# Benchmarking Debiasing Methods for LLM-based Parameter Estimates

Chalmers University of Technology

Nicolas Audinet de Pieuchon, Adel Daoud, Connor T. Jerzak, Moa Johansson, Richard Johansson

GÖTEBORGS UNIVERSITET

## Motivation

Useful to run statistical analysis to examine the relationship between variables

$$P(Y = 1 \mid X) = \sigma(\beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n)$$

e.g. coefficients of a logistic regression

In many domains data is often only available as free-form text

e.g. Computational Social Science

Expert Annotations

Generated Annotations

Expert human annotations tend to be high quality but costly to produce at scale

Annotations generated by LLMs are cheaper but risks introducing unwanted and poorly understood biases into the downstream analysis.\

**Debiasing methods:**

Use both a small subset of expert annotations to bias-correct model from generated annotations while still taking advantage of the scalability of generative methods

### Prediction-Powered Inference (PPI)*

Correct imputation estimate with rectifier:

$$\hat{\theta} = \tilde{\theta} - r_\theta$$

Compute rectifier from gradients:

$$r_\theta = \mathbb{E}[\nabla_\theta \ell_\theta(x_i, y_i) - \nabla_\theta \ell_\theta(x_i, \hat{y}_i)]$$

*Angelopoulos, Anastasios N., et al. "Prediction-powered inference." Science 382.6671 (2023): 669-674.

### Design-based Supervised Learning (DSL)†

Assume known expert labelling distribution:

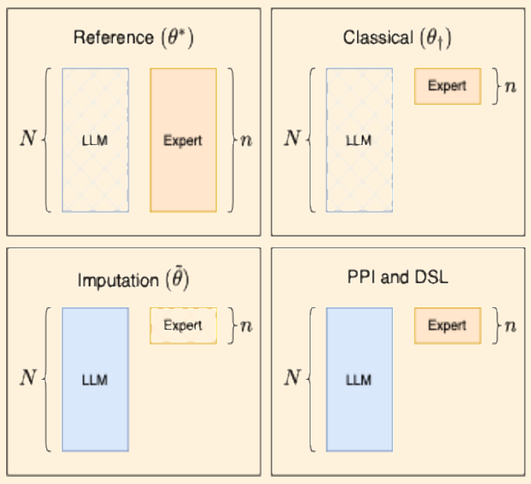
$$\pi(x_i, \hat{y}_i) = P(R_i = 1 \mid x_i, \hat{y}_i)$$

Double-robust estimate:

$$\tilde{y}_i = \hat{y}_i - \frac{R_i}{\pi_i}(\hat{y}_i - y_i) \qquad \mathbb{E}[\tilde{y}_i - y_i \mid x_i] = 0$$

†Egami, Naoki, et al. "Using large language model annotations for the social sciences: A general framework of using predicted variables in downstream analyses." Preprint from November 17 (2024): 2024.

## Setup




For all experiments:

- Logistic regression
  - 4 discrete input features
  - Binary output
- 4 different LLMs
- 4 different datasets
- Evaluation: standardised RMSE with a reference model

## Results

**RQ1:** When is it preferable to use debiasing methods over just the expert annotations?

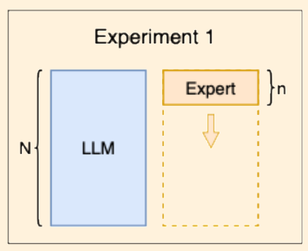
Both DSL and PPI strictly outperform using only expert annotations.

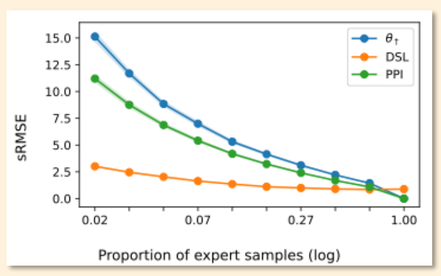
**RQ2:** What are the performance differences between debiasing methods?

DSL tends to outperform PPI, but performance is more variable.

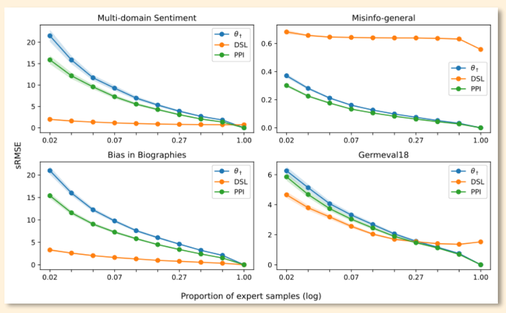
## Experiment 1

Vary the proportion of samples also annotated by human experts:



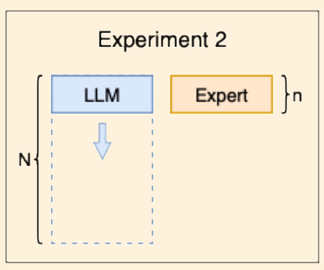

Performance per dataset:

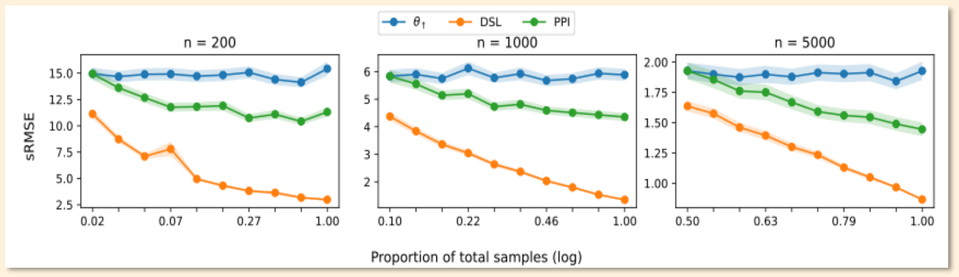



Given a fixed number of samples, how many of them should I annotate by hand to observe significant benefits from the debiasing methods?

## Experiment 2

Vary the number of samples with generated annotations:




Given a fixed number of expert-annotated samples, how many additional samples should I annotate with LLMs? Are there diminishing returns?

Scan to read the full paper!

30th ANNIVERSARY EMNLP 2025 Suzhou, China 中国苏州 November 4-9 11月4日-9日