Efficient Randomized Experiments Using Foundation Models

Fanny Yang

joint work with

Piersilvio de Bartolomeis, Javier Abad, Guanbo Wang, Konstantin Donhauser, Raymond M. Duch, Issa J. Dahabreh

ETH zürich

HARVARD UNIVERSITY

UNIVERSITY OF OXFORD

Motivation

- Randomized experiments are costly and time-consuming
  - $40,000 average cost per participant of clinical trials
  - 80% of clinical trials fail to reach enrollment targets on time

Motivation

- Randomized experiments are costly and time-consuming
  - \$40,000 average cost per participant of clinical trials
  - 80% of clinical trials fail to reach enrollment targets on time
- Can we utilize (multiple) foundation models trained on (multiple) external data sources?
  - Examples: language models trained on large corpuses, clinical models trained on observational data
  - Could be helpful if external data has relevant information
  - **But...** inferences may not be valid if model predictions are inaccurate

Motivation

- Randomized experiments are costly and time-consuming
  - \$40,000 average cost per participant of clinical trials
  - 80% of clinical trials fail to reach enrollment targets on time
- Can we utilize (multiple) foundation models trained on (multiple) external data sources?
  - Examples: language models trained on large corpuses, clinical models trained on observational data
  - Could be helpful if external data has relevant information
  - **But...** inferences may not be valid if model predictions are inaccurate
- **Our goal**: Reduce required sample size of randomized trials with externally trained models while guaranteeing valid statistical inference

Problem Setting

- **Distribution:** $\mathbb{P}$ over $(X, Y(0), Y(1), Y, A)$
  - $X \in \mathbb{R}^d$ are covariates
  - $Y \in \mathbb{R}$ is the observed outcome (bounded)
  - $Y(0), Y(1) \in \mathbb{R}$ are potential outcomes
  - $A \in \{0, 1\}$ is the treatment indicator

Problem Setting

- **Distribution:** $\mathbb{P}$ over $(X, Y(0), Y(1), Y, A)$
  - $X \in \mathbb{R}^d$ are covariates
  - $Y \in \mathbb{R}$ is the observed outcome (bounded)
  - $Y(0), Y(1) \in \mathbb{R}$ are potential outcomes
  - $A \in \{0, 1\}$ is the treatment indicator
- **Data:** Tuples $Z_i = (X_i, Y_i, A_i)_{i=1}^n$ drawn i.i.d. from $\mathbb{P}$

Problem Setting

- **Distribution:** $\mathbb{P}$ over $(X, Y(0), Y(1), Y, A)$
  - $X \in \mathbb{R}^d$ are covariates
  - $Y \in \mathbb{R}$ is the observed outcome (bounded)
  - $Y(0), Y(1) \in \mathbb{R}$ are potential outcomes
  - $A \in \{0, 1\}$ is the treatment indicator
- **Data:** Tuples $Z_i = (X_i, Y_i, A_i)_{i=1}^n$ drawn i.i.d. from $\mathbb{P}$
- **Task:** Efficiently estimate $\theta := \mathbb{E}[Y(1) - Y(0)]$

Simple unbiased mean estimators for RCT

- In RCT, the treatment probability is known: $\pi = P(A = 1)$
- Simplest difference-in-mean estimator

$$\widehat{\theta}_{\mathrm{DM}} = \frac{1}{n_1} \sum_{i:A_i=1} Y_i - \frac{1}{n_0} \sum_{i:A_i=0} Y_i, \quad \text{where} \quad n_a = |\{i : A_i = a\}|$$

Simple unbiased mean estimators for RCT

- In RCT, the treatment probability is known: $\pi = P(A = 1)$
- Simplest difference-in-mean estimator

$$\widehat{\theta}_{\mathrm{DM}} = \frac{1}{n_1} \sum_{i:A_i=1} Y_i - \frac{1}{n_0} \sum_{i:A_i=0} Y_i, \quad \text{where} \quad n_a = |\{i : A_i = a\}|$$

- basically equivalent to IPW estimator (with known π)

$$\widehat{\theta}_{IPW} = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{Y_i A_i}{\pi} - \frac{Y_i(1 - A_i)}{1 - \pi} \right] \quad \text{where} \quad n_a = |\{i : A_i = a\}|$$

Simple unbiased mean estimators for RCT

- In RCT, the treatment probability is known: $\pi = P(A = 1)$
- Simplest difference-in-mean estimator

$$\widehat{\theta}_{\mathrm{DM}} = \frac{1}{n_1} \sum_{i:A_i=1} Y_i - \frac{1}{n_0} \sum_{i:A_i=0} Y_i, \quad \text{where} \quad n_a = |\{i : A_i = a\}|$$

- basically equivalent to IPW estimator (with known π)

$$\widehat{\theta}_{IPW} = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{Y_i A_i}{\pi} - \frac{Y_i(1 - A_i)}{1 - \pi} \right] \quad \text{where} \quad n_a = |\{i : A_i = a\}|$$

Leverage **availability of covariates and multiple data sources** in practice $\rightarrow$ smaller but (asymptotically) valid confidence intervals?

Imputing missing data with predictive models

Main idea: **If we had a predictive model $\hat{h}$, we can apply it to predict the counterfactuals / unobserved outcomes for each i**

$$\widehat{\theta}_{\text{AIPW}}(\hat{h}) = \frac{1}{n}\sum_{i=1}^{n}\frac{A_i}{\pi}(Y_i - \hat{h}(X_i,1)) + \frac{1}{n}\sum_{i=1}^{n}\hat{h}(X_i,1) - \left[\frac{1}{n}\sum_{i=1}^{n}\frac{(1-A_i)}{(1-\pi)}(Y_i - \hat{h}(X_i,0)) + \frac{1}{n}\sum_{i=1}^{n}\hat{h}(X_i,0)\right]$$

- Introduced as Augmented Inverse Propensity Weighted (AIPW) estimator by Robins et al. '94 where $\hat{h}$ are **trained on RCT**

Imputing missing data with predictive models

Main idea: **If we had a predictive model $\hat{h}$, we can apply it to predict the counterfactuals / unobserved outcomes for each i**

$$\widehat{\theta}_{\mathrm{AIPW}}(\hat{h}) = \frac{1}{n}\sum_{i=1}^{n}\frac{A_i}{\pi}(Y_i - \hat{h}(X_i,1)) + \frac{1}{n}\sum_{i=1}^{n}\hat{h}(X_i,1) - \left[\frac{1}{n}\sum_{i=1}^{n}\frac{(1-A_i)}{(1-\pi)}(Y_i - \hat{h}(X_i,0)) + \frac{1}{n}\sum_{i=1}^{n}\hat{h}(X_i,0)\right]$$

- Introduced as Augmented Inverse Propensity Weighted (AIPW) estimator by Robins et al. '94 where $\hat{h}$ are **trained on RCT**
- Similar to PPI-style estimators as in Angelopoulos et al. '23 when $\hat{h}$ **can be any model**

Imputing missing data with predictive models

Main idea: **If we had a predictive model $\hat{h}$, we can apply it to predict the counterfactuals / unobserved outcomes for each i**

$$\widehat{\theta}_{\mathrm{AIPW}}(\hat{h}) = \frac{1}{n}\sum_{i=1}^{n}\frac{A_i}{\pi}(Y_i - \hat{h}(X_i,1)) + \frac{1}{n}\sum_{i=1}^{n}\hat{h}(X_i,1)$$
$$- \left[\frac{1}{n}\sum_{i=1}^{n}\frac{(1-A_i)}{(1-\pi)}(Y_i - \hat{h}(X_i,0)) + \frac{1}{n}\sum_{i=1}^{n}\hat{h}(X_i,0)\right]$$

- Introduced as Augmented Inverse Propensity Weighted (AIPW) estimator by Robins et al. '94 where $\hat{h}$ are **trained on RCT**
- Similar to PPI-style estimators as in Angelopoulos et al. '23 when $\hat{h}$ **can be any model**
- In the analysis we use the influence function ψ defined as $\widehat{\theta}_{\mathrm{AIPW}}(g) = \frac{1}{n}\sum_{i=1}^{n}\psi(Z_i; g)$ where $Z = (X, Y, A)$.

Standard AIPW using in-trial data

- In practice, standard AIPW is used with a simple outcome model $\hat{h}$ (e.g. linear) learned on RCT data

$$\hat{h}(\cdot, a) \in \arg\min_{h \in \mathcal{H}} \frac{1}{n_a} \sum_{i:A_i=a} \mathcal{L}(Y_i, h(X_i, a))$$

Standard AIPW using in-trial data

- In practice, standard AIPW is used with a simple outcome model $\hat{h}$ (e.g. linear) learned on RCT data

$$\hat{h}(\cdot, a) \in \arg\min_{h \in \mathcal{H}} \frac{1}{n_a} \sum_{i:A_i=a} \mathcal{L}(Y_i, h(X_i, a))$$

- If fit using cross-fitting instead of the whole data-set, we have both
  - unbiasedness, i.e.

$$\mathbb{E}[\widehat{\theta}_{\mathrm{AIPW}}(\hat{h})] = \theta$$

  - and if $\hat{h}$ asymptotically converges to some $h^\dagger$, we have with $V_{h^\dagger} = \mathbb{E}[\psi(Z; h^\dagger) - \theta)^2]$

$$\sqrt{n}(\widehat{\theta}_{\mathrm{AIPW}}(\hat{h}) - \theta) \rightsquigarrow \mathcal{N}(0, V_{h^\dagger})$$

Standard AIPW using in-trial data

- In practice, standard AIPW is used with a simple outcome model $\hat{h}$ (e.g. linear) learned on RCT data

$$\hat{h}(\cdot, a) \in \arg\min_{h \in \mathcal{H}} \frac{1}{n_a} \sum_{i:A_i=a} \mathcal{L}(Y_i, h(X_i, a))$$

- If fit using cross-fitting instead of the whole data-set, we have both
  - unbiasedness, i.e.

$$\mathbb{E}[\widehat{\theta}_{\mathrm{AIPW}}(\hat{h})] = \theta$$

  - and if $\hat{h}$ asymptotically converges to some $h^{\dagger}$, we have with $V_{h^{\dagger}} = \mathbb{E}[\psi(Z; h^{\dagger}) - \theta)^2]$

$$\sqrt{n}(\widehat{\theta}_{\mathrm{AIPW}}(\hat{h}) - \theta) \rightsquigarrow \mathcal{N}(0, V_{h^{\dagger}})$$

- **Limitations:**
  - Small trial sample size means limited data for learning
  - Simple models have high variance $V_{h^{\dagger}}$

Using external data and foundation models

What if we have more $\hat{h}$ candidates (e.g. trained on other data)?

Using external data and foundation models

What if we have more $\hat{h}$ candidates (e.g. trained on other data)?

Opportunity: Leverage outcome regressors trained on external data

Using external data and foundation models

What if we have more $\hat{h}$ candidates (e.g. trained on other data)?

Opportunity: Leverage outcome regressors trained on external data

- For medical applications:
  - Electronic Health Records (EHR)
  - Large observational studies
  - Historical clinical trials

Using external data and foundation models

What if we have more $\hat{h}$ candidates (e.g. trained on other data)?

Opportunity: Leverage outcome regressors trained on external data

- For medical applications:
  - Electronic Health Records (EHR)
  - Large observational studies
  - Historical clinical trials
- For social sciences (results in this paper):
  - Foundation models trained on publicly available texts

Using external data and foundation models

What if we have more $\hat{h}$ candidates (e.g. trained on other data)?

Opportunity: Leverage outcome regressors trained on external data

- For medical applications:
  - Electronic Health Records (EHR)
  - Large observational studies
  - Historical clinical trials
- For social sciences (results in this paper):
  - Foundation models trained on publicly available texts

What guarantees can we still have if we use externally trained $\hat{h}$ without requiring any assumptions on external data?

Related Work

Method	**Unbiased for finite samples**	**can be asympt. more eff. than standard AIPW**	**Asympt. no worse than standard AIPW**
Standard AIPW	✓	N/A	N/A

Related Work

Method	**Unbiased for finite samples**	**can be asympt. more eff. than standard AIPW**	**Asympt. no worse than standard AIPW**
Standard AIPW	✓	N/A	N/A
Shrinkage estimators [1]	✗	✓	✓

[1] Cheng and Cai (2021), Rosenman et al. (2023)

Related Work

Method	**Unbiased for finite samples**	**can be asympt. more eff. than standard AIPW**	**Asympt. no worse than standard AIPW**
Standard AIPW	✓	N/A	N/A
Shrinkage estimators [1]	✗	✓	✓
Prognostic scores [2]	✓	✗	✓

[1] Cheng and Cai (2021), Rosenman et al. (2023)
[2] Schuler et al. (2021), Liao et al. (2025)

Related Work

Method	Unbiased for finite samples	can be asympt. more eff. than standard AIPW	Asympt. no worse than standard AIPW
Standard AIPW	✓	N/A	N/A
Shrinkage estimators [1]	✗	✓	✓
Prognostic scores [2]	✓	✗	✓
PPI-style estimators [3]	✓	✓	✗

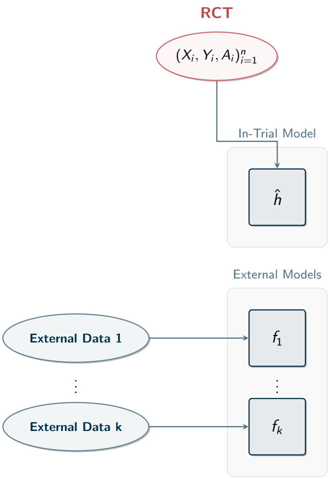
[1] Cheng and Cai (2021), Rosenman et al. (2023)
[2] Schuler et al. (2021), Liao et al. (2025)
[3] Angelopoulos et al. (2023), Poulet et al. (2025)

Related Work

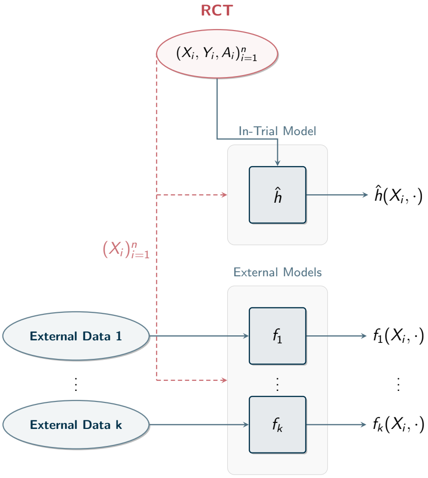
Method	Unbiased for finite samples	can be asympt. more eff. than standard AIPW	Asympt. no worse than standard AIPW
Standard AIPW	✓	N/A	N/A
Shrinkage estimators [1]	✗	✓	✓
Prognostic scores [2]	✓	✗	✓
PPI-style estimators [3]	✓	✓	✗
H-AIPW **(Ours)**	✓	✓	✓

[1] Cheng and Cai (2021), Rosenman et al. (2023)
[2] Schuler et al. (2021), Liao et al. (2025)
[3] Angelopoulos et al. (2023), Poulet et al. (2025)

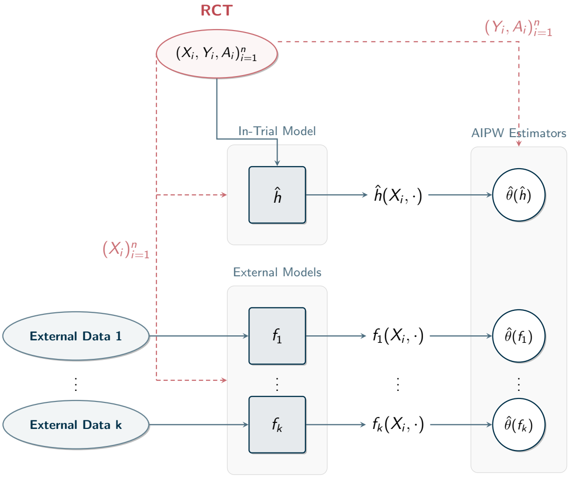
Our method: Hybrid-AIPW



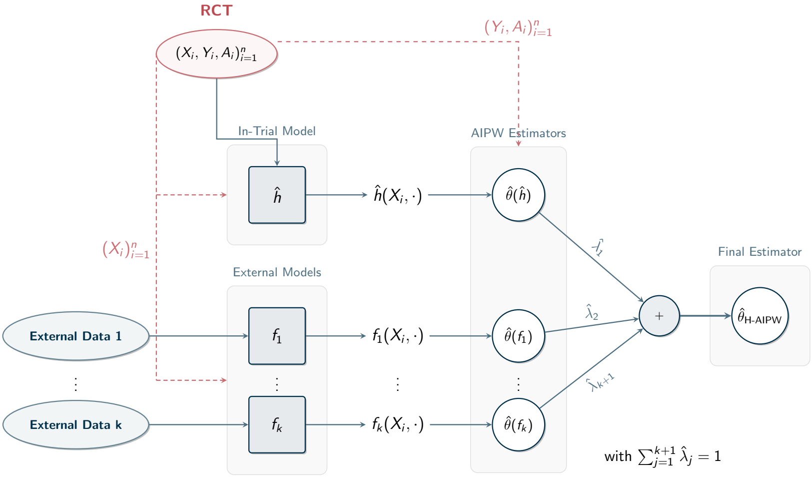
Our method: Hybrid-AIPW



Our method: Hybrid-AIPW



Our method: Hybrid-AIPW



How to choose λ?

- True optimal weights minimize the variance of the combined estimator

$$\lambda^* = \arg\min_{\lambda \in \Lambda} \lambda^T \Sigma \lambda \quad \text{subject to} \quad \sum_{j=1}^{k+1} \lambda_j = 1$$

How to choose λ?

- True optimal weights minimize the variance of the combined estimator

$$\lambda^* = \arg\min_{\lambda \in \Lambda} \lambda^T \Sigma \lambda \quad \text{subject to} \quad \sum_{j=1}^{k+1} \lambda_j = 1$$

- $\Sigma \in \mathbb{R}^{(k+1)\times(k+1)}$ is the covariance matrix with elements:

$$\Sigma_{jl} = \mathrm{Cov}(\psi(Z, g_j), \psi(Z, g_l))$$

where $\psi(Z, g)$ is the influence function corresponding to $\hat{\theta}_{AIPW}(g)$
$g_1 = \hat{h}$ is estimated from the RCT and $g_{j+1} = f_j$ for $j = 1, \ldots, k$

How to choose λ?

- True optimal weights minimize the variance of the combined estimator

$$\lambda^* = \arg\min_{\lambda \in \Lambda} \lambda^T \Sigma \lambda \quad \text{subject to} \quad \sum_{j=1}^{k+1} \lambda_j = 1$$

- $\Sigma \in \mathbb{R}^{(k+1)\times(k+1)}$ is the covariance matrix with elements:

$$\Sigma_{jl} = \mathrm{Cov}(\psi(Z, g_j), \psi(Z, g_l))$$

where $\psi(Z, g)$ is the influence function corresponding to $\hat{\theta}_{AIPW}(g)$
$g_1 = \hat{h}$ is estimated from the RCT and $g_{j+1} = f_j$ for $j = 1, \ldots, k$

- Closed-form solution:

$$\lambda^* = \frac{\Sigma^{-1}1}{1^T\Sigma^{-1}1} \quad \text{and in practice:} \quad \widehat{\lambda} = \frac{\widehat{\Sigma}^{-1}1}{1^T\widehat{\Sigma}^{-1}1}$$

Statistical Guarantees

With this choice of weights λ, we obtain the asymptotic guarantees:

Theorem (H-AIPW Guarantees) in BAWDDY**D '25:**

(a) **Consistency and Asymptotic Normality:**

$$\sqrt{n}(\widehat{\theta}_{\hat{\lambda}} - \theta) \rightsquigarrow \mathcal{N}(0, V_{\lambda^*})$$

(b) **Efficiency Guarantee:** The asymptotic variance is no greater than any individual estimator:

$$V_{\lambda^*} \leq \min_{j=1,\dots,k+1} V_j$$

where V_j is the asymptotic variance of the j-th estimator.

Statistical Guarantees

With this choice of weights λ, we obtain the asymptotic guarantees:

Theorem (H-AIPW Guarantees) in BAWDDYD '25:

(a) **Consistency and Asymptotic Normality:**

$$\sqrt{n}(\widehat{\theta}_{\hat{\lambda}} - \theta) \rightsquigarrow \mathcal{N}(0, V_{\lambda^*})$$

(b) **Efficiency Guarantee:** The asymptotic variance is no greater than any individual estimator:

$$V_{\lambda^*} \leq \min_{j=1,\ldots,k+1} V_j$$

where V_j is the asymptotic variance of the j-th estimator.

- Asymptotic efficiency never worse than standard AIPW!

Statistical Guarantees

With this choice of weights λ, we obtain the asymptotic guarantees:

Theorem (H-AIPW Guarantees) in BAWDDYD '25:

(a) **Consistency and Asymptotic Normality:**

$$\sqrt{n}(\widehat{\theta}_{\hat{\lambda}} - \theta) \rightsquigarrow \mathcal{N}(0, V_{\lambda^*})$$

(b) **Efficiency Guarantee:** The asymptotic variance is no greater than any individual estimator:

$$V_{\lambda^*} \leq \min_{j=1,\dots,k+1} V_j$$

where V_j is the asymptotic variance of the j-th estimator.

- Asymptotic efficiency never worse than standard AIPW!
- If models are accurate, may have smaller asymptotic variance!

Evaluation of effectiveness of H-AIPW on real data

Till now: social science experiments. (Plan: extend to clinical trials)

Evaluation of effectiveness of H-AIPW on real data

Till now: social science experiments. (Plan: extend to clinical trials)

- Evaluate H-AIPW on multiple randomized experiments:
  - Foreign Policy (Silverman, 2022)
  - Sociology (Melin, 2022; Kennedy, 2020; Caprariello, 2013)
  - Political Science (Fahey, 2023)
  - Psychology (Brandt, 2021)
  - Economics (Haaland, 2022)

Evaluation of effectiveness of H-AIPW on real data

Till now: social science experiments. (Plan: extend to clinical trials)

- Evaluate H-AIPW on multiple randomized experiments:
  - Foreign Policy (Silverman, 2022)
  - Sociology (Melin, 2022; Kennedy, 2020; Caprariello, 2013)
  - Political Science (Fahey, 2023)
  - Psychology (Brandt, 2021)
  - Economics (Haaland, 2022)
- Foundation models used:
  - GPT-4o, Claude 3.5 Haiku, LLaMA 3 70B
  - Multiple prompts (10 per model) to improve accuracy
- We compare against:
  - Difference in means estimator
  - Standard AIPW with (linear) outcome regression from trial data
  - PPI based PPCT (Poulet, 2025) also leveraging foundation models

Concrete Example: LLM Predictions for Political Science

- **A=0**: "protests banned due to safety concerns",
- **A=1**: "Protests banned safety concerns & cancel culture"
- **Outcome**: Degree of agreement with "Cancel culture is a problem"

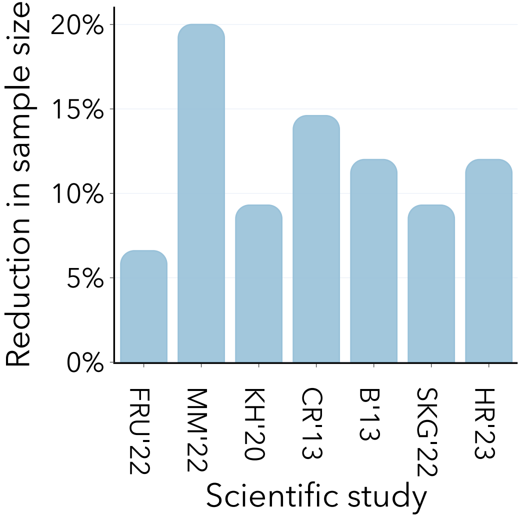
LLM Prompt (with A=1):

```
You are a 35-year-old female Democrat with liberal views and $75k
income.  A university banned an Antifa protest citing safety concerns
and that such protests contribute to cancel culture.
How much do you agree:  "Cancel culture is a big problem in today's
society"?  (1-5 scale)
```

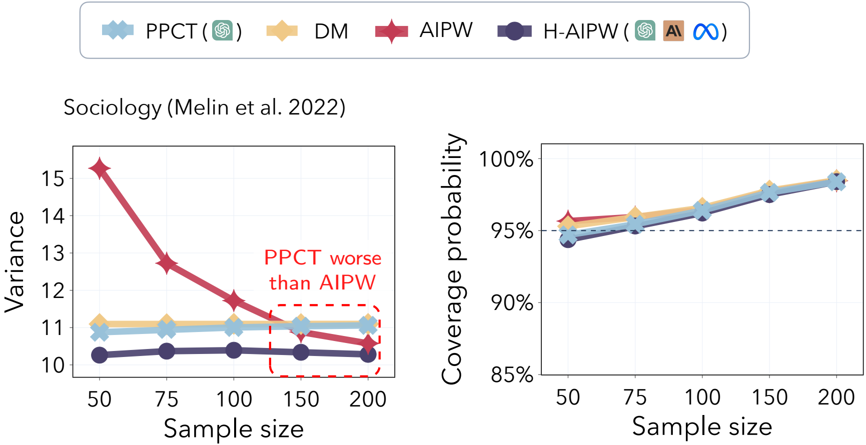
LLM Response:

As someone with liberal views, I'm skeptical of the university's framing. While I believe in maintaining safety, linking Antifa protests to "cancel culture" seems politically motivated.
I'd rate my agreement as **2** - **Disagree**.

Sample Size Reductions



Performance and Validity: Melin Study



Variance Reduction

Coverage Probability

Coverage plot shows that valid statistical inference is maintained!

Conclusion

- H-AIPW improves efficiency of randomized experiments by integrating predictions from multiple foundation models

Conclusion

- H-AIPW improves efficiency of randomized experiments by integrating predictions from multiple foundation models
- Provides substantial precision gains (up to 20% sample size reduction)

Conclusion

- H-AIPW improves efficiency of randomized experiments by integrating predictions from multiple foundation models
- Provides substantial precision gains (up to 20% sample size reduction)
- Maintains valid statistical inference without additional assumptions

Conclusion

- H-AIPW improves efficiency of randomized experiments by integrating predictions from multiple foundation models
- Provides substantial precision gains (up to 20% sample size reduction)
- Maintains valid statistical inference without additional assumptions

Conclusion

- H-AIPW improves efficiency of randomized experiments by integrating predictions from multiple foundation models
- Provides substantial precision gains (up to 20% sample size reduction)
- Maintains valid statistical inference without additional assumptions

Conclusion

- H-AIPW improves efficiency of randomized experiments by integrating predictions from multiple foundation models
- Provides substantial precision gains (up to 20% sample size reduction)
- Maintains valid statistical inference without additional assumptions

Limitations: Success depends on foundation models being well-aligned with the experimental domain

Conclusion

- H-AIPW improves efficiency of randomized experiments by integrating predictions from multiple foundation models
- Provides substantial precision gains (up to 20% sample size reduction)
- Maintains valid statistical inference without additional assumptions

Limitations: Success depends on foundation models being well-aligned with the experimental domain

GitHub repository: `https://github.com/jaabmar/HAIPW`

Thank You!

Questions?

Piersilvio De Bartolomeis

Javier Abad

Guanbo Wang

Konstantin Donhauser

Raymond Duch

Fanny Yang

Issa Dahabreh

Identification Assumptions

- **Consistency:** $Y = Y(A)$
  - Treatment is well-defined (e.g., protocol-driven interventions)
  - Observed outcome is one of the potential outcomes

Identification Assumptions

- **Consistency:** $Y = Y(A)$
  - Treatment is well-defined (e.g., protocol-driven interventions)
  - Observed outcome is one of the potential outcomes
- **Randomization:** $A \perp\!\!\!\perp (Y(0), Y(1))$
  - Directly supported by the study design
  - Treatment is independent of potential outcomes

Identification Assumptions

- **Consistency:** $Y = Y(A)$
  - Treatment is well-defined (e.g., protocol-driven interventions)
  - Observed outcome is one of the potential outcomes
- **Randomization:** $A \perp\!\!\!\perp (Y(0), Y(1))$
  - Directly supported by the study design
  - Treatment is independent of potential outcomes
- **Positivity:** $\pi_a = \mathbb{P}(A = a) > 0$ for $a \in \{0, 1\}$
  - Both treatment and control have non-zero probability
  - In (most) randomized experiments, π_a is known by design

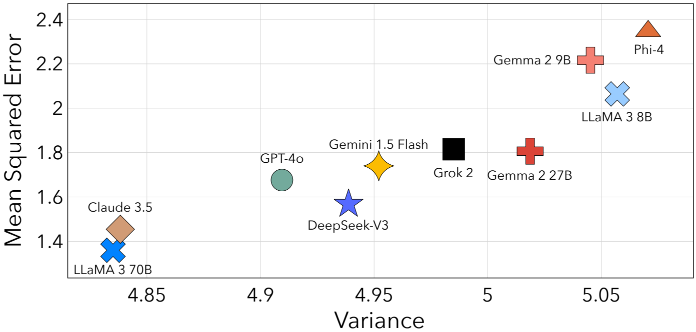
Identification Assumptions

- **Consistency:** $Y = Y(A)$
    - Treatment is well-defined (e.g., protocol-driven interventions)
    - Observed outcome is one of the potential outcomes
- **Randomization:** $A \perp\!\!\!\perp (Y(0), Y(1))$
    - Directly supported by the study design
    - Treatment is independent of potential outcomes
- **Positivity:** $\pi_a = \mathbb{P}(A = a) > 0$ for $a \in \{0, 1\}$
    - Both treatment and control have non-zero probability
    - In (most) randomized experiments, π_a is known by design

Under these assumptions:

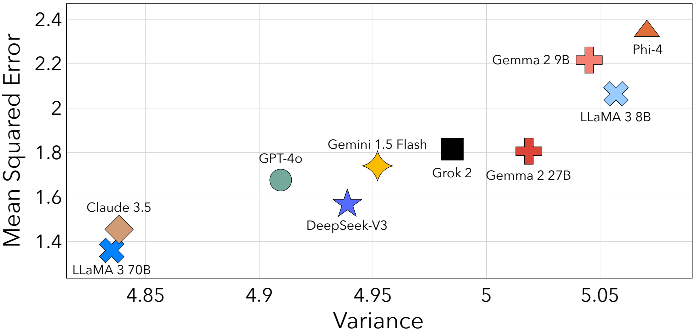
$$\theta = \mathbb{E}[Y(1) - Y(0)] = \mathbb{E}[Y|A = 1] - \mathbb{E}[Y|A = 0]$$

Impact of Model Scale



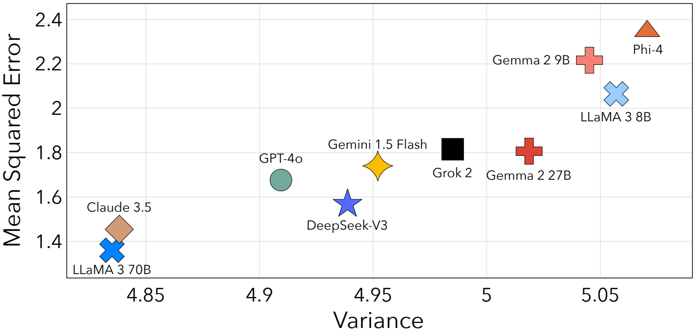
- Larger models consistently achieve lower MSE and lower variance

Impact of Model Scale



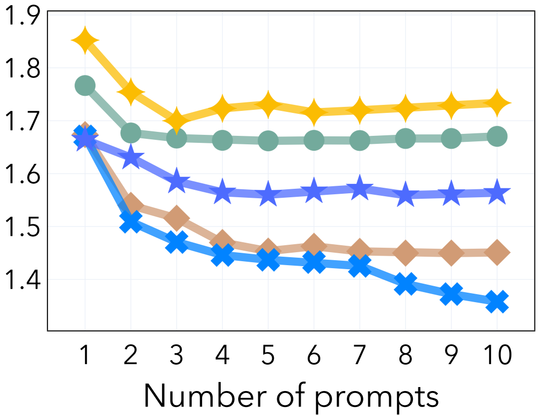
- Larger models consistently achieve lower MSE and lower variance
- LLaMA 3 70B performs exceptionally well despite fewer parameters than some competitors

Impact of Model Scale



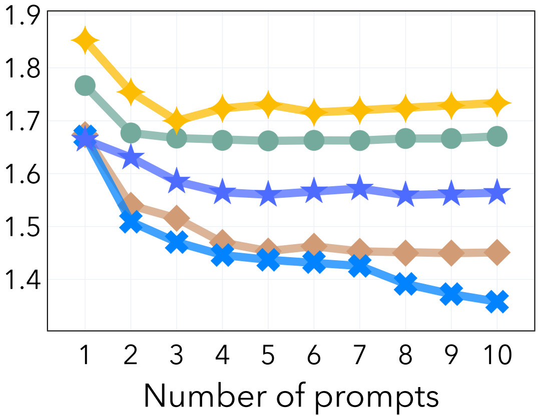
- Larger models consistently achieve lower MSE and lower variance
- LLaMA 3 70B performs exceptionally well despite fewer parameters than some competitors
- Clear relationship between prediction accuracy and variance reduction

Impact of Inference-Time Compute



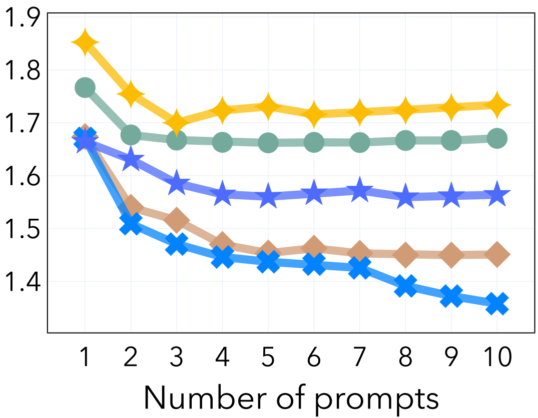
- Averaging over multiple prompts consistently reduces MSE

Impact of Inference-Time Compute



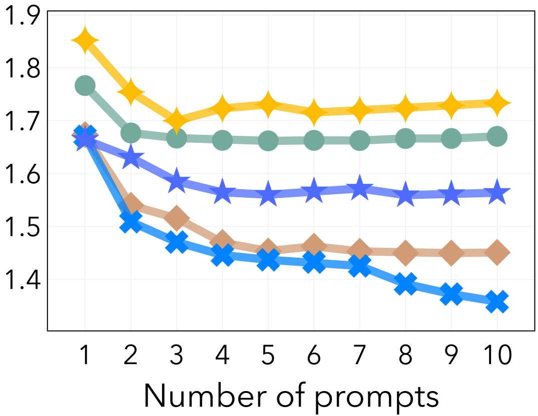
- Averaging over multiple prompts consistently reduces MSE
- Lower MSE is associated with higher precision

Impact of Inference-Time Compute



- Averaging over multiple prompts consistently reduces MSE
- Lower MSE is associated with higher precision
- Using more prompts improves H-AIPW precision across studies

Impact of Inference-Time Compute



- Averaging over multiple prompts consistently reduces MSE
- Lower MSE is associated with higher precision
- Using more prompts improves H-AIPW precision across studies
- Diminishing returns after approximately 10-15 prompts