Challenges and an Empirical Evaluation Framework

for text-based confounding adjustment

Galen Weld*, Peter West*, Maria Glenski, David Arbour, Ryan A. Rossi, Tim Althoff

#CDSM20 - Causal Data Science Meeting, November 11, 2020

* Equal contribution

PAUL G. ALLEN SCHOOL
OF COMPUTER SCIENCE & ENGINEERING

1. Why Causal Inference with Text?

a. Background & Recent Work
b. Common Representations, Models and Estimators

2. Framework for Evaluation

a. Five Challenges for Causal Inference with Text
b. Method for Generation of Semi-Synthetic Datasets

3. Evaluation of Common Methods

a. Text Representations, Models and Estimators
b. Results and Areas for Future Improvement

Why causal inference with text?

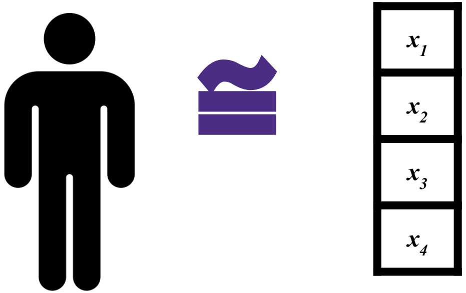
Causal Inference with Text

In much causal inference literature: people are represented with *structured* covariates (age, gender)

Natural language can contain this information in an *unstructured* form

We can represent people with text, *e.g.* social media histories

As long as confounders are encoded in text, we can adjust for them - in *theory*...
How well does this work in practice?



Exciting Recent Applications

Many recent papers have applied causal inference methods to text - too many to list! Keith et. al present an excellent review[1].

Areas of applications include:

- Mental Health[2]
- Gender in Social Media[3]
- many more...

[1]Katherine A. Keith, David Jensen, and Brendan O'Connor. *Text and Causal Inference: A Review of Using Text to Remove Confounding from Causal Estimates.* (ACL '20)
[2]M. De Choudhury, E. Kiciman, M. Dredze, G. Coppersmith, and M. Kumar. 2016. *Discovering Shifts to Suicidal Ideation from Mental Health Content in Social Media* (CHI '16)
[3]V. Veitch, D. Sridhar, and D.M. Blei. 2020 *Adapting Text Embeddings for Causal Inference*. arXiv:1905.12741

Methods for Text-Based Confounding Adjustment

Method

Text Representation

- N-grams
- LDA
- Word embeddings[1,2]

+

Propensity Score Model

- Logistic Regression
- Neural Methods[2]

+

ATE Estimator

- IPTW
- Stratification
- Matching

These methods are not the only methods, but they're the most commonly used.

[1] F. Johansson, U. Shalit, and D. Sontag. 2016. *Learning representations for counterfactual inference.* In ICML.
[2] V. Veitch, D. Sridhar, and D.M. Blei. 2020 *Adapting Text Embeddings for Causal Inference*. arXiv:1905.12741
[3] N. Kallus, X. Mao, and M. Udell. 2018. *Causal inference with noisy and missing covariates via matrix factorization.* In NeurIPS.
[4] M.E. Roberts, B.M. Stewart, R.A. Nielsen. 2020. *Adjusting for Confounding with Text Matching.* In AJPS.

Evaluation of Methods for Causal Inference with Text

- Evaluation is difficult without ground truth
- Methods are often used without clear justification
- No benchmark exists: how should practitioners choose?

Current Methods Disagree

How much of a problem is the lack of evaluation techniques?

- Conducted experiments inspired by 2 previously published papers[1,2]
- Computed ATE estimates using 11 different methods
- On both datasets, methods disagree! At most one can be correct.

[1] V. Veitch, D. Sridhar, and D.M. Blei. 2019. *Using Text Embeddings for Causal Inference*. arXiv:1905.12741
[2] M. De Choudhury, E. Kiciman, M. Dredze, G. Coppersmith, and M. Kumar. 2016. *Discovering Shifts to Suicidal Ideation from Mental Health Content in Social Media* (CHI '16).

Problem Statement:

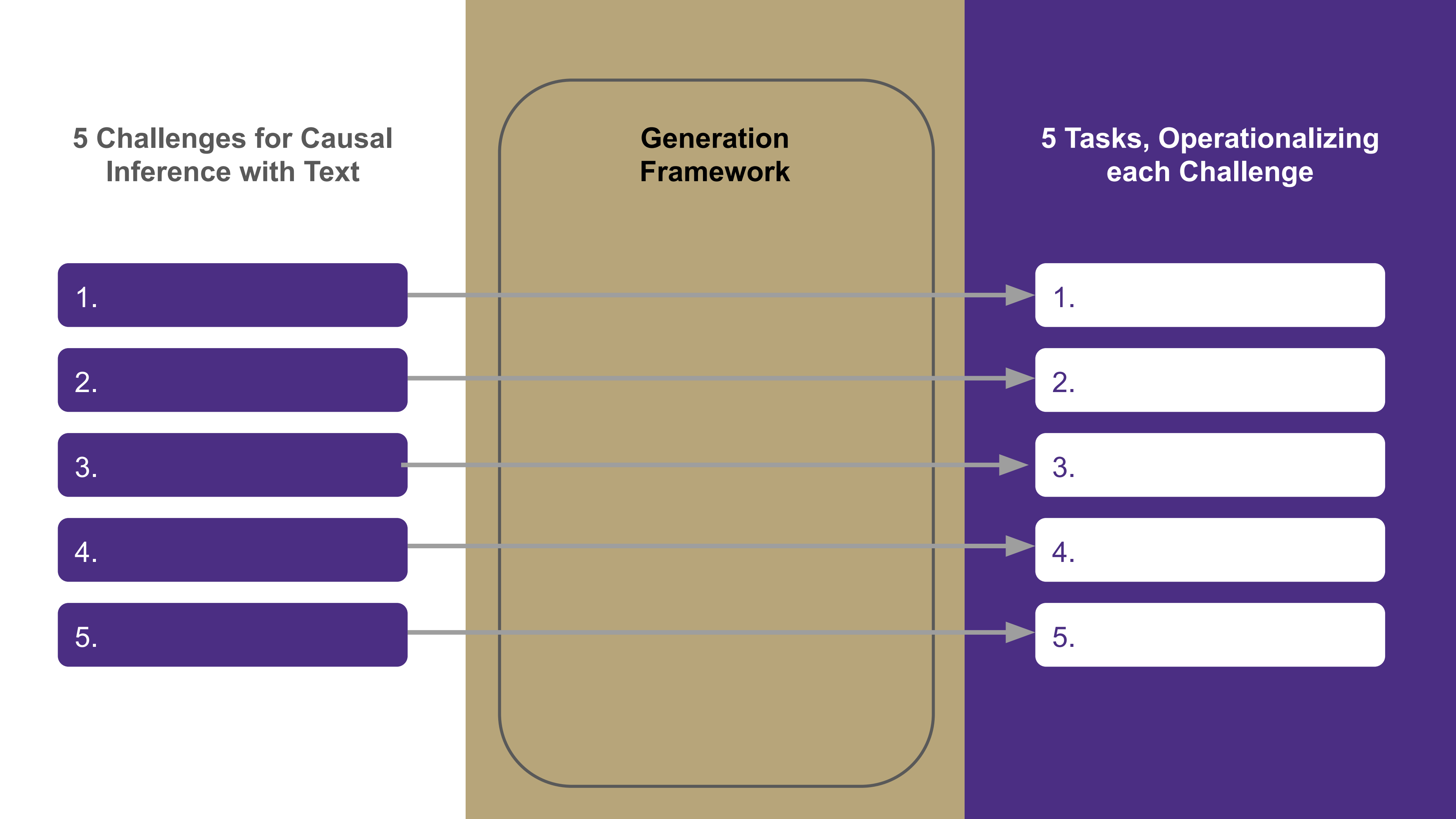
How do we evaluate methods for adjusting for confounding with text?

1. Why Causal Inference with Text?

a. Background & Recent Work
b. Common Representations, Models and Estimators

2. Framework for Evaluation

a. Five Challenges for Causal Inference with Text
b. Method for Generation of Semi-Synthetic Tasks


5 Challenges for Causal Inference with Text
Generation Framework
5 Tasks, Operationalizing each Challenge
1.
2.
3.
4.
5.
1.
2.
3.
4.
5.

1. Linguistic Complexity

Ideally, methods should be able to recognize the importance of different phrases

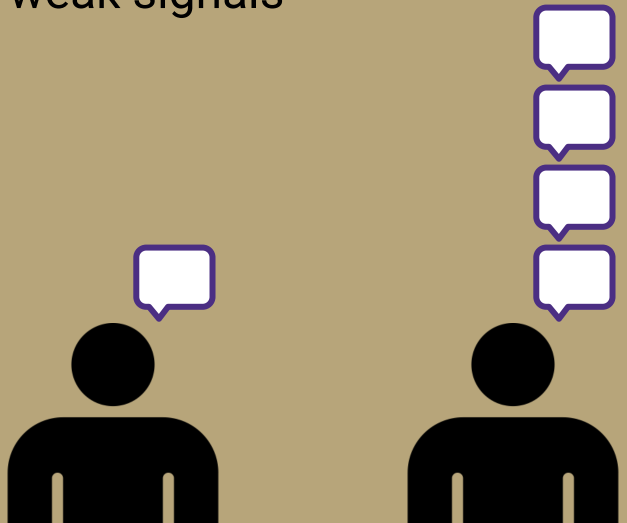
Indicative of depression

I feel depressed

I am isolated from my peers

2. Signal Intensity

Methods should be able to detect weak signals

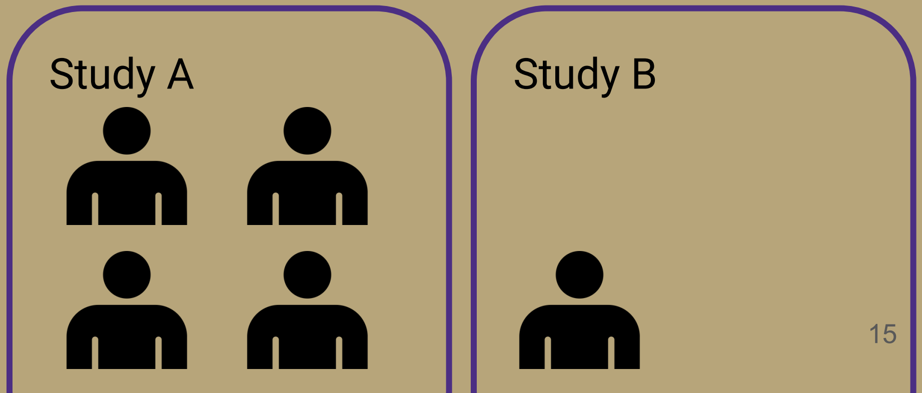


3. Strength of Selection Effect

Methods should be able to adjust for confounding, even when there is limited overlap in the distribution of language between treated and untreated users

4. Sample Size

Ideally, methods would be able to perform well even with limited observations



5. Placebo Test

Methods should not predict a causal effect when none is present

True ATE = 0

5 Challenges

for Causal Inference with Text

1. Linguistic Complexity
2. Signal Intensity
3. Strength of Selection Effect
4. Sample Size
5. Placebo Test

Generation of Semi-Synthetic Tasks

- Counterfactuals are almost never known in real life,
  - Both synthetic and semi-synthetic datasets are used for evaluation
- Use semi-synthetic data to generate each task
  - Start with the same real-world text: Reddit user profiles
  - Perturb the text to make a dataset with a known true ATE
  - Can then empirically evaluate the bias of model
- Synthetic component enables evaluation, while real component preserves realism
  - Best of both worlds!
- For each challenge, generate tasks with levels of increasing difficulty
  - Challenges form an “axis” along which we can vary the difficulty

Generative Method

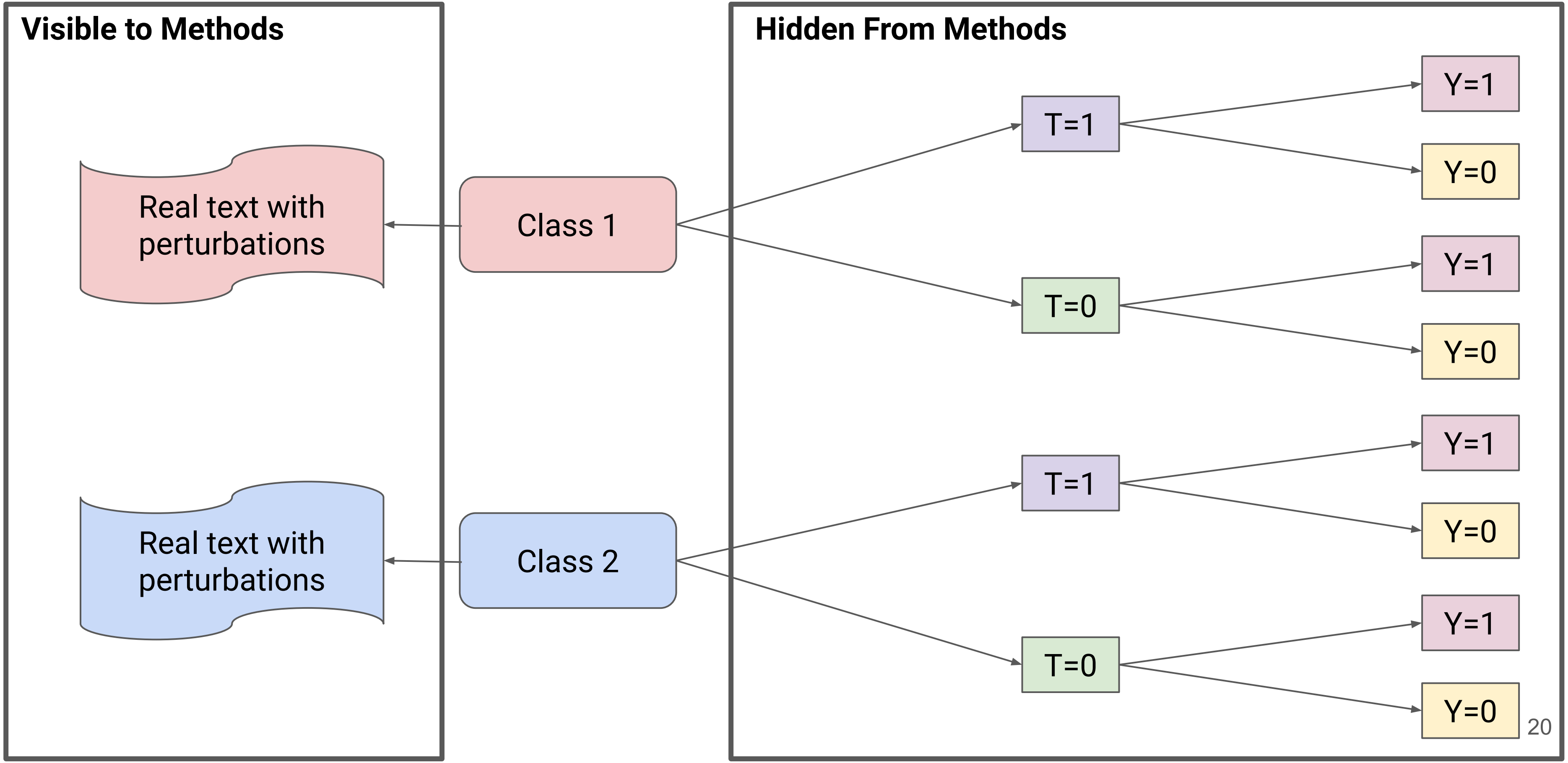
Simplified model of the world, with only two kinds of people:

- Class 1 (*e.g.* people who struggle with depression)
- Class 2 (*e.g.* people who don't)

This is an clear simplification

However, if methods fail here, unlikely they will do better in the real world

Generative Method



Task 1: Linguistic Complexity

Can methods recognize different phrases as indicative of the same treatment?

Increasing Difficulty

Level 1: Append the same synthetic post

Level 2: Append a random post mentioning sickness

Level 3: Append a random post mentioning sickness or isolation

Level 4: Append a random post on sickness, social isolation, or death

Task 2: Signal Intensity

Can methods detect weak signals?

Increasing Difficulty

Level 1:

Signal to Noise Ratio is infinitely high (10:0)

Level 2:

Signal to Noise Ratio is 3:1

Task 3: Strength of Selection Effect

How does performance diminish as the overlap between treated and control groups decrease?

Increasing Difficulty

Weak Selection Effect (easier):

.9/.1 split for class 1 to be treated, class 2 untreated

Strong Selection Effect (harder):

.95/.05 split for class 1 to be treated, class 2 untreated

Task 4: Sample Size

Can methods perform well with limited training data?

Increasing Difficulty

Level 1:

Train on all 3,200 users

Level 2:

Train on a random subset of 1,600 users

Level 3:

Train on a random subset of 800 users

Task 5: Placebo Test

Do methods falsely predict causal effects when none are present?

ATE for class 1 set to +.9

ATE for class 2 set to -.9

As classes are balanced, overall ATE is 0

1. Why Causal Inference with Text?

a. Background & Recent Work
b. Common Methods

2. Framework for Evaluation

a. Five Challenges for Causal Inference with Text
b. Method for Generation of Semi-Synthetic Datasets

3. Evaluation of Common Methods

a. Text Representations, Models and Estimators
b. Results and Areas for Future Improvement

What methods do we evaluate?

Text Representation

- Unigrams (binary)
- Bigrams (binary)
- Bigrams (counted)
- LDA features

×

Propensity Score Model

- Logistic Regression
- Simple Neural Net
  (1 fully connected hidden layer)

×

ATE Estimator

- IPTW (Hajek[2])
- Stratification
- 1:1 Matching

SHERBERT, extending Causal BERT[1] to long text histories

×

[1]V. Veitch, D. Sridhar, and D.M. Blei. 2019. *Using Text Embeddings for Causal Inference*. arXiv:1905.12741

[1]Hajek, J. 1970. *A characterization of limiting distributions´ of regular estimates.* Zeitschrift fur Wahrscheinlichkeitstheorie und Verwandte Gebiete

Key Findings

Text representations and propensity score models matter more than ATE estimators.

Key Findings

Many models fail a placebo test - this is greatly concerning!

Key Findings

Transformer-based representations and models offer a promising path for improvement.

Key Findings

However, transformer-based models have limitations.

- Struggle with counting
- Require more data to be trained effectively.

Conclusion & Future Work

Every model has room for improvement - more work is needed

Our framework is not “complete” - no framework can be!

We contribute:

- the first evaluation framework in this space, consisting consisting of 5 tasks
- an evaluation of 27 common methods

Hope to spark a continued conversation on how best to evaluate causal inference methods for text.

Thank You. Questions?

https://behavioral-data.github.io/CausalInferenceChallenges/

@galenweld gweld@cs.washington.edu

Real World Experiments: Gender and Moderation

Gender Experiment[1]	
n	90,000 posts
Observation (O)	Posts from 3 subreddits in 2018
Treatment (T)	Author's flair is 'male' or 'female'
Outcome (Y)	Post's final score: # upvotes - # downvotes
Features (X)	The text of the post

Moderation Experiment[2]	
n	13,786 user histories
Observation (O)	Users' post history from /r/science, 2015-2017
Treatment (T)	User has a post removed by a moderator in 2018
Outcome (Y)	Number of posts a user makes in 2019
Features (X)	Users' post histories

[1]V. Veitch, D. Sridhar, and D.M. Blei. 2019. *Using Text Embeddings for Causal Inference*. arXiv:1905.12741
[2]M. De Choudhury, E. Kiciman, M. Dredze, G. Coppersmith, and M. Kumar. 2016. *Discovering Shifts to Suicidal Ideation from Mental Health Content in Social Media* (CHI '16).

Real World Experiments: Models

Compared 9 models for each experiment

2 commonly used models:

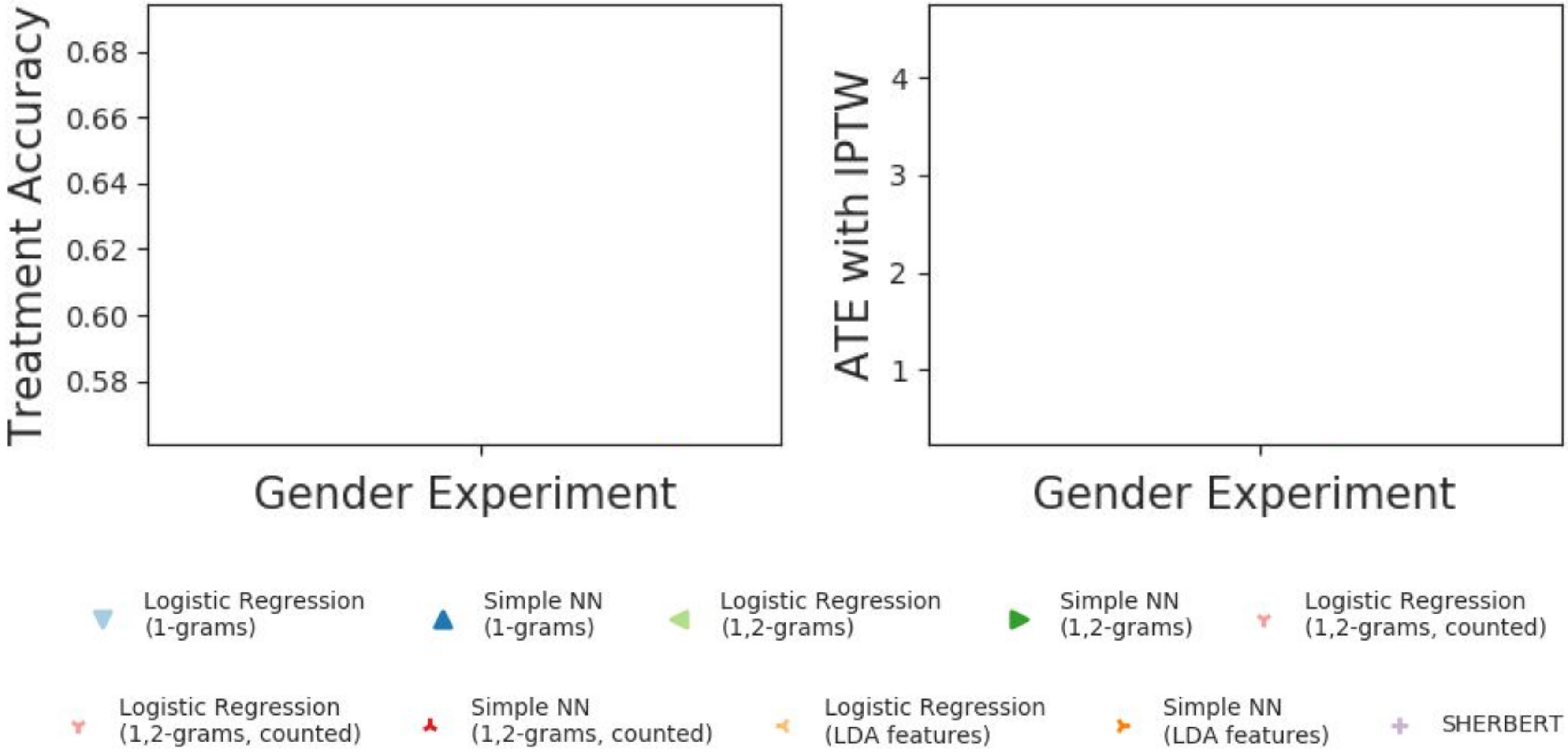
- Logistic Regression
- Simple Neural Network

3 kinds of features:

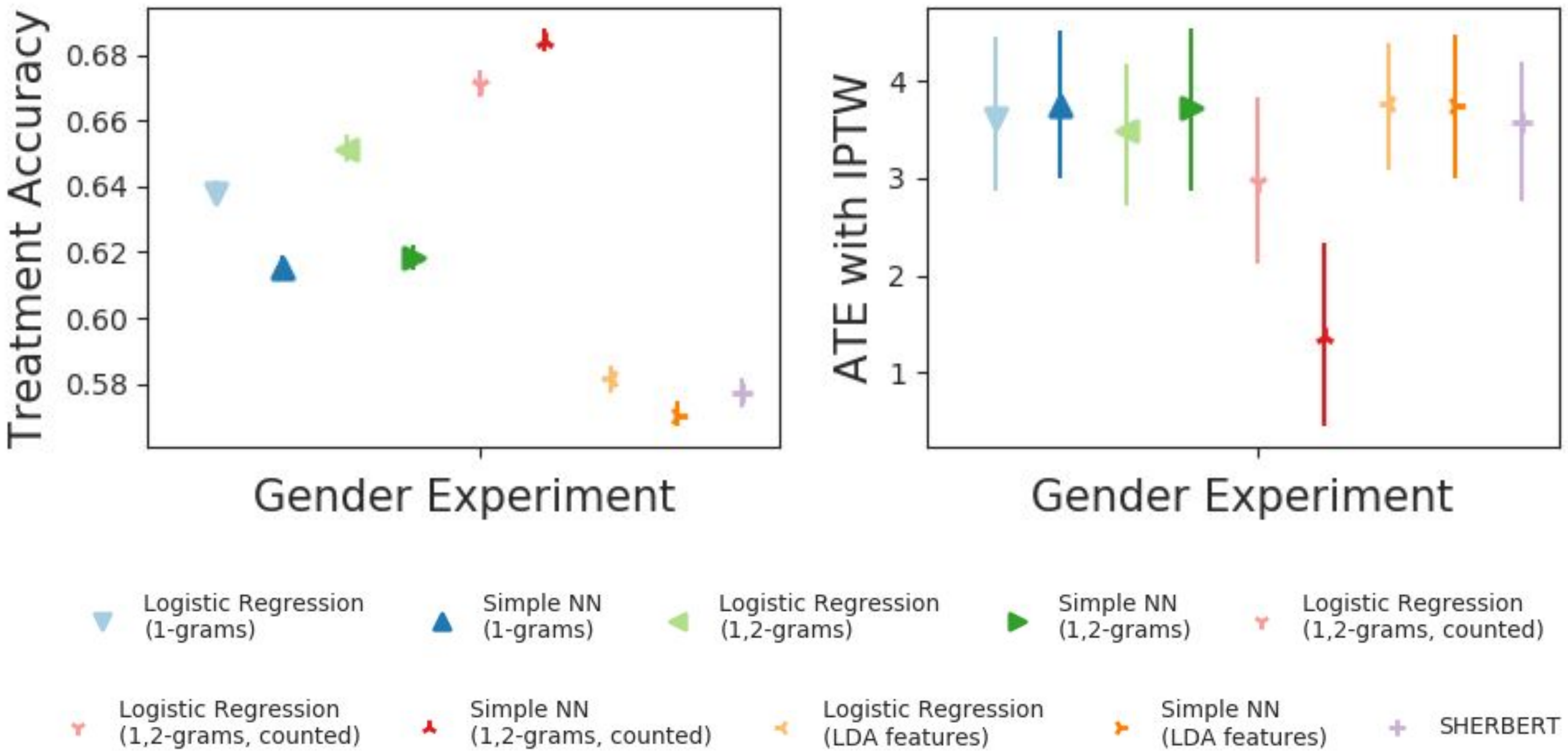
- Unigrams (binary)
- Bigrams (binary and counted)
- Latent Dirichlet allocation (LDA)

SHERBERT, our BERT-derived hierarchical model

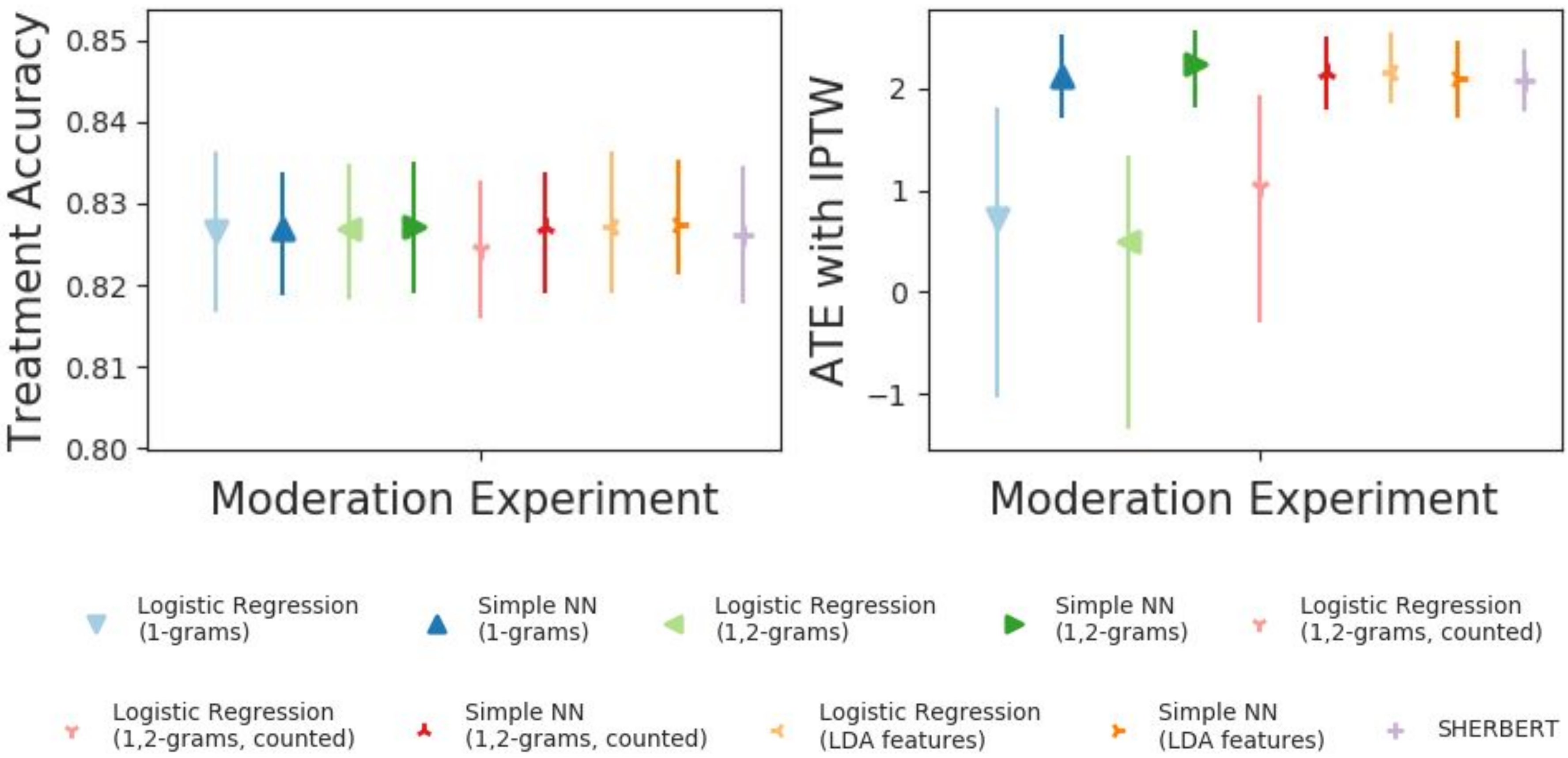
Real World Experiments: Gender Results



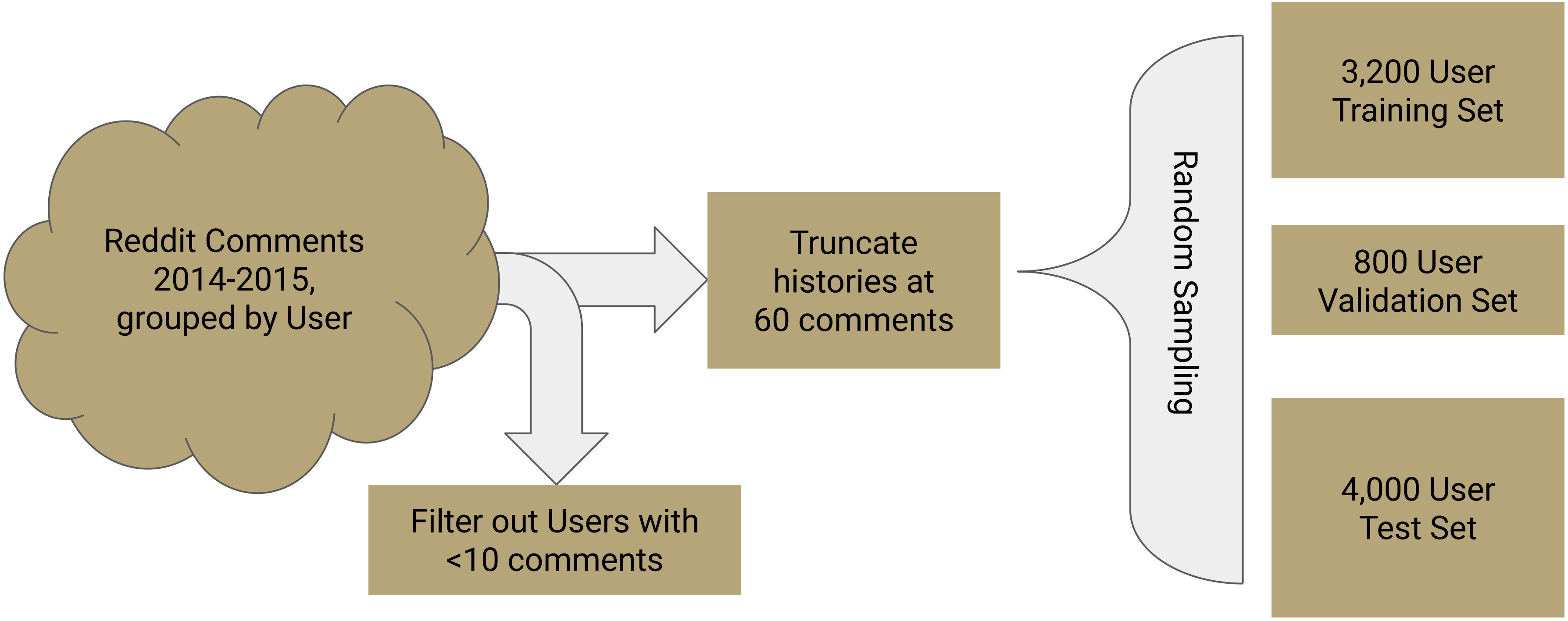
Real World Experiments: Gender Results



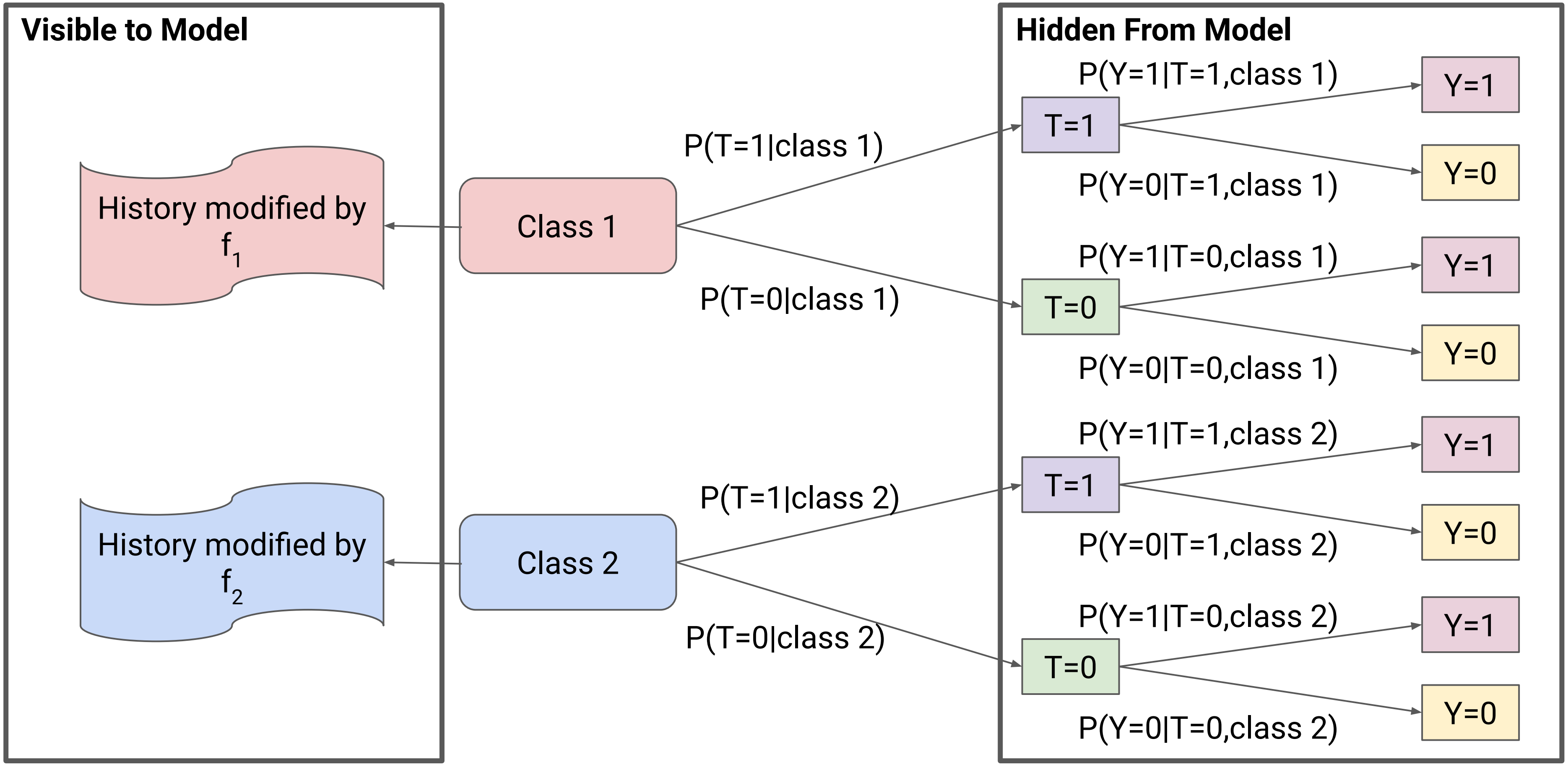
Real World Experiments: Moderation Results



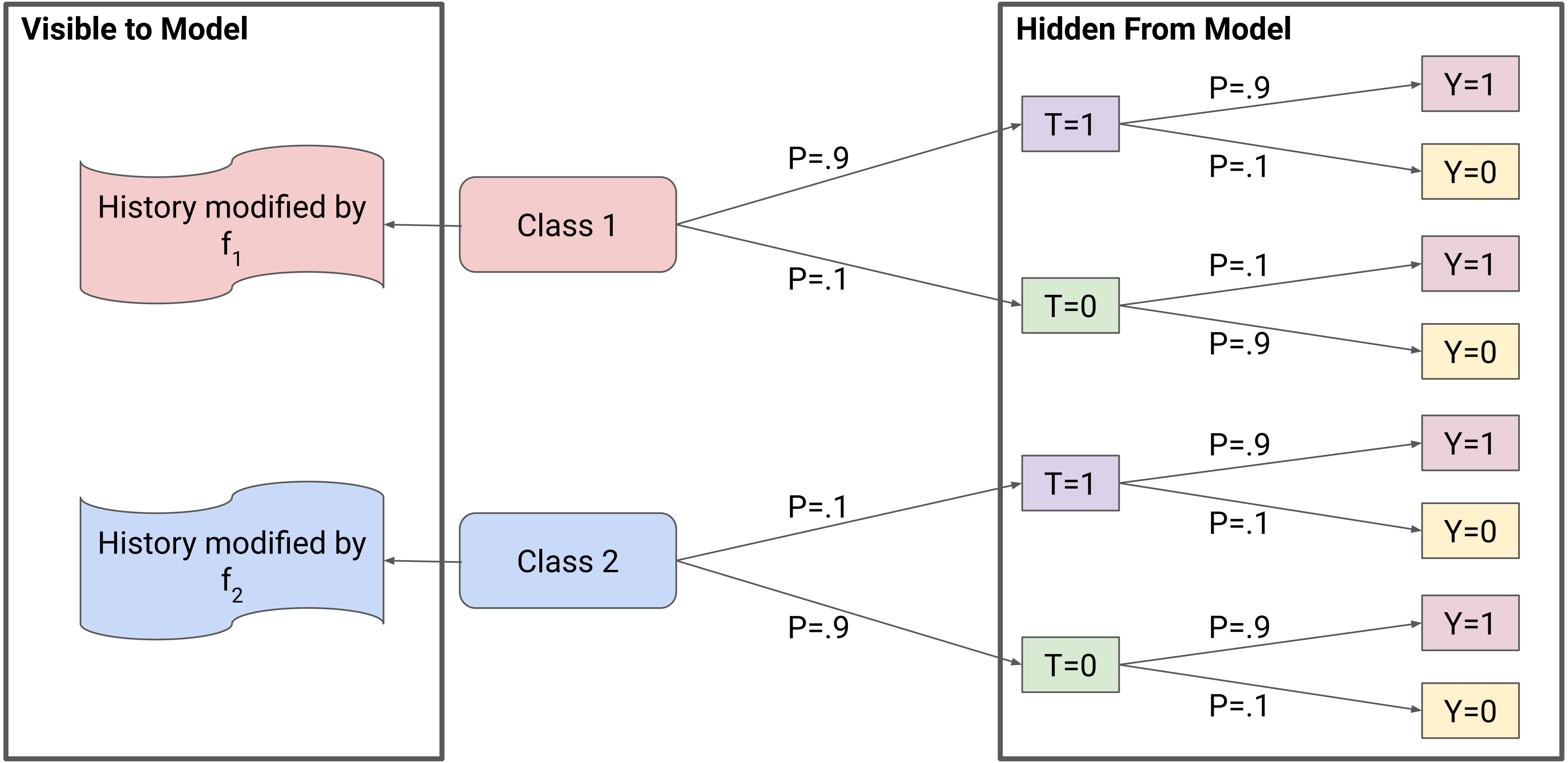
Reddit Data


Reddit Comments
2014-2015,
grouped by User
Truncate
histories at
60 comments
Filter out Users with
<10 comments
Random Sampling
3,200 User
Training Set
800 User
Validation Set
4,000 User
Test Set

Generative Method



Generative Method



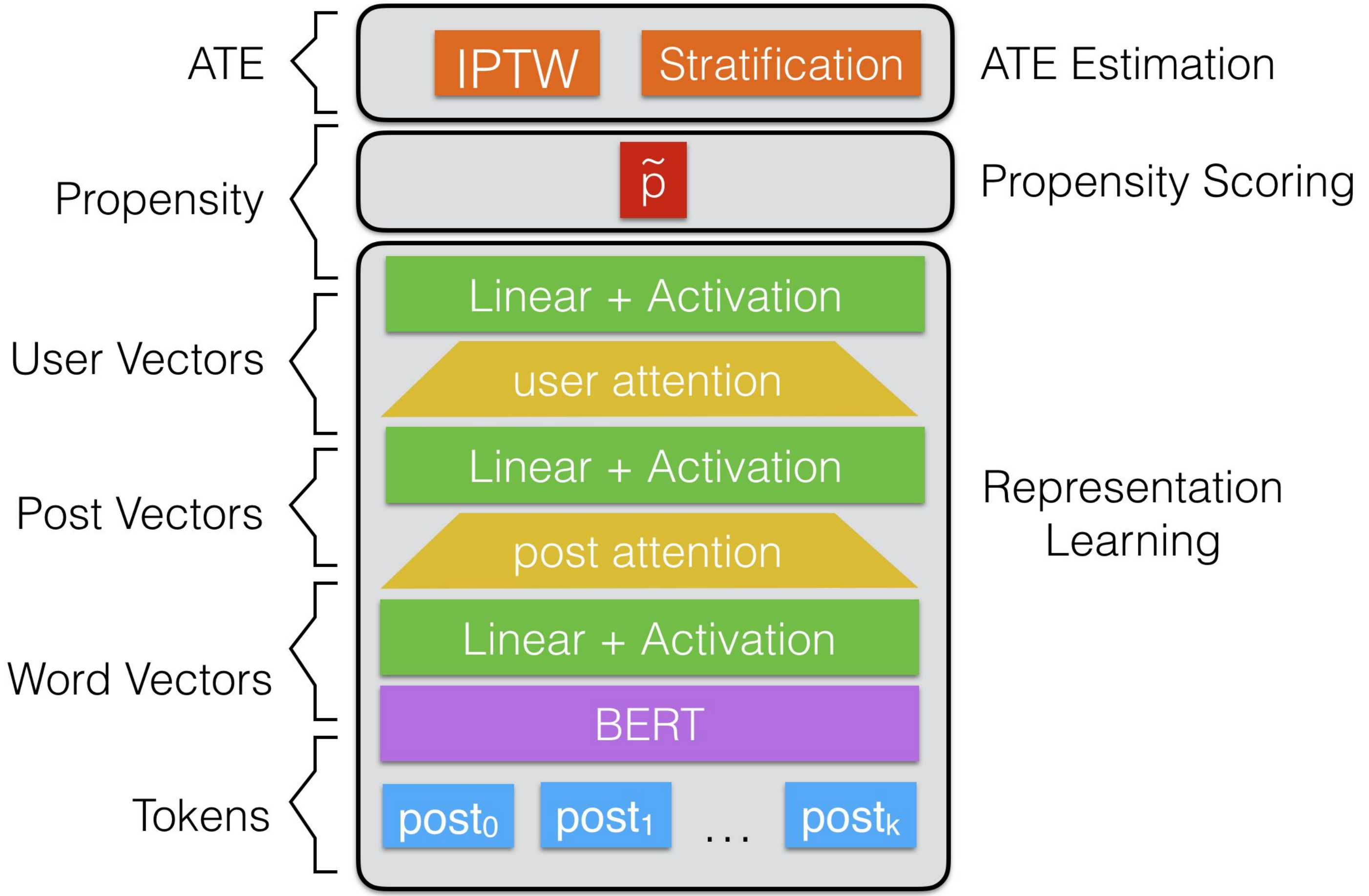
SHERBERT Model

cauSal HiERarchical variant of BERT

Expands upon Causal BERT from Veitch, *et. al,*[*] with better scalability

[*]V. Veitch, D. Sridhar, and D.M. Blei. 2019. *Using Text Embeddings for Causal Inference*. arXiv:1905.12741

SHERBERT Model



3 kinds of Synthetic Posts

used in f_1 and f_2 to insert into users' histories

1. Sickness Posts

e.g.
"The doctor told me I have AIDS"
"I got diagnosed with cancer last week"

56 different posts

2. Social Isolation Posts

e.g.
"I feel so alone, my last friend said they needed to stop seeing me."

12 different posts

3. Death Posts

e.g.

“I just found out my mom died”

“My girlfriend passed away recently”

135 different posts

Baselines: Theoretical Comparison Points

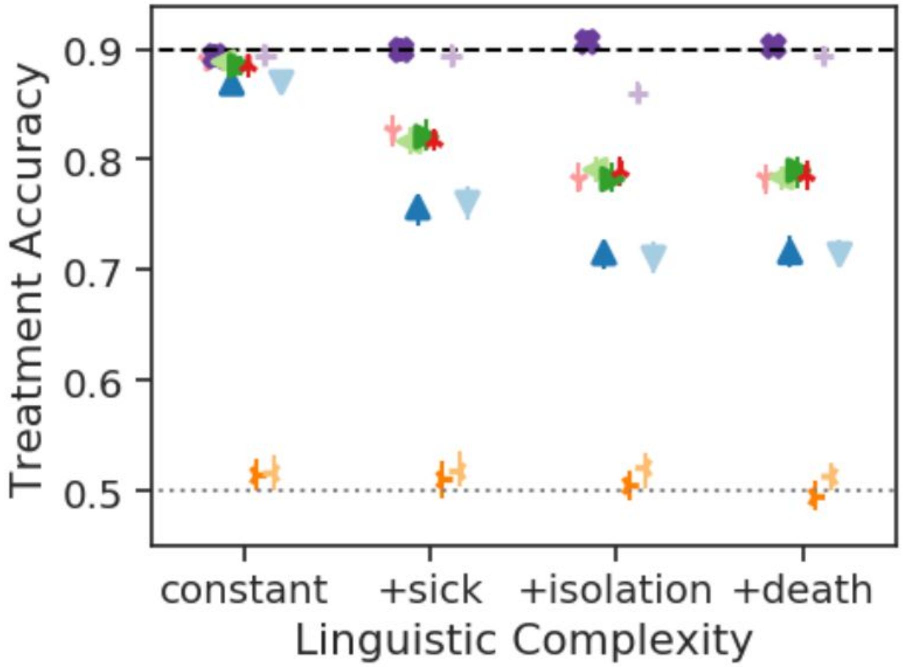
Unadjusted Estimator (lower bound)

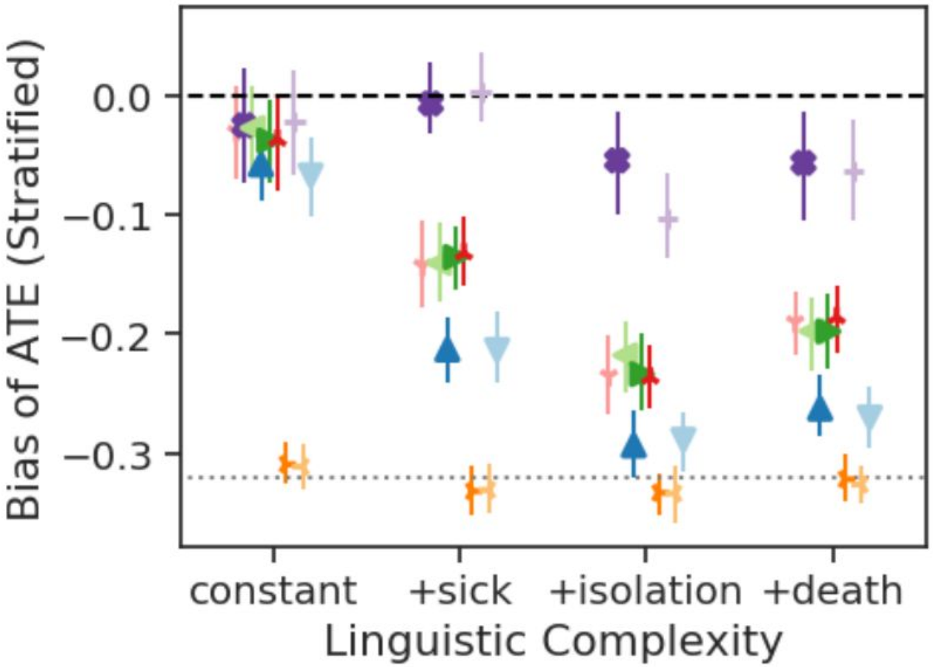
- Outputs propensity score estimate of .5 for every observation
- Effectively does not adjust for confounding

Oracle (upper bound)

- Outputs the true propensity score
- Differs only from the theoretically optimal performance due to finite sample effects

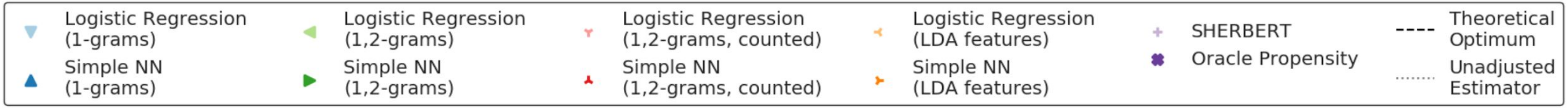
Results: Linguistic Complexity



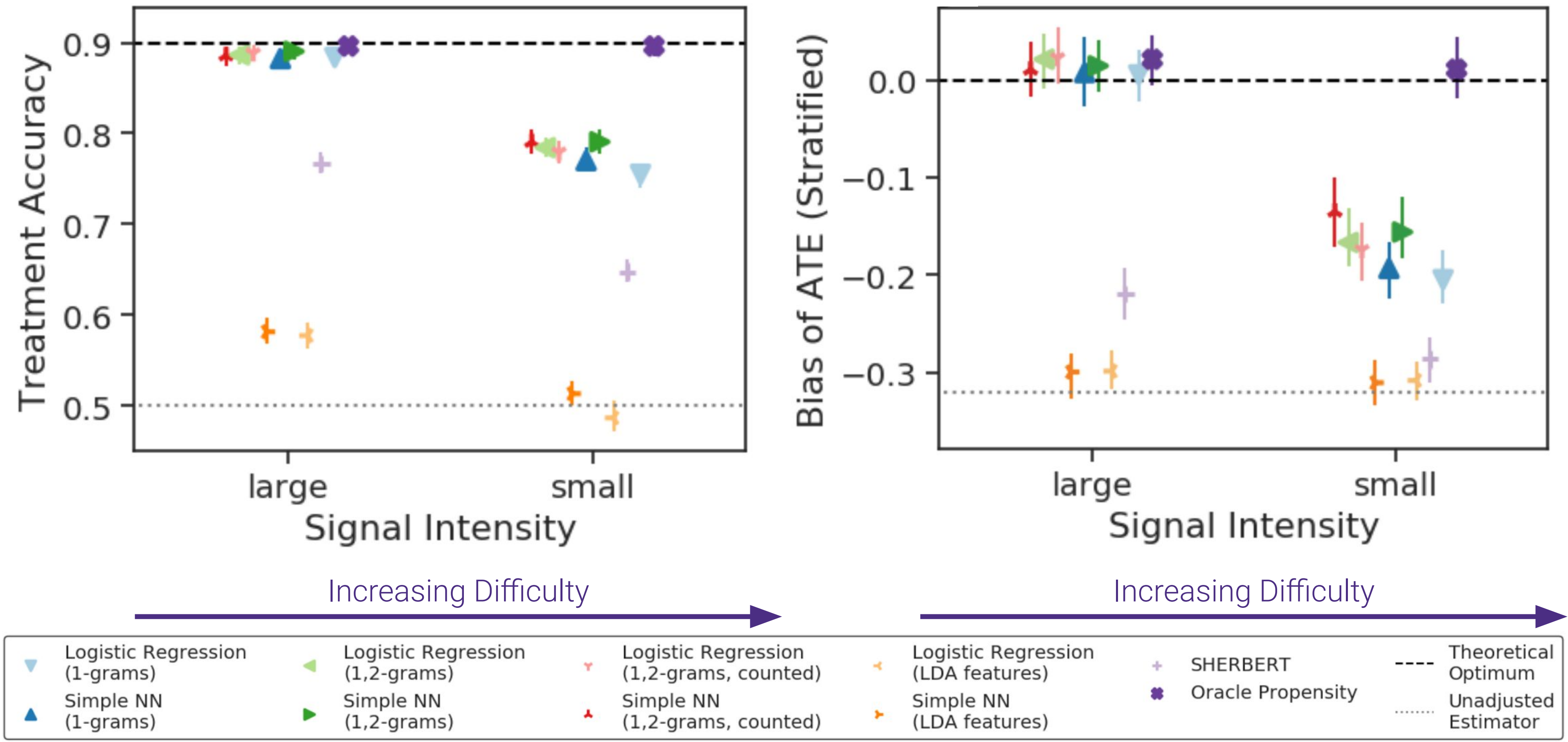


Increasing Difficulty

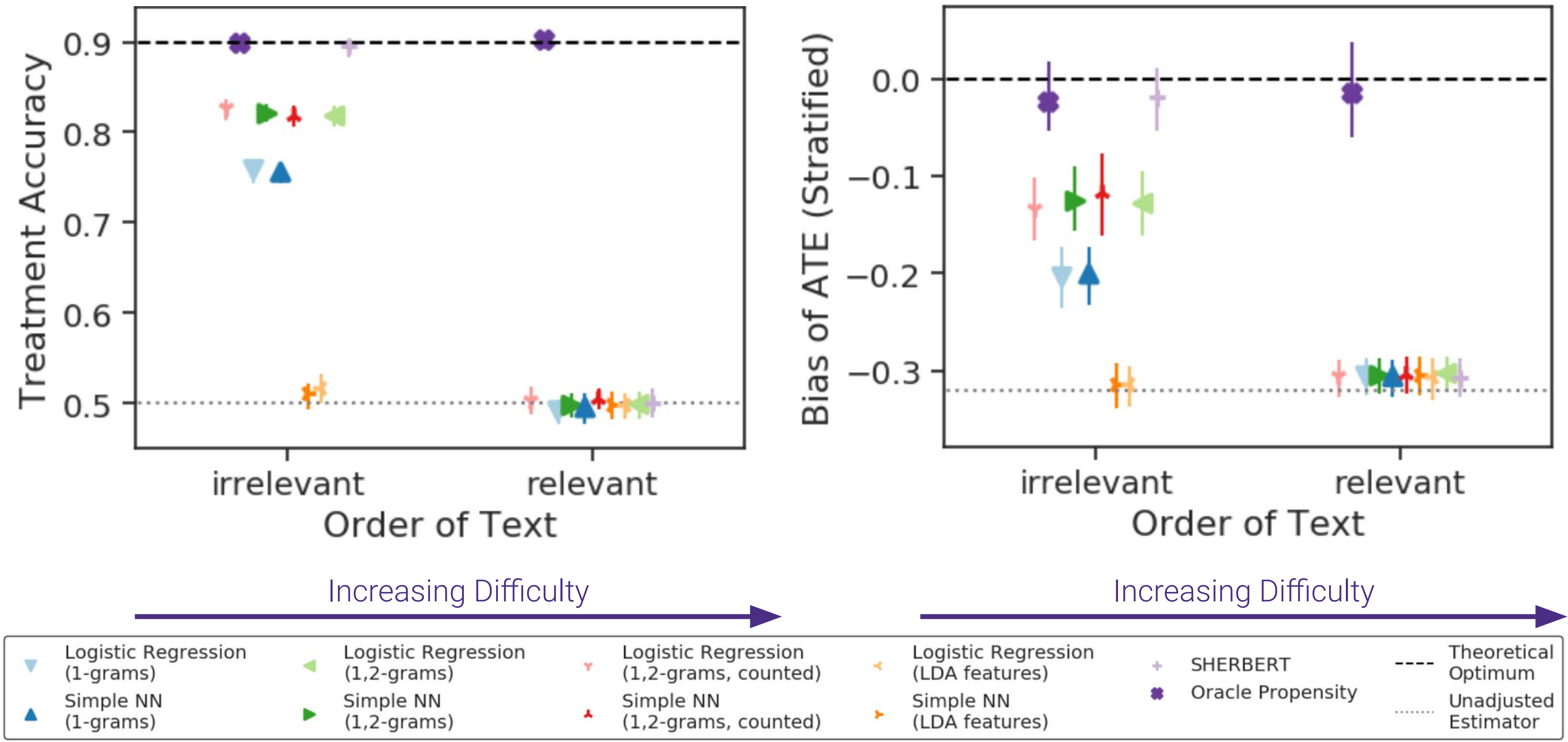
Increasing Difficulty



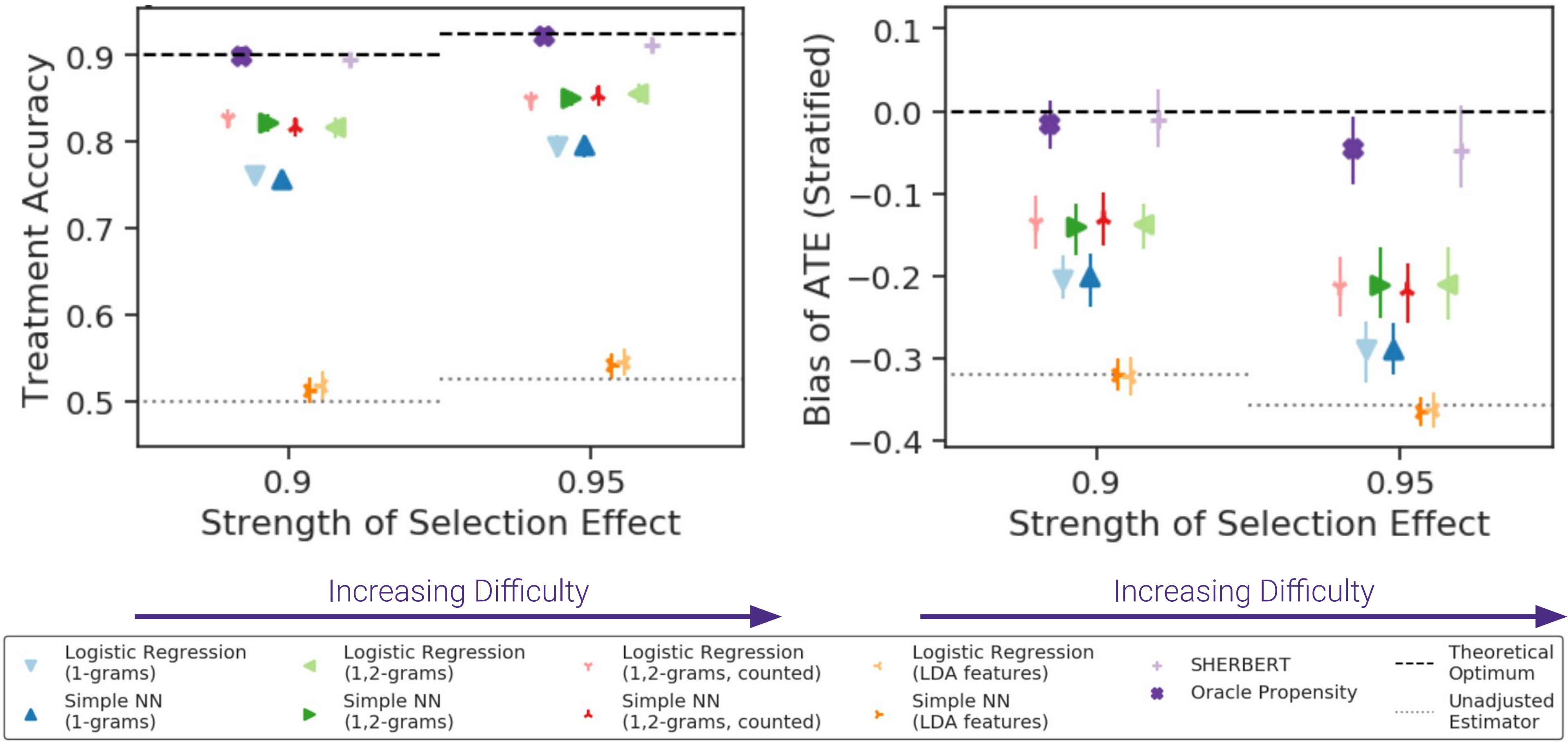
Results: Signal Intensity



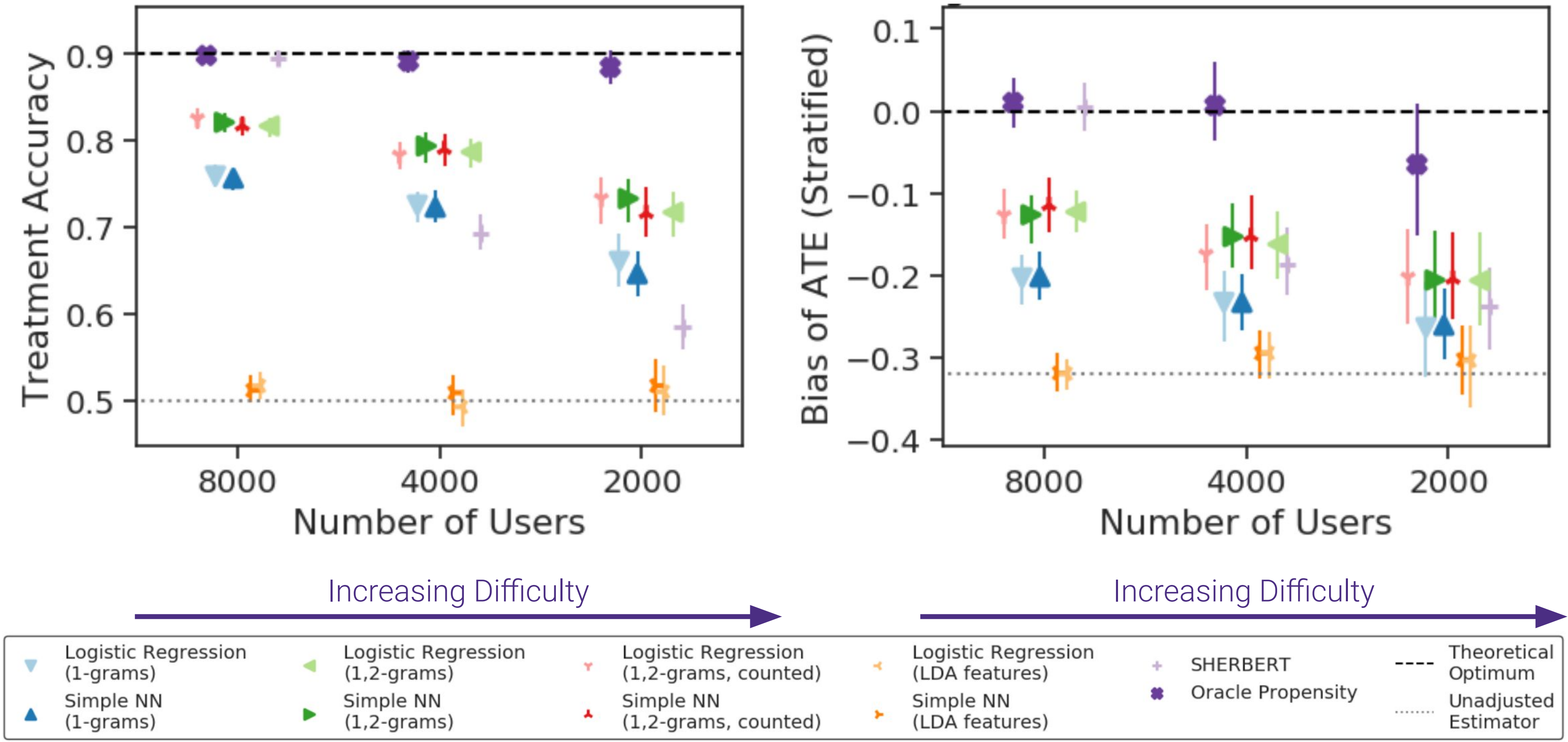
Results: Order of Text



Results: Strength of Selection Effect



Results: Number of Users



Results: Absence of (Non-Zero) Treatment Effect