Efficient Randomized Experiments Using Foundation Models

Piersilvio De Bartolomeis

joint work with

Javier Abad, Guanbo Wang, Konstantin Donhauser,
Raymond M. Duch, Fanny Yang, 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 leverage (multiple) foundation models trained on external data sources?
  - Examples: language models trained on large text 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 leverage (multiple) foundation models trained on external data sources?
  - Examples: language models trained on large text 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 $(X_i, Y_i, A_i)_{i=1}^n$ 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 $(X_i, Y_i, A_i)_{i=1}^n$ drawn i.i.d. from $\mathbb{P}$
- **Task:** Efficiently estimate $\theta := \mathbb{E}[Y(1) - Y(0)]$

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:** $\pi = \mathbb{P}(A = 1) > 0$
  - Both treatment and control have non-zero probability
  - In (most) randomized experiments, π 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:** $\pi = \mathbb{P}(A = 1) > 0$
  - Both treatment and control have non-zero probability
  - In (most) randomized experiments, π is known by design

Under these assumptions:

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

Difference in means estimator

- The simplest approach for randomized experiments:

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

Difference in means estimator

- The simplest approach for randomized experiments:

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

- Consistent and asymptotically normal:

$$\sqrt{n}(\widehat{\theta}_{\mathrm{DM}} - \theta) \rightsquigarrow \mathcal{N}(0, V_{\mathrm{DM}})$$

Difference in means estimator

- The simplest approach for randomized experiments:

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

- Consistent and asymptotically normal:

$$\sqrt{n}(\widehat{\theta}_{\mathrm{DM}} - \theta) \rightsquigarrow \mathcal{N}(0, V_{\mathrm{DM}})$$

- **Is this the most efficient estimator?** No, covariates are ignored!

Leverage **availability of covariates** → smaller confidence intervals

Imputing missing data with predictive models

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

Imputing missing data with predictive models

Main idea: **If we had a predictive model $\hat{h}$, we can use it to predict the counterfactuals 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 $\hat{h}$ is trained on RCT

Imputing missing data with predictive models

Main idea: **If we had a predictive model $\hat{h}$, we can use it to predict the counterfactuals 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 $\hat{h}$ is trained on RCT
- Similar to PPI-style estimators as in Angelopoulos et al. '23 where $\hat{h}$ can be any external model

Standard AIPW using in-trial data

- In practice, standard AIPW uses 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 uses 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 $h^\dagger$, we have

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

Standard AIPW using in-trial data

- In practice, standard AIPW uses 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 $h^\dagger$, we have

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

- Variance $V_{h^\dagger}$ is minimized when $h^\dagger = \mathbb{E}[Y|X, A]$, achieving the lowest possible variance among all regular estimators

AIPW limitations and new opportunities

- In RCTs, sample size is too small.
  - Unlikely to learn a good outcome regression from $(X_i, Y_i, A_i)_{i=1}^n$.
  - A simple function class $\mathcal{H}$ (e.g., linear), yields limited gains.
  - Achieving efficiency requires a good estimate of $\mathbb{E}[Y|X, A]$

AIPW limitations and new opportunities

- In RCTs, sample size is too small.
  - Unlikely to learn a good outcome regression from $(X_i, Y_i, A_i)_{i=1}^n$.
  - A simple function class $\mathcal{H}$ (e.g., linear), yields limited gains.
  - Achieving efficiency requires a good estimate of $\mathbb{E}[Y|X, A]$

AIPW limitations and new opportunities

- In RCTs, sample size is too small.
  - Unlikely to learn a good outcome regression from $(X_i, Y_i, A_i)_{i=1}^n$.
  - A simple function class $\mathcal{H}$ (e.g., linear), yields limited gains.
  - Achieving efficiency requires a good estimate of $\mathbb{E}[Y|X, A]$
- **Opportunity:** Leverage external data to learn better outcome models
  - 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

Leveraging external data

- **Challenge:** External models may not generalize to trial population
  - Distribution shift between external data and trial data
  - Naively using external models could yield *worse* efficiency than standard AIPW

Leveraging external data

- **Challenge:** External models may not generalize to trial population
    - Distribution shift between external data and trial data
    - Naively using external models could yield *worse* efficiency than standard AIPW

- What guarantees can we still have if we use an external model without requiring any additional assumptions?
    - Need to fall back to trial data when external models perform poorly

Related Work

Method	Unbiased	Can be asympt. better than standard AIPW	Asympt. no worse than standard AIPW
Standard AIPW	✓	N/A	N/A

Related Work

Method	Unbiased	Can be asympt. better 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**	**Can be asympt. better than standard AIPW**	**Asympt. no worse than standard AIPW**
Standard AIPW	✓	N/A	N/A
Shrinkage estimators [1]	✗	✓	✓
PROCOVA [2]	✓	✗	✓

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

Related Work

Method	Unbiased	Can be asympt. better than standard AIPW	Asympt. no worse than standard AIPW
Standard AIPW	✓	N/A	N/A
Shrinkage estimators [1]	✗	✓	✓
PROCOVA [2]	✓	✗	✓
PPI-style estimators [3]	✓	✓	✗

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

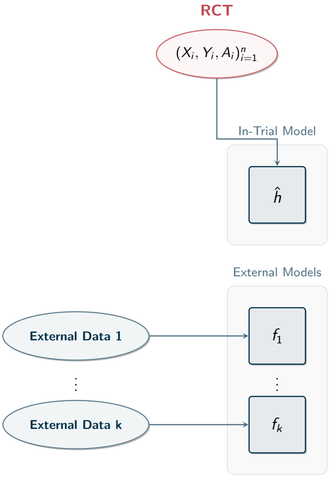
[3] Angelopoulos et al. (2023), Poulet et al. (2025)

Related Work

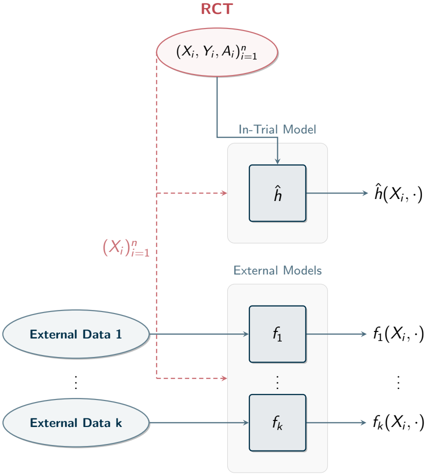
Method	**Unbiased**	**Can be asympt. better than standard AIPW**	**Asympt. no worse than standard AIPW**
Standard AIPW	✓	N/A	N/A
Shrinkage estimators [1]	✗	✓	✓
PROCOVA [2]	✓	✗	✓
PPI-style estimators [3]	✓	✓	✗
H-AIPW **(Ours)**	✓	✓	✓

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

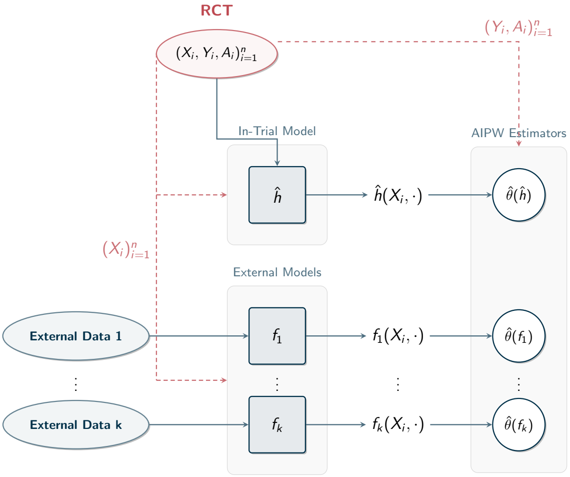
Our method: H-AIPW



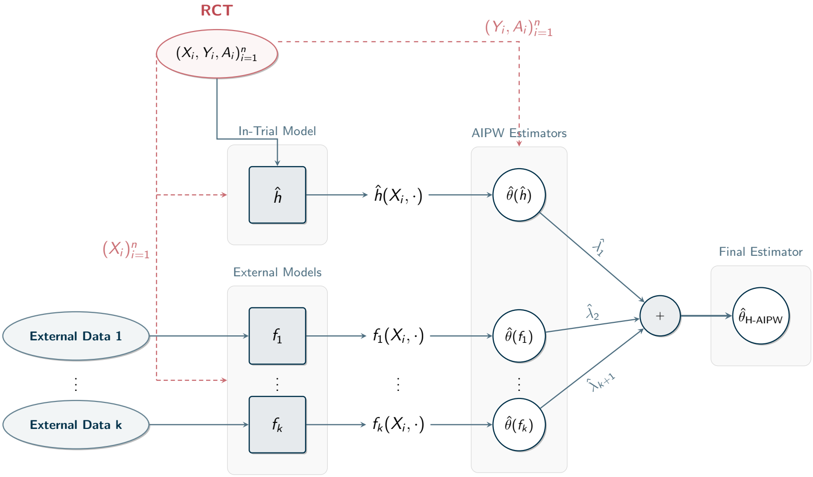
Our method: H-AIPW



Our method: H-AIPW



Our method: H-AIPW



Hybrid augmented inverse probability weighting

- **Foundation Models:**
  - Access to multiple pre-trained foundation models $f_1, f_2, \ldots, f_k$
  - Models trained on external data, potentially more accurate than $\hat{h}$

Hybrid augmented inverse probability weighting

- **Foundation Models:**
  - Access to multiple pre-trained foundation models $f_1, f_2, \ldots, f_k$
  - Models trained on external data, potentially more accurate than $\hat{h}$
- Include AIPW estimators using each model: $\widehat{\theta}_{\mathrm{AIPW}}(f_j)$
- Include the standard AIPW estimator with $\hat{h}$ estimated from trial data

Hybrid augmented inverse probability weighting

- **Foundation Models:**
  - Access to multiple pre-trained foundation models $f_1, f_2, \ldots, f_k$
  - Models trained on external data, potentially more accurate than $\hat{h}$

- Include AIPW estimators using each model: $\widehat{\theta}_{\mathrm{AIPW}}(f_j)$
- Include the standard AIPW estimator with $\hat{h}$ estimated from trial data

H-AIPW Estimator

$$\widehat{\theta}_\lambda = \lambda_1 \widehat{\theta}_{\mathrm{AIPW}}(\hat{h}) + \sum_{j=1}^{k} \lambda_{j+1} \widehat{\theta}_{\mathrm{AIPW}}(f_j)$$

where $\lambda \in \mathbb{R}^{k+1}$ such that $\sum_{j=1}^{k+1} \lambda_j = 1$

Why weights must sum up to 1

- The constraint $\sum_{j=1}^{k+1} \lambda_j = 1$ is crucial for unbiasedness

Why weights must sum up to 1

- The constraint $\sum_{j=1}^{k+1} \lambda_j = 1$ is crucial for unbiasedness
- With this constraint, H-AIPW is in the class of AIPWs with a combined outcome model:

$$\begin{aligned}\widehat{\theta}_\lambda &= \lambda_1 \widehat{\theta}_{\mathrm{AIPW}}(\hat{h}) + \sum_{j=1}^{k} \lambda_{j+1} \widehat{\theta}_{\mathrm{AIPW}}(f_j) \\ &= \widehat{\theta}_{\mathrm{AIPW}} \left(\lambda_1 \hat{h} + \sum_{j=1}^{k} \lambda_{j+1} f_j \right)\end{aligned}$$

Why weights must sum up to 1

- The constraint $\sum_{j=1}^{k+1} \lambda_j = 1$ is crucial for unbiasedness
- With this constraint, H-AIPW is in the class of AIPWs with a combined outcome model:

$$\begin{aligned}\widehat{\theta}_\lambda &= \lambda_1 \widehat{\theta}_{\mathrm{AIPW}}(\hat{h}) + \sum_{j=1}^{k} \lambda_{j+1} \widehat{\theta}_{\mathrm{AIPW}}(f_j) \\ &= \widehat{\theta}_{\mathrm{AIPW}} \left(\lambda_1 \hat{h} + \sum_{j=1}^{k} \lambda_{j+1} f_j \right)\end{aligned}$$

- H-AIPW inherits all the nice theoretical properties of AIPW

How to choose λ?

- True optimal weights minimize the variance of the combined estimator

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

How to choose λ?

- True optimal weights minimize the variance of the combined estimator

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

- $\Sigma \in \mathbb{R}^{(k+1)\times(k+1)}$ 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

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

- $\Sigma \in \mathbb{R}^{(k+1)\times(k+1)}$ 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 $j = 1, \dots, k$

- Closed-form solution:

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

Statistical Guarantees

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

Theorem (H-AIPW Guarantees) in DAWDDYD '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.

Statistical Guarantees

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

Theorem (H-AIPW Guarantees) in DAWDDYD '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 DAWDDYD '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!
- If models are accurate, may have smaller asymptotic variance!

Empirical evaluation on real data

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

Empirical evaluation on real data

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

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

Empirical evaluation on real data

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

- Evaluate H-AIPW on multiple survey 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):

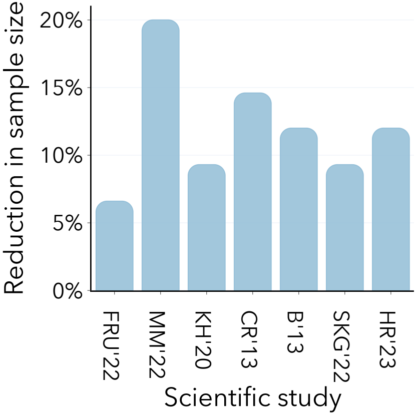
```
You are a 35-year-old female Democrat with liberal views and $75k
income.  A university banned an Antifa protest citing safety concerns
and that such protests contribute to cancel culture.
How much do you agree:  "Cancel culture is a big problem in today's
society"?  (1-5 scale)
```

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**.

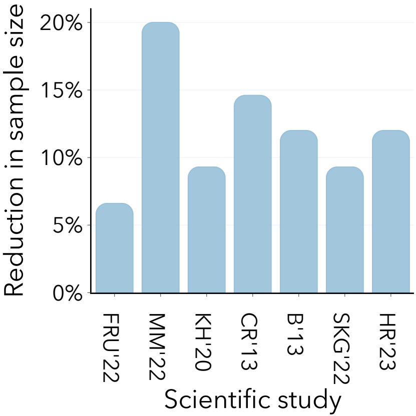
Sample size reductions

- Each study is subsampled to $n = 75$ participants



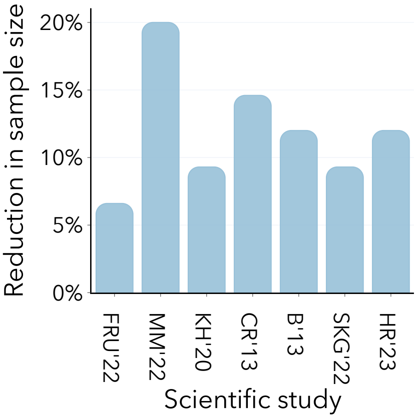
Sample size reductions

- Each study is subsampled to $n = 75$ participants
- The bars show the percentage sample size reduction to match confidence interval width of standard AIPW



Sample size reductions

- Each study is subsampled to $n = 75$ participants
- The bars show the percentage sample size reduction to match confidence interval width of standard AIPW
- H-AIPW achieves the same precision as standard AIPW with up to 20% fewer samples

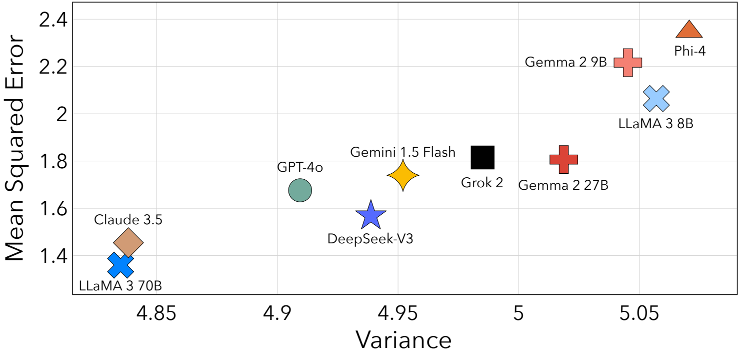


Variance reduction

	Melin et al. (2022)		Silverman et al. (2022)		Kennedy et al. (2020)		Fahey et al. (2023)	
Estimator	$n = 100$	$n = 200$	$n = 100$	$n = 200$	$n = 100$	$n = 200$	$n = 100$	$n = 200$
H-AIPW	**10.39**	**10.28**	**2.10**	**2.14**	**17.09**	**17.47**	**4.87**	4.94
PPCT	11.00	11.06	2.25	2.26	17.87	17.97	4.88	**4.91**
PROCOVA	11.81	10.62	2.24	2.22	18.38	18.11	5.18	5.09
AIPW (boosting)	12.82	12.44	2.82	2.83	23.09	23.12	6.31	6.37
AIPW (standard)	11.72	10.57	2.22	2.20	18.09	17.95	5.09	5.04
DM	11.10	11.10	2.30	2.30	18.07	18.08	5.61	5.62

	Caprariello et al. (2013)		Brandt (2013)		Haaland et al. (2023)		Shuman et al. (2024)	
Estimator	$n = 100$	$n = 200$	$n = 100$	$n = 200$	$n = 100$	$n = 200$	$n = 100$	$n = 200$
H-AIPW	**5.88**	**5.96**	**11.86**	**11.90**	**4.49**	**4.44**	**8.46**	**8.91**
PPCT	5.99	6.01	12.07	12.12	4.50	4.52	9.08	9.14
PROCOVA	6.41	6.13	12.77	12.25	4.73	**4.44**	9.12	9.55
AIPW (boosting)	7.79	7.60	15.20	14.70	5.39	5.22	10.53	10.67
AIPW (standard)	6.39	6.18	12.55	12.13	4.82	4.55	9.20	10.31
DM	6.15	6.15	12.81	12.80	5.72	5.71	13.83	13.83

Impact of model scale



Larger models tend to provide better predictions, leading to smaller variance and better efficiency gains

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

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! Any Questions?

Piersilvio De Bartolomeis

Javier Abad

Guanbo Wang

Konstantin Donhauser

Raymond Duch

Fanny Yang

Issa Dahabreh