Estimation

- Longitudinal neural network architecture with mean independence



- Use longitudinal influence function for estimation

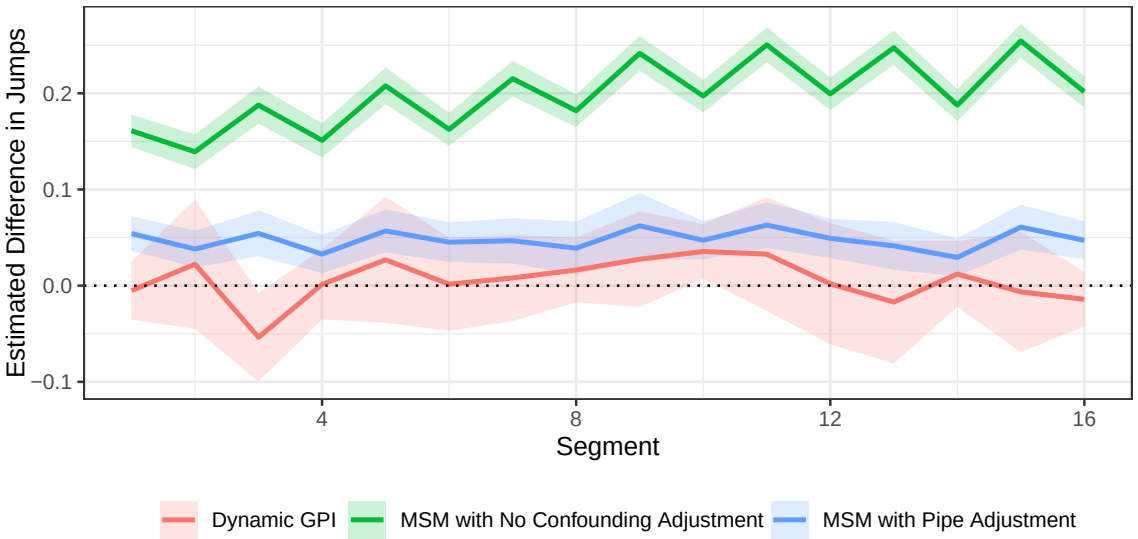
Super Mario Bros.™ Causal Benchmark Dataset

- Setup
  - Treatment: appearance of Princess Peach
  - Confounder: number of pipes
  - Outcome: number of jumps by Mario
- Confounding: Princess Peach appears more frequently when there are more pipes in each segment
- Mario is played by a reinforcement learning agent
  - realistic outcome model (complex black-box policy)
  - Princess Peach does not affect Mario's jumps
  - True causal effect is zero
- $N = 10,000$ videos with equal length and autoscroll

Analysis Setup

- 1-second segments
- Regenerate each video segment by Cosmos Tokenizer
  - original video dimension: 460,800 = (320, 480, 3)
  - dimension of internal representations: 38,400 = (16, 40, 60)
- Estimate the trajectory of the average number of jumps
- contrast between $\delta = 5.0$ (more Peach appearance) and $\delta = 0.5$ (less Peach appearance)
- Three estimators
  1. Dynamic GPI (proposed)
  2. Marginal Structural Models (MSM) adjusting for the number of pipes
  3. Marginal Structural Models without confounding adjustment
- Same interventions and estimation strategies: only difference is confounding adjustment

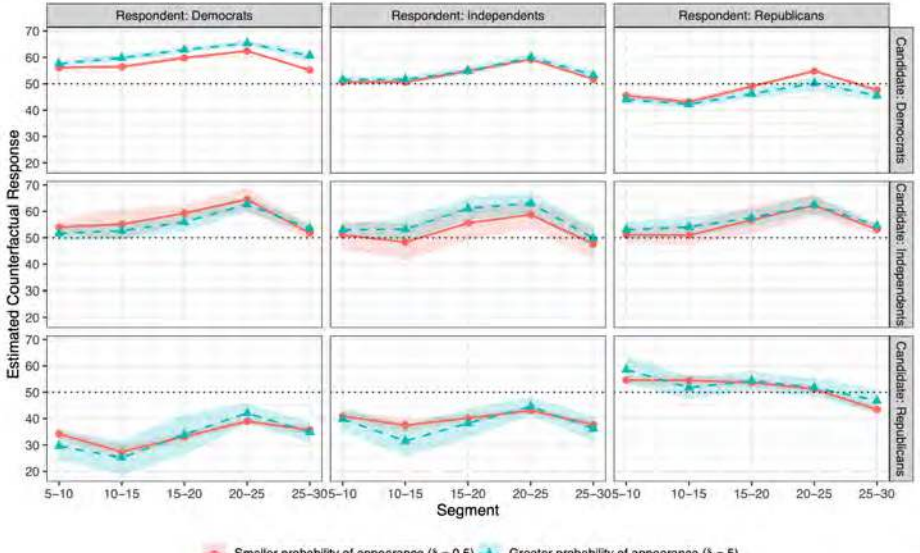
Validation Results



Empirical Application: Setup

- 5-second segment, regenerated by Cosmos tokenizer as before
- For audio data,
  1. use Whisper for auto-transcription with time stamps
  2. use LLaMa3.1 (8B) to regenerate transcripts and obtain internal representations
- Dynamic GPI: $\delta = 0.5$ (less appearance) and $\delta = 5.0$ (more appearance)

Empirical Application: Results



Concluding Remarks

- Generative AI can be used to improve causal inference
  - GenAI can (re)generate high-quality videos
  - Its internal representations contain all the information in a low-dimensional format
  - We can use them to estimate the causal effects of specific video features in a dynamic setting
- Papers:
  1. overview: "GenAI Powered Inference" (*PNAS*, Forthcoming)
  2. technical paper: "Causal Inference with Generative Artificial Intelligence: Application to Texts as Treatments" (*JASA*, Forthcoming)
  3. dynamic GPI: https://arxiv.org/abs/2607.06126
- Open-source software GPI (GenAI Powered Inference):

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

- Further extensions
  - better use of speech / audio information
  - fully continuous time (no segmentation)
  - policy learning with unstructured treatments