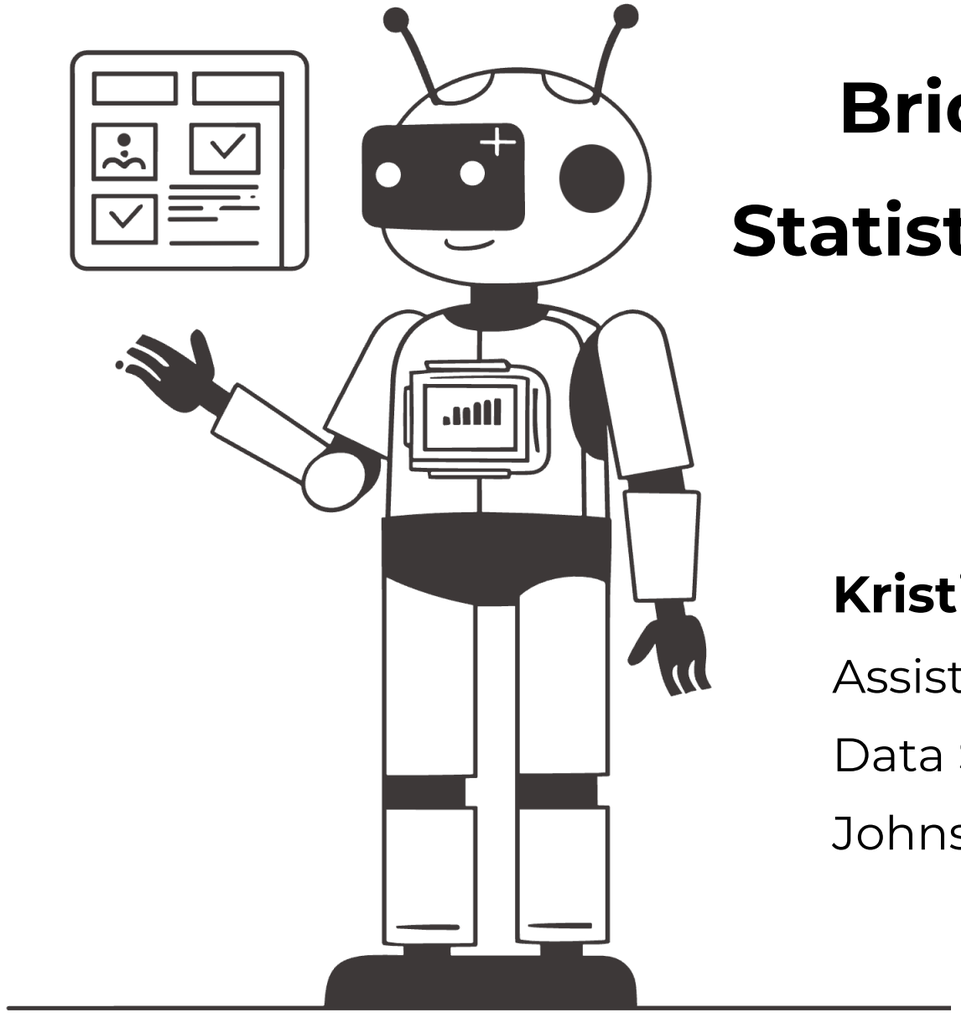


Bridging Human and LLM Annotations for Statistically Valid Computational Social Science

Kristina Gligorić

Assistant Professor, Computer Science Department

Data Science Institute, Center for Language and Speech Processing

Johns Hopkins University

JOHNS HOPKINS
UNIVERSITY

Yay data!

Explosion of Text Data

Opportunities for New Insights

Unlocking New Knowledge

Real-world social science questions

- What is the impact of COVID-19 vaccine online misinformation on vaccination intent?
- Does air pollution lower people’s expressed happiness on social media?
- Does negativity influence online news consumption?

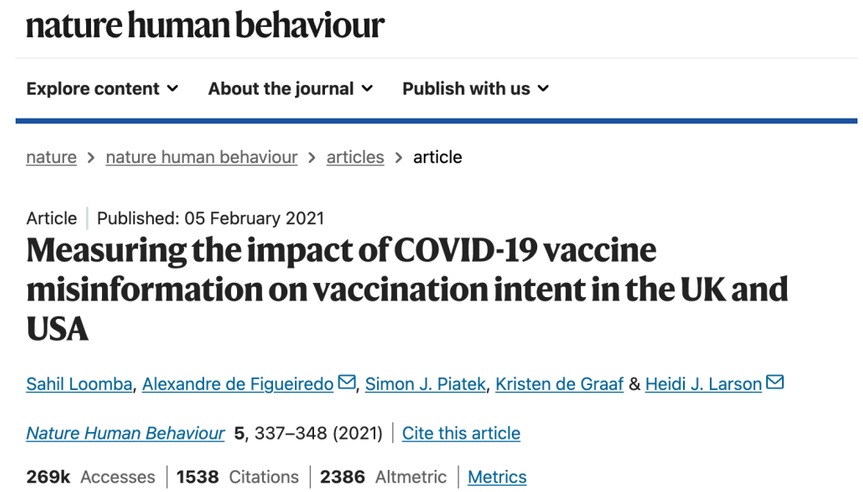
Real-world social science questions

- What is the impact of COVID-19 vaccine online misinformation on vaccination intent?
- Does air pollution lower people's expressed happiness on social media?
- Does negativity influence online news consumption?

Finding an answer requires using large textual datasets (e.g., social media posts) and needs **data labeling**.

Real-world social science questions

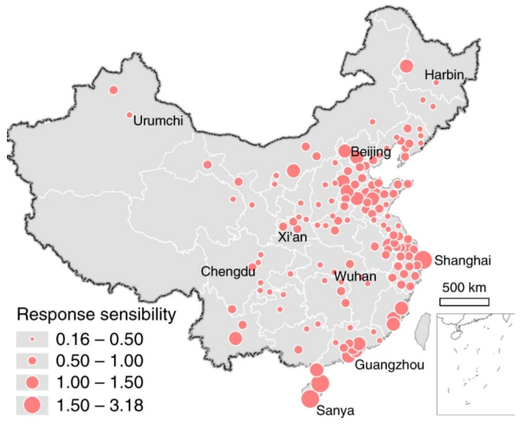
- What is the impact of COVID-19 vaccine online misinformation on vaccination intent?
  - Annotation: “Does this social media post contain misleading claims?”
  - To estimate: whether people who see misinformation online report lower intent to vaccinate

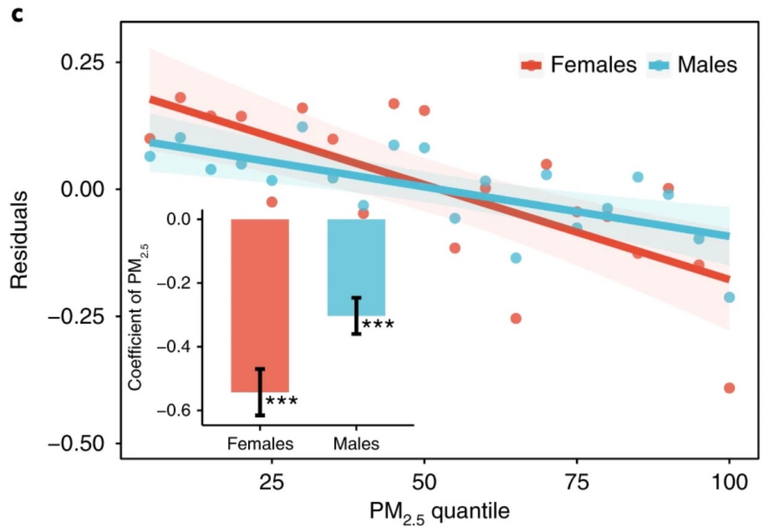


Loomba, S., De Figueiredo, A., Piatek, S. J., De Graaf, K., & Larson, H. J. (2021). Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. *Nature human behavior*, *5*(3), 337-348.

Real-world social science questions

- Does air pollution lower people's expressed happiness on social media?
  - Annotation: "Does this social media post contain high or low positive affect?"
  - To estimate: whether people living in a more polluted environment express less happiness on social media

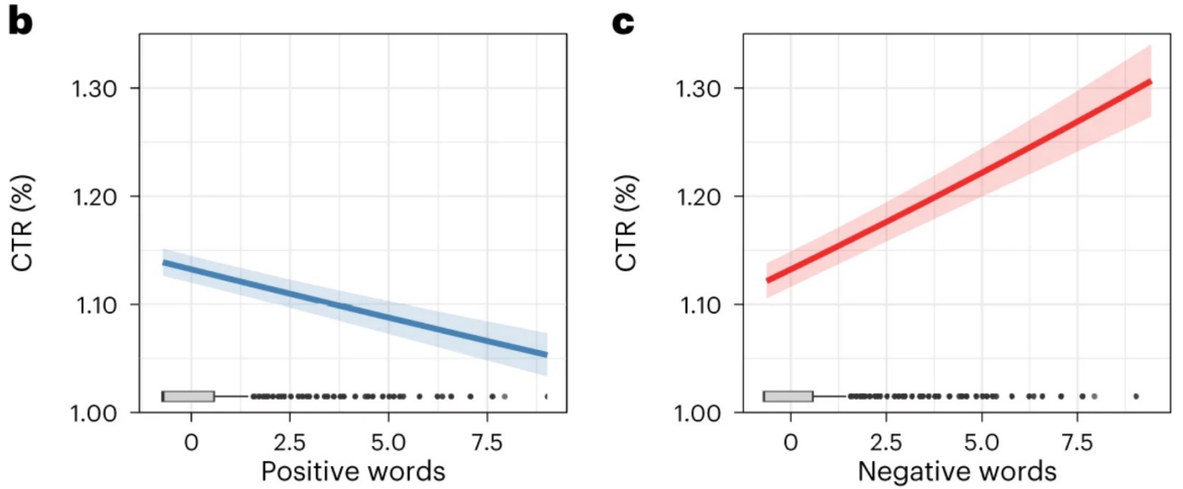




Zheng, S., Wang, J., Sun, C., Zhang, X., & Kahn, M. E. (2019). Air pollution lowers Chinese urbanites' expressed happiness on social media. Nature human behaviour, 3(3), 237-243.

Real-world social science questions

- Does negativity influence online news consumption?
  - Annotation: “Does this online news contain high or low negative affect?”
  - To estimate: whether the negativity of online news predict consumption

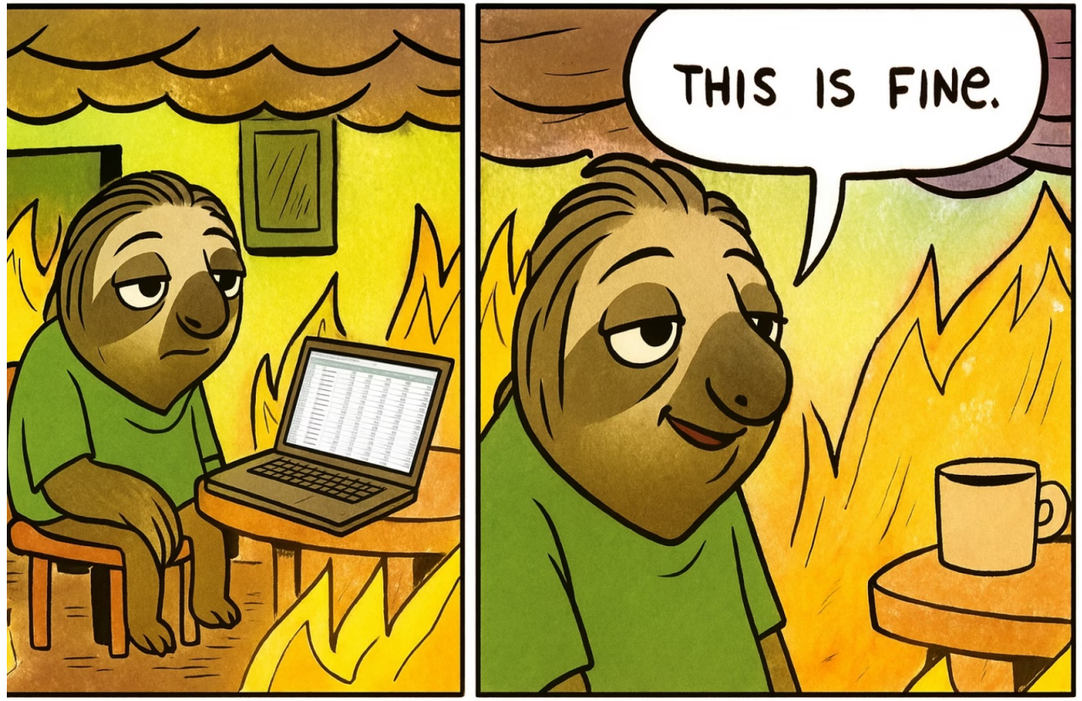


Robertson, C. E., Pröllochs, N., Schwarzenegger, K., Pärnamets, P., Van Bavel, J. J., & Feuerriegel, S. (2023). Negativity drives online news consumption. Nature human behaviour, 7(5), 812-822.

Challenges with Human Annotation

While human annotations are the gold standard for quality and nuance, they are also **slow** and **expensive**

- require aggregating judgments from many annotators
- very costly, especially if coming from experts



Can LLMs Replace Human Annotators?

BRIEF REPORT | POLITICAL SCIENCES |

ChatGPT outperforms crowd workers for text-annotation tasks

Fabrizio Gilardi, Meysam A

March 01 2024

Can Large Language Models Transform Computational Social Science?

In Special Collection: CogNet

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, Diyi Yang

Out of One, Many: Using Language Models to Simulate Human Samples

Published online by Cambridge University Press: **21 February 2023**

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting and David Wingate

PERSPECTIVE | SOCIAL SCIENCES |

Can Generative AI improve social science?

Christopher A. Bail | Authors Info & Affiliations

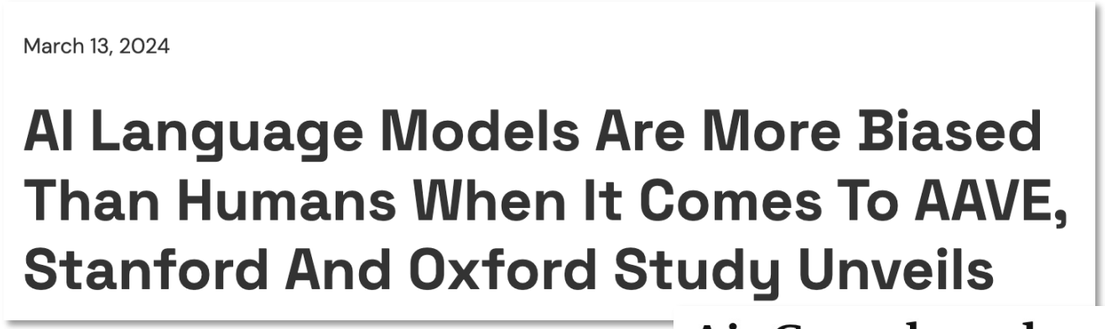
Edited by David Lazer, Northeastern University, Boston, MA; received September 7, 2023; accepted April 5, 2024, by Editorial Board Member Mark Granovetter

May 9, 2024 | 121 (21) e2314021121 | https://doi.org/10.1073/pnas.2314021121

Can LLMs Replace Human Annotators?

LLM annotations are fast & affordable!

... But they do not always align with human judgment (e.g., biases, factual inaccuracies, inconsistency)



This Lecture

Methods for trustworthy social science with possibly untrustworthy LLMs

Basic idea: LLMs shouldn't replace real human data; they should complement it

Best of both words: leverage power of AI + retain scientific rigor

We will be following notation from **Confidence-Driven Inference (CDI)** (Gligoric*, Zrnic*, Lee*, Candes, Jurafsky [NAACL, 2025])
Method as presented will combine ideas from several works, which we will reference along the way

Stanford Data Science

Stanford SPARQ

Stanford | NLP

Kristina Gligorić*

Tijana Zrnić*

Cinoo Lee*

Emmanuel Candes

Dan Jurafsky

Setup

T_1	H_1
T_2	H_2
⋮	⋮
T_{N-1}	H_{N-1}
T_N	H_N

N text instances T_i

missing human annotations H_i

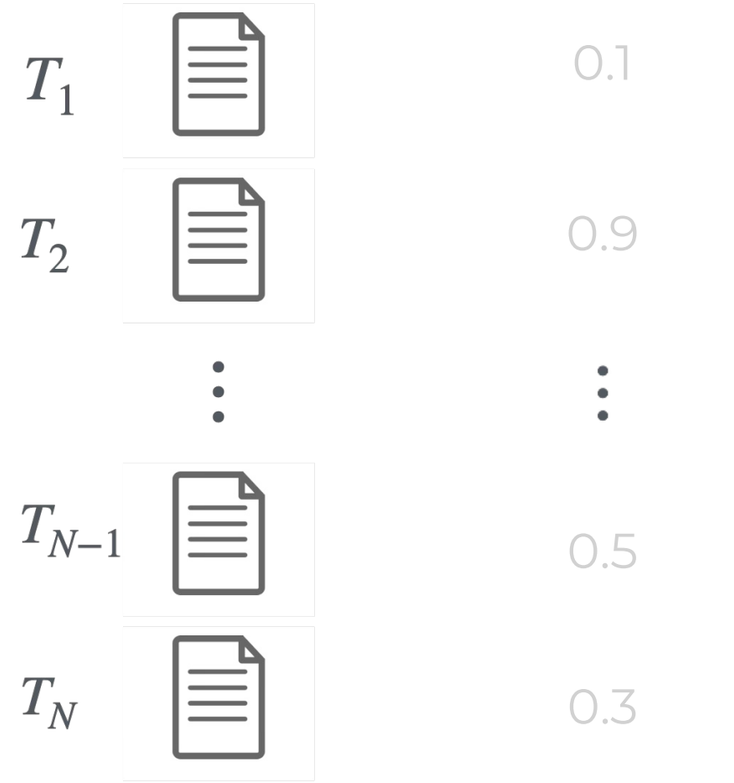
media article	political leaning
T_1 	left
T_2 	right
⋮	⋮
T_{N-1} 	right
T_N 	center

Setup

T_1	H_1
T_2	H_2
⋮	⋮
T_{N-1}	H_{N-1}
T_N	H_N

N text instances T_i

missing human annotations H_i

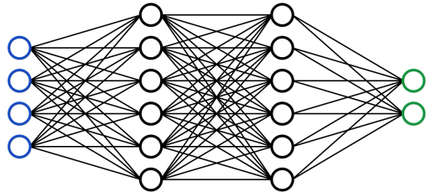


Setup

T_1	H_1
T_2	H_2
⋮	⋮
T_{N-1}	H_{N-1}
T_N	H_N

N text instances T_i

missing human annotations H_i



large language model

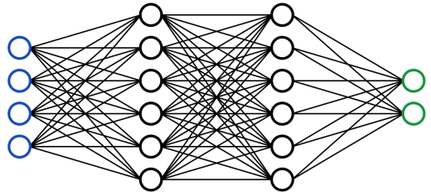
can be used to produce $\hat{H}_i$ that approximate human annotations H_i

Setup

T_1	$\hat{H}_1$
T_2	$\hat{H}_2$
⋮	⋮
T_{N-1}	$\hat{H}_{N-1}$
T_N	$\hat{H}_N$

N text instances T_i

missing human annotations H_i



large language model

issue: $\hat{H}_i$ are potentially biased annotations!

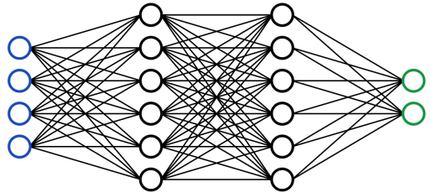
Unless we are willing to assume that the LLM is accurate, there is no hope of reaching valid conclusions without any human annotations!

Setup

T_1	$\hat{H}_1$
T_2	$\hat{H}_2$
⋮	⋮
T_{N-1}	$\hat{H}_{N-1}$
T_N	$\hat{H}_N$

N text instances T_i

missing human annotations H_i



large language model

budget: can collect at most $n \ll N$ human annotations

goal: estimate quantity of interest θ^*

examples of θ^*:

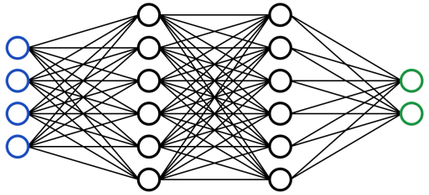
- change in political leaning on X after Elon Musk acquisition
- effect of certain linguistic devices on perceived sentiment
- whatever we care about learning once we have human annotations!

Setup

T_1	$\hat{H}_1$
T_2	$\hat{H}_2$
⋮	⋮
T_{N-1}	$\hat{H}_{N-1}$
T_N	$\hat{H}_N$

N text instances T_i

missing human annotations H_i



large language model

important note: $\mathrm{H_i}$ do not necessarily correspond to annotations from a *single* human
They are “gold” annotations; e.g., obtained by aggregating annotations from multiple annotators.

A Special Case: $\boldsymbol{\theta^* = \mathbf{mean}(H_i)}$

Example: $H_i \in \{0,1\}$ indicates if article has right leaning; θ^* = prevalence of right leaning

$$\theta^* = \mathrm{mean}(H_i) = \frac{1}{\mathrm{N}}(1 + 1 + 0 + 1 + \cdots + 0) = \text{fraction of right-leaning articles}$$

right-leaning left-leaning

A Special Case: $\boldsymbol{\theta^* = \mathbf{mean}(H_i)}$

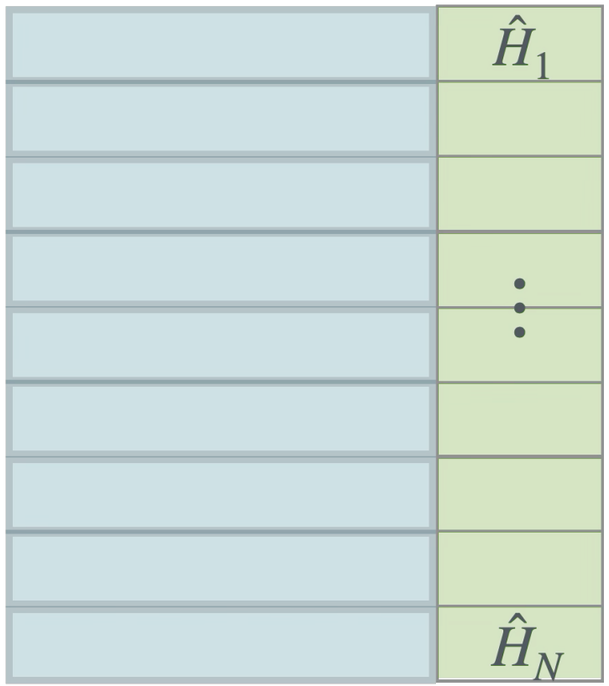
Example: $H_i \in \{0,1\}$ indicates if article has right leaning; θ^* = prevalence of right leaning

Step 1: Collect LLM annotations for all texts

A Special Case: $\boldsymbol{\theta^* = \mathbf{mean}(H_i)}$

Example: $H_i \in \{0,1\}$ indicates if article has right leaning; θ^* = prevalence of right leaning

Step 1: Collect LLM annotations for all texts

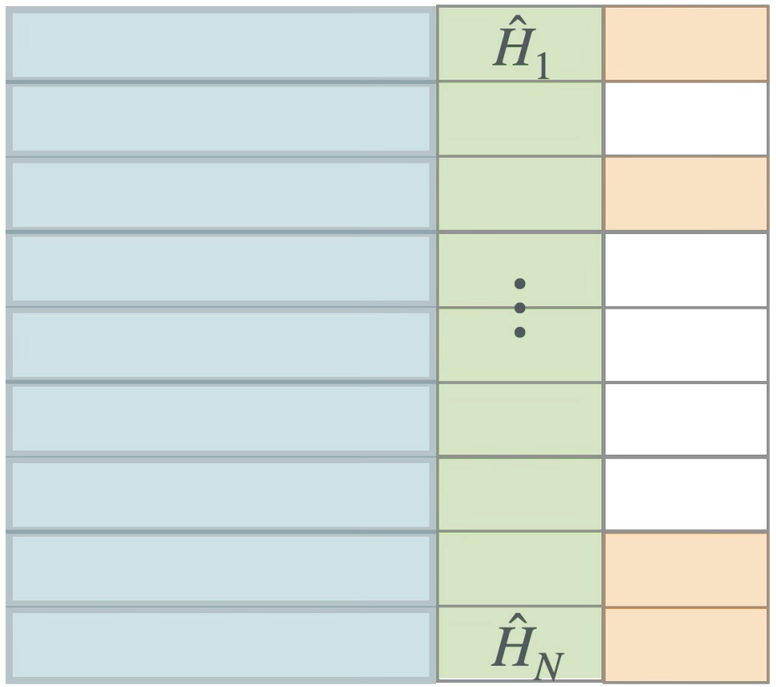


A Special Case: $\boldsymbol{\theta^* = \mathbf{mean}(H_i)}$

Example: $H_i \in \{0,1\}$ indicates if article has right leaning; θ^* = prevalence of right leaning

Step 1: Collect LLM annotations for all texts

Step 2: Collect n human annotations uniformly at random

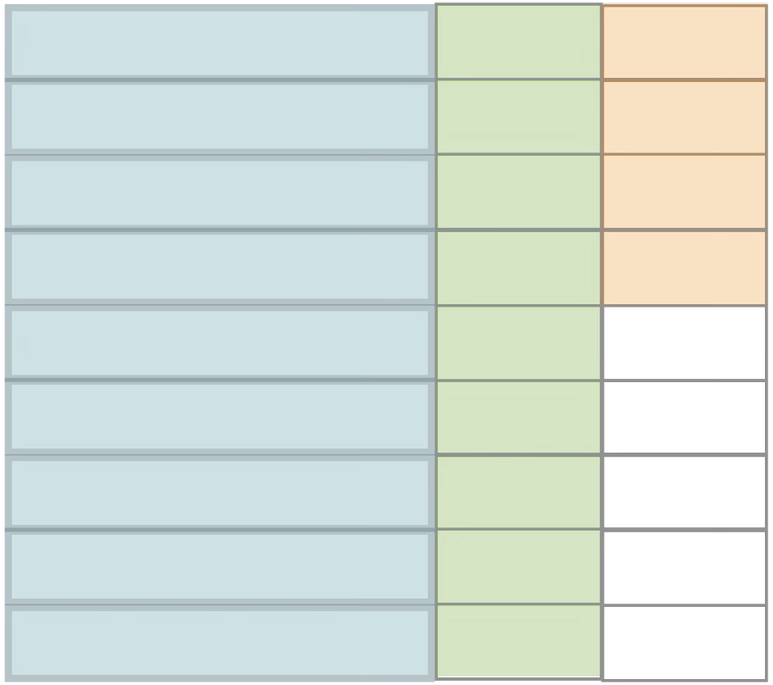


A Special Case: $\boldsymbol{\theta^* = \mathbf{mean}(H_i)}$

Example: $H_i \in \{0,1\}$ indicates if article has right leaning; θ^* = prevalence of right leaning

Step 1: Collect LLM annotations for all texts

Step 2: Collect n human annotations uniformly at random



A Special Case: $\theta^* = \mathbf{mean}(H_i)$

Example: $H_i \in \{0,1\}$ indicates if article has right leaning; θ^* = prevalence of right leaning

Step 1: Collect LLM annotations for all texts

Step 2: Collect n human annotations uniformly at random

T_1	$\hat{H}_1$	H_1
⋮	⋮	⋮
T_n	$\hat{H}_n$	H_n
T_{n+1}	$\hat{H}_{n+1}$	H_{n+1}
⋮	⋮	⋮
T_N	$\hat{H}_N$	H_N

House of Representatives …	0	1
The Pentagon accidentally …	0	0
Democrats clash over …	1	1
Gun lobby may emerge …	0	0
Senate confirms FBI …	0	
Senate Coronavirus Bill …	1	
What does climate change …	1	
Bipartisan Harvard panel …	1	
Elon Musk has idea to …	0	

A Special Case: $\theta^* = \mathbf{mean}(H_i)$

Example: $H_i \in \{0,1\}$ indicates if article has right leaning; θ^* = prevalence of right leaning

Step 1: Collect LLM annotations for all texts

Step 2: Collect n human annotations uniformly at random

Step 3: Given $(H_1, \hat{H}_1), \ldots, (H_n, \hat{H}_n), \hat{H}_{n+1}, \ldots, \hat{H}_N$, compute estimate of θ^*

$$\hat{\theta}^{\text{PPI}} = \underbrace{\text{mean}(\hat{H}_{n+1}, \ldots, \hat{H}_N)}_{\text{naïve estimate}} - \underbrace{\text{mean}(\hat{H}_1 - H_1, \ldots, \hat{H}_n - H_n)}_{\text{bias}}$$

Prediction-powered inference. Angelopoulos, Bates, Fannjiang, Jordan, Zrnic [Science, 2023]

Design-based supervised learning. Egami, Hinck, Stewart, Wei [NeurIPS, 2023]

A Special Case: $\theta^* = \mathbf{mean}(H_i)$

$$\hat{\theta}^{\mathrm{PPI}} = \mathrm{mean}\left(\hat{H}_{n+1}, \ldots, \hat{H}_N\right) - \mathrm{mean}(\hat{H}_1 - H_1, \ldots, \hat{H}_n - H_n)$$

Theorem. For *any* data, $\hat{\theta}^{\mathrm{PPI}}$ is:

- accurate: $\hat{\theta}^{\mathrm{PPI}} \to \theta^*$ as the data size grows
- well-behaved: $\hat{\theta}^{\mathrm{PPI}} \approx N(\theta^*, \sigma^2)$

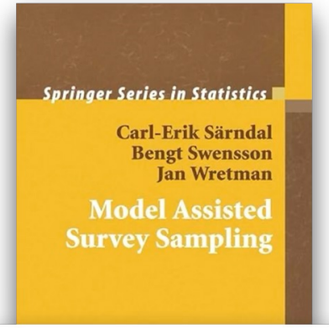
⇒ can form a confidence interval $(\hat{\theta}^{\mathrm{PPI}} \pm r)$ via bootstrap or normal approximation

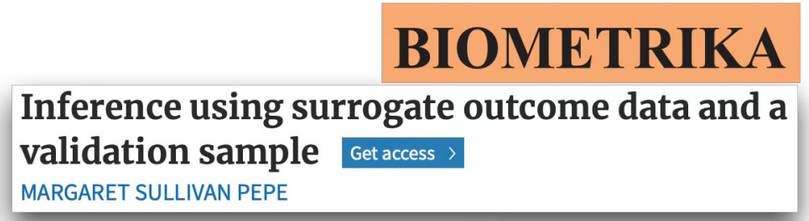
Prediction-powered inference. Angelopoulos, Bates, Fannjiang, Jordan, Zrnic [Science, 2023]

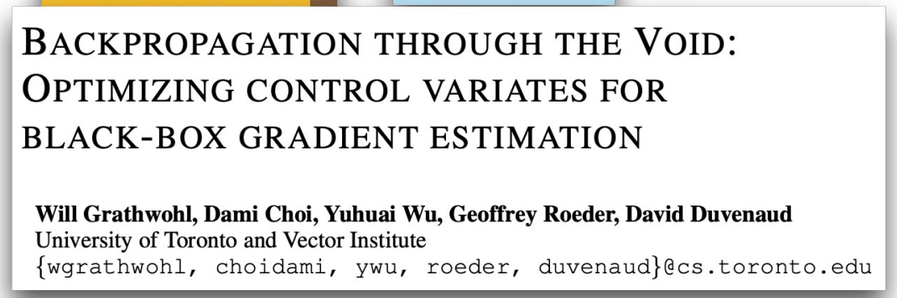
Design-based supervised learning. Egami, Hinck, Stewart, Wei [NeurIPS, 2023]

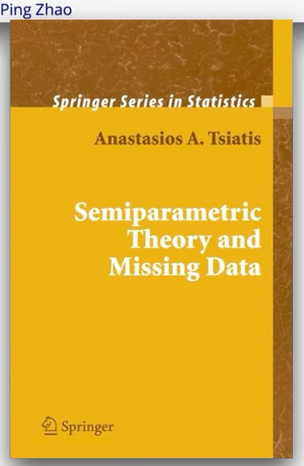
A Special Case: $\theta^* = \mathbf{mean}(H_i)$

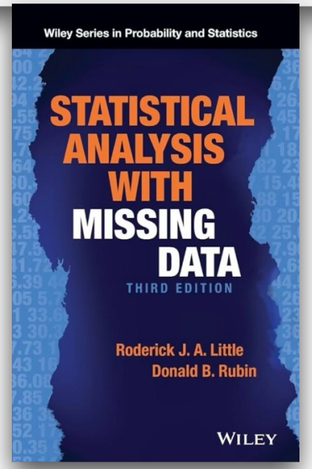
$$\hat{\theta}^{\text{PPI}} = \text{mean}\left(\hat{H}_{n+1}, \ldots, \hat{H}_N\right) - \text{mean}(\hat{H}_1 - H_1, \ldots, \hat{H}_n - H_n)$$

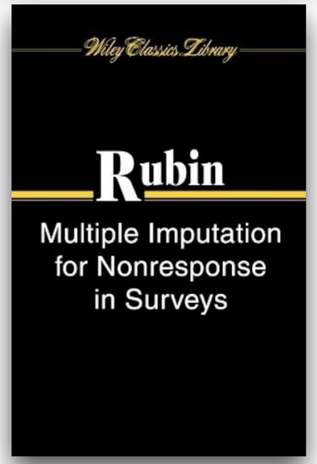


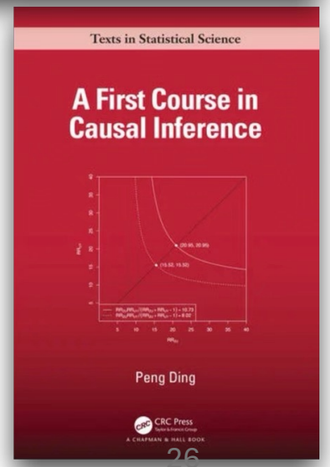












Political Leaning

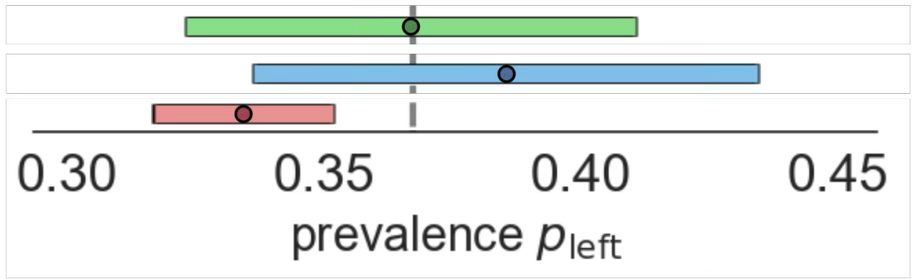
T_i — media articles*

H_i — human annotations of political leaning

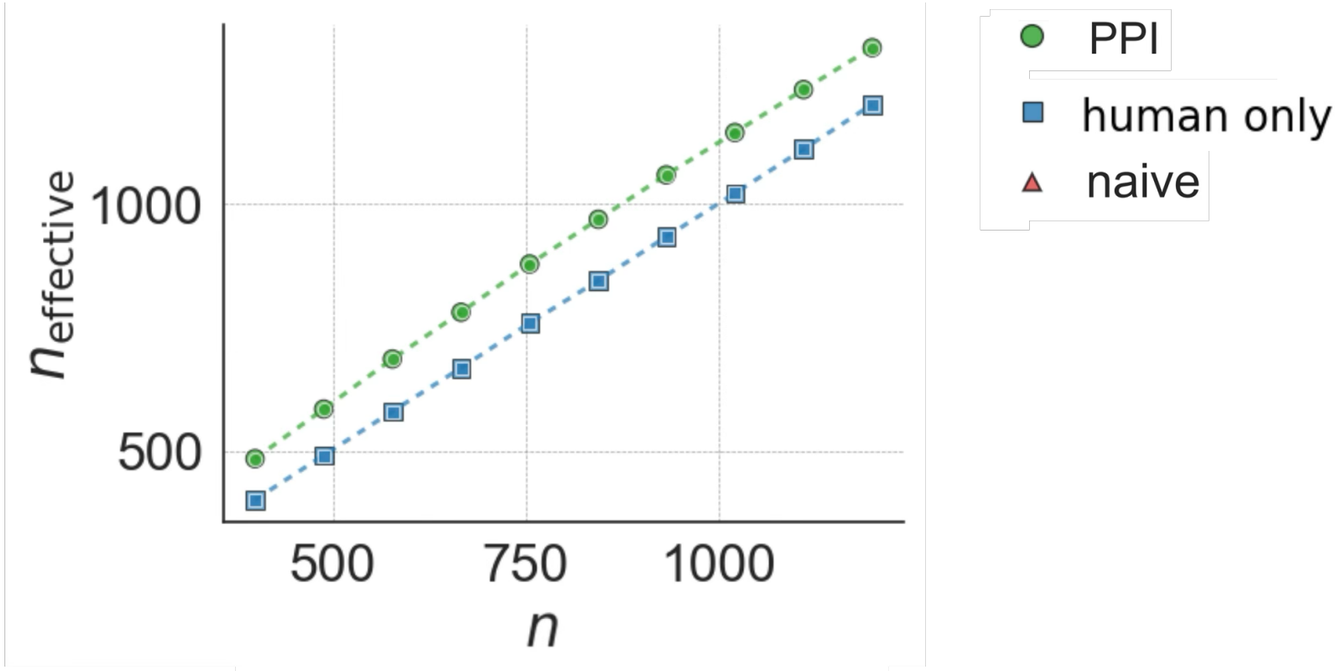
θ^* — fraction of left-leaning articles

LLM — GPT-4o

Naively relying on LLMs is risky!



$$\hat{\theta}^{\text{naive}} = \text{mean}(\hat{H}_1, \ldots, \hat{H}_N)$$



*Baly et al. EMNLP, 2020

What are we missing?

1) That was only mean estimation. I want to run regressions (e.g., compute causal effects) and compute other more complex statistics (e.g. correlations, odds ratios, etc).

All these points are addressed by **Confidence-Driven Inference (CDI)**

General Quantities of Interest

1)

We want to learn $\theta^* = \hat{\theta}\left((X_i, H_i)_{i=1}^N\right)$, where X_i are (optionally) additional side covariates

θ^* = logistic regression coef. of $H \sim X$

is polite?

contains gratitude words?

General Quantities of Interest

1)

We want to learn $\theta^* = \hat{\theta}\left((X_i, H_i)_{i=1}^N\right)$, where X_i are (optionally) additional side covariates

- e.g. X_i indicates whether T_i contains gratitude words, or which media source the article comes from

General estimator:

$$\hat{\theta}^{\mathrm{CDI}} = \underbrace{\hat{\theta}\left(\left(X_i, \widehat{H}_i\right)_{i=n+1}^N\right)}_{\text{naïve estimate}} - \underbrace{\left(\hat{\theta}\left(\left(X_i, \widehat{H}_i\right)_{i=1}^n\right) - \hat{\theta}((X_i, H_i)_{i=1}^n)\right)}_{\text{bias}}$$

Theorem. For *any* data, $\hat{\theta}^{\mathrm{CDI}}$ is:

- accurate: $\hat{\theta}^{\mathrm{CDI}} \to \theta^*$ as the data size grows
- well-behaved: $\hat{\theta}^{\mathrm{CDI}} \approx N(\theta^*, \sigma^2)$

$\Rightarrow$ can form a confidence interval $(\hat{\theta}^{\mathrm{CDI}} \pm r)$ via bootstrap or normal approximation

Active Data Collection

2)

Human expertise should be reserved for “hard” problems; want $\mathbf{Prob}(\text{collect } H_i)$ large for difficult T_i

It is optimal to have large $\mathbf{Prob}(\text{collect } H_i)$ for instances where $\mathbf{err}(H_i, \widehat{H}_i)$ is the largest

We have to include inverse probabaility weights in the estimator if we sample adaptively:

$$\hat{\theta}^{\text{CDI}} = \hat{\theta}\left((X_i, \widehat{H}_i; \overline{W}_i)_{i=n+1}^{N}\right) - \left(\hat{\theta}\left((X_i, \widehat{H}_i; W_i)_{i=1}^{n}\right) - \hat{\theta}((X_i, H_i; W_i)_{i=1}^{n})\right)$$

$$W_i = \frac{1\{\text{collect } H_i\}}{\text{Prob}(\text{collect } H_i)}$$

$$\overline{W}_i = \frac{1 - 1\{\text{collect } H_i\}}{1 - \text{Prob}(\text{collect } H_i)}$$

Zrnic, Candes [ICML, 2024]
Gligoric, Zrnic, Lee, Candes, Jurafsky [NAACL, 2025]
Kluger, Lu, Zrnic, Wang, Bates [2025]

Active Data Collection

2)

Human expertise should be reserved for “hard” problems; want $\mathrm{Prob}(\text{collect } H_i)$ large for difficult T_i

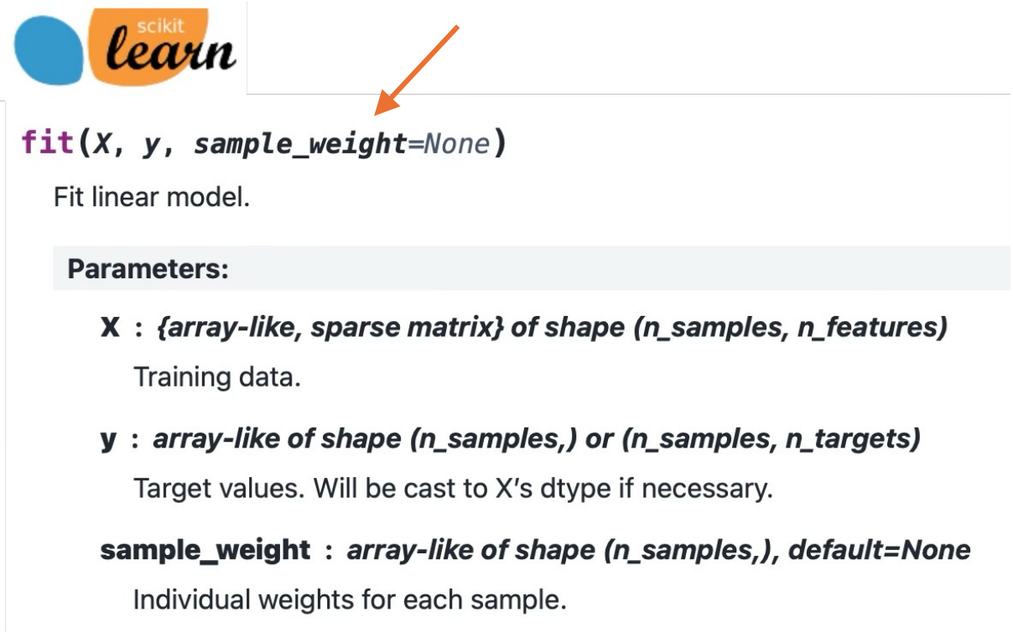
It is statistically optimal to have large $\mathrm{Prob}(\text{collect } H_i)$ for instances where $\mathrm{err}(H_i, \widehat{H}_i)$ is the largest

We have to include inverse probabaility weights in the estimator if we sample adaptively:

$$\hat{\theta}^{\mathrm{CDI}} = \hat{\theta}\left((X_i, \widehat{H}_i; \overline{W}_i)_{i=n+1}^{N}\right) - \left(\hat{\theta}\left((X_i, \widehat{H}_i; W_i)_{i=1}^{n}\right) - \hat{\theta}((X_i, H_i; W_i)_{i=1}^{n})\right)$$

$$W_i = \frac{1\{\text{collect } H_i\}}{\mathrm{Prob}(\text{collect } H_i)}$$

$$\overline{W}_i = \frac{1 - 1\{\text{collect } H_i\}}{1 - \mathrm{Prob}(\text{collect } H_i)}$$



Active Data Collection

2)

Human expertise should be reserved for “hard” problems; want $\mathbf{Prob}(\mathbf{collect}\ H_i)$ large for difficult T_i

It is statistically optimal to have large $\mathbf{Prob}(\mathbf{collect}\ H_i)$ for instances where $\mathbf{err}(H_i, \widehat{H}_i)$ is the largest

We have to include inverse probabaility weights in the estimator if we sample adaptively:

$$\hat{\theta}^{\mathrm{CDI}} = \hat{\theta}\left((X_i, \widehat{H}_i; \overline{W}_i)_{i=n+1}^{N}\right) - \left(\hat{\theta}\left((X_i, \widehat{H}_i; W_i)_{i=1}^{n}\right) - \hat{\theta}((X_i, H_i; W_i)_{i=1}^{n})\right)$$

$$W_i = \frac{1\{\mathrm{collect}\ H_i\}}{\mathrm{Prob}(\mathrm{collect}\ H_i)}$$

$$\overline{W}_i = \frac{1 - 1\{\mathrm{collect}\ H_i\}}{1 - \mathrm{Prob}(\mathrm{collect}\ H_i)}$$

Theorem. For *any* data, $\hat{\theta}^{\mathrm{CDI}}$ is:

- accurate: $\hat{\theta}^{\mathrm{CDI}} \to \theta^*$ as the data size grows
- well-behaved: $\hat{\theta}^{\mathrm{CDI}} \approx N(\theta^*, \sigma^2)$

⇒ can form a confidence interval $(\hat{\theta}^{\mathrm{CDI}} \pm r)$
via bootstrap or normal approximation

Confidence-Driven Inference

2)

To approximately sample where $\mathrm{err}(H_i, \hat{H}_i)$ is the largest, we look at LLM uncertainty

In our experiments, the most useful uncertainties were based on verbalized confidence (Tian et al., 2023)

Stage 1	What is the political bias of the following article? Output either A,B, or C. Output a letter only. A) Left B) Center C) Right Article: <text> Answer:
Stage 2	How likely is it that the following article has a <previously provided answer: left-leaning, centrist, or right-leaning> political bias? Output the probability only (a number between 0 and 1). Text: <text> Probability:

Tian, Mitchell, Zhou, Sharma, Rafailov, Yao, Finn, Manning. EMNLP, 2023

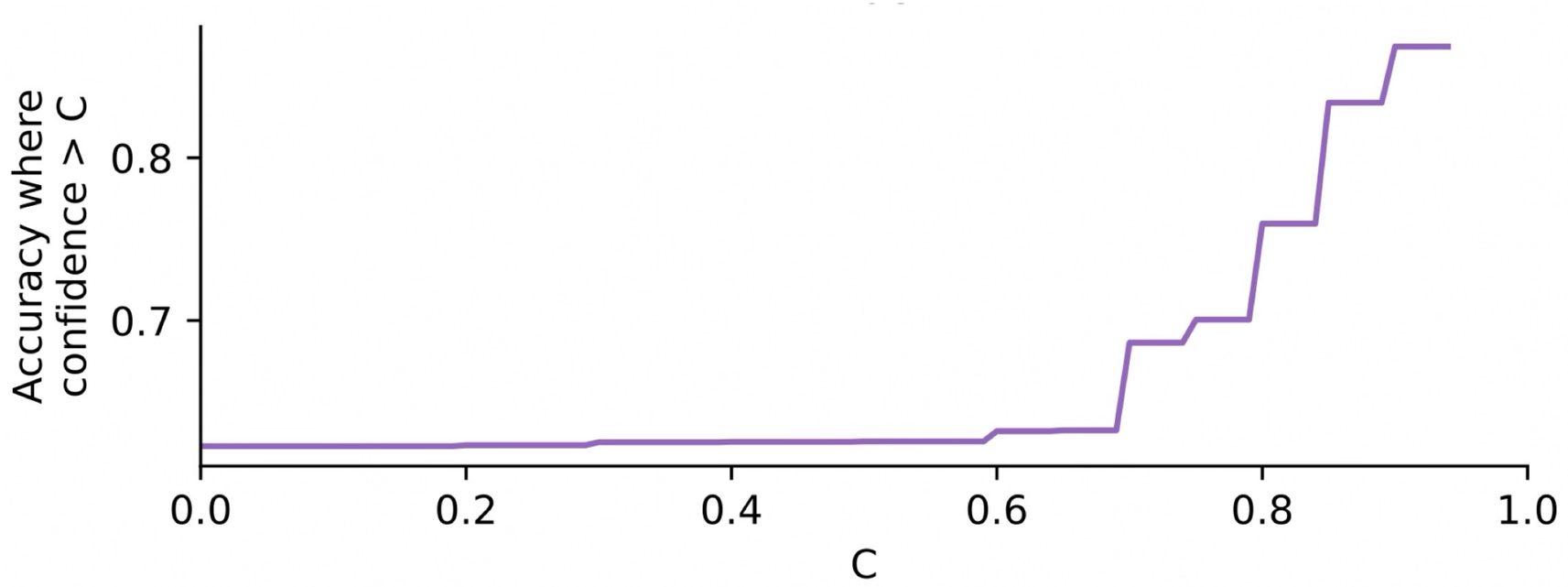
Confidence-Driven Inference

2)

To approximately sample where $\mathrm{err}(H_i, \widehat{H}_i)$ is the largest, we look at LLM uncertainty

In our experiments, the most useful uncertainties were based on verbalized confidence (Tian et al., 2023)

Confidence reflects accuracy!



Gligoric, Zrnic, Lee, Candes, Jurafsky. NAACL, 2025

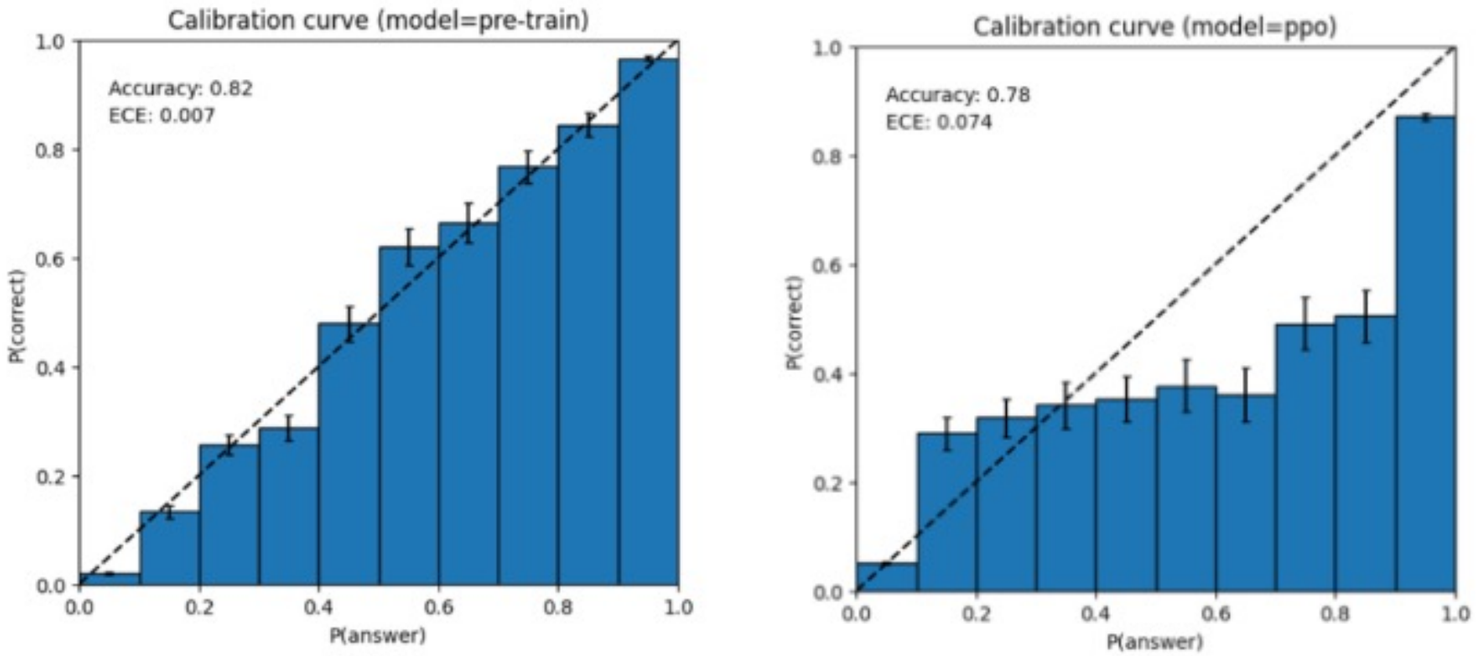
Confidence-Driven Inference

2)

To approximately sample where $\mathrm{err}(H_i, \hat{H}_i)$ is the largest, we look at LLM uncertainty

In our experiments, the most useful uncertainties were based on verbalized confidence (Tian et al., 2023)

Confidence reflects accuracy!



Left: Calibration plot of the pre-trained GPT-4 model on an MMLU subset. The model's confidence in its prediction closely matches the probability of being correct. The dotted diagonal line represents perfect calibration. Right: Calibration plot of post-trained PPO GPT-4 model on the same MMLU subset. Our current process hurts the calibration quite a bit.

https://openai.com/index/gpt-4-research/

Confidence-Driven Inference

2)

To approximately sample where $\mathrm{err}(H_i, \widehat{H}_i)$ is the largest, we look at LLM uncertainty

In our experiments, the most useful uncertainties were based on verbalized confidence (Tian et al., 2023)

Confidence reflects accuracy!

We fit a mapping $\widehat{\mathrm{err}}$ from confidence C_i to $\mathrm{err}(H_i, \widehat{H}_i)$ as we collect data and set $\mathrm{Prob}(\mathrm{collect}\ H_i) \propto \widehat{\mathrm{err}}(C_i)$ (normalized to meet the budget constraint)

Safeguard Against Poor LLM Annotations

3)

Power tuning interpolates between using and not using LLM annotations

$$\hat{\theta}^{\lambda} = \lambda \cdot \hat{\theta}\left((X_i, \widehat{H}_i; \overline{W}_i)_{i=n+1}^{N}\right) - (\lambda \cdot \hat{\theta}\left((X_i, \widehat{H}_i; W_i)_{i=1}^{n}\right) - \hat{\theta}((X_i, H_i; W_i)_{i=1}^{n}))$$

$\lambda = 0$

human-only

Safeguard Against Poor LLM Annotations

3)

Power tuning interpolates between using and not using LLM annotations

$$\hat{\theta}^{\lambda} = \lambda \cdot \hat{\theta}\left(\left(X_i, \widehat{H}_i; \widehat{W}_i\right)_{i=n+1}^{N}\right) - \left(\lambda \cdot \hat{\theta}\left(\left(X_i, \widehat{H}_i; W_i\right)_{i=1}^{n}\right) - \hat{\theta}((X_i, H_i; W_i)_{i=1}^{n})\right)$$

$\lambda = 0$ human-only

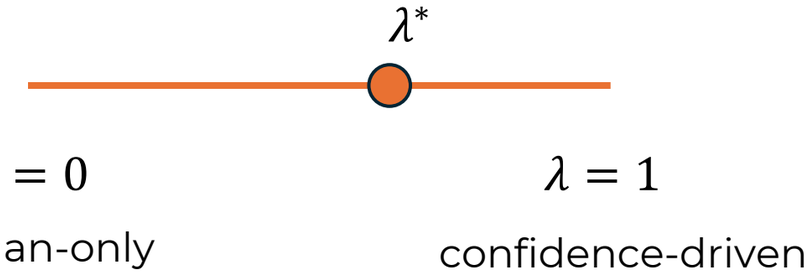
$\lambda = 1$ confidence-driven

Safeguard Against Poor LLM Annotations

3)

Power tuning interpolates between using and not using LLM annotations

$$\hat{\theta}^{\lambda} = \lambda \cdot \hat{\theta}\left((X_i, \widehat{H}_i; \widehat{W}_i)_{i=n+1}^{N}\right) - (\lambda \cdot \hat{\theta}\left((X_i, \widehat{H}_i; W_i)_{i=1}^{n}\right) - \hat{\theta}((X_i, H_i; W_i)_{i=1}^{n}))$$



Optimal tuning λ^* is proportional to how well H and $\widehat{H}$ correlate and can be computed explicitly

Theorem. After tuning, $\hat{\theta}^{\lambda}$ achieves lower MSE than both $\hat{\theta}^{\mathbf{CDI}}$ and $\hat{\theta}^{\mathbf{human}}$.

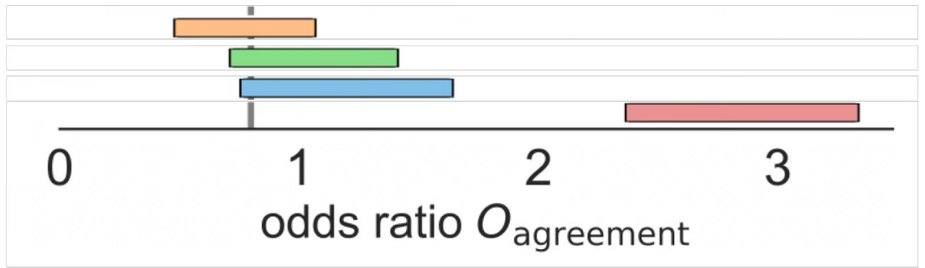
Media Stance on Global Warming

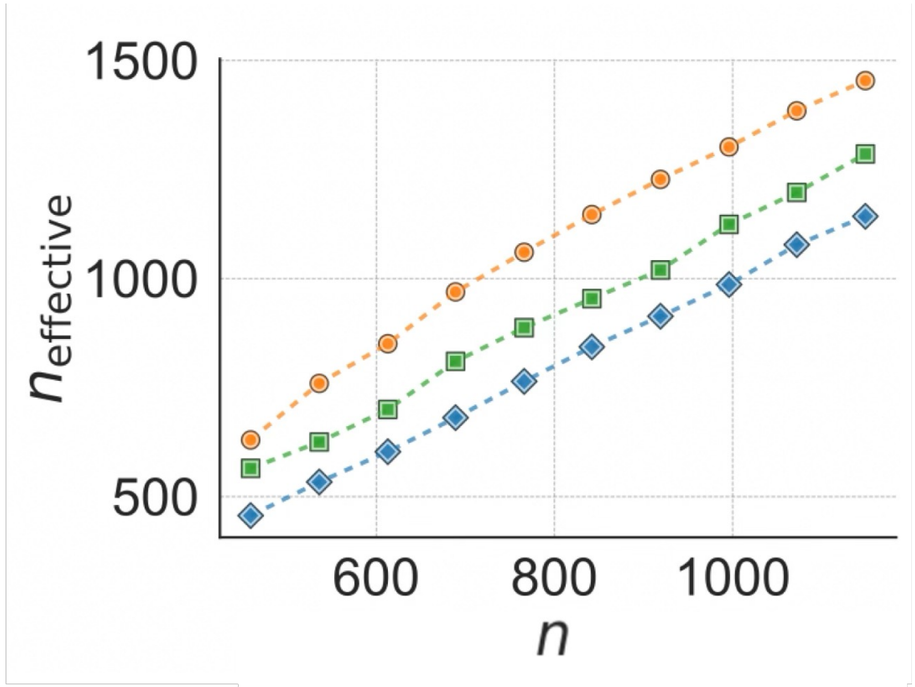
T_i — news headlines*

H_i — human annotations of article stance on global warming

θ^* — odds ratio quantifying relationship between affirming devices (e.g. “expert”, “award-winning scientist”) and stance on global warming

LLM — GPT-4o





*Luo et al. EMNLP, 2020

Confidence-Driven Inference: Step by Step

Input: (shuffled) texts T_i, LLM API, human annotation API, estimator of interest $\hat{\theta}$

Step 1: Collect LLM annotations $\widehat{H}_i$ and confidence scores C_i for all texts T_i

Step 2: Collect human annotations H_i for texts $i = 1, \ldots, n_{\text{init}}$; fit mapping $\widehat{\text{err}}$ from C_i to $\text{err}(H_i, \widehat{H}_i)$

Step 3: Set $\text{Prob}(\text{collect } H_i) = \frac{n}{N} \frac{\widehat{\text{err}}(C_i)}{\text{mean}(\widehat{\text{err}}(C_i))}$ for next n_{batch} texts; make sampling decisions $\xi_i \in \{0,1\}$ with $\text{Prob}(\text{collect } H_i)$

Step 4: Collect human annotations H_i for texts with $\xi_i = 1$; refit $\widehat{\text{err}}$ from all collected data so far

Step 5: Repeat Steps 3-4 until pass through all N texts is finished

Step 6: Compute tuning parameter λ and final estimate

Step 7: Compute confidence interval $(\hat{\theta}^{\lambda} \pm r)$ via bootstrap

Output: estimate $\hat{\theta}^{\lambda}$ and confidence interval $(\hat{\theta}^{\lambda} \pm r)$

A practical example

Questions so far?

Politeness

Oxford **Learner's Dictionaries**

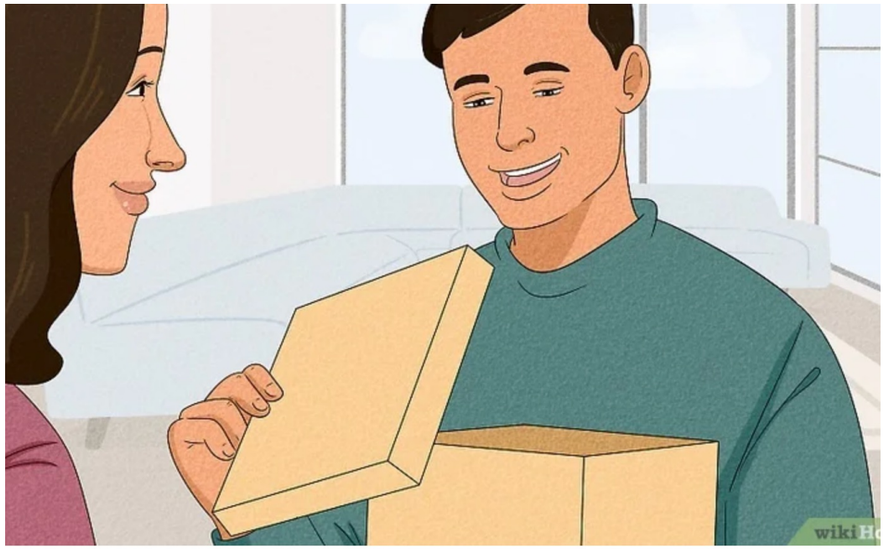
politeness *noun*

1 ★ good manners and respect for the feelings of others

SYNONYM **courtesy** (1)

wikiHow to do anything... PRO

What to Do When You Get a Gift You Don't Like



1 Act naturally.

You don't need to feign excitement. Instead, summon up a positive feeling by thinking how nice it was that someone is giving you a gift.

- Try to react immediately. If you pause after you open the gift, you might seem disappointed.
- Smile if you can. It might help to remind yourself that they were trying to make you happy!

What can politeness annotation tell us?

Surprisingly a lot!

Politeness annotation can inform various types of research questions:

e.g., gender, race, status and power

What can politeness annotation tell us?

Is there a gender difference in language use?

Gender Differences in Language Use:
An Analysis of 14,000 Text Samples

Matthew L. Newman
Department of Social and Behavioral Sciences
Arizona State University

Carla J. Groom*
Department of Psychology
The University of Texas at Austin

Lori D. Handelman
Oxford University Press
New York

James W. Pennebaker
Department of Psychology
The University of Texas at Austin

Women use polite forms and

hedging more than men

(“Would you mind if…”, “I guess…”)

What can politeness annotation tell us?

Do police talk to White and Black drivers differently? If yes, how?

What can politeness annotation tell us?

Do police talk to White and Black drivers differently? If yes, how?

RESEARCH ARTICLE | PSYCHOLOGICAL AND COGNITIVE SCIENCES |

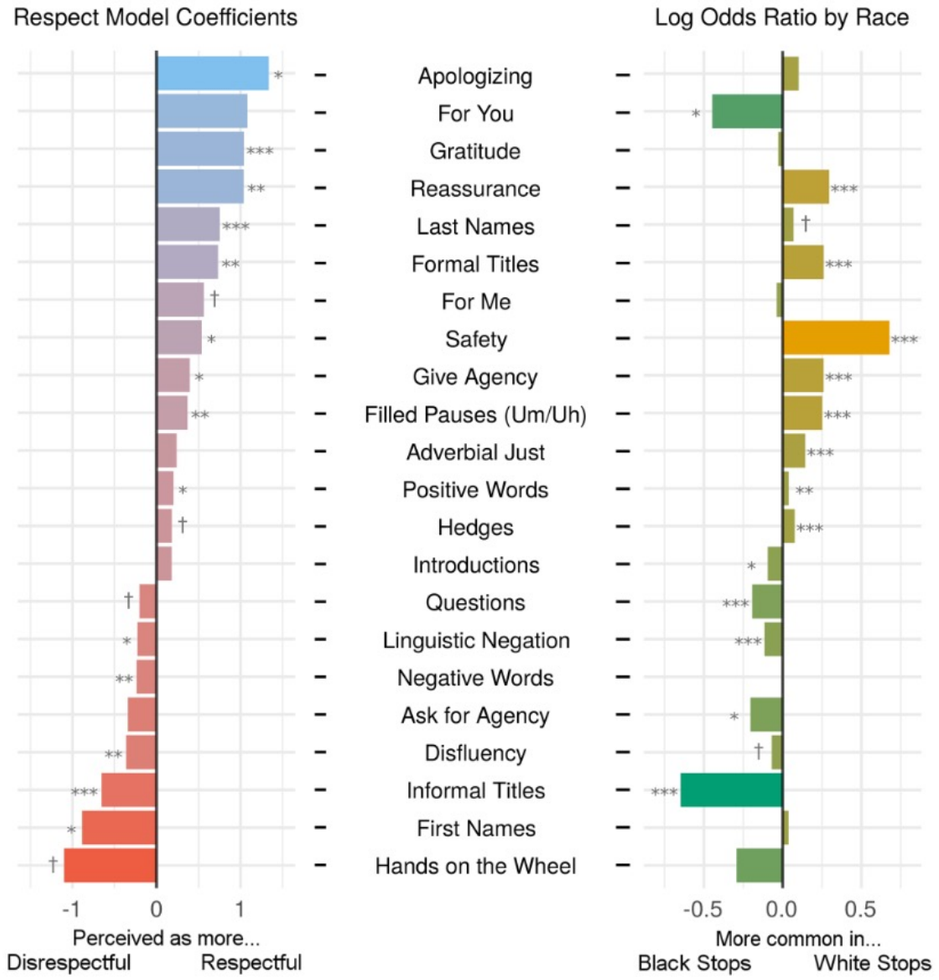
Language from police body camera footage shows racial disparities in officer respect

Rob Voigt, Nicholas P. Camp, Vinodkumar Prabhakaran, +5, and Jennifer L. Eberhardt Authors Info & Affiliations

> “Officers speak with consistently less respect/politeness toward black vs. white community members, even after controlling for the race of the officer, the severity of the infraction, the location of the stop, and the outcome of the stop.”

What can politeness annotation tell us?

Do police talk to White and Black drivers differently? If yes, how?



What can politeness annotation tell us?

Does power corrupt?

A Computational Approach to Politeness with Application to Social Factors

Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, Christopher Potts

What can politeness annotation tell us?

Does power corrupt?

Yes!

Wikipedia editors who would eventually be elected to administrator roles were significantly more polite than other users **before** their promotion.

However, **after** election (higher-status position), they became less polite.

What can politeness annotation tell us?

Does power corrupt?

Yes!

Similarly, on Stack Exchange, users with the highest reputation scores were found to be less polite than users with low or middle-level reputations.

Politeness data

A Computational Approach to Politeness with Application to Social Factors

Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, Christopher Potts

Dataset comes from the Wikipedia community of editors and the Stack Exchange question-answer community

Politeness data

The Stack Exchange question-answer community

The best way to get from Stockholm to Kolmården zoo by public transport

Asked 12 years, 2 months ago Modified 5 years, 2 months ago Viewed 6k times

7

I'm going to visit Stockholm with kids and wonder what is the best option to get to the Kolmården zoo. My first thought was to take the train as the fastest transport, but as far as I understand, trains station is not that close to the Zoo.

So, what will be better: to get to the train station and go by local bus or to take the bus from the very beginning?

public-transport sweden stockholm zoos

3 Answers Sorted by: Highest score (default)

7

For this kind of question, resrobot is an excellent tool.

It appears the quickest transport is a combination of train and bus:

- Train Stockholm – Norrköping; for example, 07:59 - 09:23, 08:21 - 09:33, or 09:40 - 11:11.
- Bus 432 or 433 Norrköping – Kolmården. For example, 10:00 - 10:09, or 11:47 - 12:00.

Kolmården Djurpark is 27 km from the Norrköping central station. Waiting times between train and bus appear rather long (more than 30 minutes). If you're many, you could consider taking a taxi. When booked via SJ, a taxi costs roughly 600 SEK, but a look at the prices for Taxikurir Norrköping suggests it's probably less than 400 SEK there (I get 330 SEK based on their prices). This can be compared to the bus, which costs 74 SEK for an adult and 51 SEK for youth or senior; so for a family of 2 parents, 2 children, the bus would be 250 SEK.

Politeness data

Wikipedia community of editors

Johns Hopkins University

74 languages

Article Talk | Read Edit View history Tools

From Wikipedia, the free encyclopedia

Coordinates: 39°19′44″N 76°37′13″W

"JHU" redirects here. For the Sri Lankan political party, see Jathika Hela Urumaya.

Johns Hopkins University (often abbreviated as **Johns Hopkins**, **Hopkins**, or **JHU**) is a private research university in Baltimore, Maryland, United States. Founded in 1876 based on the European research institution model, Johns Hopkins is considered to be the first research university in the U.S.[8][9]

The university was named for its first benefactor, the American entrepreneur and Quaker philanthropist Johns Hopkins.[10] Hopkins's $7 million bequest (equivalent to $166 million in 2024)[11] to establish the university and the affiliated Johns Hopkins Hospital in Baltimore was the largest philanthropic gift in U.S. history up to that time.[12][13] Daniel Coit Gilman, who was inaugurated as Johns Hopkins's first president on February 22, 1876,[14] led the university to revolutionize higher education in the U.S. by integrating teaching and research.[15] In 1900, Johns Hopkins became a founding member of the Association of American Universities.[16] The university has led all U.S. universities in annual research and development expenditures for over four consecutive decades.[17][18] The School of Medicine, established in 1893, has achieved international recognition for its pioneering biomedical research.

Johns Hopkins University

Latin: *Universitas Hopkinsiensis*[1][2]

Motto	*Veritas vos liberabit* (Latin)
Motto in English	"The truth will set you free"
Type	Private research university
Established	February 22, 1876; 149 years ago
Accreditation	MSCHE

Politeness data

“talk page”

- (cur | prev) ○ 23:15, 26 April 2023 ElKevbo (talk | contribs) . . (7,797 bytes) (+271) . . *(→Inclusion of "consistently ranked among the top and most prestigious universities in the United States and the world" in the lede: If this is information that is important enough to be included in the lede, editors should be able to provide sources that explicitly support it)* (undo)
- (cur | prev) ○ 17:56, 25 April 2023 Sauzer (talk | contribs) . . (7,526 bytes) **(+642)** . . *(→Inclusion of "consistently ranked among the top and most prestigious universities in the United States and the world" in the lede: Reply)* (undo) *(Tag: Reply)*

Politeness data

Dataset: *requests* from Wikipedia editors & Stack Exchange question-answer community.

Politeness data

Dataset: *requests* from Wikipedia editors & Stack Exchange question-answer community.

For each **request**, the annotator has to indicate how polite they perceived the request to be by using a slider with values ranging from “very impolite” to “very polite.”

For sake of simplicity, we’ll be using 0 (not polite) and 1 (polite).

Politeness data

Our goal is to estimate **two target statistics**:

***mean*(H):** prevalence of politeness, i.e., the fraction of texts in the corpus that are polite.

Politeness data

Our goal is to estimate **two target statistics**:

***mean*(H):** prevalence of politeness, i.e., the fraction of texts in the corpus that are polite.

β_{hedge}: the impact of linguistic features of hedging (X) on the perceived politeness (H), estimated with a logistic regression.

- Essentially measuring whether a request having “I suggest…” influences the politeness rating
- E.g., we could estimate that hedging (e.g., “I suggest…”) would make the text 20% more likely to be perceived as polite.

Politeness data

https://github.com/kristinagligoric/cdi-tutorial

Import libraries

In [1]:

```
import numpy as np
from scipy.stats import norm, bernoulli
import pprint
import pandas as pd
from tqdm import tqdm
from tqdm.notebook import tqdm
import time
from ppi_py.utils import bootstrap
import re
import openai
import requests
import json
from datetime import datetime, timezone
import time
import zipfile
import io
import random
from sklearn.linear_model import LogisticRegression
%load_ext autoreload
%autoreload 2

from utils import llms, qualtrics, prolific, mturk, inference
from utils.llms import annotate_texts_with_llm, collect_llm_confidence, get_llm_annotations
from utils.qualtrics import create_and_activate_surveys
from utils.prolific import run_prolific_annotation_pipeline
from utils.mturk import run_mturk_annotation_pipeline
from utils.inference import train_sampling_rule, sampling_rule_predict, confidence_driven_inference, collect_initial_human_annotations, run_adaptive_sampling
```

Politeness data

https://github.com/kristinagligoric/cdi-tutorial

Setup credentials

```
def load_credentials(file_path="credentials.txt"):
    credentials = {}
    with open(file_path, "r") as f:
        for line in f:
            if '=' in line:
                key, value = line.strip().split('=', 1)
                credentials[key.strip()] = value.strip()
    return credentials

# Load credentials, or put your key here in plain text
creds = load_credentials()
AWS_ACCESS_KEY_ID = creds.get("AWS_ACCESS_KEY_ID")
AWS_SECRET_ACCESS_KEY = creds.get("AWS_SECRET_ACCESS_KEY")
OPENAI_API_KEY = creds.get("OPENAI_API_KEY")
QUALTRICS_API_KEY = creds.get("QUALTRICS_API_KEY")
QUALTRICS_API_URL = creds.get("QUALTRICS_API_URL")
PROLIFIC_API_KEY = creds.get("PROLIFIC_API_KEY")
```

Politeness data

https://github.com/kristinagligoric/cdi-tutorial

Set parameters for Confidence-Driven Inference (CDI)

```
# if true, we collect LLM annotations and human annotations from scratch, if false, load pre-collected ones
COLLECT_LLM = False
COLLECT_HUMAN = False

# if COLLECT_HUMAN = True, specify whether to use "Prolific" or "MTURK"
HUMAN_SOURCE = "MTURK"

alpha = 0.1  # desired error level for confidence interval
burnin_steps = 5  # we collect the first burnin_steps points to initialize sampling rule

n_batches = 2

n_human = 15 # budget on number of human annotations (including burnin_steps)

N = 100 # corpus size, or the size of the random subset of the corpus that will be annotated with an LLM

random_state = 42

#define the estimator function
#mask for valid labels and specify how to use weights
def mean_estimator(y, weights):
    y, weights = y[~np.isnan(y)], weights[~np.isnan(y)]
    return np.sum(y * weights) / np.sum(weights)
```

Politeness data

https://github.com/kristinagligoric/cdi-tutorial

Step 1: Collect LLM annotations for all the texts

```
data = get_llm_annotations(df=df,
    text_based_feature=text_based_feature,
    COLLECT_LLM=COLLECT_LLM,
    llm_parameters = llm_parameters,
    N=N,
    random_state=random_state)
```

Step 2: Collect warmup human annotations (initial set)

```
data = collect_initial_human_annotations(
    data=data,
    df=df,
    burnin_steps=burnin_steps,
    COLLECT_HUMAN=COLLECT_HUMAN,
    HUMAN_SOURCE=HUMAN_SOURCE,
    N=N,
    random_state=random_state,
    human_annotation_parameters = human_annotation_parameters)
```

Step 3: Strategically collect human annotations

```
data = run_adaptive_sampling(
    data=data,
    df=df,
    burnin_steps=burnin_steps,
    n_human=n_human,
    n_batches=n_batches,
    COLLECT_HUMAN=COLLECT_HUMAN,
    HUMAN_SOURCE=HUMAN_SOURCE,
    human_annotation_parameters = human_annotation_parameters)
```

```
Collecting batch 1/2...
Collecting 4 human annotations.

Collecting batch 2/2...
Collecting 8 human annotations.
17 human datapoints collected in total.
```

Politeness data

https://github.com/kristinagligoric/cdi-tutorial

Step 4: Compute the CDI estimate and confidence intervals

We showcase estimation of $\text{mean}(H)$: prevalence of the politeness, i.e., the fraction of texts in the corpus that are polite

```
estimate, (lower_bound, upper_bound) = confidence_driven_inference(
    estimator = mean_estimator,
    Y = data['human'].values,
    Yhat = data['llm'].values,
    sampling_probs =  data['sampling_probs'].values,
    sampling_decisions = data['sampling_decisions'].values,
    alpha = alpha)

print("CDI estimate of the target statistic (mean(H): prevalence of politeness):")
print('point estimate:',estimate.round(4))
print('confidence intervals:', lower_bound.round(4), upper_bound.round(4))

print("Ground truth mean(H) estimate (if we had access to human annotations on the full text corpus):")
print(np.mean(pd.read_csv('data/politeness_dataset.csv').sample(n = N, random_state = random_state)['Prediction_human'].values))
```

```
CDI estimate of the target statistic (mean(H): prevalence of politeness):
point estimate: 0.4901
confidence intervals: 0.2563 0.7032
Ground truth mean(H) estimate (if we had access to human annotations on the full text corpus):
0.58
```

Politeness data

https://github.com/kristinagligoric/cdi-tutorial

β estimation

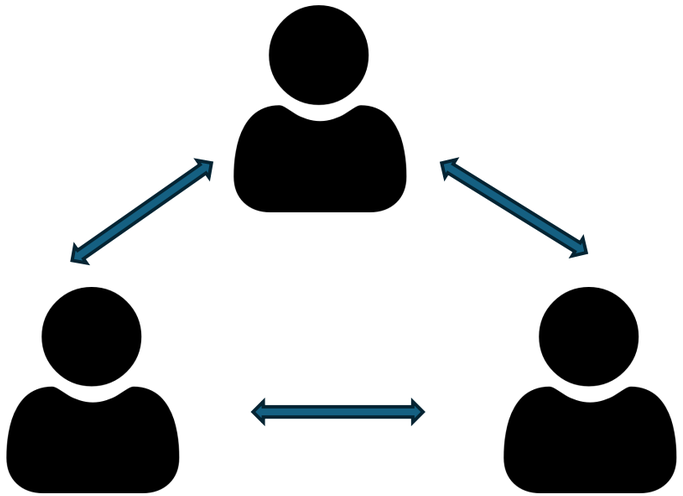
```
estimate, (lower_bound, upper_bound) = confidence_driven_inference(
    estimator = log_reg_estimator,
    Y = data['human'].values,
    Yhat = data['llm'].values,
    X = data['X'].values.reshape(-1, 1),
    sampling_probs =  np.ones(len(data))/len(data),
    sampling_decisions = data['sampling_decisions'].values,
    alpha = alpha)

print("CDI estimate of the target statistic (β: effect of X on H):")
print('point estimate:',estimate.round(4))
print('confidence intervals:', lower_bound.round(4), upper_bound.round(4))
```

```
CDI estimate of the target statistic (β: effect of X on H):
point estimate: 0.4433
confidence intervals: 0.2734 0.5991
```

Alternative approaches vs the ”debiasing route”

The “alternative annotator test”

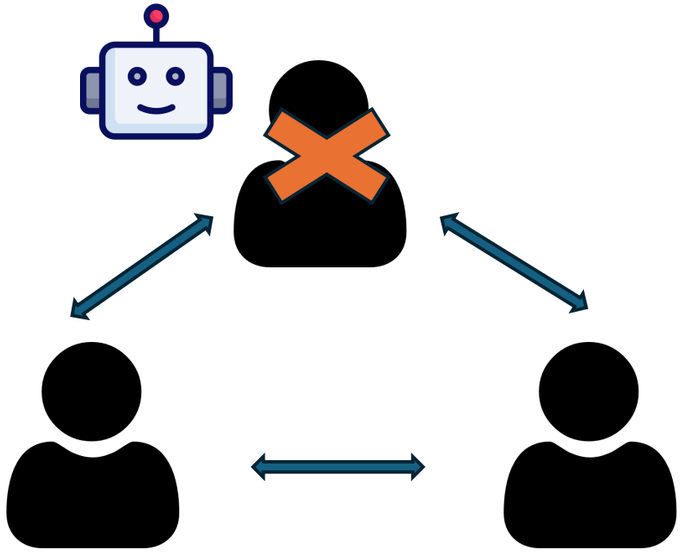


What is the level of disagreement?

Calderon, Nitay, Roi Reichart, and Rotem Dror. "The alternative annotator test for llm-as-a-judge: How to statistically justify replacing human annotators with llms." *arXiv preprint arXiv:2501.10970* (2025).

Alternative approaches vs the ”debiasing route”

The “alternative annotator test”



What is the level of disagreement **now**?

Practical argument of an upper limit, but no bounds on validity

Calderon, Nitay, Roi Reichart, and Rotem Dror. "The alternative annotator test for llm-as-a-judge: How to statistically justify replacing human annotators with llms." *arXiv preprint arXiv:2501.10970* (2025).

Further Problems & References

Statistical factuality guarantees for language models.
Mohri, Hashimoto (2024), Cherian, Gibbs, Candes (2024), Rubin-Toles, Gambhir, Ramji, Roth, Goel (2025)

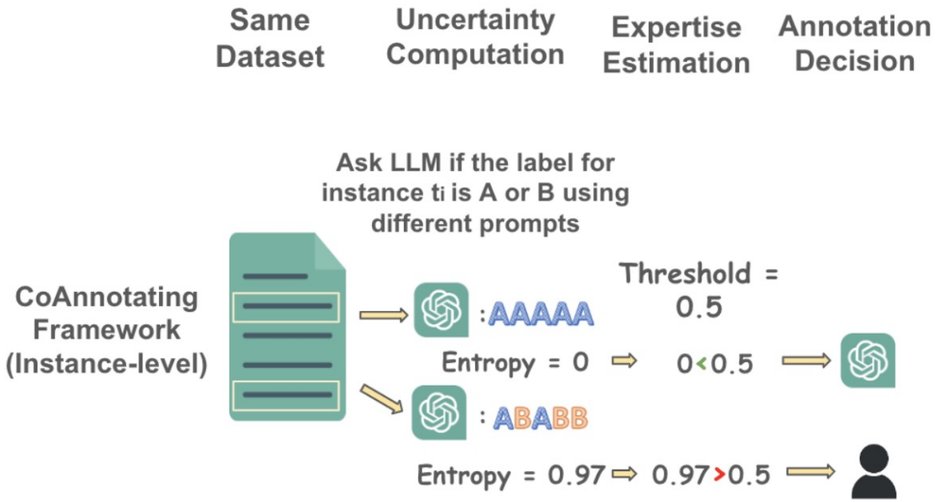
Further Problems & References

Statistical factuality guarantees for language models.
Mohri, Hashimoto (2024), Cherian, Gibbs, Candes (2024), Rubin-Toles, Gambhir, Ramji, Roth, Goel (2025)

Combining human and LLM annotations for approximately correct annotations.
Li, Shi, Ziems, Kan, Chen, Liu, Yang (2023), Kim, Mitra, Chen, Rahman, Zhang (2024), Candes, Ilyas, Zrnic (2025)

CoAnnotating: Uncertainty-Guided Work Allocation between Human and Large Language Models for Data Annotation

Minzhi Li [†§] Taiwei Shi [‡] Caleb Ziems [¶]
Min-Yen Kan [†] Nancy F. Chen [§] Zhengyuan Liu [§] Diyi Yang [¶]
[†]National University of Singapore [§]Institute for Infocomm Research (I²R), A*STAR
[‡]University of Southern California [¶]Stanford University
li.minzhi@u.nus.edu taiweish@usc.edu cziems@stanford.edu
nfychen@i2r.a-star.edu.sg liu_zhengyuan@i2r.a-star.edu.sg
kanmy@comp.nus.edu.sg diyiy@cs.stanford.edu



Further Problems & References

Statistical factuality guarantees for language models.
Mohri, Hashimoto (2024), Cherian, Gibbs, Candes (2024), Rubin-Toles, Gambhir, Ramji, Roth, Goel (2025)

Combining human and LLM annotations for approximately correct annotations.
Li, Shi, Ziems, Kan, Chen, Liu, Yang (2023), Kim, Mitra, Chen, Rahman, Zhang (2024), Candes, Ilyas, Zrnic (2025)

Valid evaluation of LLMs with synthetic data.
Chatzi, Straitouri, Thejaswi, Gomez Rodriguez (2024), Boyeau, Angelopoulos, Yosef, Malik, Jordan (2025)

What we’re missing: Opportunities for future research

LLM annotations with multi-modal inputs

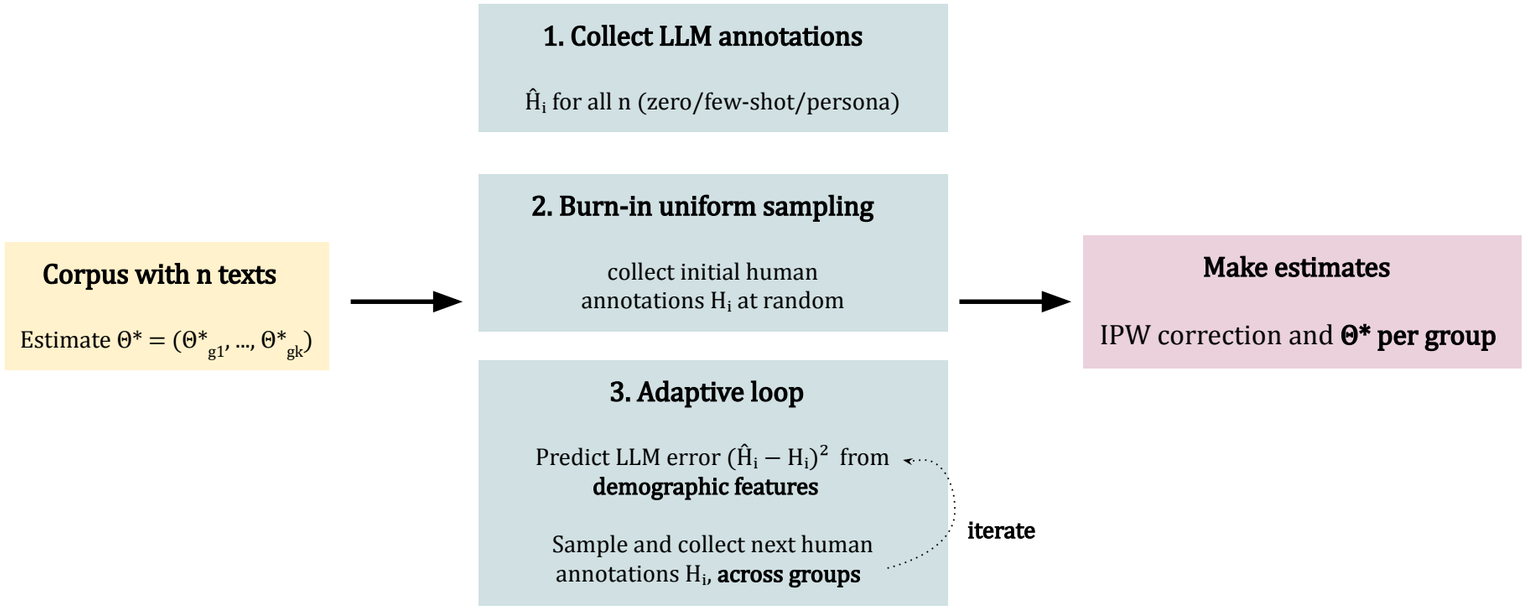
How to annotate videos? E.g., how do we find the most informative frames?

What we're missing: Opportunities for future research

LLM annotations with multi-modal inputs

What if there is no ground truth?

We want to estimate a vector of θ^* s



https://arxiv.org/pdf/2603.21404

Navya Mehrotra

Adam Visokay

Kristina Gligorić

What we're missing: Opportunities for future research

LLM annotations with multi-modal inputs

What if there is no ground truth?

What if LLM predictions are not calibrated?

How do we train distilled models, prioritizing calibration?

What we’re missing:
Opportunities for future research

LLM annotations with multi-modal inputs

What if there is no ground truth?

What if LLM predictions are not calibrated?

What if we want to do fine-tuning?

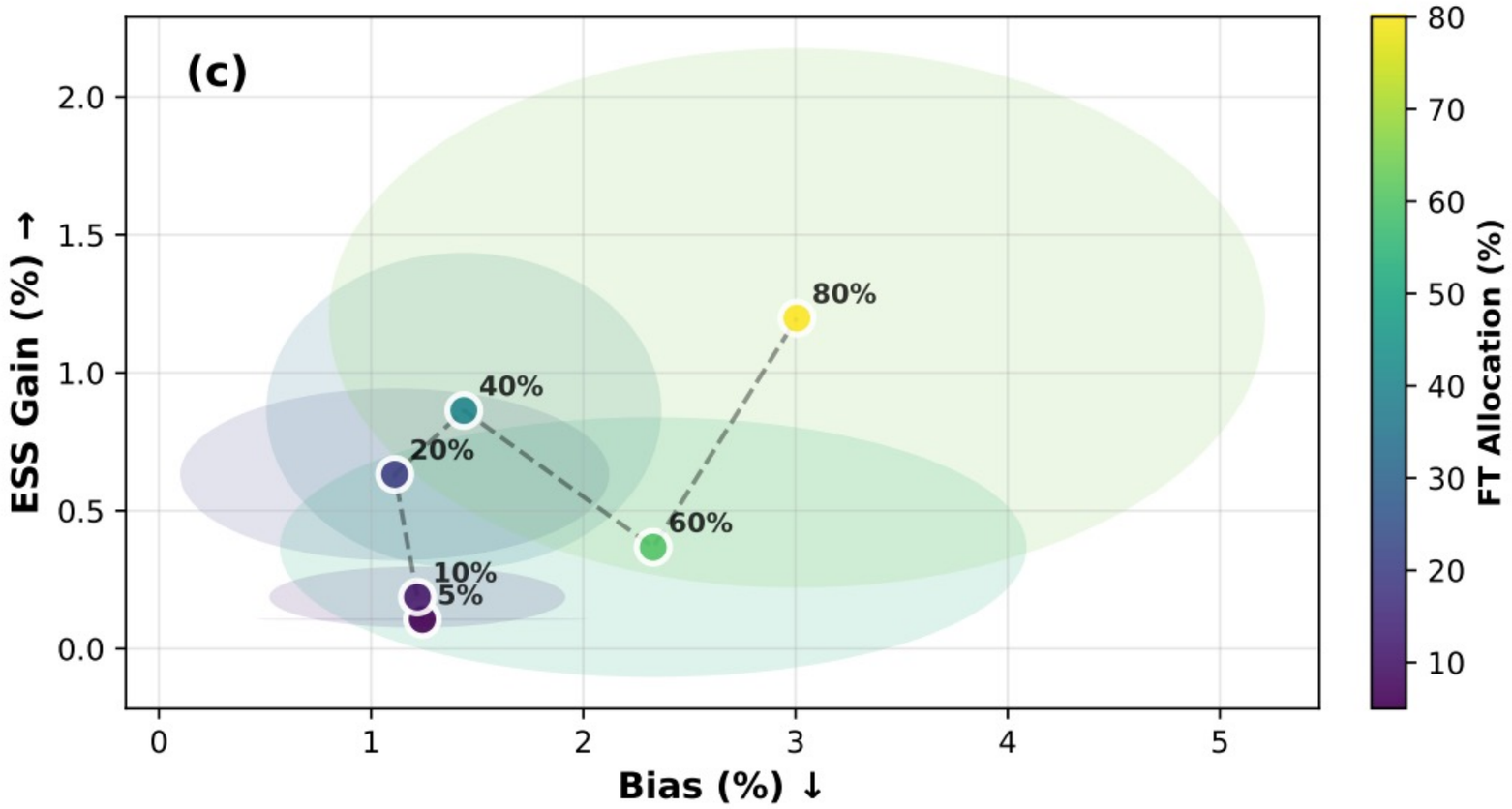
What we're missing: Opportunities for futu

Valid Survey Simulations with Limited Human Data: The Roles of Prompting, Fine-Tuning, and Rectification

Stefan Krsteski[1*], Giuseppe Russo[1,2*], Serina Chang[3], Robert West[1], Kristina Gligorić[4]
[1]EPFL
[2]Stanford University
[3]University of California, Berkeley
[4]Johns Hopkins University

To appear @ ACL 2026

What if we want to do fine-tuning?



What we're missing:
Opportunities for future research

LLM annotations with multi-modal inputs

What if there is no ground truth?

What if LLM predictions are not calibrated?

What if we want to do fine-tuning?

What are good scientific practices?

The problem of “LLM hacking”

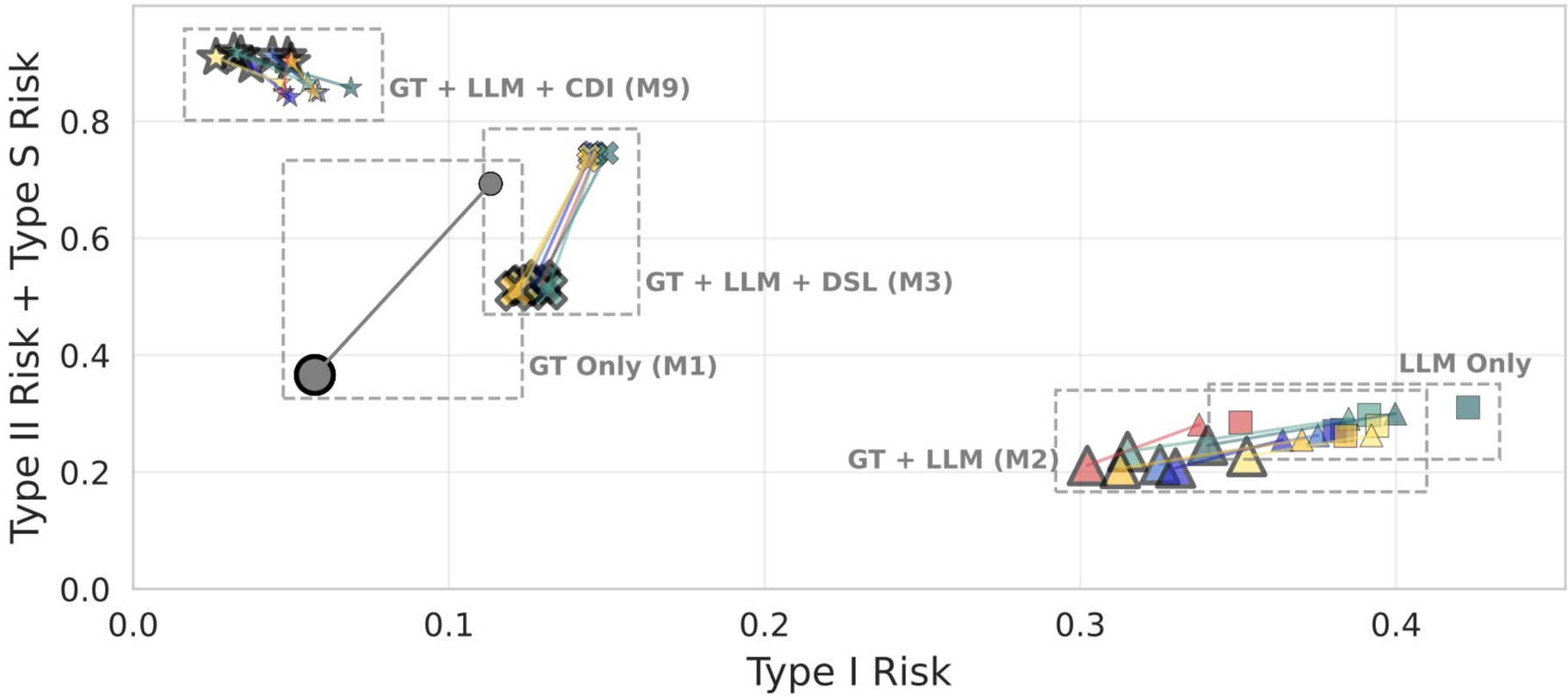
Every LLM-based annotation requires researchers to make numerous configuration choices, including:

- which model to use
- how to formulate the prompt
- which decoding parameters to set
- how to map outputs to categories
- …

These choices become a “garden of forking paths”

Baumann, Joachim, et al. "Large language model hacking: Quantifying the hidden risks of using llms for text annotation." *arXiv preprint arXiv:2509.08825* (2025).

The problem of “LLM hacking”



Baumann, Joachim, et al. "Large language model hacking: Quantifying the hidden risks of using llms for text annotation." *arXiv preprint arXiv:2509.08825* (2025).

contact:
gligoric@jhu.edu