Language Models as Opinion Models

Techniques and Applications

William Brannon
wbrannon@mit.edu

Dissertation Defense
Massachusetts Institute of Technology

May 13, 2025

Dissertation Committee

Deb Roy, Ph.D.

Professor of Media Arts and Sciences

Massachusetts Institute of Technology

Jacob Andreas, Ph.D.

Associate Professor of EECS

Massachusetts Institute of Technology

John Horton, Ph.D.

Associate Professor of Information Technologies

Massachusetts Institute of Technology

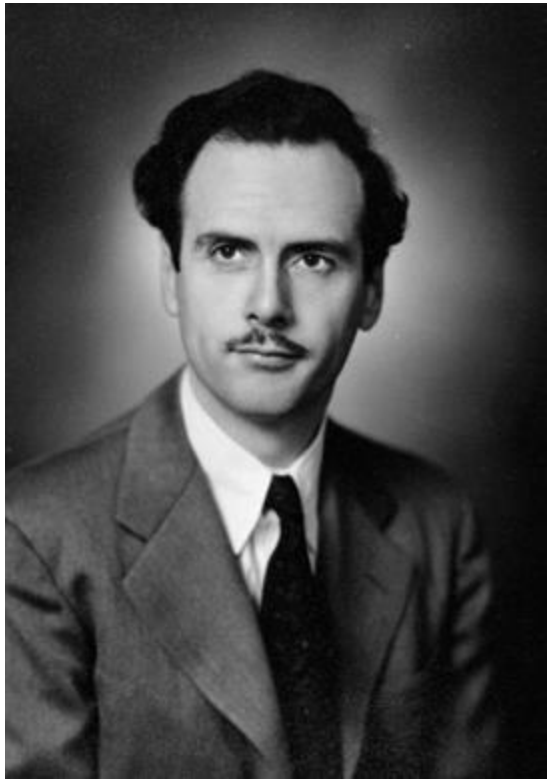
The Many Influences on Public Opinion

> The real environment is altogether too big, too complex, and too fleeting for direct acquaintance.
>
> -- Walter Lippmann
>
> Lippmann (1922)

Media Matters!

We shape our tools, and
thereafter they shape us.

-- John Culkin, summarizing McLuhan



McLuhan (1964)

So Does Psychology

Group Identity

Figure 1 Personal and social identities.

From Brewer (1991)

Social Learning, Conformism

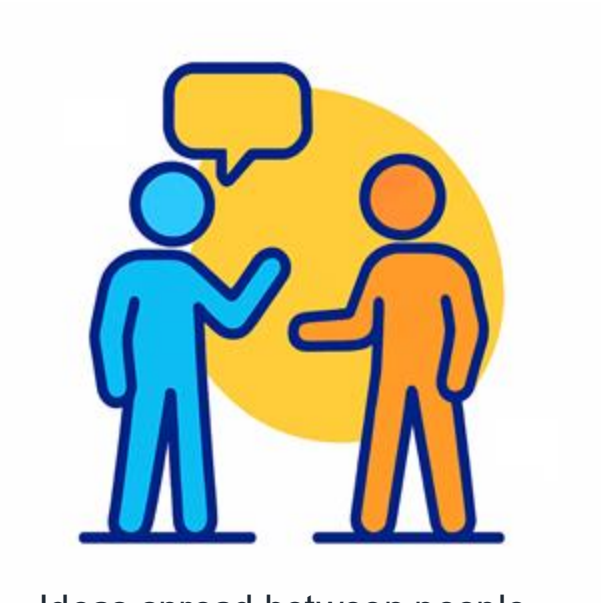
From Milgram et al (1969)

Cross-Pressure

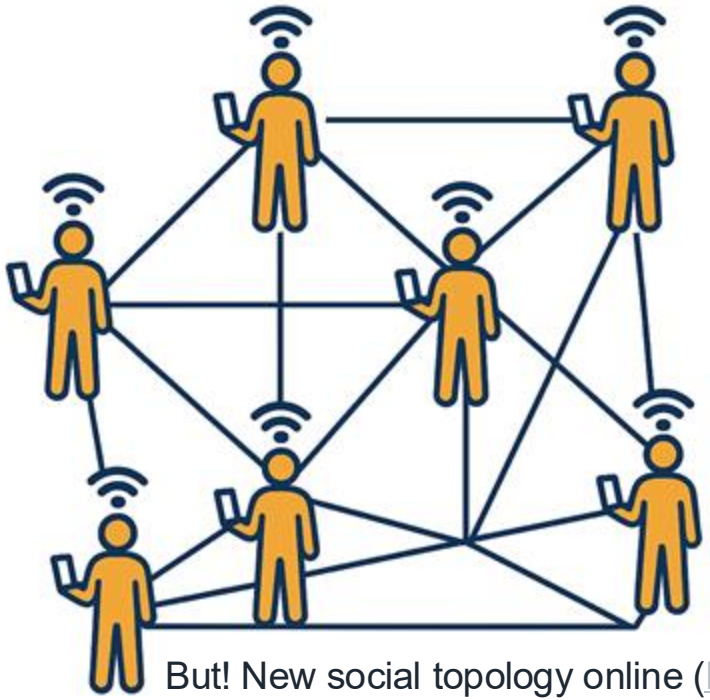
Berelson et al (1955), Zaller (1992)
Image from VSG

And more! Opinion formation is complex

The Internet Has Changed This Process



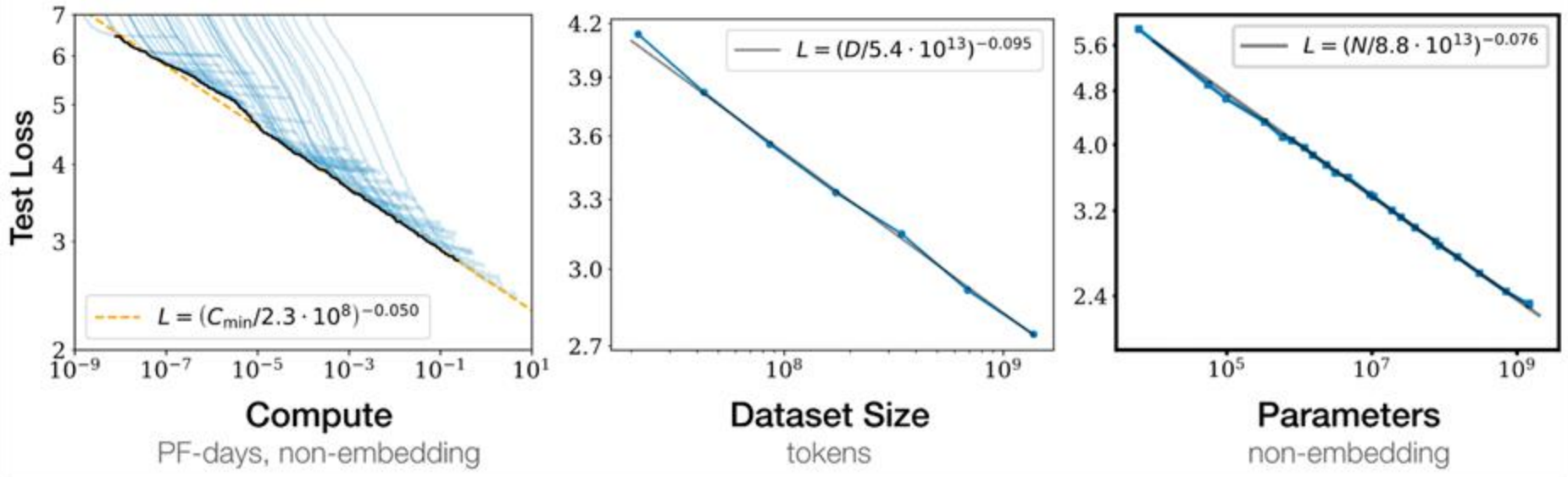
Ideas spread between people (see Katz and Lazarsfeld 1955)



But! New social topology online (boyd 2011), less media/social life distinction

Images via ChatGPT

...and Made It Easier to Study



Kaplan et al (2020)

They don't call it "web"-scale data for nothing!

LLMs: Really useful!

Llama 3

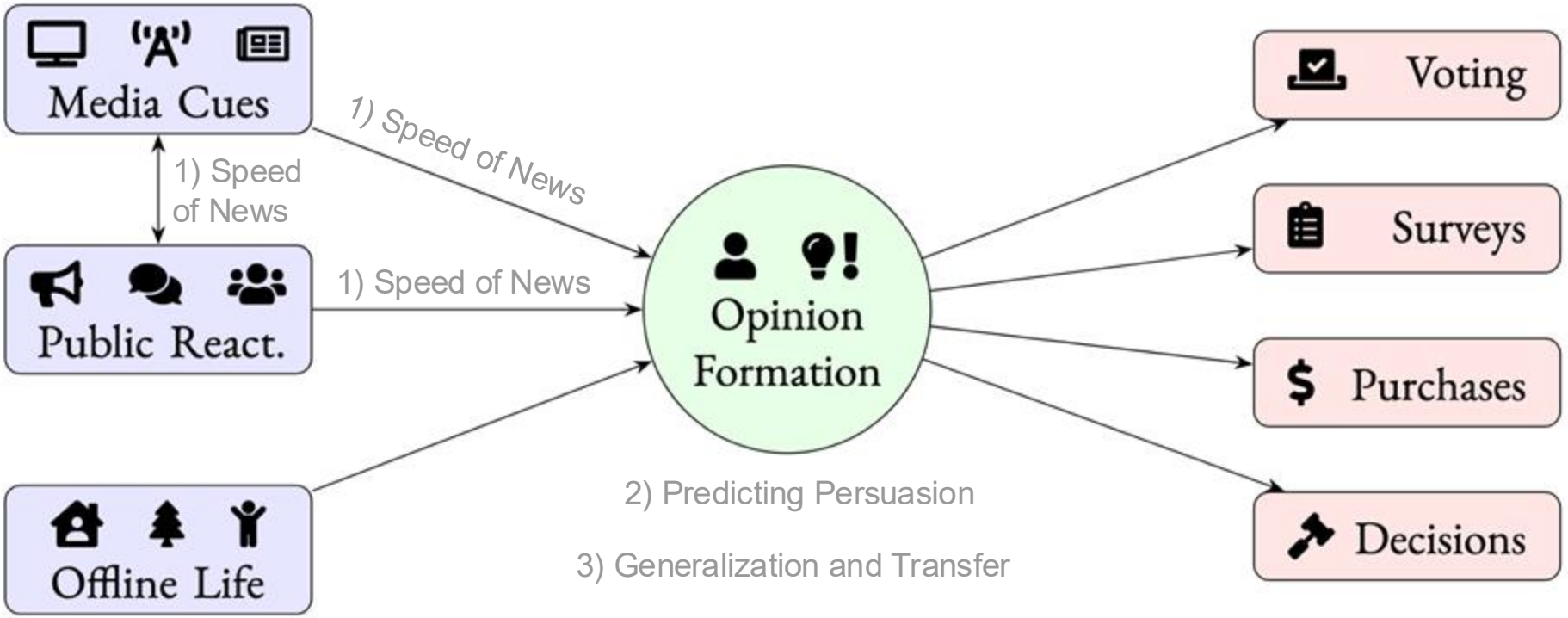
Claude

Image (appropriately)
via ChatGPT

OpenAI

Gemini

Putting It All Together



Speed and Sentiment of News

Can we identify differences in news cycle speed and sentiment, esp. outrage, between media? (SciRep ‘24)

(RQ1)

Media Is Changing

The news cycle has sped up and gotten more negative.

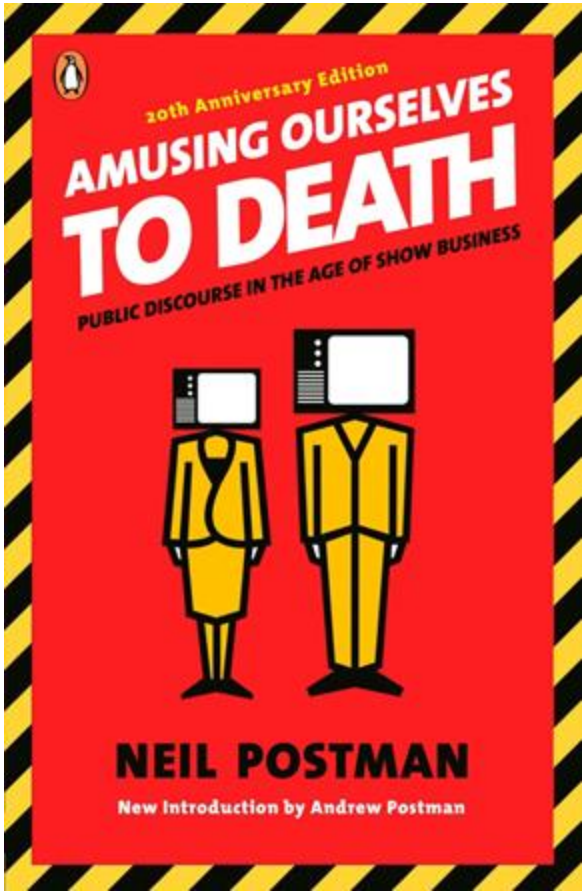
New media are involved, but how? What are their effects?

John Rose, Ogden Newspapers

Why Does This Matter?

Shorter news cycle + greater negativity

= weaker democracy?



From 1985: This is a longstanding concern

Why Does This Matter?

Does social media have a **first-mover advantage** that lets its biases influence journalism and discourse overall?

Media of Interest

Twitter

+

Radio

Specific Data Sources

Elite Twitter

2,834 national VIPs

Firehose
Random Sample

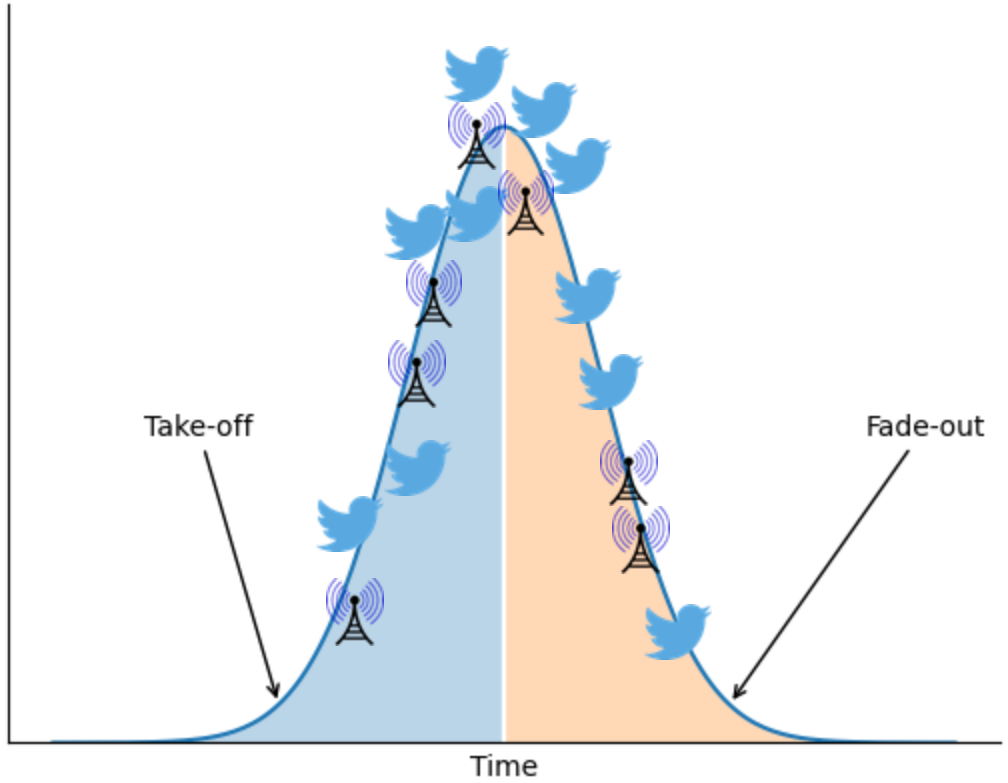
US Talk Radio

228 stations: talk and public, 518k hours

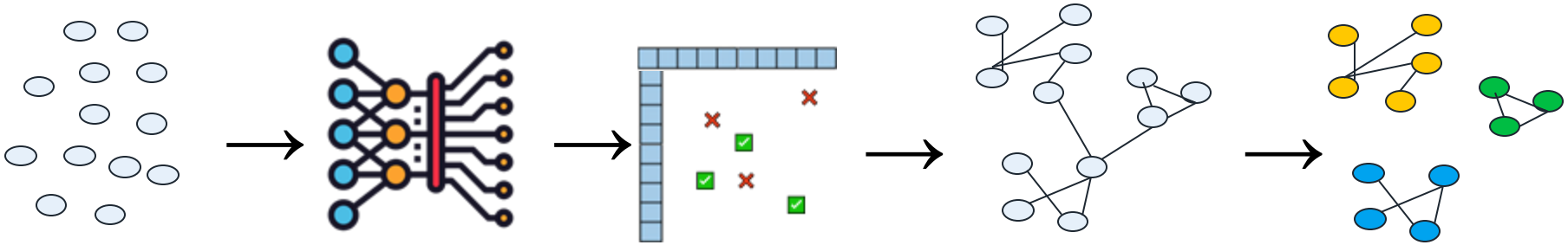
Three separate periods:
Sep/Oct 2019, Mar/Apr 2020, Jan/Feb 2021

What's An Event?

Media events are defined via media: a group of related tweets or radio statements



Methodology: Detecting Events



Twitter or Radio Statement Batch

Pretrained LM Embeddings

Filter by Cosine Similarity + Time

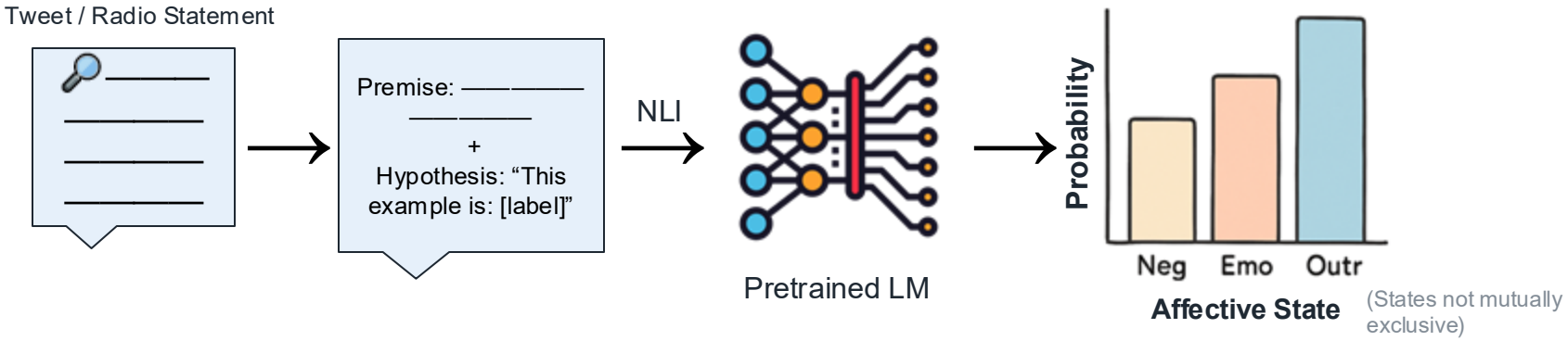
Form Graph, Weight by Sim.

Louvain Community Detection

Finally, **filter out non-news events**: threshold centroids' cosine similarity to elite-Twitter events. → 1,694 events

See Laban et al (2017), Staykovski et al (2019)

Methodology: Identifying Affect

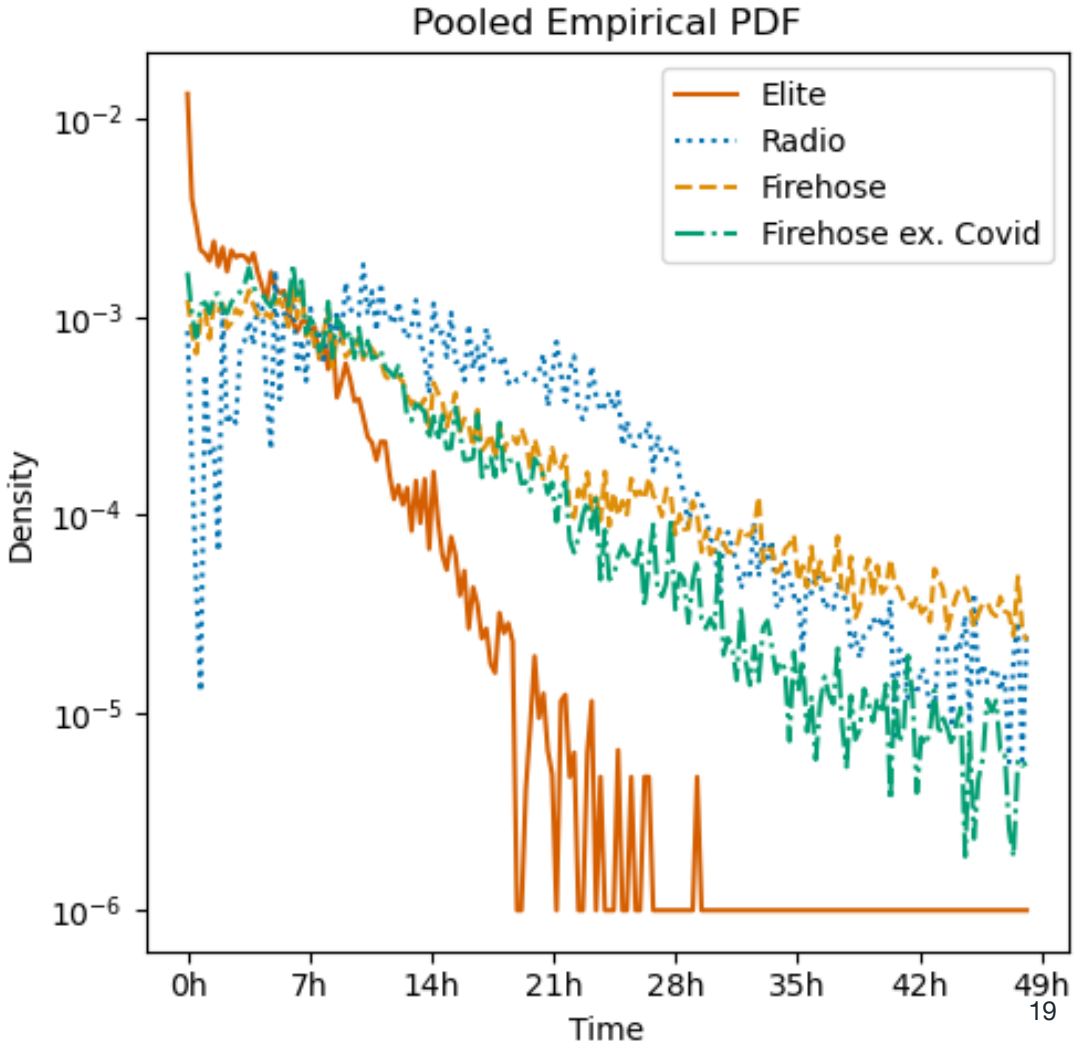


Yin et al (2019)

Lifecycle of An Event

Elite Twitter rises **and falls** faster than radio!

Firehose too, **but**: unusual stuff during 2020 Covid discourse

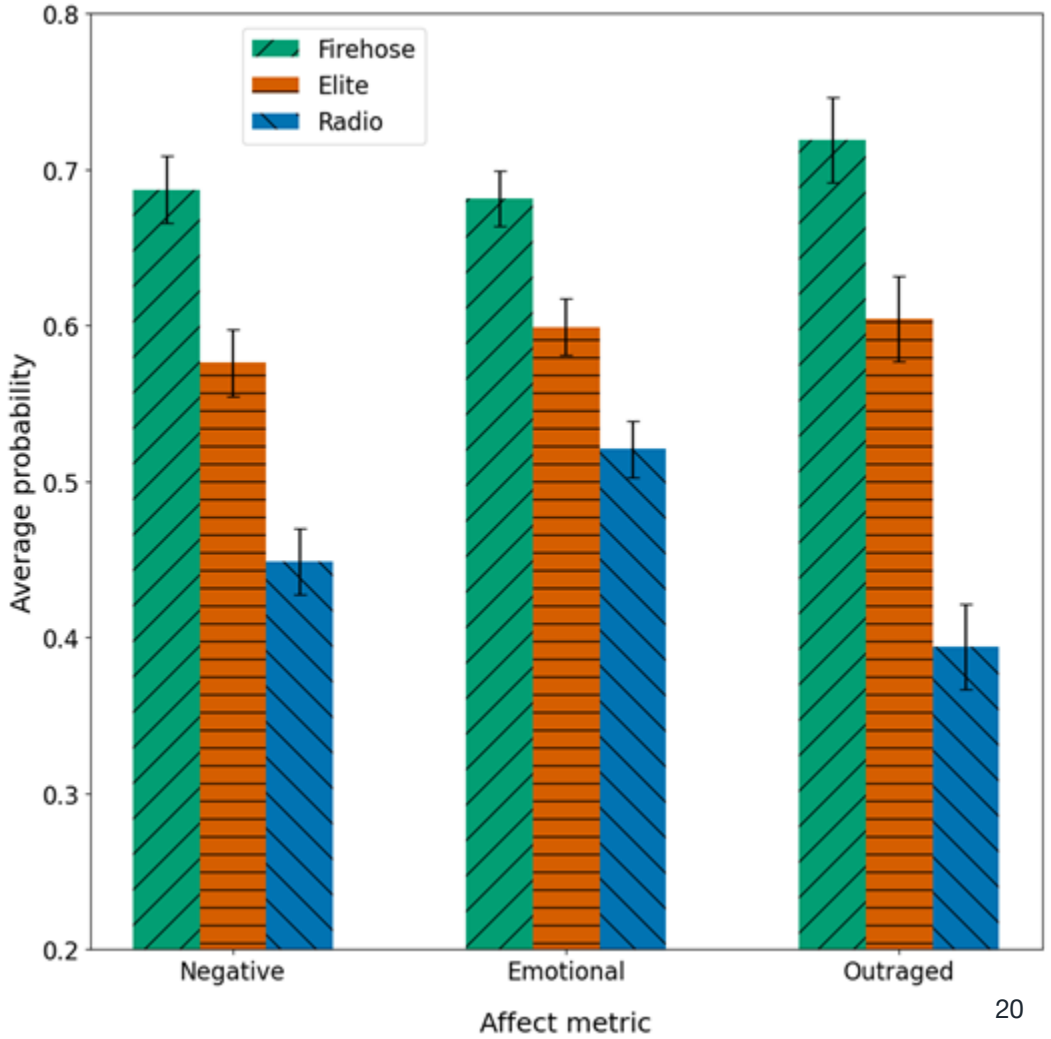


Affective Biases

Consistently: Firehose > Elite > Radio

Medium effect: Elite + firehose are both more negative, etc., than radio

Audience capture? Is the audience rewarding negativity and thus encouraging more of it?



Conclusions

Faster news cycles on Twitter,
at start and end

More negativity and outrage
(systematic bias)

→ Shrinking attention span,
compressed discourse

Conclusions

Faster news cycles on Twitter,
at start and end

Shrinking attention span,
compressed discourse

More negativity and outrage
(systematic bias)

First large-scale data-driven comparison of outrage
between Twitter and traditional media!

Predicting Persuasion

Heterogeneous Treatment Effect Estimation via LLM

Can we use LMs to predict persuasion by messages in experimental settings?

(RQ2)

What problem are we trying to solve?

The Challenge

People are always trying to persuade other people

But persuasion is complicated, contextual, hard to understand

(See Hewitt 2024)

The Research Question

LLMs have knowledge about the world and people – **can they predict persuasion response**?

RCTs: randomization *identifies* the effect

Clip art via ChatGPT

Where Are We?

What's been done?

- Political science persuasion literature
- Survey + message-test experiments
- Classic experimental methodology (Fisher, Neyman, Rubin)
- Statistical HTE / TEE estimation - (meta-learners, causal forests)
- Recent neural models (DragonNet)

Where Are We Going?

💡 **But**: Little work on pretraining! 💡

- Can LLMs let us run fast *in silico* trials?
- Not a replacement for studying people: Like animal studies, for social science.

Formalism: Treatment Effect Estimation (TEE)

Consider a randomized experiment:

- Subjects i = 1, ..., N
- Treatment status T = 0, 1
- Potential outcomes Y_i(0), Y_i(1)

We usually consider the ATE, but individual effects matter too.

$$\tau = \mathbb{E}[Y(1) - Y(0)]$$

$$\hat{\tau} = \frac{1}{N} \sum_{i=1}^{N} Y_i(1) - Y_i(0)$$

Average treatment effect

$$\tau_i = \mathbb{E}[Y_i(1) - Y_i(0) \mid X_i]$$

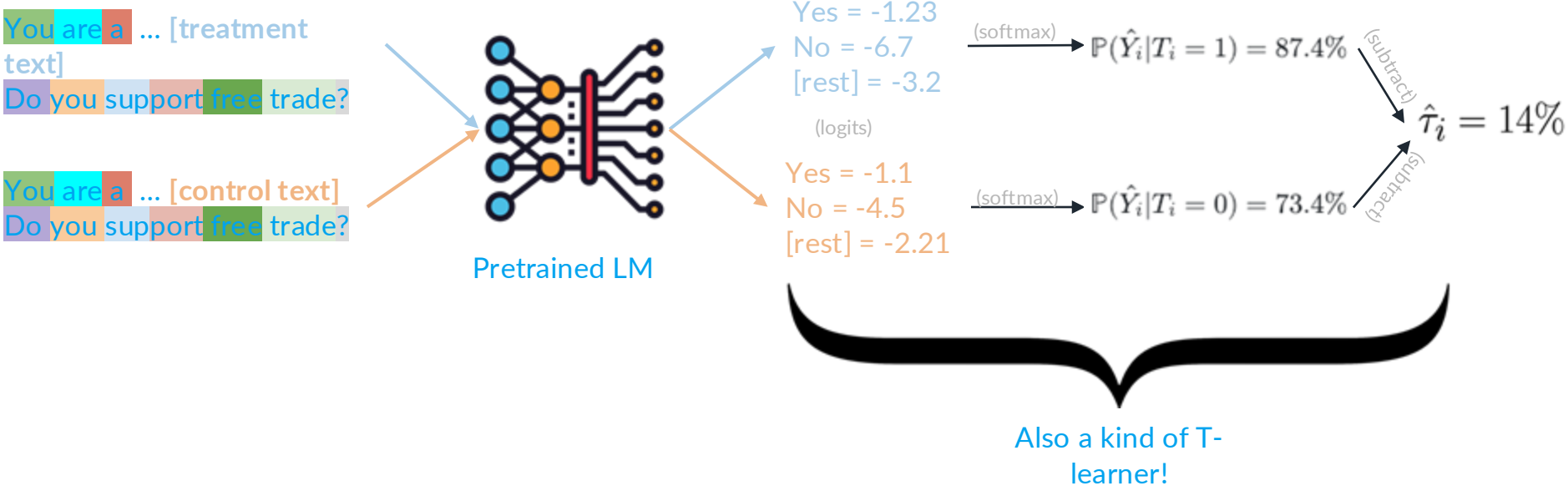
Individual treatment effect ('CATE function')

If treatment is textual (e.g. a survey experiment), can we use LLMs to estimate τ_i?

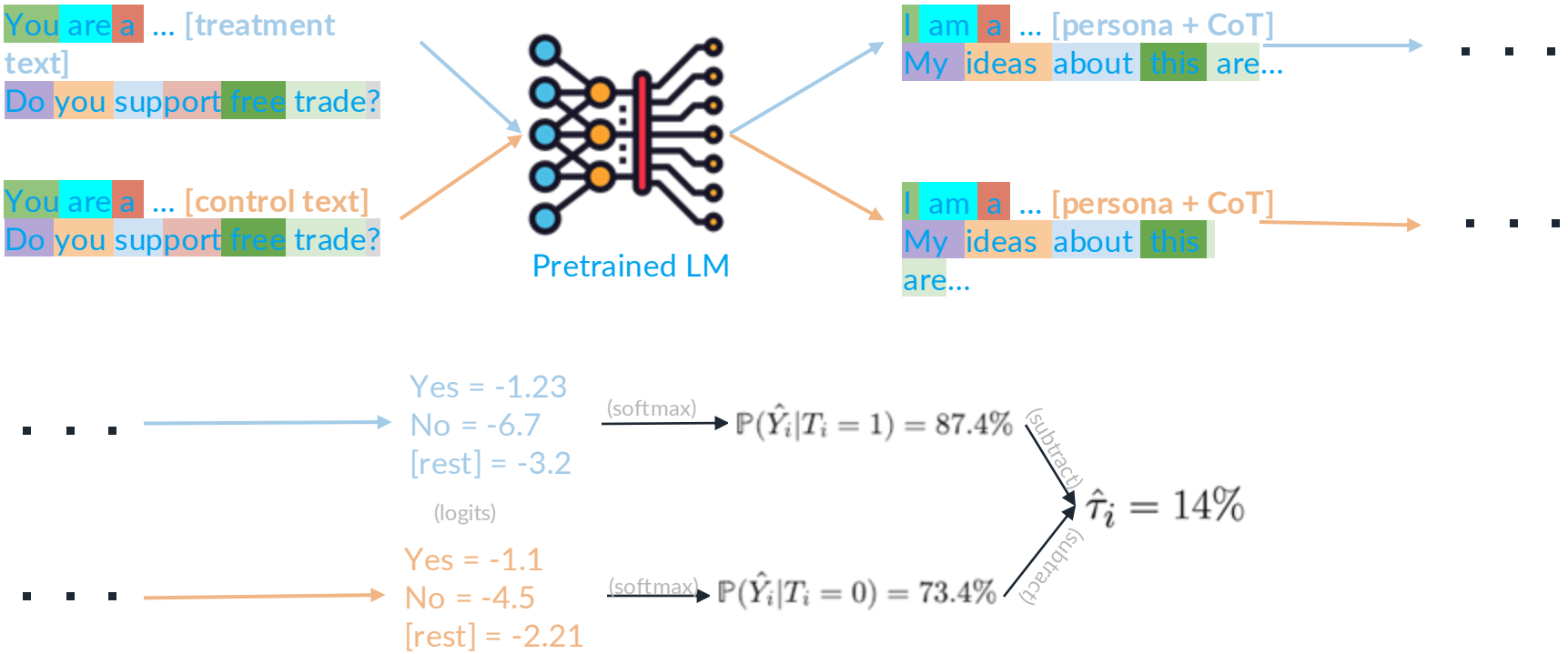
See Rubin (2005) for an overview of P.O. framework

Methods

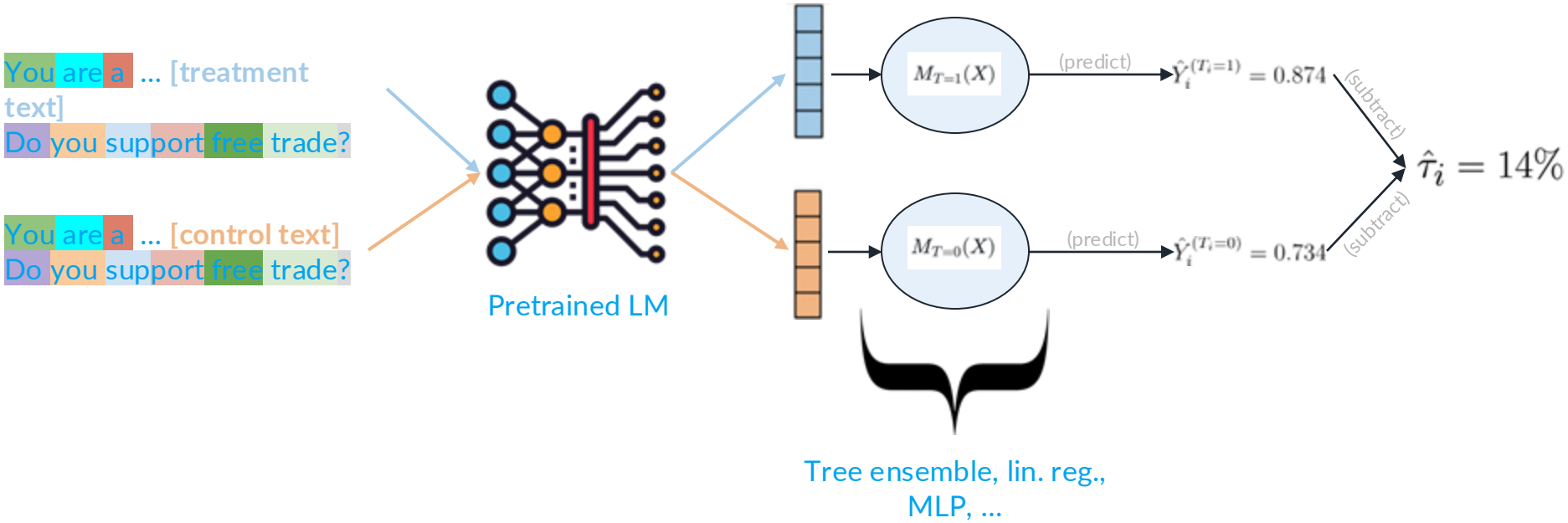
Approach 1: Text2Text



Approach 2: Persona-Based / Inference-Time Compute



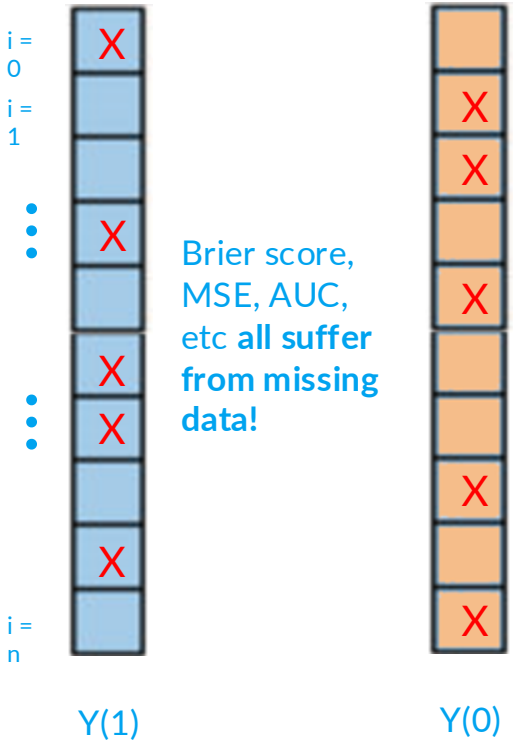
Approach 3: Representational Regression



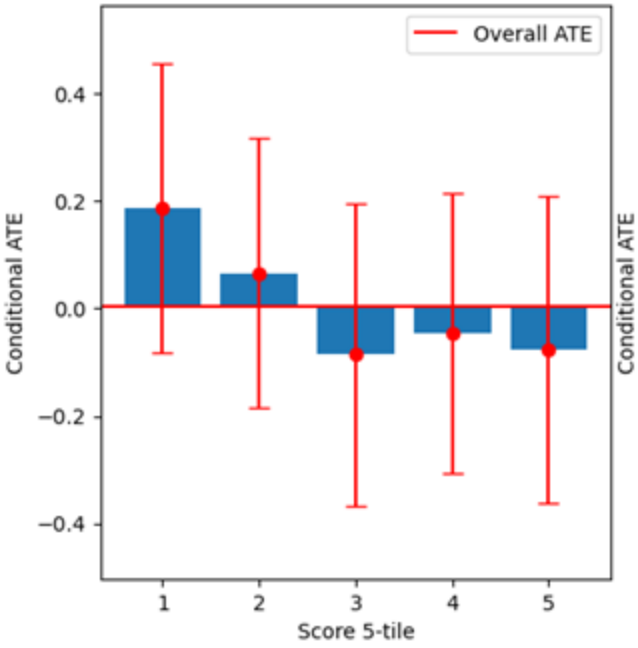
Evaluation

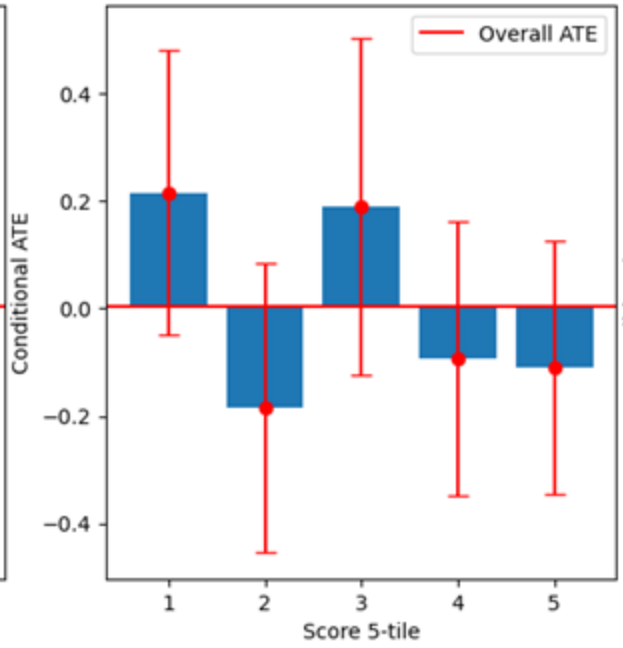
Evaluating these models is not straightforward

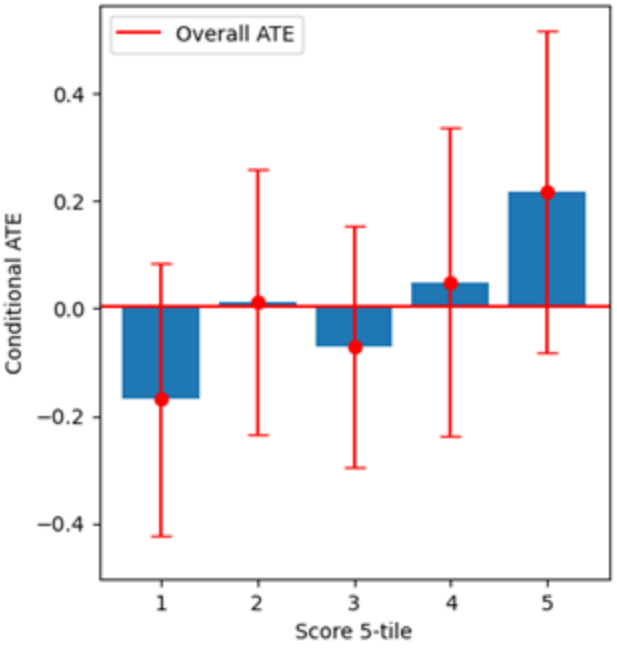
Why? The “**fundamental problem of causal inference**”: we don’t observe both potential outcomes



What does good performance look like?





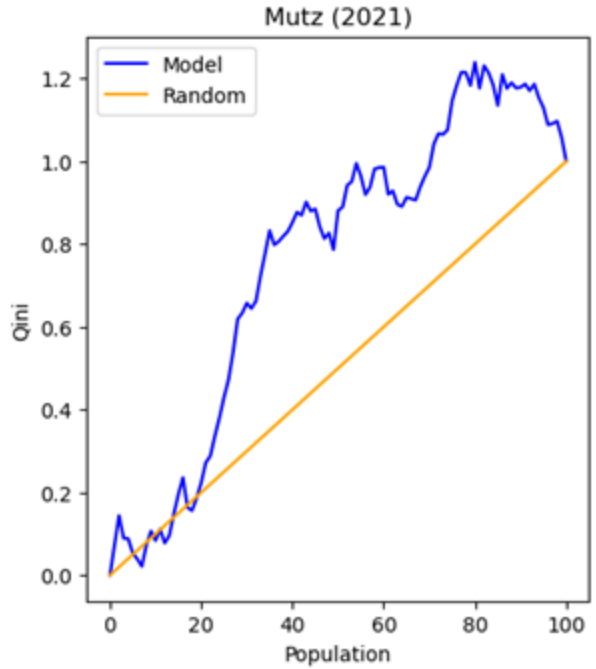


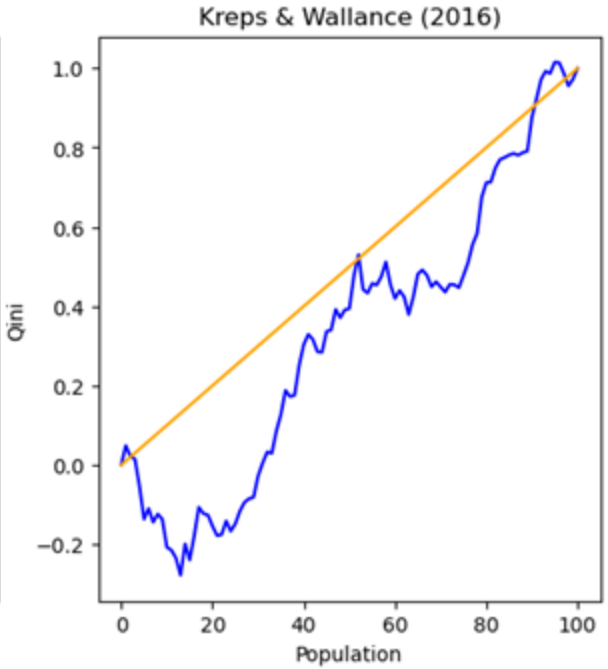
❌ ❌ ✅

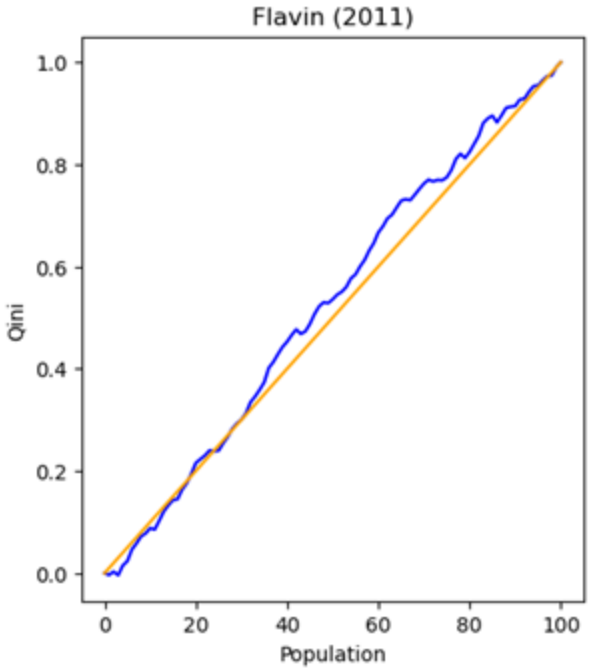
Conditional ATEs

Easy to understand but high-variance, loses information, not usual anymore

What does good performance look like?







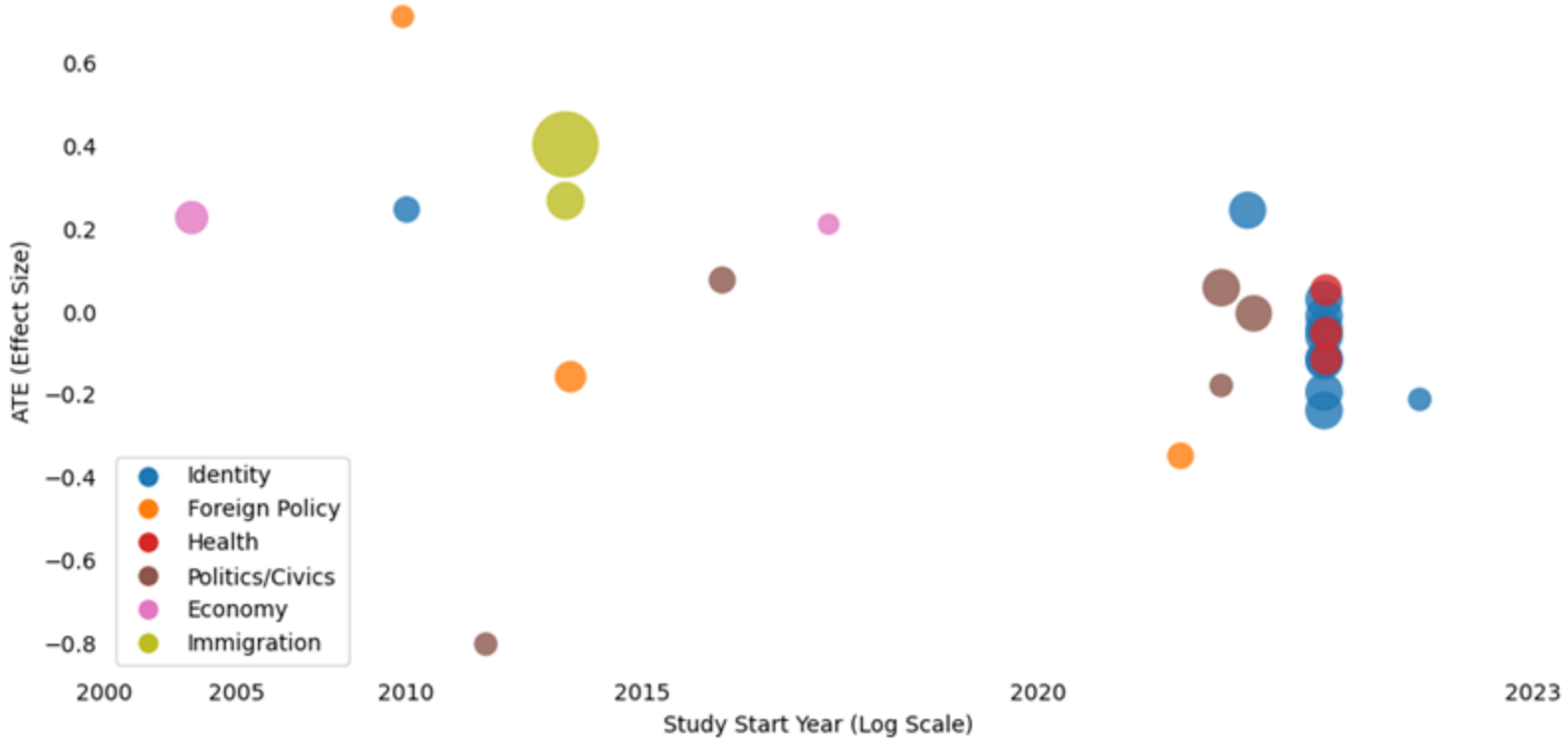
The area between the model curve and idealized random curve, suitably normalized, is called the *Qini coefficient*

Qini curves

Sort exp. subjects descending by predicted responsiveness; at each subject, “how much” treatment effect have we had?

Datasets

28 studies on a variety of topics over more than 19 years, with a range of effect sizes



What do these datasets look like?

Intro script

TABLE A.17 **Mutz (2017): Treatments and outcomes**

Introductory text: When Michael Morrison took a job at the steel mill in the center of Granite City, Ill., in 1999, he assumed his future was ironclad. He was 38, a father with three young children.
"I felt like I had finally gotten into a place that was so reliable I could retire there," he said.
Although it had changed hands, the mill had been there since the end of the 19th century. For those willing to sweat, the mill was a reliable means of supporting a family.
Mr. Morrison began by shoveling slag out of the furnaces, working his way up to crane driver. From inside a cockpit tucked in the rafters of the building, he manned the controls, guiding a 350-ton ladle that spilled molten iron.
It was a difficult job requiring perpetual focus, and he was paid accordingly.

Job loss due to trade	Job loss due to automation
Now his job has been eliminated due to trade with China. Chinese workers now man the same machine that Mr. Morrison once operated. As the company website describes, "Of the 74 machines that were operating in the factory, 63 are now operating in China."	Now his job has been eliminated due to automation. Robots now man the same machine that Mr. Morrison once operated. As the company website describes, "Of the 74 machines that were operating in the factory, 63 now run on their own with no human intervention."
Mr. Morrison has not been able to find other work, and he has no idea how he will pay for his children's college education. "When they don't need me anymore," he said, "I'm nothing."	Mr. Morrison has not been able to find other work, and he has no idea how he will pay for his children's college education. "When they don't need me anymore," he said, "I'm nothing."

Treatment text

Control text

Opinion question

"Do you favor or oppose the federal government in Washington negotiating more free trade agreements?" [1: Strongly oppose, 4: Strongly favor]

What do we know about the respondents?

We only have basic demographic data about each respondent!

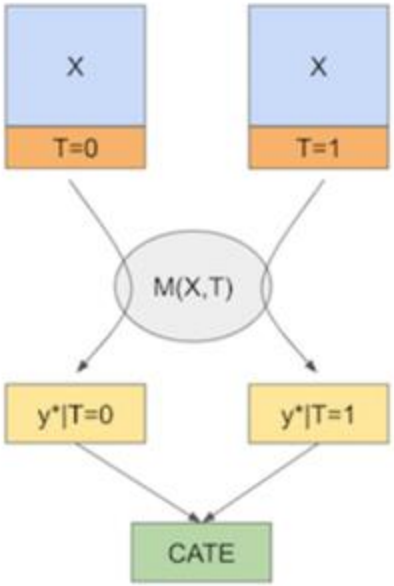
(Usually slightly more than shown.)

```
You are a {feats['age']} year old, {feats['race']} {feats['sex']}. {feats['educ']} You are a {feats['ideo']} and a {feats['pid_7']}. You are taking a survey about politics; we will read you an article and then ask your opinion about an issue.
```

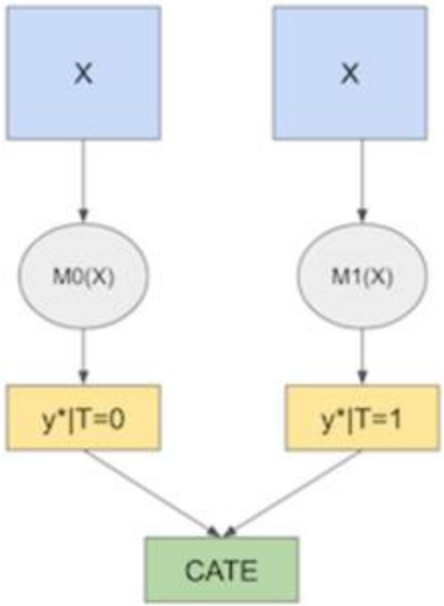
This problem is hard!

We can beat baselines only if **LLMs have transferable knowledge** about how these variables **relate to the treatment**.

Baselines



S-learner
(w/ random forest)



T-learner
(w/ random forest)

Several others, more complicated:

- X-learner + random forest
- Causal forest
- DragonNet

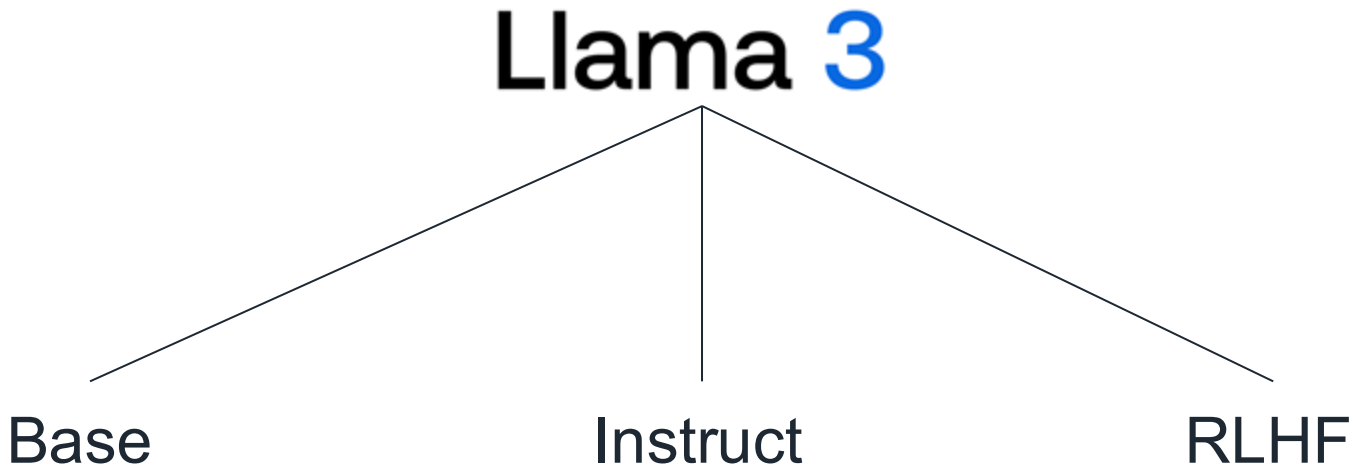
Note that we also use the T-learner in one of the LLM-based approaches described prior

Image source

Meta-learners from Künzel et al (2019)

Models

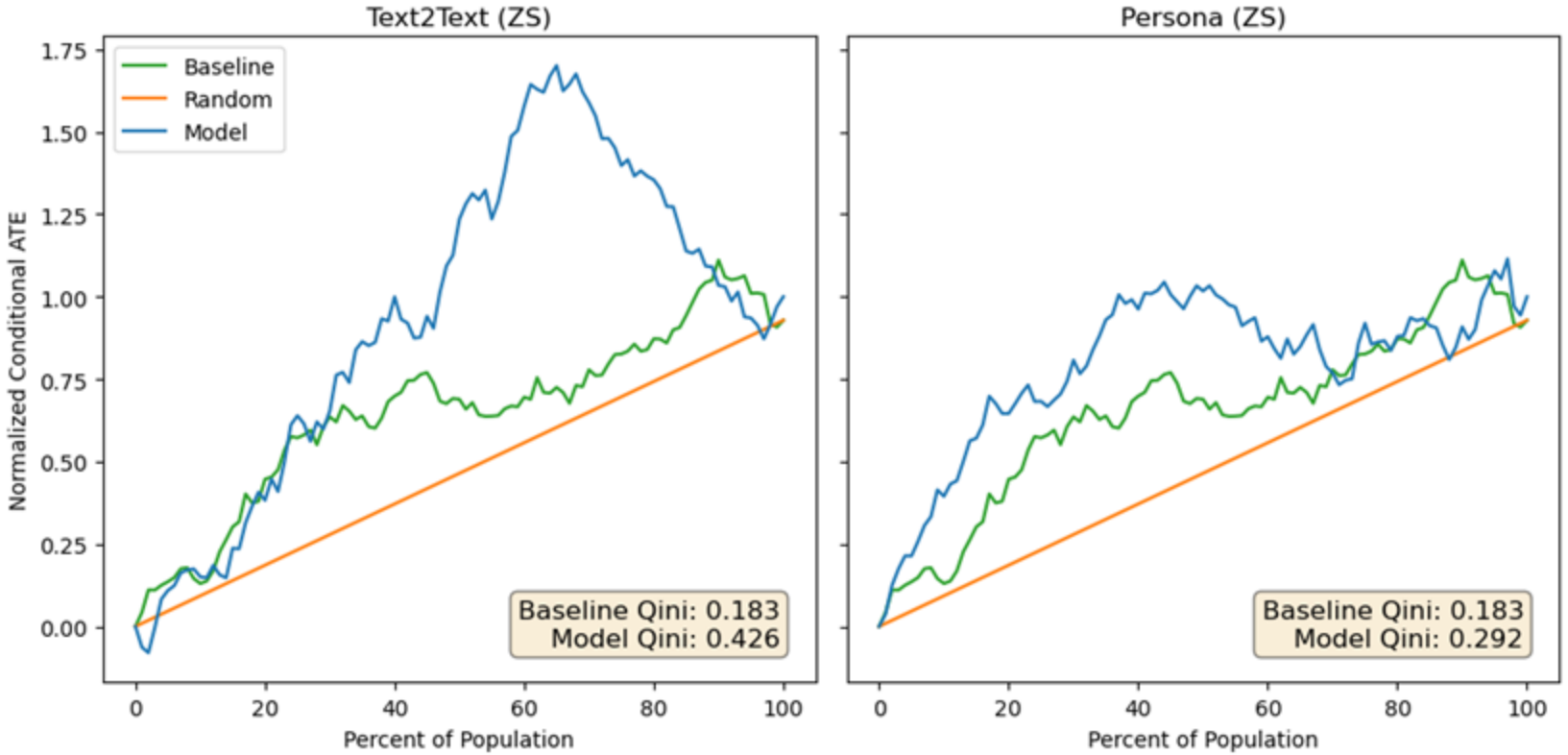
Local models: more control over experiments + can compare w/ and w/out post-training



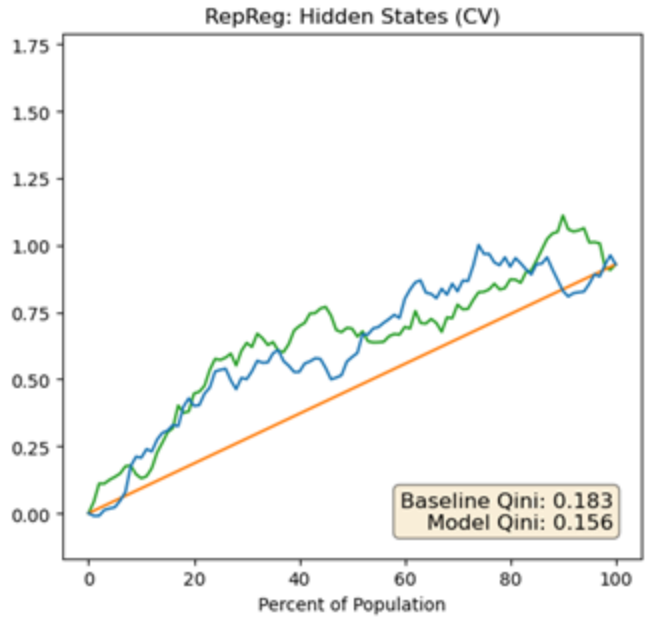
(See also scaling experiments with OpenAI models in dissertation)

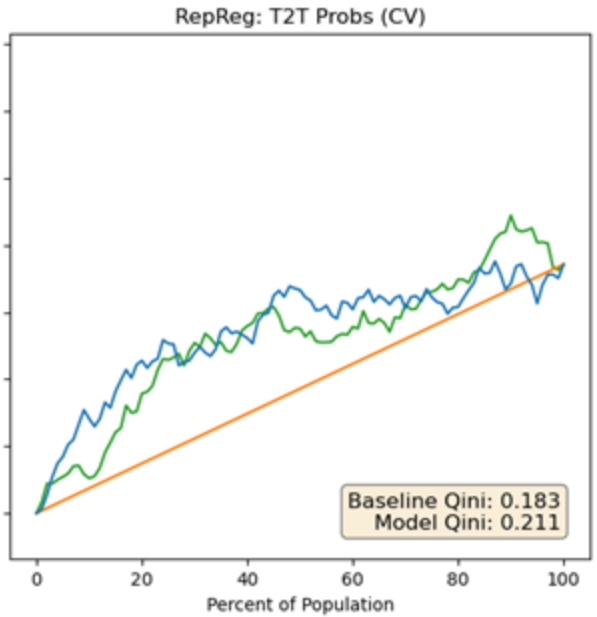
Results

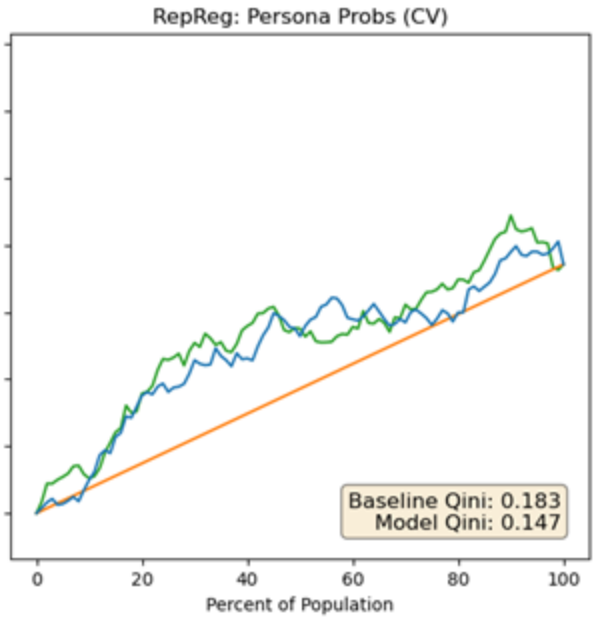
Zero-Shot Methods Work Well!



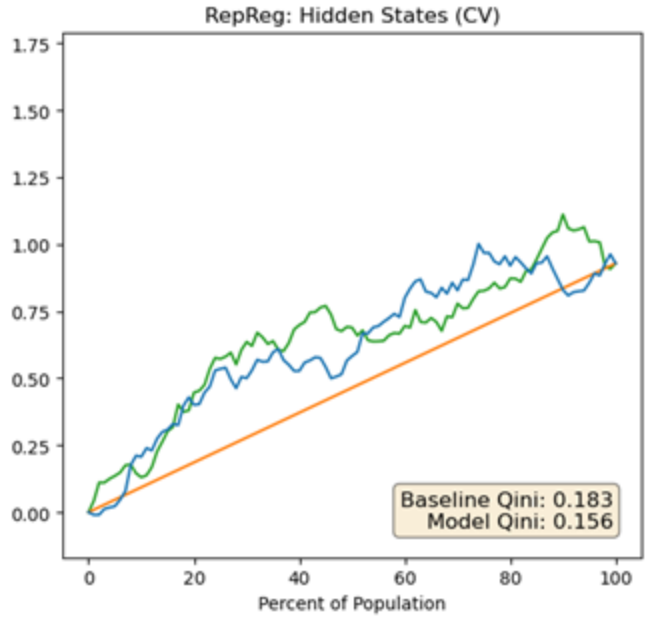
...RepReg Less So

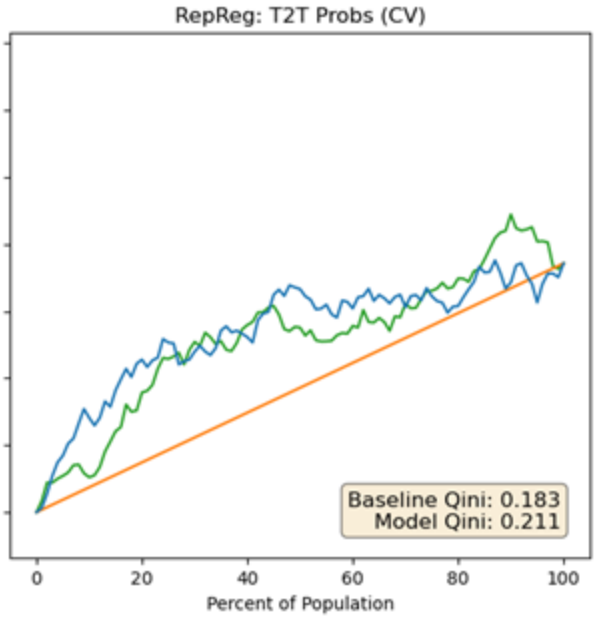


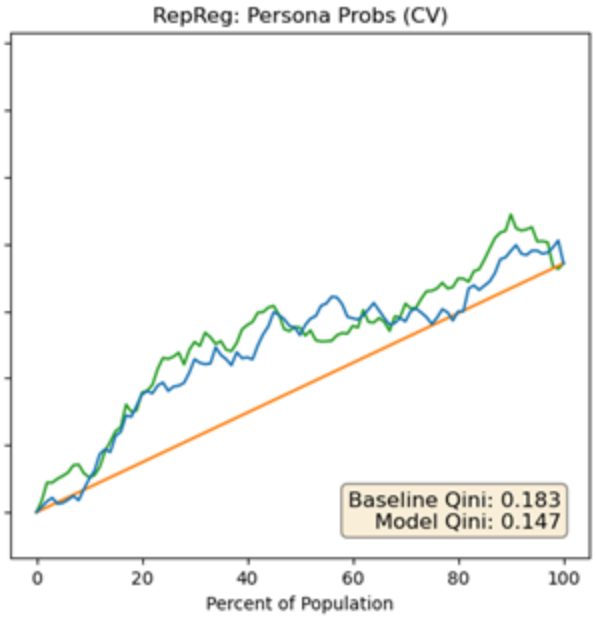




...RepReg Less So

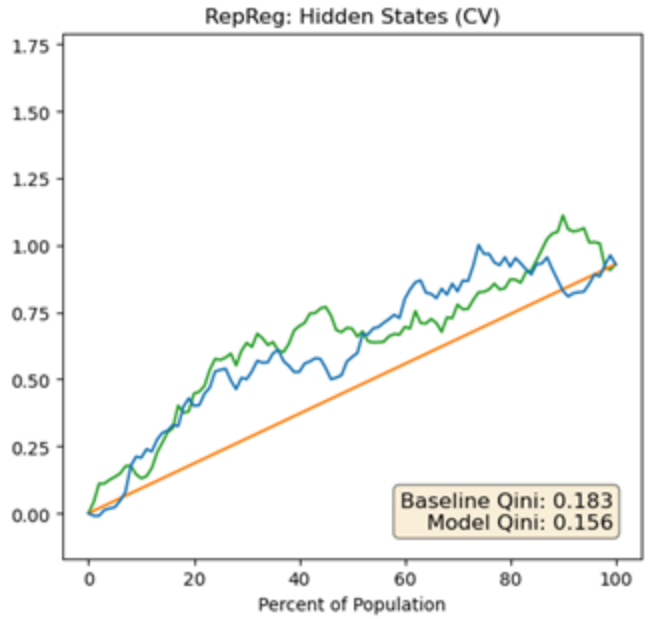


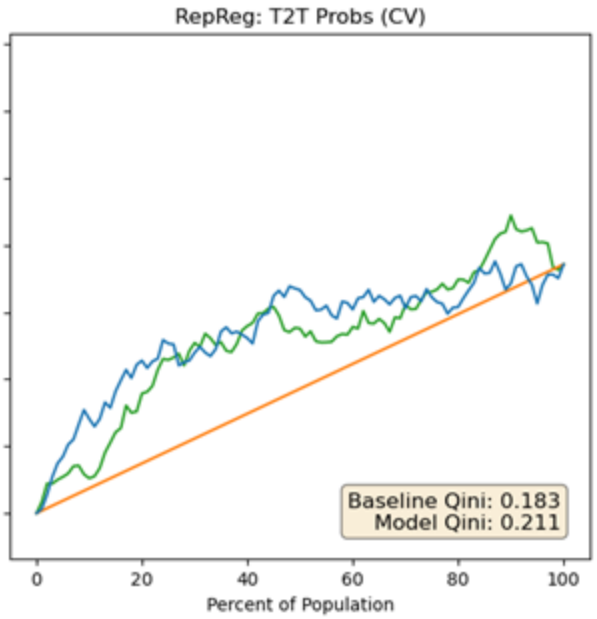


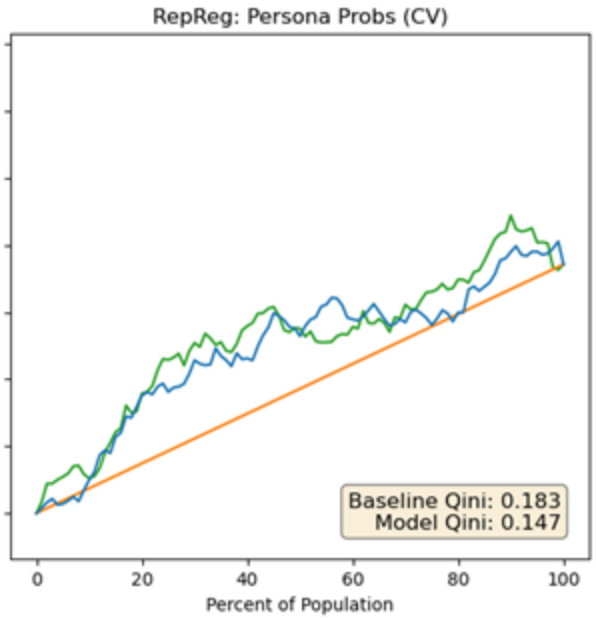


Above baseline!

…RepReg Less So





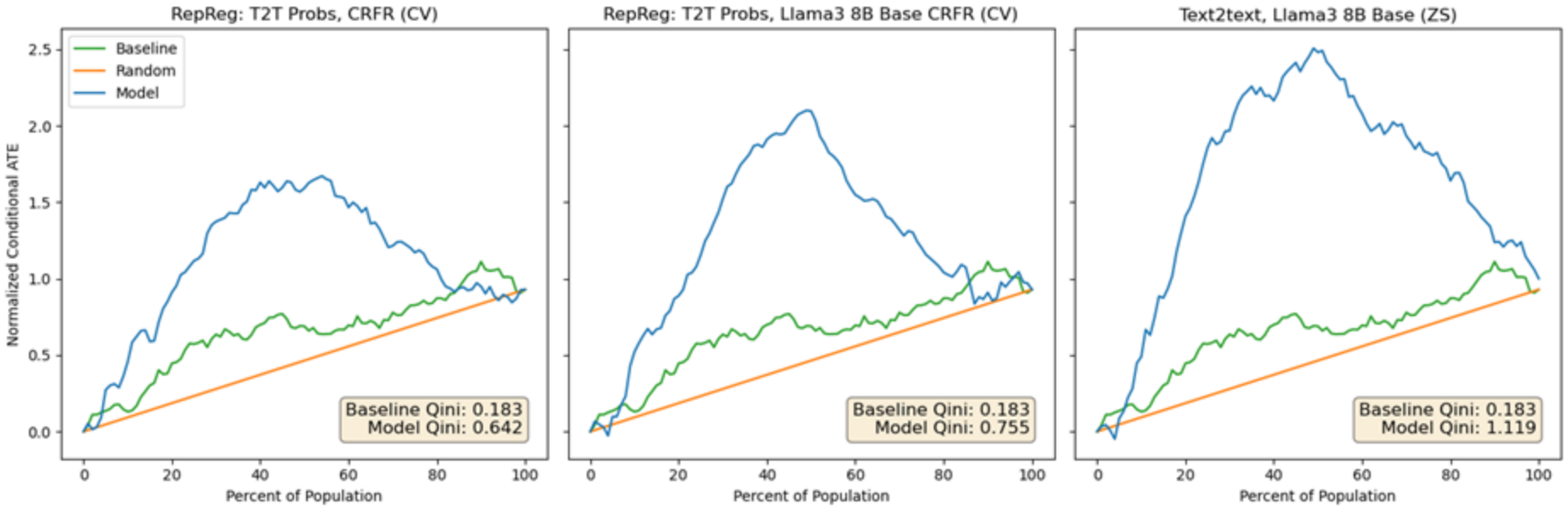


Not always…

Above baseline!

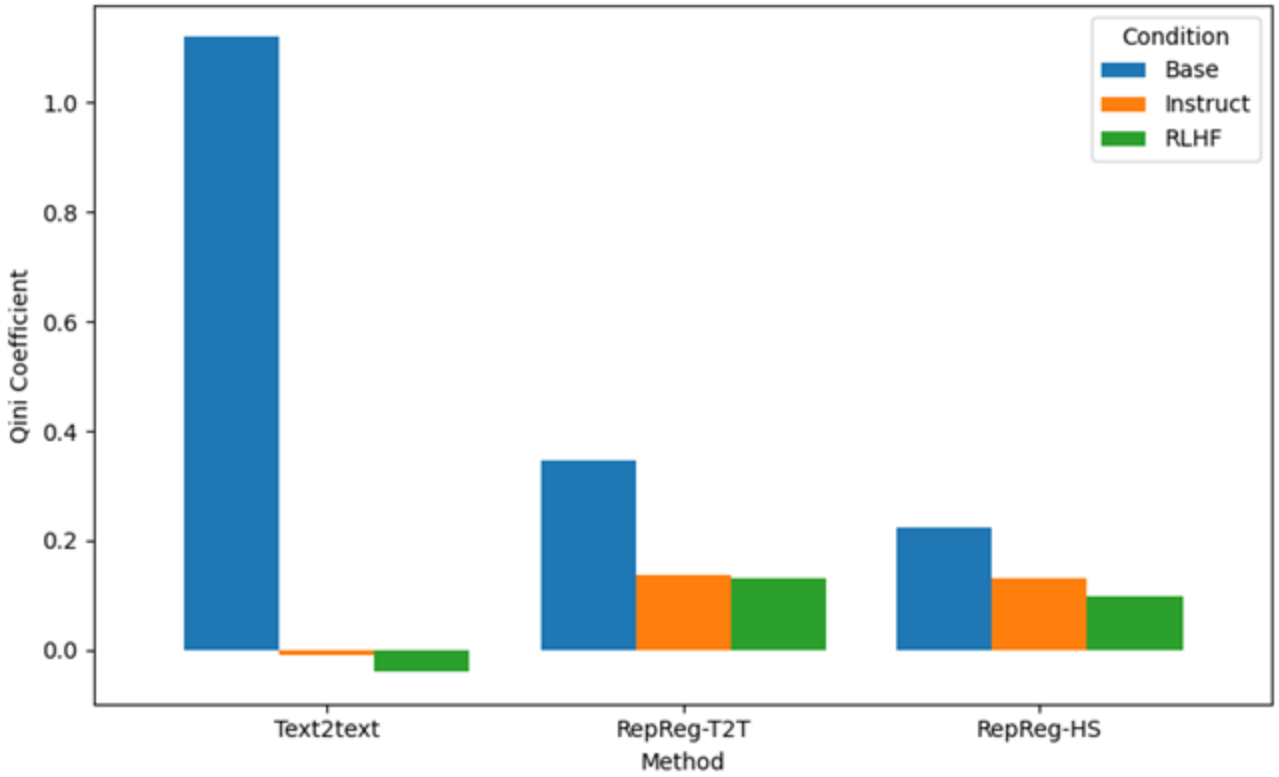
Not always…

What's The Best We Can Do?



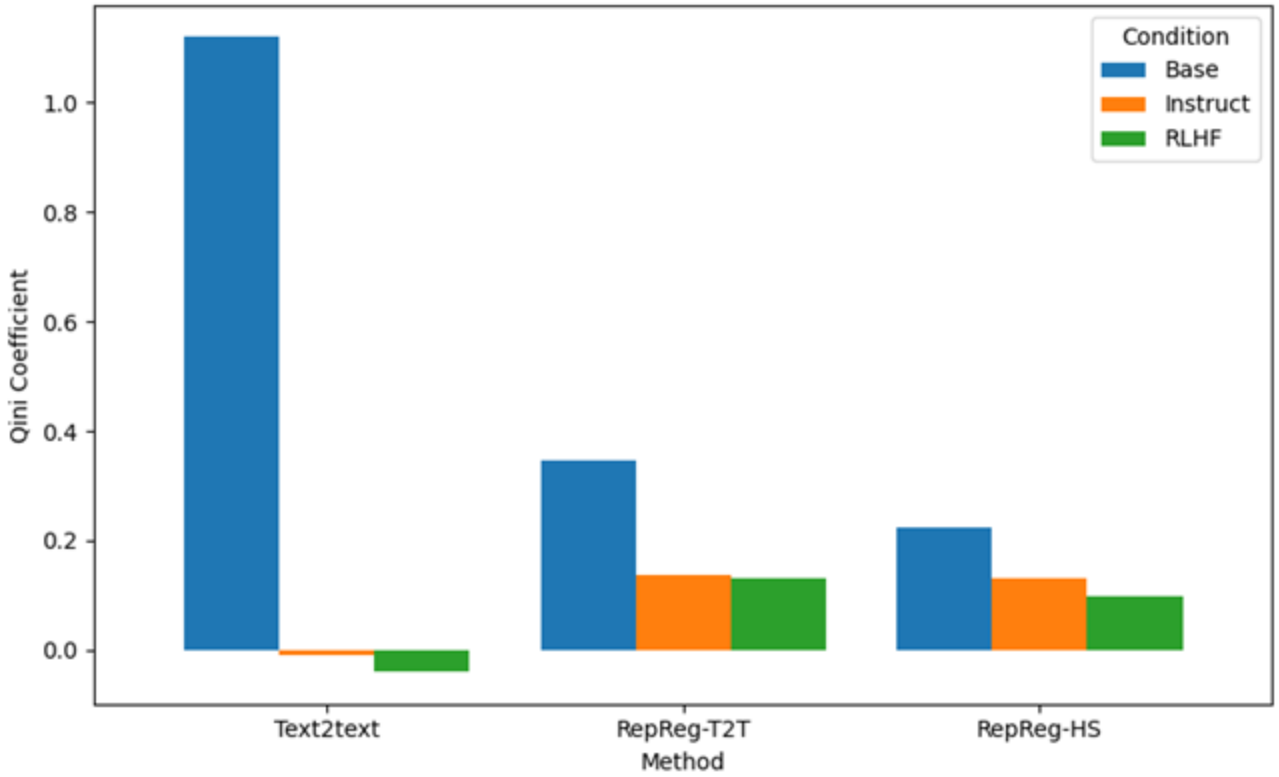
Note: not prespecified!

Post-Training Hurts Performance! A Lot!



Post-Training Hurts Performance! A Lot!

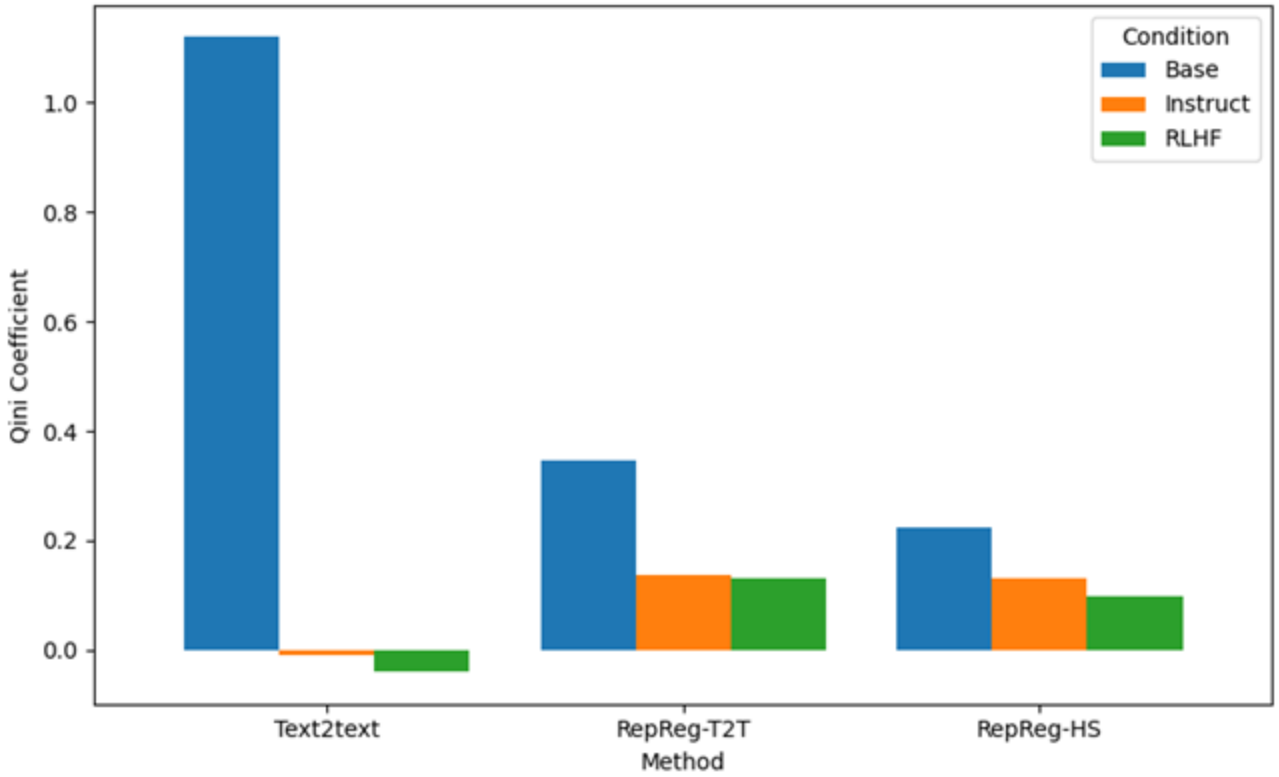
This is probably about calibration: instruction tuning and RLHF hurt it (Zhu et al, 2023)



Post-Training Hurts Performance! A Lot!

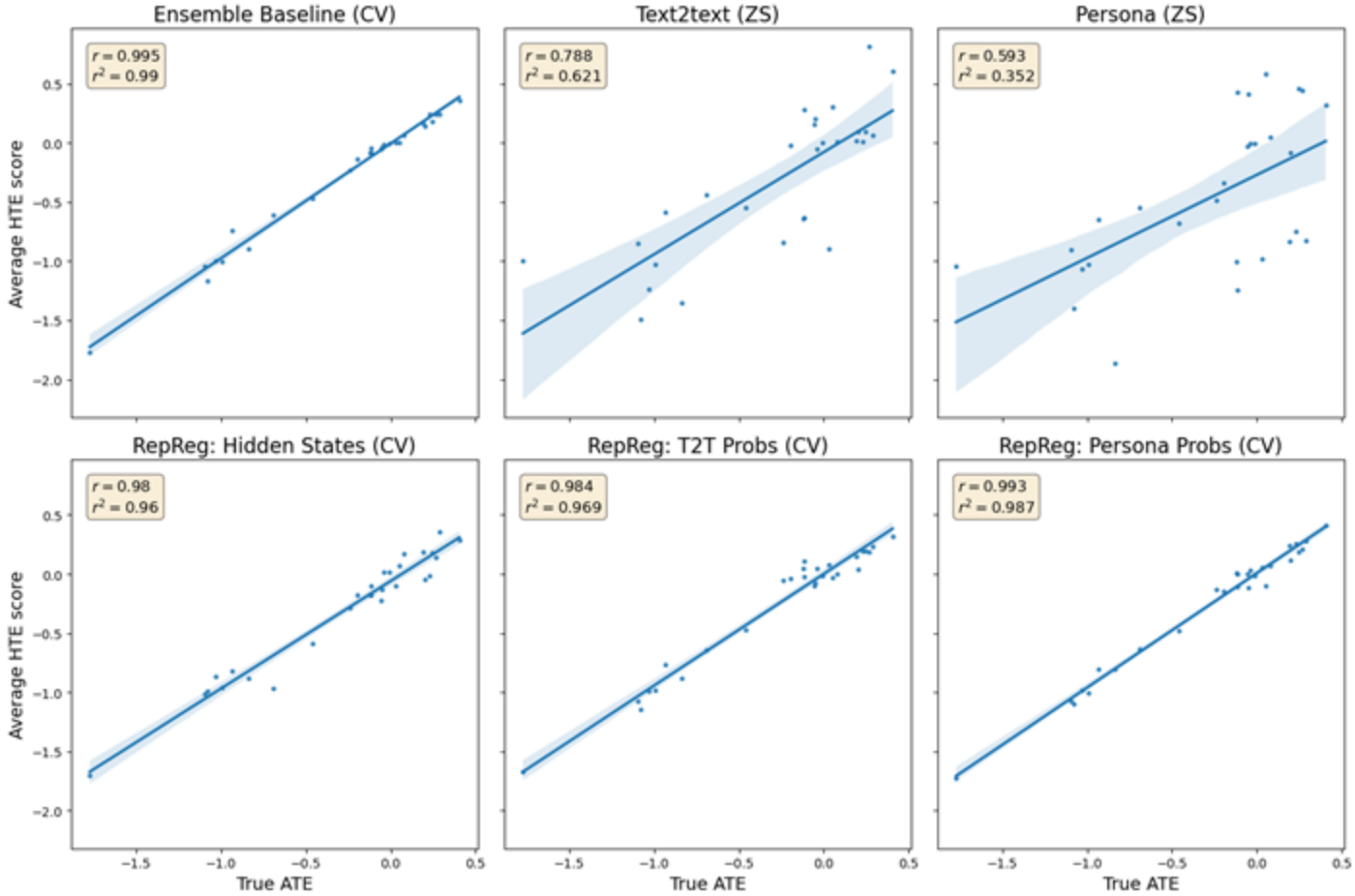
This is probably about calibration: instruction tuning and RLHF hurt it (Zhu et al, 2023)

Note how instruct/RLHF + second-stage model outperforms random: there's signal, but not right for expected value



ATE Prediction: Quite Good Zero-Shot

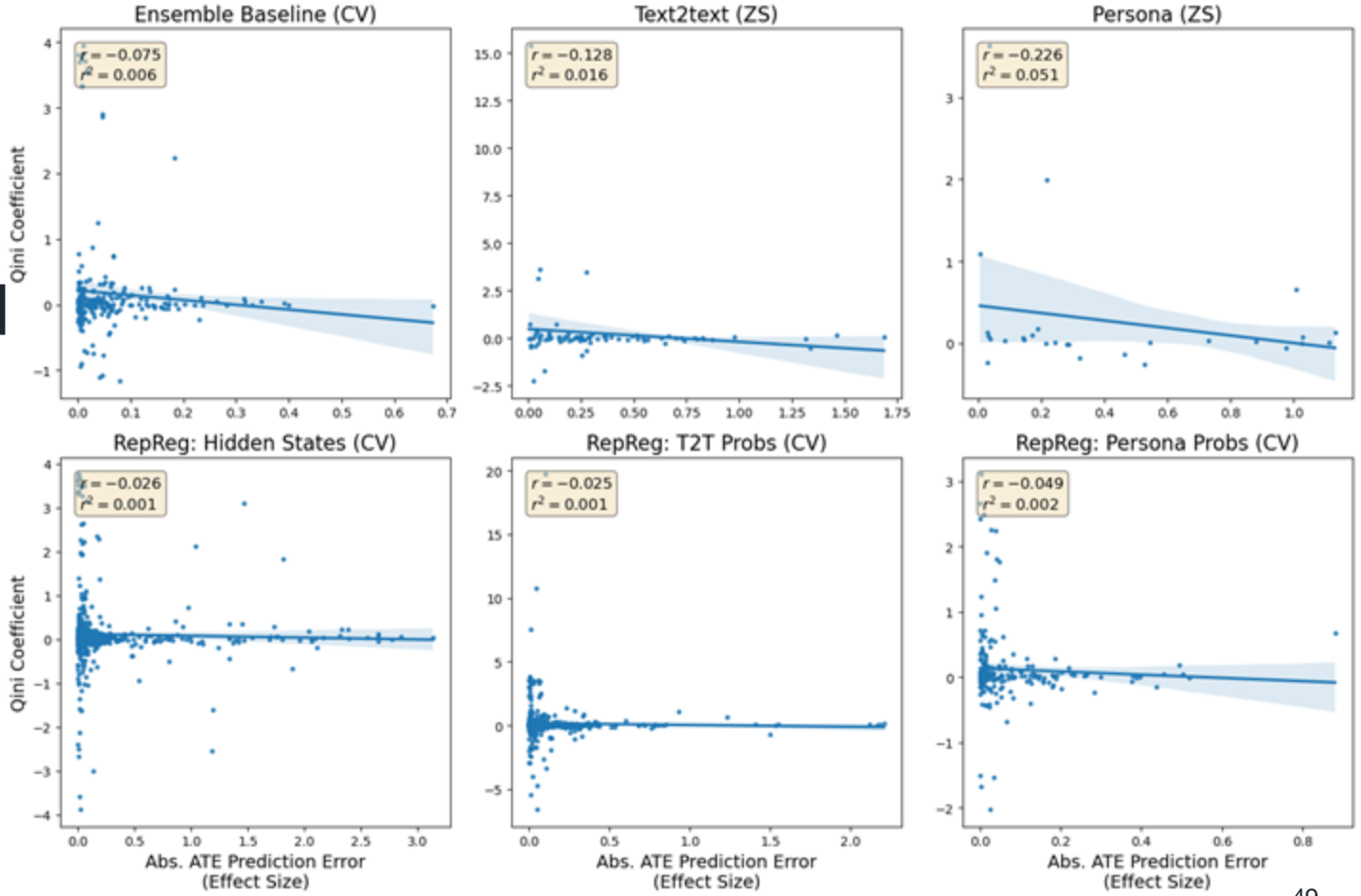
Note: This task isn't hard for cross-validated methods with training data



ATE + HTE Prediction Just Aren't Very Related

This is very consistent across specifications, subgroups, etc.

Surprising! Drawing on different model capabilities?



Is Persuasion Generalizable?

Is there transfer learning from a) news or b) other experiments to the task of predicting persuasion?

(RQ3)

News Tuning

Datasets

All four stations for the entire period

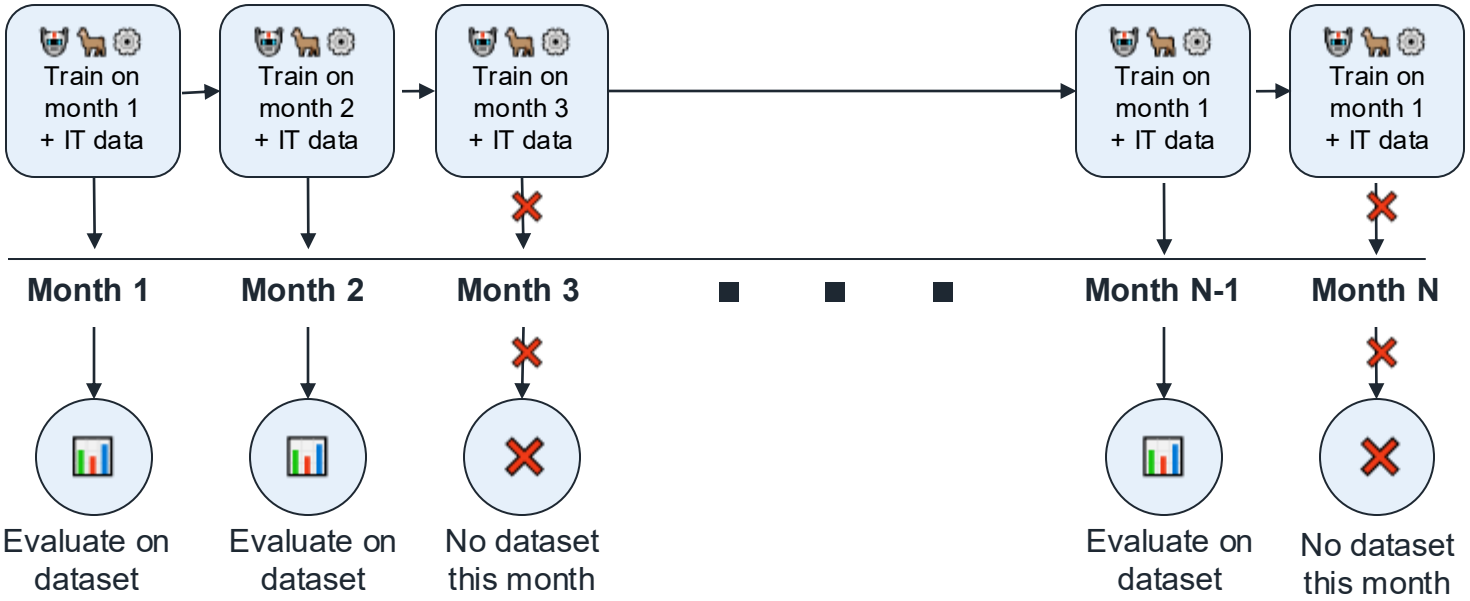
But: only 17 / 28 experiments in this window to evaluate with

4.84 GB of cleaned text

Experiment Setup

We use:

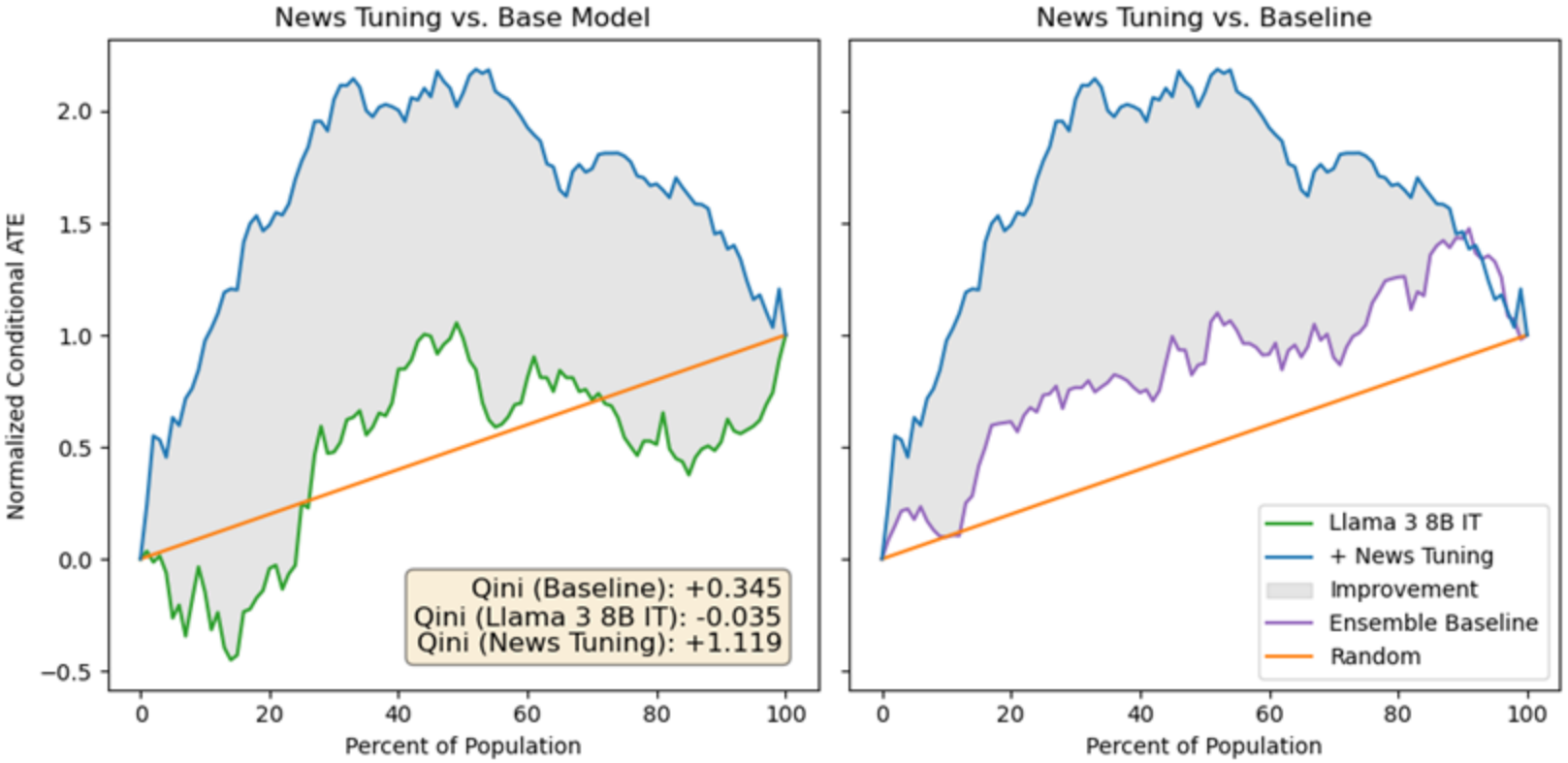
- Llama3 8B IT only for cost
- Full fine-tuning
- Standard AR SFT training



News Tuning Is Extremely Helpful

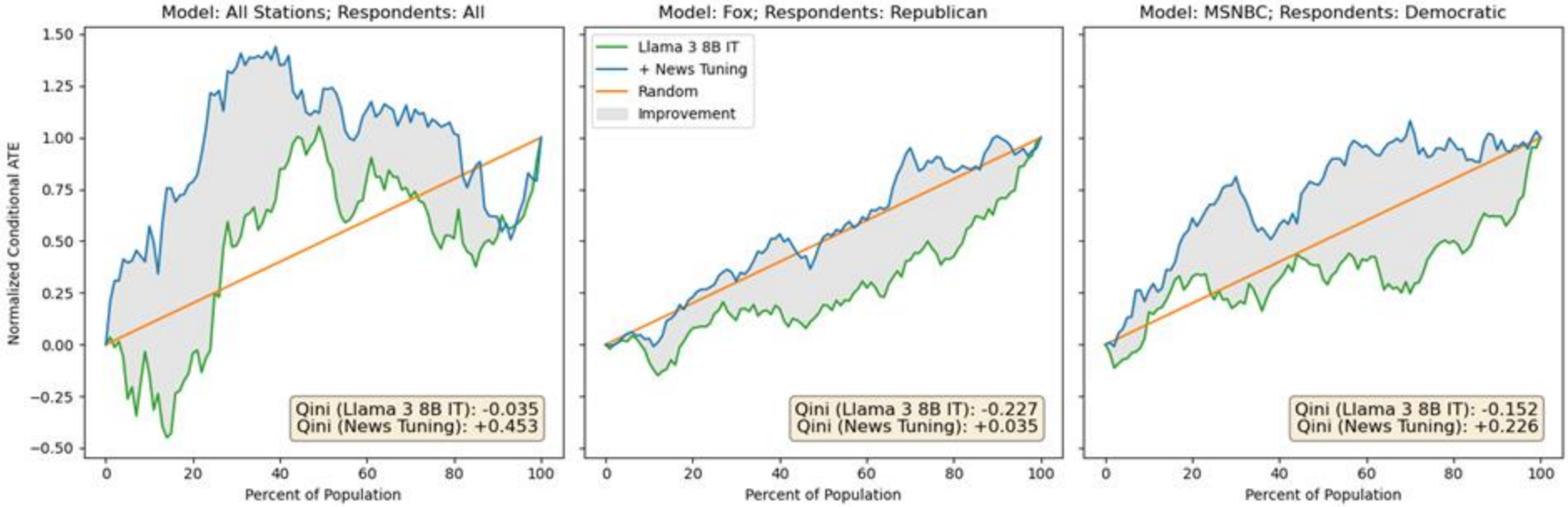
Shown is ensemble of three models:

- All-data
- MSNBC-only
- Fox-only



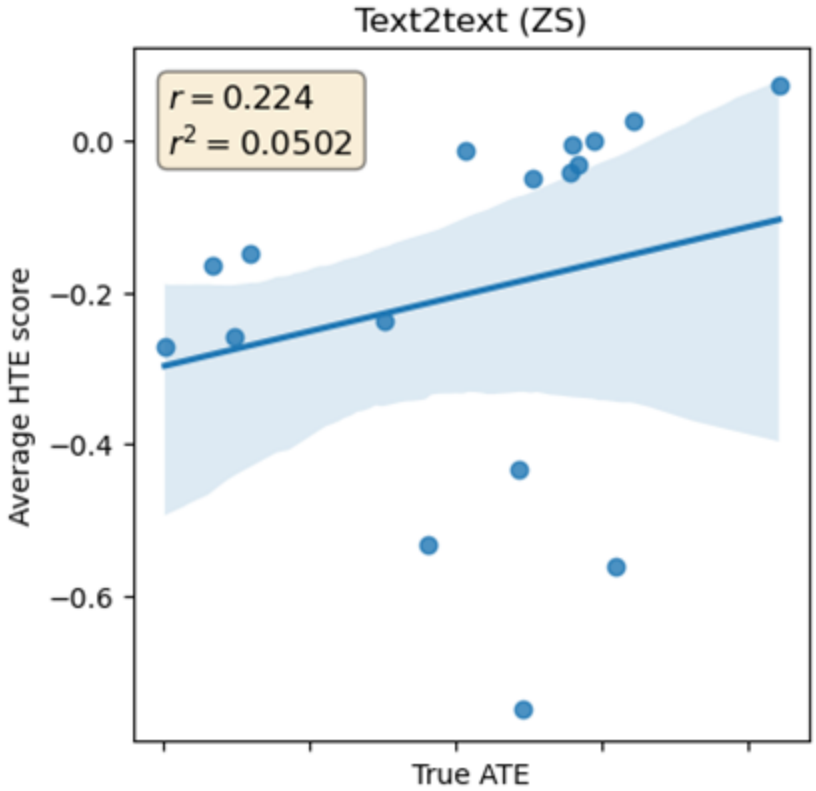
Breaking Those Out: Still Improves

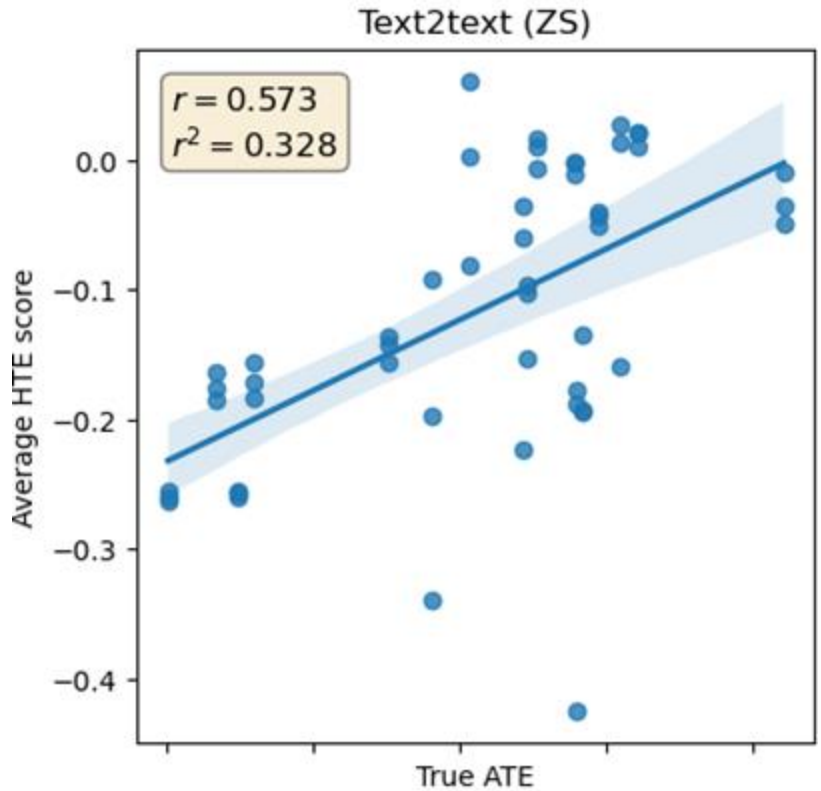
Matched sets of respondents: all-data on all respondents, MSNBC on Dems, Fox on Reps



ATE Performance: Before And After

Persona (not shown) gets worse, because training has degraded instruction-following capabilities





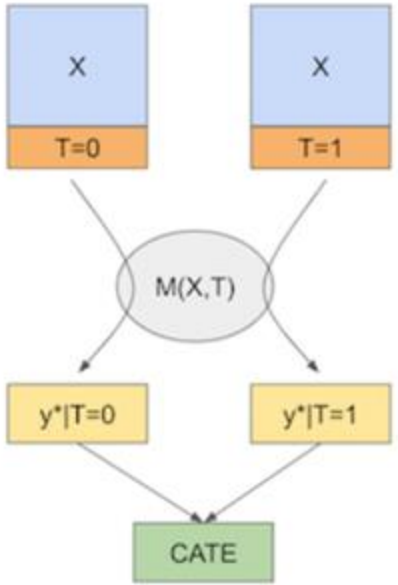
Generalization: In-Distribution

Random Splits of Experiments

Model Architecture: S-learner Recap

- Train one model with all predictors + treatment indicator T
- For new examples, predict twice: once each with T = 1 and T = 0
- The final prediction is the difference

Standard training loss (treating Likert data as continuous) is MSE

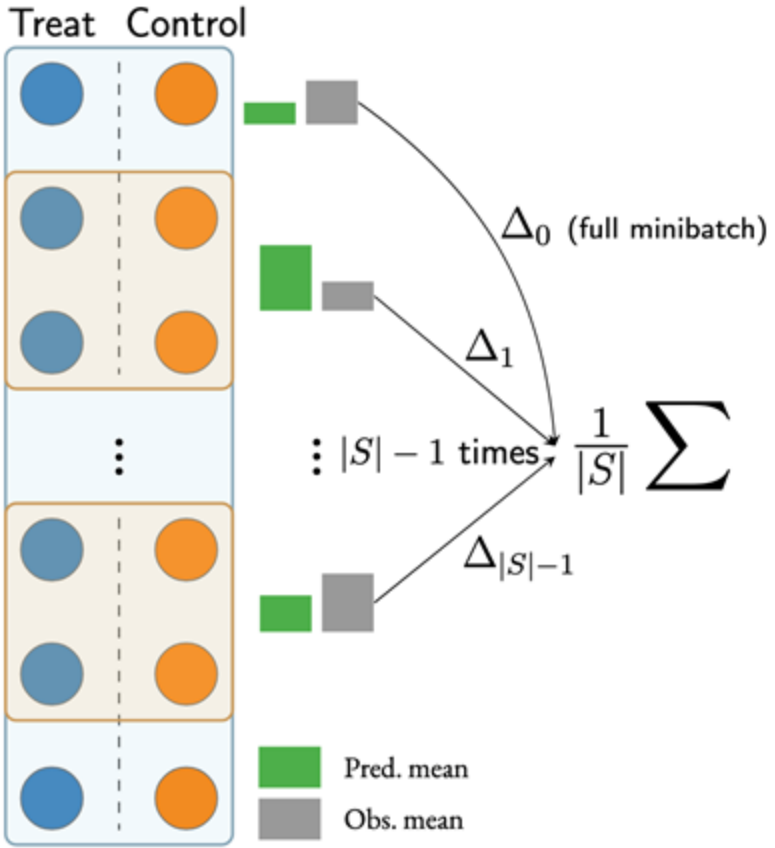


S-learner

Model Architecture: New M3 Loss

Final loss is MSE + M3

We use Pythia-410m rather than Llama3 for GPU memory reasons

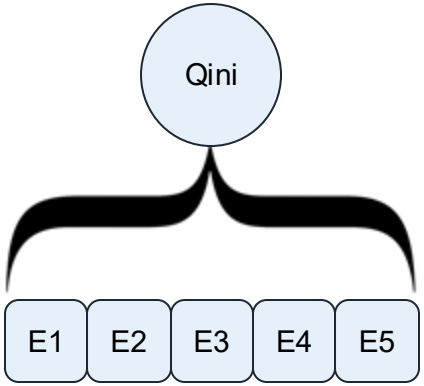


$$\frac{1}{|S|}\sum$$

$$\mathcal{L}_{M3} = \frac{1}{|S|}\sum_{s \in S}\Big(D_{obs}(s) - D_{pred}(s)\Big)^2$$

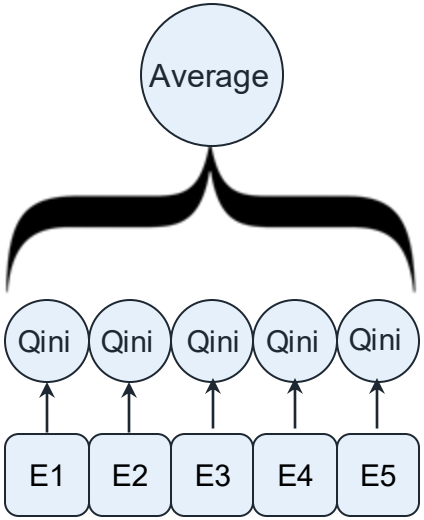
Evaluation Metrics

Pooled Qini:



Treat as one dataset

Individual Qini:

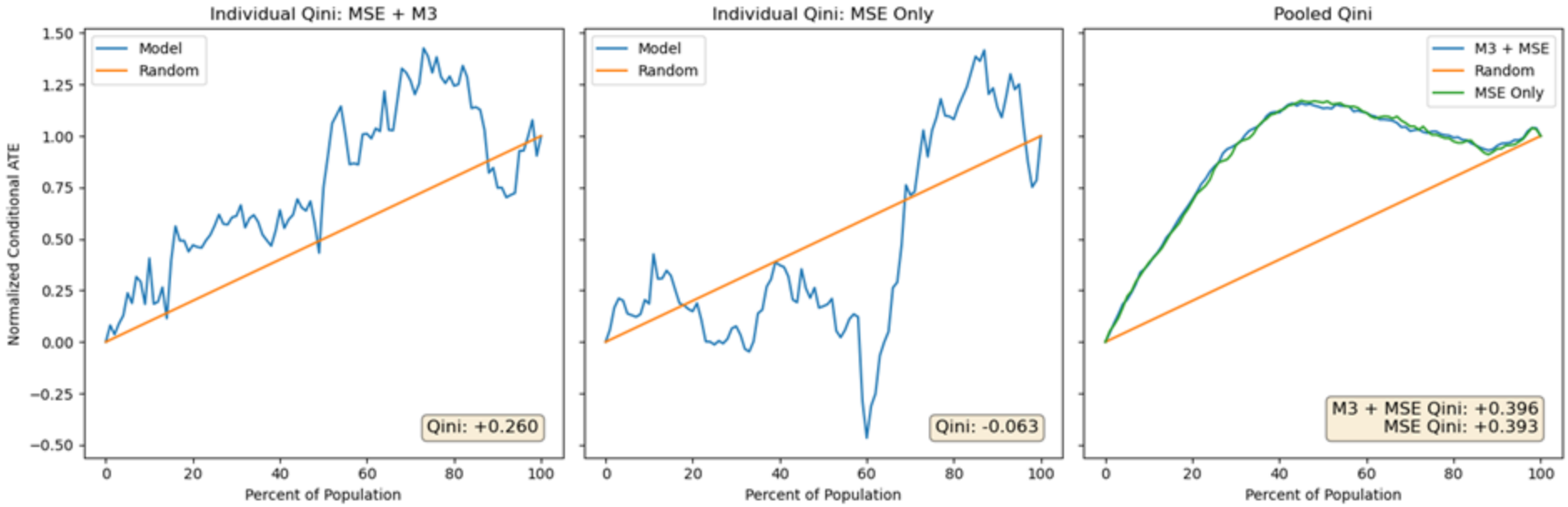


Average performance on each dataset (compute curves, average curves, compute Qini coef.)

Results

M3 objective helps!

- Performance ~ Llama3 8B IT w/ test-time persona.
- Without new objective, worse than chance.

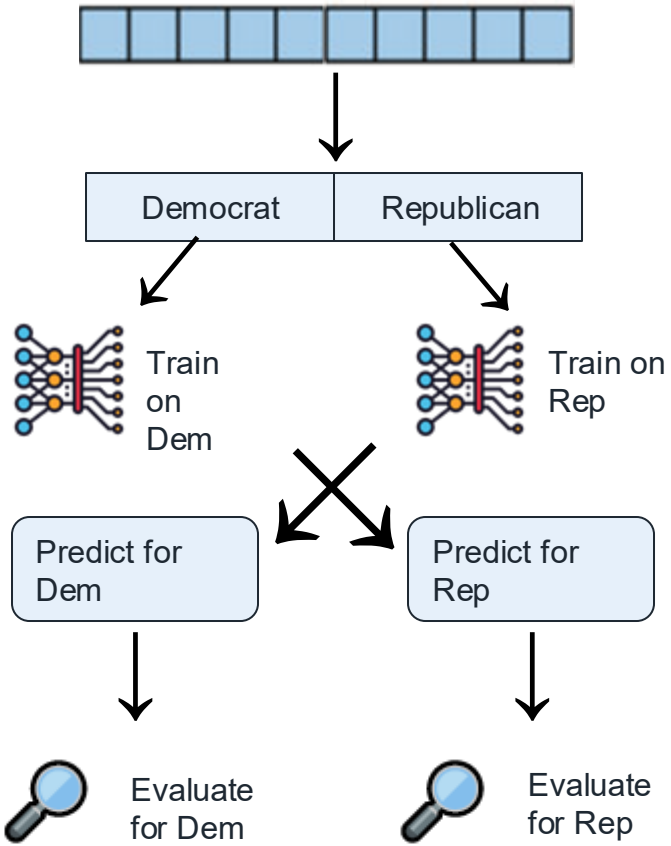


Generalization: Out of Distribution

Held-Out Demographics + Experiments

Demographic Splits

Three variables:
- Race: White / Nonwhite
- Gender: Male / Female
- Party: Democrat / Republican



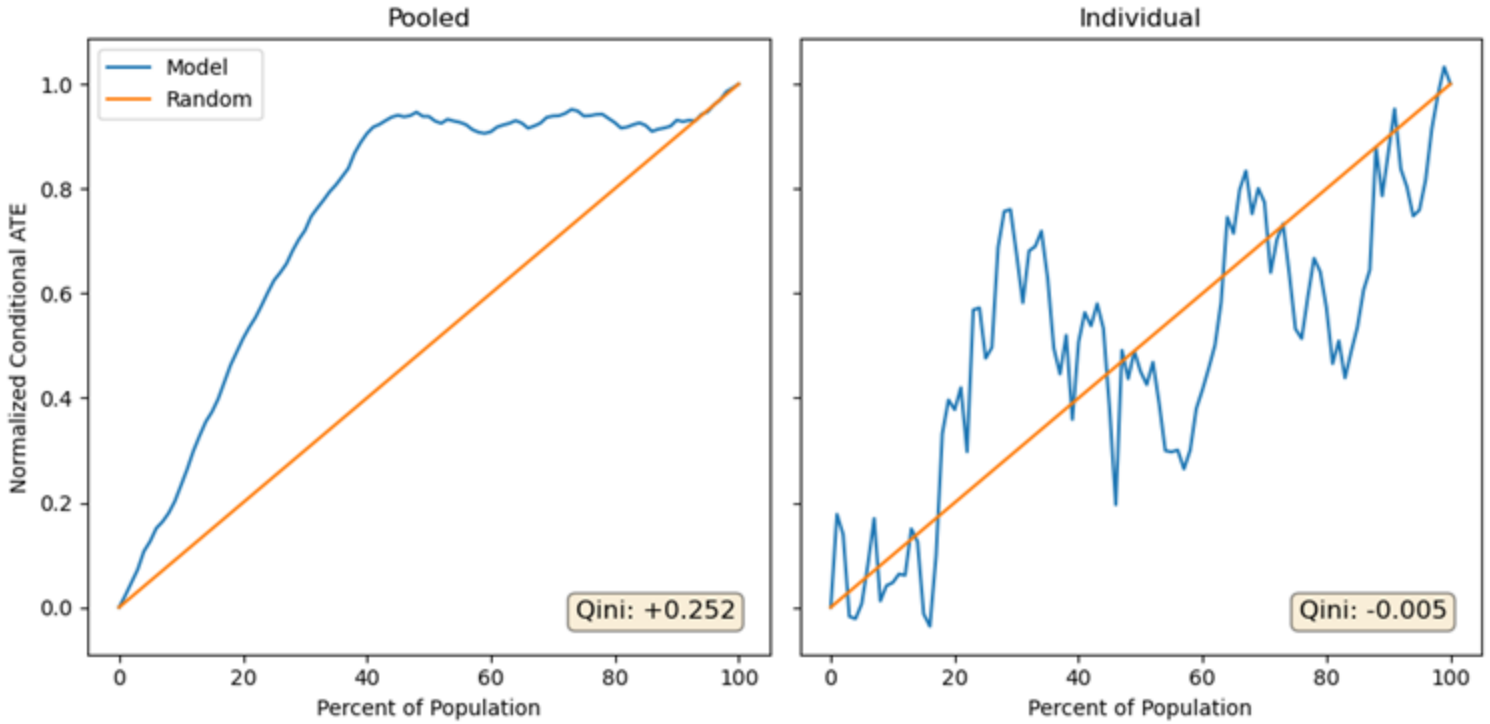
All data, experiments pooled

Demographic split

(Similarly for gender + race)

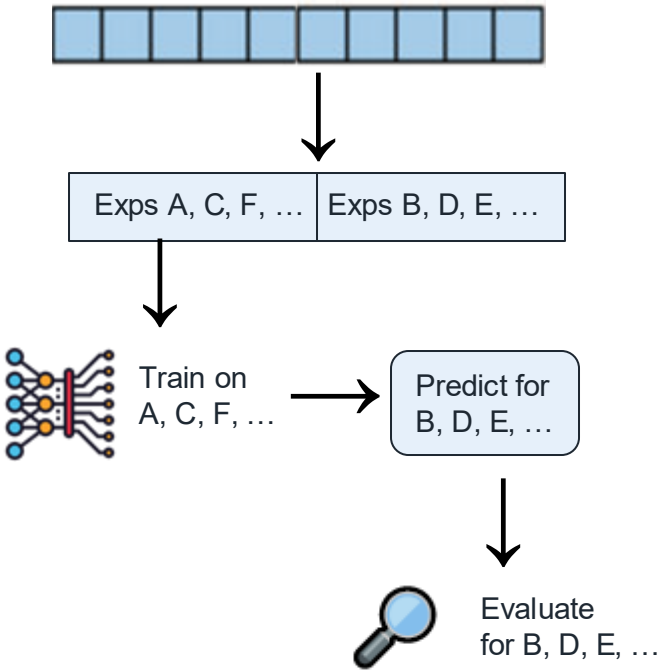
Model has never seen this data before

Results: Mixed



Cross-Dataset Splits

Repeat 3x over different random splits of the datasets + average

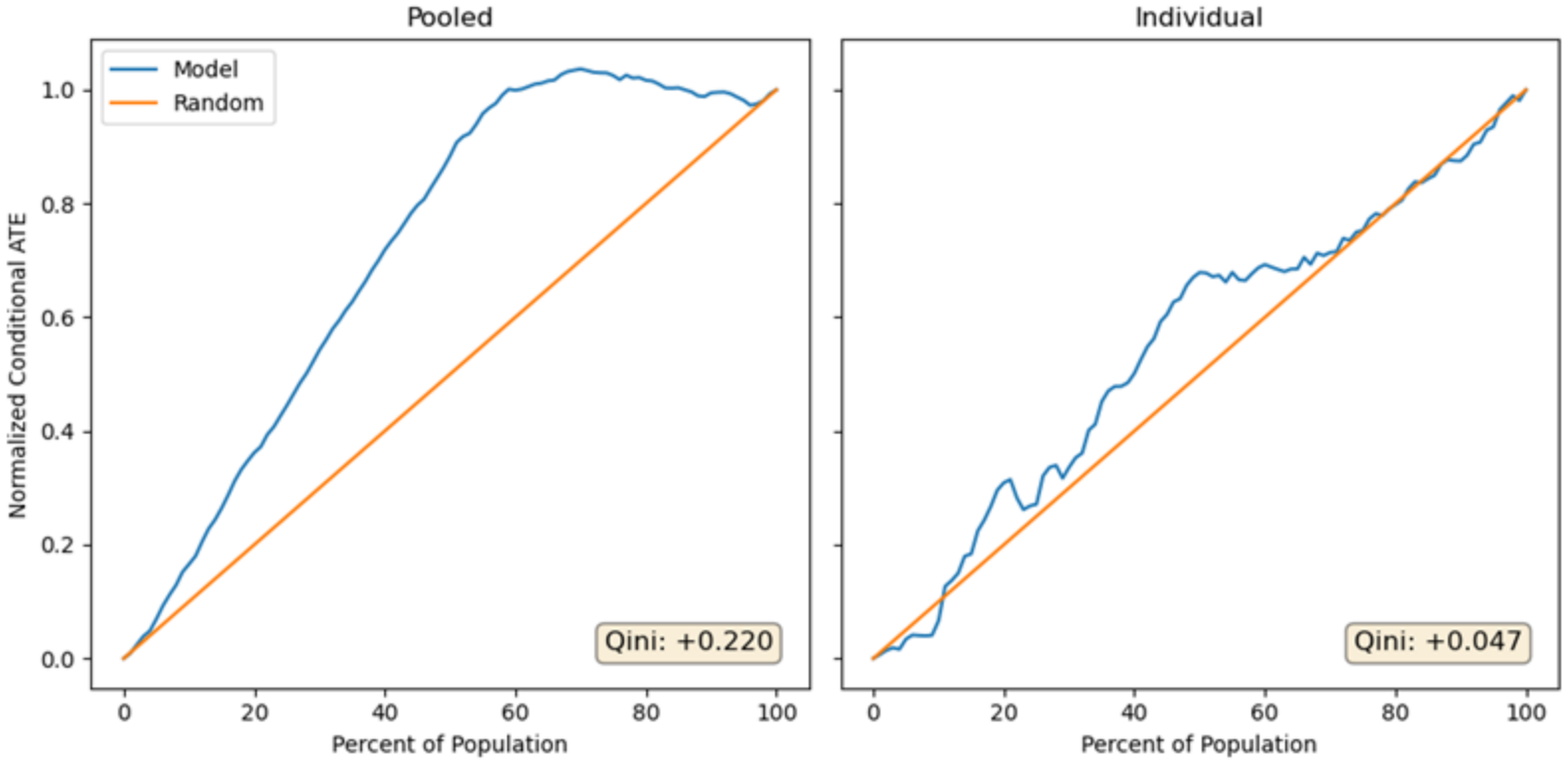


All data, experiments kept separate

Randomly split datasets

Model has never seen this data before

Results: Also Mixed



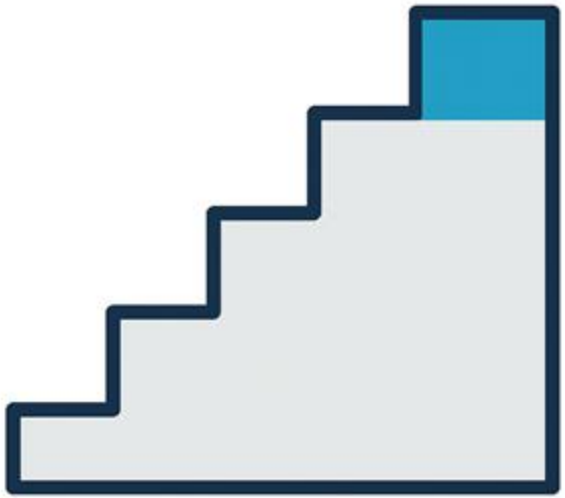
Ethics

This Is Not a Mind Control Device

Persuasive effects are typically small

Just because we can predict an effect well, doesn't mean the effect is large

Extension of Existing Practice

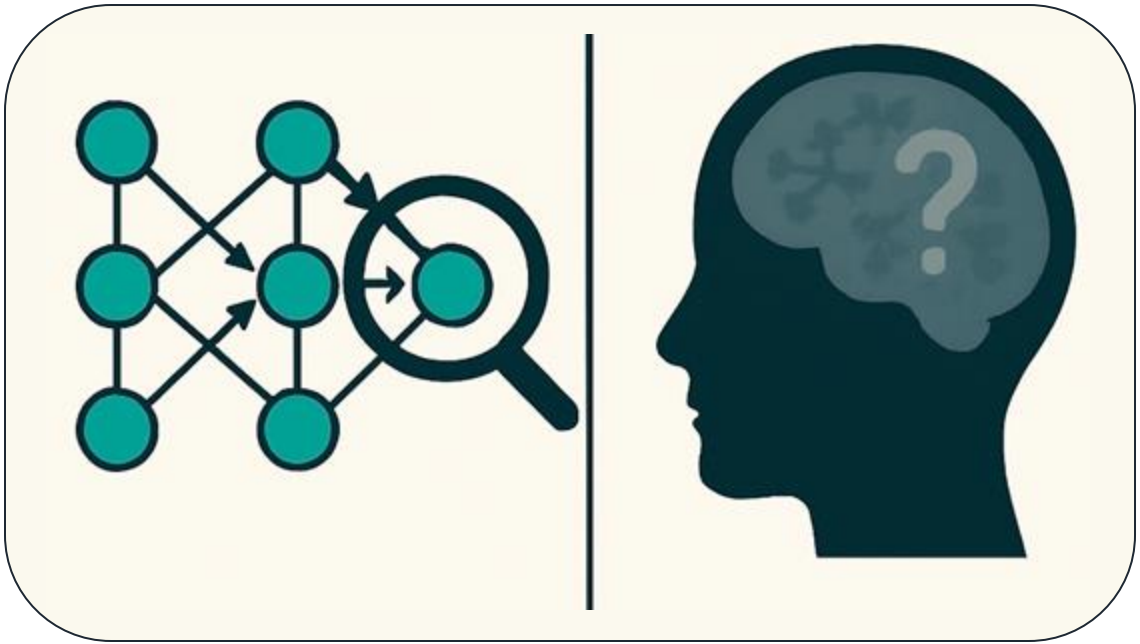


HTE estimation is well established

Commonly used in industry practice already

Improvement, not radical departure

Much Greater Scientific Value



Most useful for studying opinion change

Unlike with an LLM, you don't get to see how every neuron fires when someone changes their mind

Thank you!

William Brannon
wbrannon@mit.edu

MIT center for constructive communication

mit media lab