Leveraging Large Language Models to Improve Precision in Randomized Controlled Trials

Jaylin Lowe, Adam Sales, Johann Gagnon-Bartsch

MSSISS 2026

March 27th, 2026

Motivation

- RCTs are generally considered the "gold standard" in causal inference, but they can be small and expensive

Motivation

- RCTs are generally considered the "gold standard" in causal inference, but they can be small and expensive
- Many methods exist to improve precision in RCTs

Motivation

- RCTs are generally considered the "gold standard" in causal inference, but they can be small and expensive
- Many methods exist to improve precision in RCTs
- Some of these methods leverage external data

Data Integration Approach (Gagnon-Bartsch et. al, 2023)

One option is to use observational data to improve precision. This works by:

Data Integration Approach (Gagnon-Bartsch et. al, 2023)

One option is to use observational data to improve precision. This works by:

1. Training a model on the observational data.

Data Integration Approach (Gagnon-Bartsch et. al, 2023)

One option is to use observational data to improve precision. This works by:

1. Training a model on the observational data.
2. Generating predictions for the RCT observations based on this model.

Data Integration Approach (Gagnon-Bartsch et. al, 2023)

One option is to use observational data to improve precision. This works by:

1. Training a model on the observational data.
2. Generating predictions for the RCT observations based on this model.
3. Augmenting the RCT covariates with these predictions, now considered an additional covariate.

Data Integration Approach (Gagnon-Bartsch et. al, 2023)

One option is to use observational data to improve precision. This works by:

1. Training a model on the observational data.
2. Generating predictions for the RCT observations based on this model.
3. Augmenting the RCT covariates with these predictions, now considered an additional covariate.
4. Obtain leave-one-out predictions for all observations in the RCT using a model trained on the RCT covariates plus this additional covariate.

Data Integration Approach (Gagnon-Bartsch et. al, 2023)

One option is to use observational data to improve precision. This works by:

1. Training a model on the observational data.
2. Generating predictions for the RCT observations based on this model.
3. Augmenting the RCT covariates with these predictions, now considered an additional covariate.
4. Obtain leave-one-out predictions for all observations in the RCT using a model trained on the RCT covariates plus this additional covariate.
5. Incorporate these leave-one-out predictions into an estimator of the average treatment effect.

Data Integration Approach (Gagnon-Bartsch et. al, 2023)

One option is to use observational data to improve precision. This works by:

1. Training a model on the observational data.
2. Generating predictions for the RCT observations based on this model.
3. Augmenting the RCT covariates with these predictions, now considered an additional covariate.
4. Obtain leave-one-out predictions for all observations in the RCT using a model trained on the RCT covariates plus this additional covariate.
5. Incorporate these leave-one-out predictions into an estimator of the average treatment effect.

If the predictions from the observational dataset are correlated with the RCT outcome, then this can lead to substantial gains in precision.

What if we used a LLM instead of an observational dataset?

Intuition for LLMs

- LLMs are trained on large amounts of data and may be able to leverage their training data to make reasonable predictions.

Intuition for LLMs

- LLMs are trained on large amounts of data and may be able to leverage their training data to make reasonable predictions.
- A useful observational dataset is not always available.

Intuition for LLMs

- LLMs are trained on large amounts of data and may be able to leverage their training data to make reasonable predictions.
- A useful observational dataset is not always available.
- LLMs may have more or different information than an observational dataset.

Intuition for LLMs

- LLMs are trained on large amounts of data and may be able to leverage their training data to make reasonable predictions.
- A useful observational dataset is not always available.
- LLMs may have more or different information than an observational dataset.
- The LLM predictions may be inconsistent, biased, or hallucinated but the estimator will remain unbiased and precision cannot be harmed

Intuition for LLMs

- LLMs are trained on large amounts of data and may be able to leverage their training data to make reasonable predictions.
- A useful observational dataset is not always available.
- LLMs may have more or different information than an observational dataset.
- The LLM predictions may be inconsistent, biased, or hallucinated but the estimator will remain unbiased and precision cannot be harmed
- We focus on how to do this in practice.

Case Studies

1: Sentencing of Defendants and Recidivism (Green and Winik, 2010)

- Natural experiment looking at effect of sentencing and probation on recidivism.

1: Sentencing of Defendants and Recidivism (Green and Winik, 2010)

- Natural experiment looking at effect of sentencing and probation on recidivism.
- Treatment is judge assignment.

1: Sentencing of Defendants and Recidivism (Green and Winik, 2010)

- Natural experiment looking at effect of sentencing and probation on recidivism.
- Treatment is judge assignment.
- Outcome is whether or not they were arrested again within 4 years.

1: Sentencing of Defendants and Recidivism (Green and Winik, 2010)

- Natural experiment looking at effect of sentencing and probation on recidivism.
- Treatment is judge assignment.
- Outcome is whether or not they were arrested again within 4 years.
- Dataset includes demographics, prior history, and details of the crime for 1,0003 defendants arrested in DC.

2: Cognitive Tutor Algebra (Pane et. al, 2014)

- Efficacy trial looking at effect of new algebra curriculum. Treatment is new curriculum or standard curriculum.

2: Cognitive Tutor Algebra (Pane et. al, 2014)

- Efficacy trial looking at effect of new algebra curriculum. Treatment is new curriculum or standard curriculum.
- Outcome is score on algebra proficiency exam.

2: Cognitive Tutor Algebra (Pane et. al, 2014)

- Efficacy trial looking at effect of new algebra curriculum. Treatment is new curriculum or standard curriculum.
- Outcome is score on algebra proficiency exam.
- Student level data with demographics and pretest score on algebra readiness exam.

3: Open Access Paper Citations (Davis, 2011)

- RCT investigating the impact of open access policies on number of citations.

3: Open Access Paper Citations (Davis, 2011)

- RCT investigating the impact of open access policies on number of citations.
- Treatment is open access or typical access.

3: Open Access Paper Citations (Davis, 2011)

- RCT investigating the impact of open access policies on number of citations.
- Treatment is open access or typical access.
- Outcome is number of citations 3 years after publication.

3: Open Access Paper Citations (Davis, 2011)

- RCT investigating the impact of open access policies on number of citations.
- Treatment is open access or typical access.
- Outcome is number of citations 3 years after publication.
- 1,248 papers across five journals with only four basic covariates.

3: Open Access Paper Citations (Davis, 2011)

- RCT investigating the impact of open access policies on number of citations.
- Treatment is open access or typical access.
- Outcome is number of citations 3 years after publication.
- 1,248 papers across five journals with only four basic covariates.
- We obtain abstracts through PubMed.

Methods

Overview

- Goal: obtain prediction from LLM for each observation that is highly correlated with RCT outcome.

Overview

- Goal: obtain prediction from LLM for each observation that is highly correlated with RCT outcome.
- Asking the LLM to make a prediction for each observation gives overly similar results.

Overview

- Goal: obtain prediction from LLM for each observation that is highly correlated with RCT outcome.
- Asking the LLM to make a prediction for each observation gives overly similar results.
- Instead, we pair observations together and ask the LLM to compare them.

Question Formulation

- Give covariates for both observations.

Question Formulation

- Give covariates for both observations.
- If there are missing values, we tell the LLM this value is unknown.

Question Formulation

- Give covariates for both observations.
- If there are missing values, we tell the LLM this value is unknown.
- LLM is asked to predict which observation is more likely to exhibit a specific quality. Typically, this is the outcome variable, but other qualities may be used.

Covariate Construction

- Goal is to construct a covariate from the LLM predictions that can explain more of the variance than the RCT covariates we already have.

Covariate Construction

- Goal is to construct a covariate from the LLM predictions that can explain more of the variance than the RCT covariates we already have.
- In some cases, it makes sense not to consider all possible pairs. In this scenario, we stratify on a covariate and consider only pairs where both observations are in the same stratum.
    1. Observations are not directly comparable.
    2. Improving precision.
    3. Improving computation time.

Covariate Construction

- Goal is to construct a covariate from the LLM predictions that can explain more of the variance than the RCT covariates we already have.
- In some cases, it makes sense not to consider all possible pairs. In this scenario, we stratify on a covariate and consider only pairs where both observations are in the same stratum.
  1. Observations are not directly comparable.
  2. Improving precision.
  3. Improving computation time.
- Selection frequency is used as a covariate.

Modeling and Evaluation

To evaluate the LLM predictions, we:

1. Train a model on the RCT data, complete with the new LLM covariate.

Modeling and Evaluation

To evaluate the LLM predictions, we:

1. Train a model on the RCT data, complete with the new LLM covariate.

2. Create two versions of each observation, one assigned to control and one assigned to treatment. Obtain leave-one-out predictions for each version.

Modeling and Evaluation

To evaluate the LLM predictions, we:

1. Train a model on the RCT data, complete with the new LLM covariate.
2. Create two versions of each observation, one assigned to control and one assigned to treatment. Obtain leave-one-out predictions for each version.
3. Use these imputations to calculate an estimate of the average treatment effect and its variance.

Results

1: Sentencing of Defendants and Recidivism

- LLM predictions do not help reduce precision.

1: Sentencing of Defendants and Recidivism

- LLM predictions do not help reduce precision.
- Asking for explanations reveals LLM bases predictions on prior history.

1: Sentencing of Defendants and Recidivism

- LLM predictions do not help reduce precision.
- Asking for explanations reveals LLM bases predictions on prior history.
- Random forest reveals age is the strongest predictor for this dataset.

2: Cognitive Tutor Algebra

- LLM predictions are significant when included in a linear model.

2: Cognitive Tutor Algebra

- LLM predictions are significant when included in a linear model.
- Does not translate into a meaningful reduction in variance.

2: Cognitive Tutor Algebra

- LLM predictions are significant when included in a linear model.
- Does not translate into a meaningful reduction in variance.
- Many other useful covariates were already available.

3: Open Access Papers

- For each pair, we construct 11 covariates: one from asking about which paper will have higher citations, and 10 from additional qualities of interest.

3: Open Access Papers

- For each pair, we construct 11 covariates: one from asking about which paper will have higher citations, and 10 from additional qualities of interest.
- We compare four types of models: one with the base covariates, one with the base covariates plus the citation based covariate, one with the base covariates plus the 10 quality covariates, and one with all of these.

3: Open Access Papers

- For each pair, we construct 11 covariates: one from asking about which paper will have higher citations, and 10 from additional qualities of interest.
- We compare four types of models: one with the base covariates, one with the base covariates plus the citation based covariate, one with the base covariates plus the 10 quality covariates, and one with all of these.
- LLM predictions significantly improve precision.

3: Open Access Papers

- For each pair, we construct 11 covariates: one from asking about which paper will have higher citations, and 10 from additional qualities of interest.
- We compare four types of models: one with the base covariates, one with the base covariates plus the citation based covariate, one with the base covariates plus the 10 quality covariates, and one with all of these.
- LLM predictions significantly improve precision.
- Improvements are largest when citation base covariate is combined with the 10 qualities.

Estimator Standard Errors for Open Access Dataset

Journal	Base Covariates	Base + 11 LLM Covariates
Science	0.1227	0.0977
Neurophysiology	0.1300	0.1080
Genetics	0.1110	0.0999
Applied Physiology	0.1707	0.1680
FASEB	0.1066	0.0912

Key Takeaways

- Case Study 1: follows different pattern than similar RCTs

Key Takeaways

- Case Study 1: follows different pattern than similar RCTs
- Case Study 2: Other covariates are too useful

Key Takeaways

- Case Study 1: follows different pattern than similar RCTs
- Case Study 2: Other covariates are too useful
- Case Study 3: Works well since we can leverage abstracts, which traditional methods cannot

Key Takeaways

- Case Study 1: follows different pattern than similar RCTs
- Case Study 2: Other covariates are too useful
- Case Study 3: Works well since we can leverage abstracts, which traditional methods cannot
- Method works best with RCTs that are small, follow similar patterns to RCTs in the same contexts, have few covariates, and contain text or image data.

Questions?

Estimator

Our estimator is:

$$\hat{\tau} = \frac{1}{N}\sum_{i=1}^{n} Z_i * \frac{Y_i - \hat{m}_i}{p} - \frac{1}{N}\sum_{i=1}^{n}(1 - Z_i) * \frac{Y_i - \hat{m}_i}{1 - p} \tag{1}$$

where $\hat{m}_i = p\hat{y}^c(\tilde{\mathbf{x}}_i) + (1-p)\hat{y}^t(\tilde{\mathbf{x}}_i)$
The estimated variance is:

$$\widehat{\mathrm{Var}}(\hat{\tau}) = \frac{1}{N}\left[\frac{p}{1-p}\hat{E}_c^2 + \frac{1-p}{p}\hat{E}_t^2 + 2\sqrt{\hat{E}_c^2\hat{E}_t^2}\right] \tag{2}$$

where $\hat{E}_c^2 = \frac{1}{n_c}\sum_{i=1}^{N}(1 - Z_i) * [\hat{y}_{-i}^c(\mathbf{x}_i) - y_i^c]^2$ and $\hat{E}_t^2$ is defined analogously