Countering Bias in AI Methods for Text Data

Nicolas Audinet de Pieuchon

6th of February 2026

Bias in AI Methods for Text Data

Bias in AI Methods for Text Data

Many types of bias: statistical bias, model bias, representational bias, social biases of all kinds

Bias in AI Methods for Text Data

Many types of bias: statistical bias, model bias, representational bias, social biases of all kinds

Many sources of bias: training data, data pre-processing, model architecture, learning algorithm

Bias in AI Methods for Text Data

Many types of bias: statistical bias, model bias, representational bias, social biases of all kinds

Many sources of bias: training data, data pre-processing, model architecture, learning algorithm

Bias in AI methods leads to **real harm!**

Bias in AI Methods for Text Data

Many types of bias: statistical bias, model bias, representational bias, social biases of all kinds

Many sources of bias: training data, data pre-processing, model architecture, learning algorithm

Bias in AI methods leads to **real harm!**

My focus: removing unwanted information from text representations

Bias in AI Methods for Text Data

Many types of bias: statistical bias, model bias, representational bias, social biases of all kinds

Many sources of bias: training data, data pre-processing, model architecture, learning algorithm

Bias in AI methods leads to **real harm!**

My focus: removing unwanted information from text representations

Why?

1) To protect people
2) Applications to causal inference
3) To see if we can!

Text Rewriting

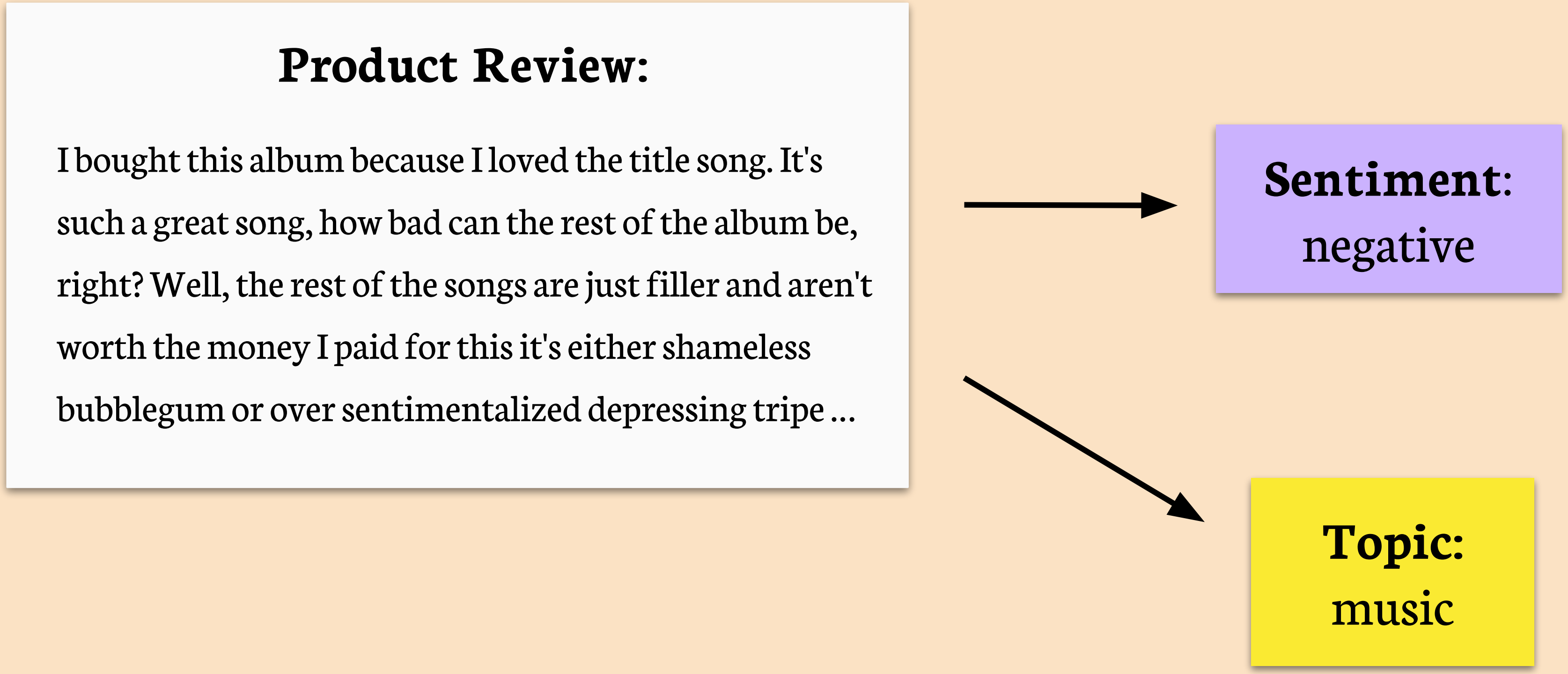
Product Review:

I bought this album because I loved the title song. It's such a great song, how bad can the rest of the album be, right? Well, the rest of the songs are just filler and aren't worth the money I paid for this it's either shameless bubblegum or over sentimentalized depressing tripe ...

Product Review:

I bought this album because I loved the title song. It's such a great song, how bad can the rest of the album be, right? Well, the rest of the songs are just filler and aren't worth the money I paid for this it's either shameless bubblegum or over sentimentalized depressing tripe ...

Sentiment:
negative



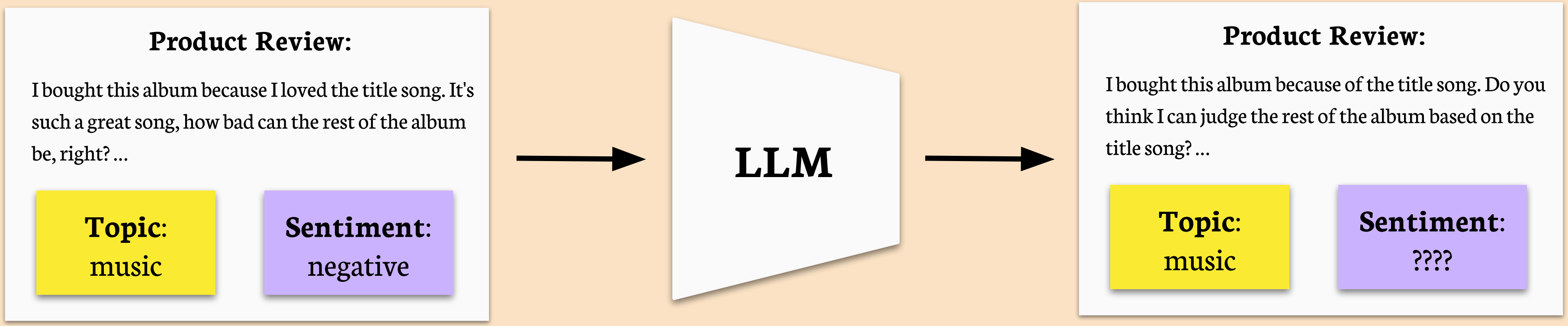
Product Review:

I bought this album because I loved the title song. It's such a great song, how bad can the rest of the album be, right? Well, the rest of the songs are just filler and aren't worth the money I paid for this it's either shameless bubblegum or over sentimentalized depressing tripe ...

Sentiment: negative

Topic: music

Can we use an LLM to rewrite the review to make it neutral?



Product Review:

I bought this album because I loved the title song. It's such a great song, how bad can the rest of the album be, right? ...

Topic: music

Sentiment: negative

LLM

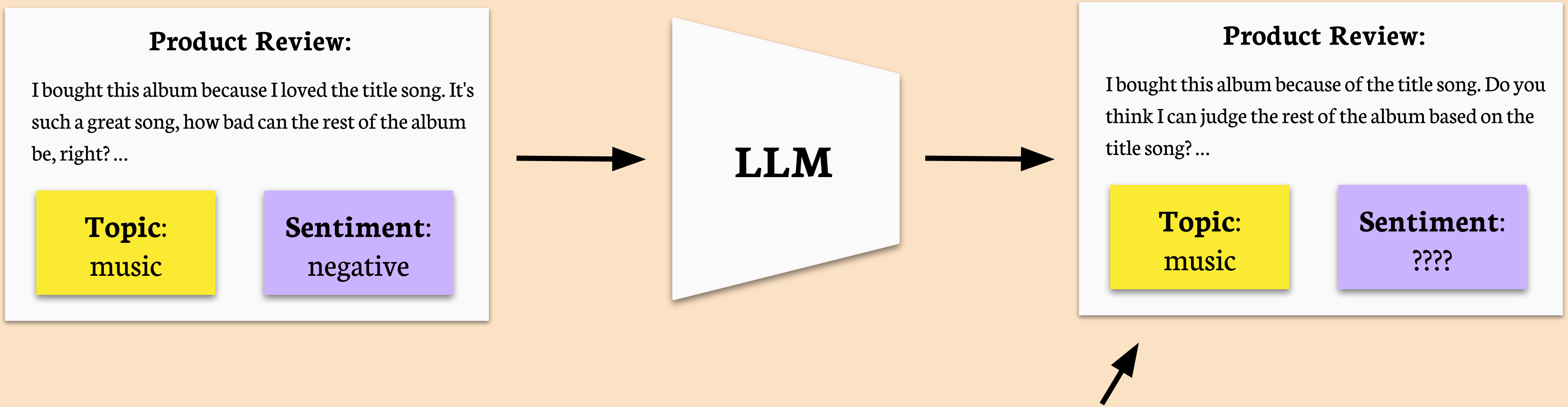
Product Review:

I bought this album because of the title song. Do you think I can judge the rest of the album based on the title song? ...

Topic: music

Sentiment: ????

Can we use an LLM to rewrite the review to make it neutral?



Product Review:

I bought this album because I loved the title song. It's such a great song, how bad can the rest of the album be, right? ...

Topic: music

Sentiment: negative

Product Review:

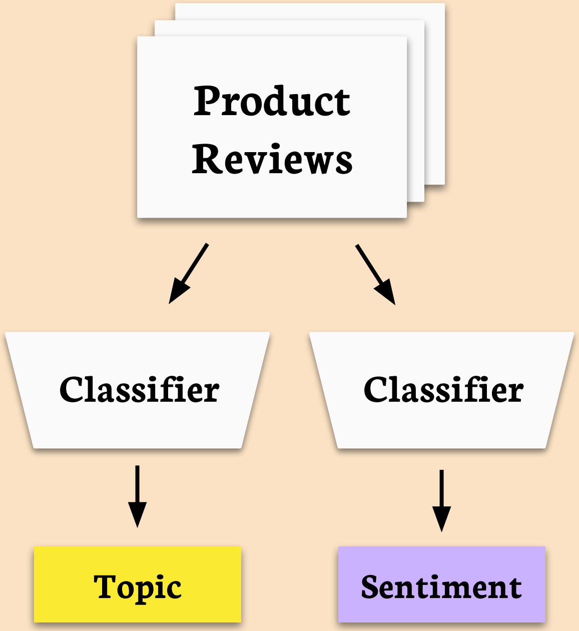
I bought this album because of the title song. Do you think I can judge the rest of the album based on the title song? ...

Topic: music

Sentiment: ????

Minimal intrusion: change the text as little as possible

Product Reviews



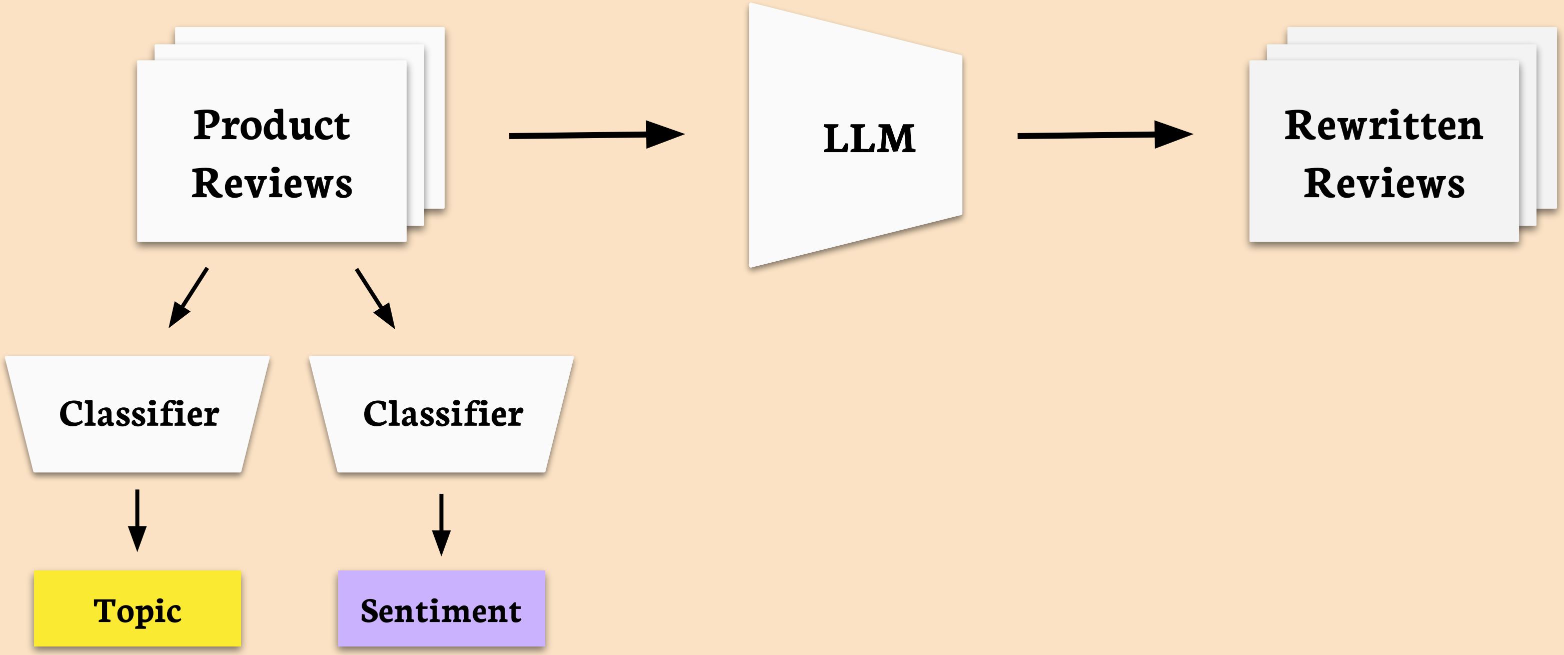
Product Reviews

Classifier

Classifier

Topic

Sentiment



Product Reviews

LLM

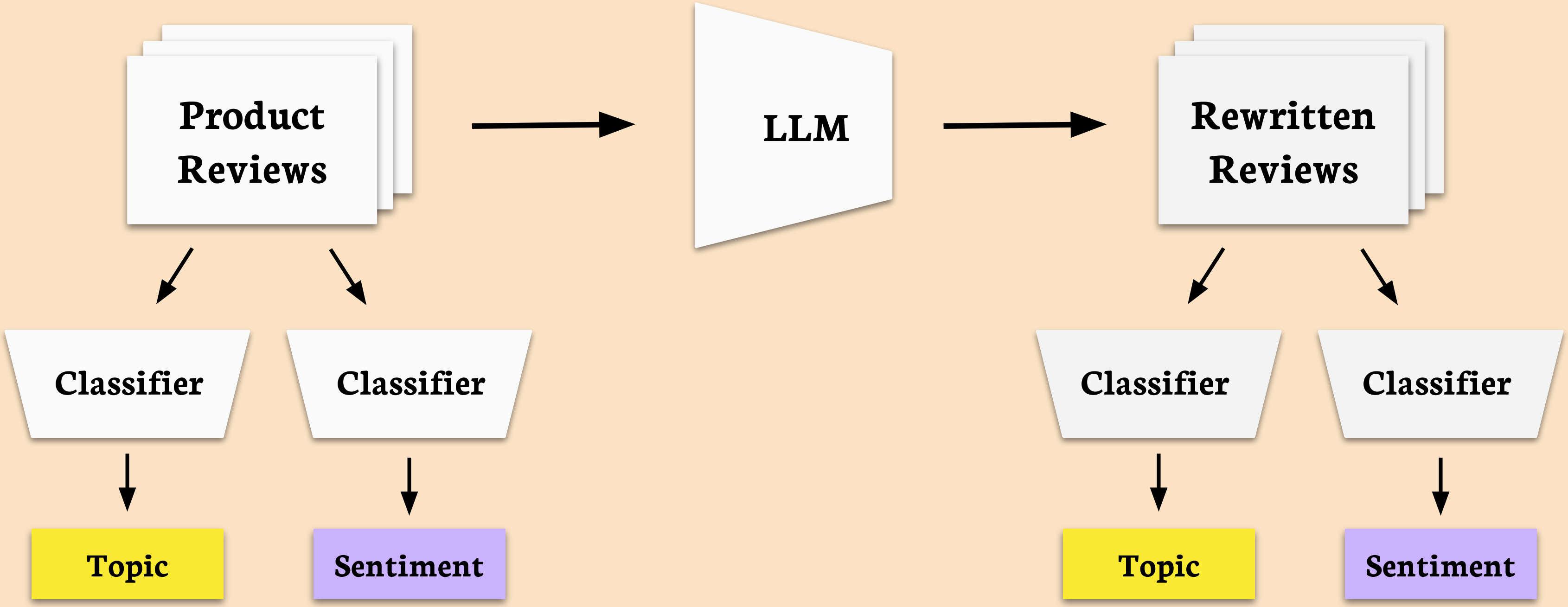
Rewritten Reviews

Classifier

Classifier

Topic

Sentiment



Product Reviews

LLM

Rewritten Reviews

Classifier

Classifier

Classifier

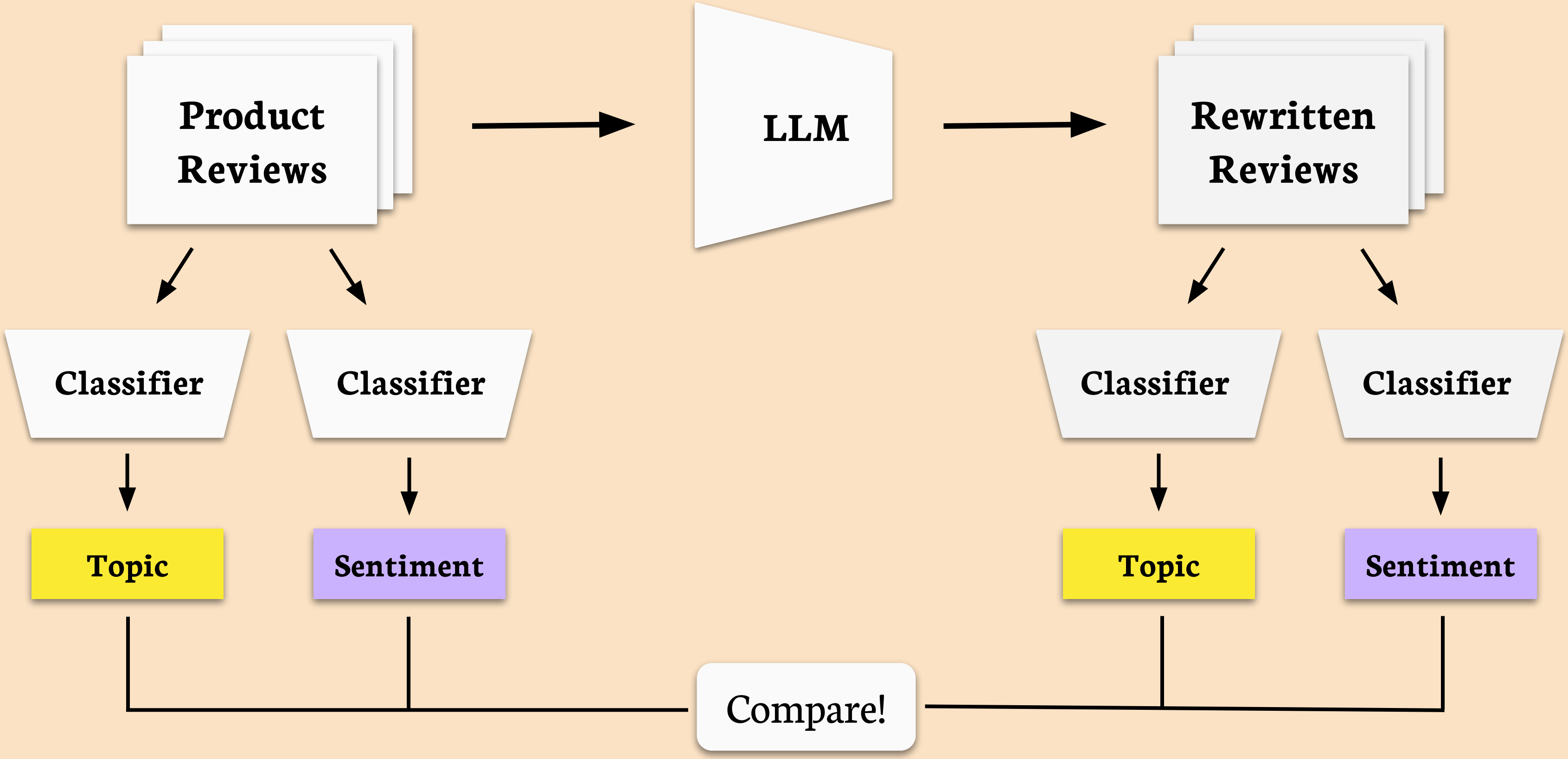
Classifier

Topic

Sentiment

Topic

Sentiment



Product Reviews

LLM

Rewritten Reviews

Classifier

Classifier

Classifier

Classifier

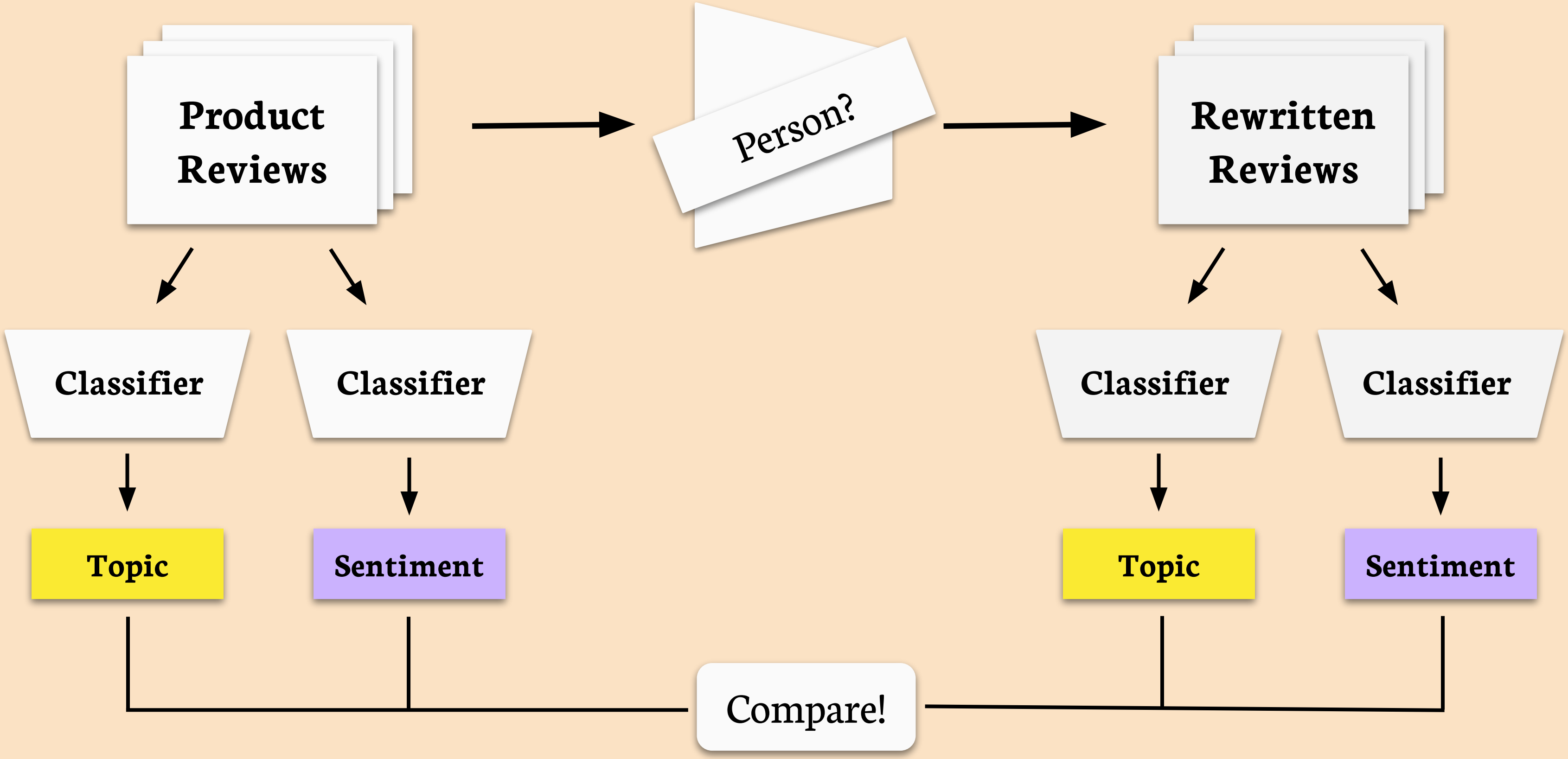
Topic

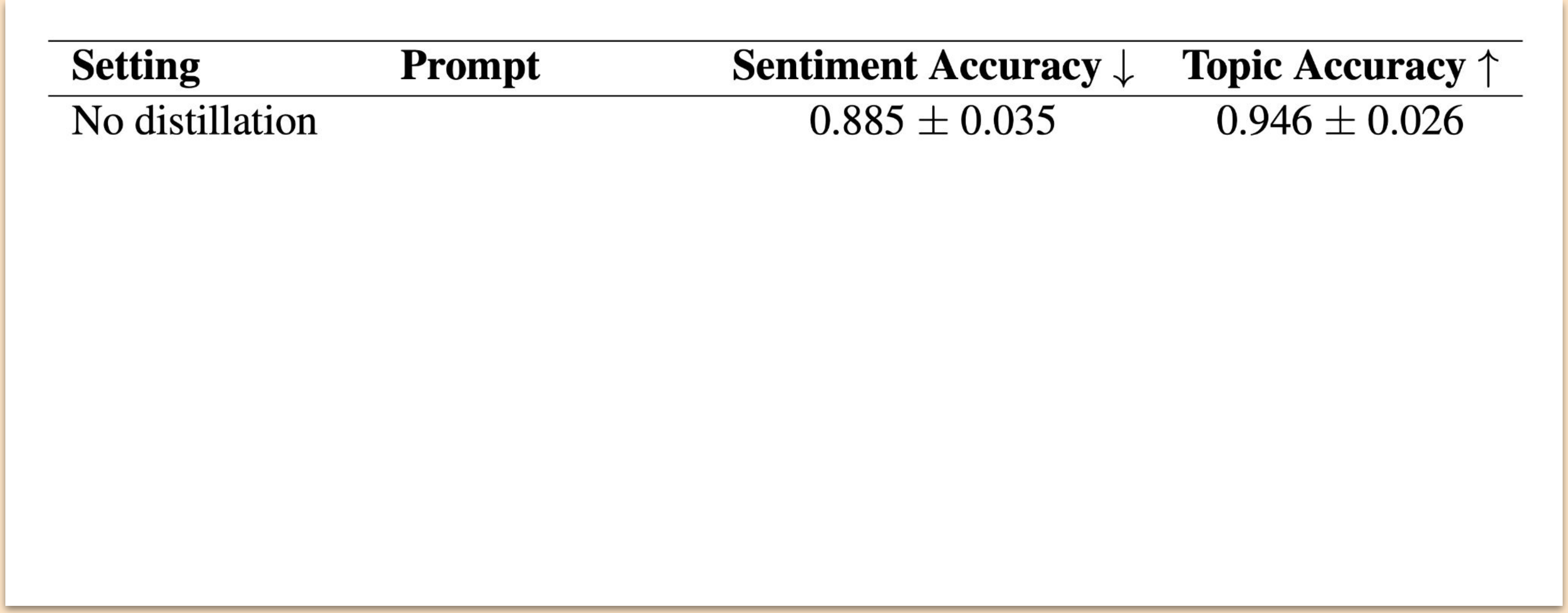
Sentiment

Topic

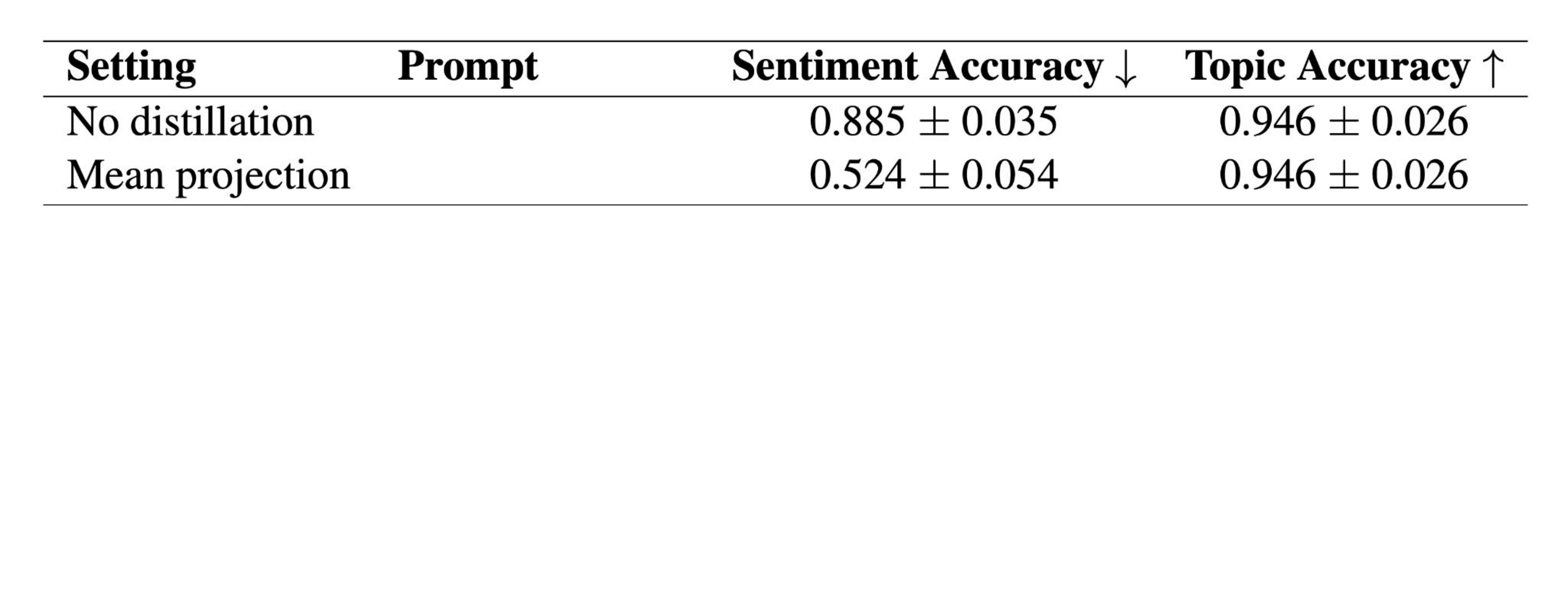
Sentiment

Compare!

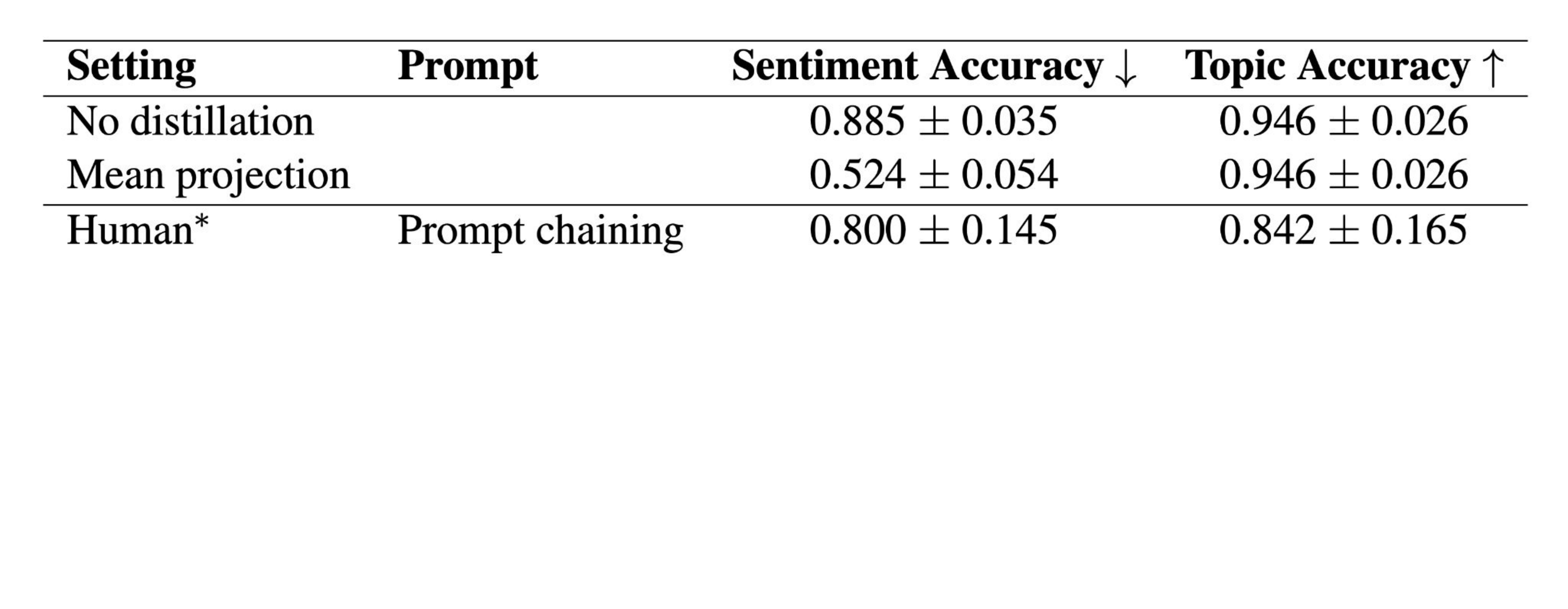

Product Reviews
Person?
Rewritten Reviews
Classifier
Classifier
Classifier
Classifier
Topic
Sentiment
Topic
Sentiment
Compare!



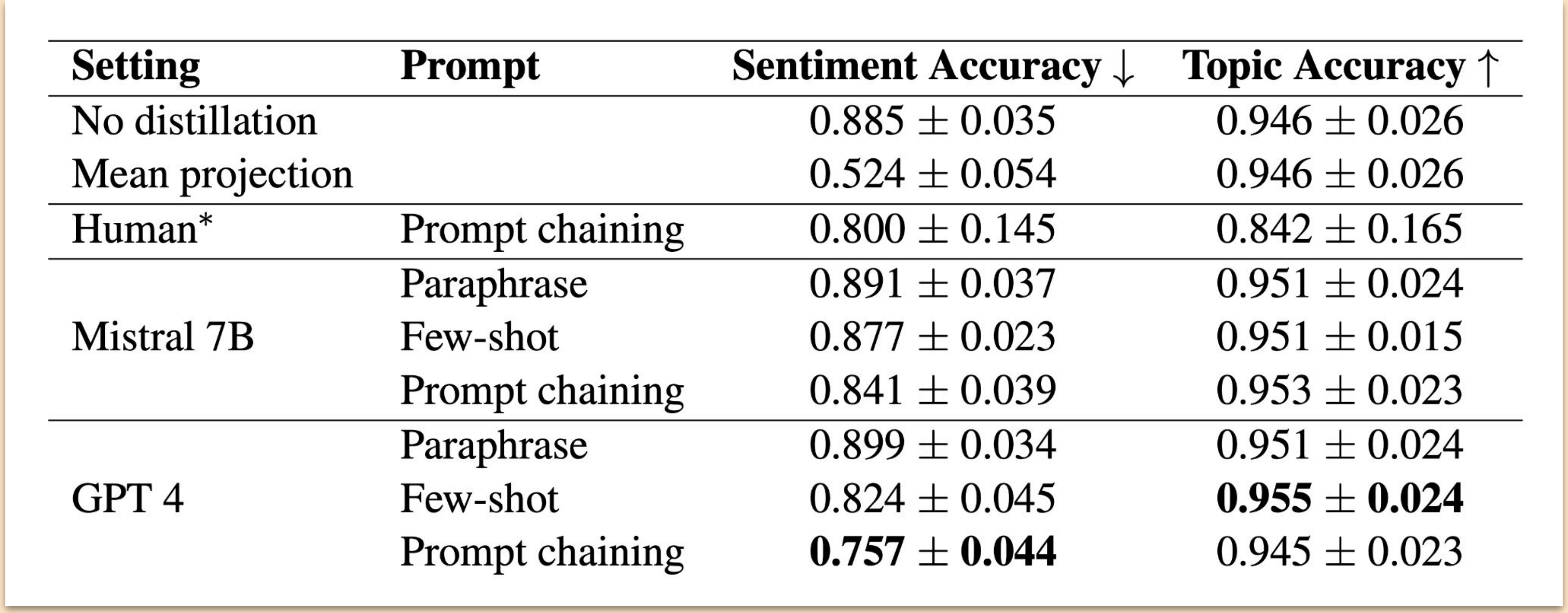
Setting	Prompt	Sentiment Accuracy ↓	Topic Accuracy ↑
No distillation		0.885 ± 0.035	0.946 ± 0.026



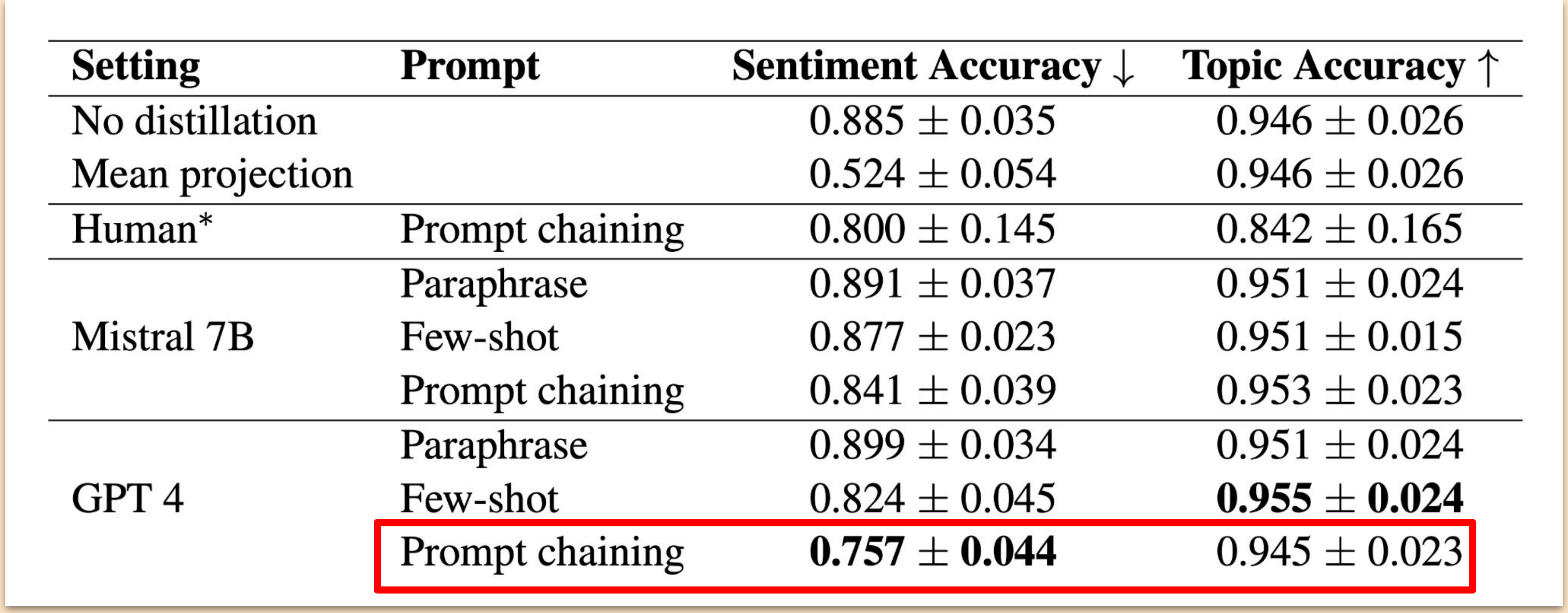
Setting	**Prompt**	**Sentiment Accuracy** ↓	**Topic Accuracy** ↑
No distillation		0.885 ± 0.035	0.946 ± 0.026
Mean projection		0.524 ± 0.054	0.946 ± 0.026



Setting	**Prompt**	**Sentiment Accuracy ↓**	**Topic Accuracy ↑**
No distillation		0.885 ± 0.035	0.946 ± 0.026
Mean projection		0.524 ± 0.054	0.946 ± 0.026
Human*	Prompt chaining	0.800 ± 0.145	0.842 ± 0.165



Setting	Prompt	Sentiment Accuracy ↓	Topic Accuracy ↑
No distillation		0.885 ± 0.035	0.946 ± 0.026
Mean projection		0.524 ± 0.054	0.946 ± 0.026
Human*	Prompt chaining	0.800 ± 0.145	0.842 ± 0.165
Mistral 7B	Paraphrase	0.891 ± 0.037	0.951 ± 0.024
	Few-shot	0.877 ± 0.023	0.951 ± 0.015
	Prompt chaining	0.841 ± 0.039	0.953 ± 0.023
GPT 4	Paraphrase	0.899 ± 0.034	0.951 ± 0.024
	Few-shot	0.824 ± 0.045	**0.955 ± 0.024**
	Prompt chaining	**0.757 ± 0.044**	0.945 ± 0.023



Setting	Prompt	Sentiment Accuracy ↓	Topic Accuracy ↑
No distillation		0.885 ± 0.035	0.946 ± 0.026
Mean projection		0.524 ± 0.054	0.946 ± 0.026
Human*	Prompt chaining	0.800 ± 0.145	0.842 ± 0.165
Mistral 7B	Paraphrase	0.891 ± 0.037	0.951 ± 0.024
	Few-shot	0.877 ± 0.023	0.951 ± 0.015
	Prompt chaining	0.841 ± 0.039	0.953 ± 0.023
GPT 4	Paraphrase	0.899 ± 0.034	0.951 ± 0.024
	Few-shot	0.824 ± 0.045	**0.955** ± **0.024**
	Prompt chaining	**0.757** ± **0.044**	0.945 ± 0.023

Can LLMs Disentangle Text?

1.

Some success, but nowhere near baselines

2.

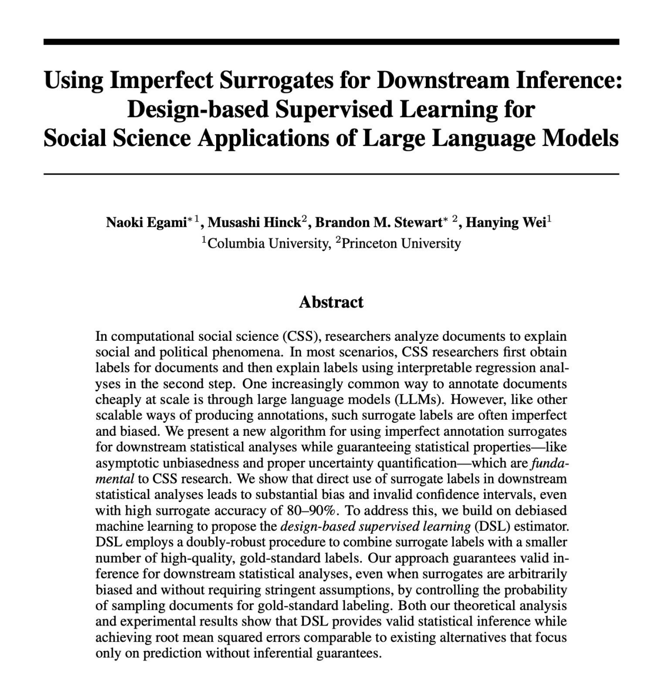
Difficult task for both humans and LLMs

3.

Performance might be task-dependent

Benchmarking Debiasing Methods

Advances in post-hoc debiasing methods …



Using Imperfect Surrogates for Downstream Inference: Design-based Supervised Learning for Social Science Applications of Large Language Models

Naoki Egami*[1], Musashi Hinck[2], Brandon M. Stewart*[2], Hanying Wei[1]
[1]Columbia University, [2]Princeton University

Abstract

In computational social science (CSS), researchers analyze documents to explain social and political phenomena. In most scenarios, CSS researchers first obtain labels for documents and then explain labels using interpretable regression analyses in the second step. One increasingly common way to annotate documents cheaply at scale is through large language models (LLMs). However, like other scalable ways of producing annotations, such surrogate labels are often imperfect and biased. We present a new algorithm for using imperfect annotation surrogates for downstream statistical analyses while guaranteeing statistical properties—like asymptotic unbiasedness and proper uncertainty quantification—which are *fundamental* to CSS research. We show that direct use of surrogate labels in downstream statistical analyses leads to substantial bias and invalid confidence intervals, even with high surrogate accuracy of 80–90%. To address this, we build on debiased machine learning to propose the *design-based supervised learning* (DSL) estimator. DSL employs a doubly-robust procedure to combine surrogate labels with a smaller number of high-quality, gold-standard labels. Our approach guarantees valid inference for downstream statistical analyses, even when surrogates are arbitrarily biased and without requiring stringent assumptions, by controlling the probability of sampling documents for gold-standard labeling. Both our theoretical analysis and experimental results show that DSL provides valid statistical inference while achieving root mean squared errors comparable to existing alternatives that focus only on prediction without inferential guarantees.

RESEARCH

MACHINE LEARNING

Prediction-powered inference

Anastasios N. Angelopoulos*†, Stephen Bates*†, Clara Fannjiang*†, Michael I. Jordan*†, Tijana Zrnic*†

Prediction-powered inference is a framework for performing valid statistical inference when an experimental dataset is supplemented with predictions from a machine-learning system. The framework yields simple algorithms for computing provably valid confidence intervals for quantities such as means, quantiles, and linear and logistic regression coefficients without making any assumptions about the machine-learning algorithm that supplies the predictions. Furthermore, more accurate predictions translate to smaller confidence intervals. Prediction-powered inference could enable researchers to draw valid and more data-efficient conclusions using machine learning. The benefits of prediction-powered inference were demonstrated with datasets from proteomics, astronomy, genomics, remote sensing, census analysis, and ecology.

Imagine a scientist has a machine-learning system that can supply accurate predictions about a phenomenon far more cheaply than any gold-standard experimental technique. The scientist may wish to use these predictions as evidence in drawing scientific conclusions. For example, accurate predictions of three-dimensional structures have been made for a vast catalog of known protein sequences (*1*, *2*) and are now being used in proteomics studies (*3*, *4*). Such machine-learning systems are increasingly common in modern scientific inquiry, in domains ranging from cancer prognosis to microclimate modeling. Predictions are not perfect, however, and this may lead to incorrect conclusions. Moreover, as predictions beget other predictions, the cumulative effect can amplify the imperfections. How can modern science leverage machine-learning predictions in a statistically principled way?

One way to use predictions is to follow the imputation approach: Proceed as if they are gold-standard measurements. Although this lets the scientist draw conclusions cheaply and quickly owing to the high-throughput nature of the machine-learning system, the conclusions may be invalid because the predictions may have biases.

Another possibility is to apply the classical approach: Ignore the machine-learning predictions and only use the available gold-standard measurements, which are typically far less abun-

validity of the resulting conclusions. Prediction-powered inference provides a protocol for combining predictions, which are abundant but not always trustworthy, with gold-standard data, which are trusted but scarce, to compute confidence intervals and *P* values. The resulting confidence intervals and *P* values are statistically valid, as in the classical approach, but also leverage the information contained in the predictions, as in the imputation approach, to make the confidence intervals smaller and the *P* values more powerful.

Prediction-powered inference applies to any machine-learning system; as such, it absolves the need for case-by-case analyses dependent on the machine-learning algorithm on hand. The proposed protocol thereby could enable researchers to report on and assess the evidence for their conclusions in a fully standardized way.

Protocol for prediction-powered inference

The protocol for prediction-powered inference proceeds as follows. The scientist wishes to construct a confidence interval for a quantity θ^*, such as the mean outcome or a regression coefficient quantifying the statistical association between the outcome and a feature. Toward this goal, they have access to a small gold-standard dataset of features paired with outcomes, $(X, Y) = ((X_1, Y_1), ..., (X_n, Y_n))$, as

Prediction-powered inference uses the gold-standard dataset to quantify and correct for the errors made by the machine-learning algorithm on the unlabeled dataset, thereby enabling researchers to reliably incorporate predictions when constructing confidence intervals. The three-step protocol is outlined below and visualized in Fig. 1.

1) Estimand. The first step is to select an estimand θ^*. The estimand is the quantity the scientist is interested in knowing—for example, the mean outcome $E[Y_i]$, median outcome $\mathrm{median}(Y_i)$, a linear regression coefficient obtained by regressing Y onto X, etc.

2) Measure of fit and rectifier. The key step is to identify the right measure of fit m_θ and rectifier Δ_θ for the selected estimand. For every candidate value of the estimand θ, the measure of fit m_θ is computed on the unlabeled dataset imputed with predictions, $(X', \hat{Y}')$ and quantifies how likely θ^* is to be equal to θ on the basis of the imputed data. The closer m_θ is to zero, the more plausible it is for θ^* to be equal to θ.

The rectifier Δ_θ is a notion of prediction error that is relevant for the estimand of interest. It is defined as the difference of the measure of fit m_θ computed on the labeled data, (X, Y), and the labeled data when the true outcomes are replaced with predicted ones, $(X, \hat{Y})$. If the predictions are perfect, the rectifier is equal to zero.

Table 1 states the appropriate measure of fit and rectifier for common estimands of interest: the mean outcome, median outcome, q-quantile of the outcome, and linear and logistic regression coefficients when regressing Y onto X. A general recipe for deriving the right measure of fit and corresponding rectifier for a broad class of other estimands is provided in the SM.

3) Prediction-powered confidence interval. Finally, the measure of fit and rectifier are carefully combined to form a prediction-powered confidence interval for θ^*. This process is called rectifying the confidence interval. The prediction-powered confidence interval is constructed as $C^{\mathrm{PP}} = \{\theta \text{ such that } |m_\theta + \Delta_\theta| \leq w_\theta(\alpha)\}$ and is guaranteed to contain the esti-

X1	X2	...	Y
1	T	...	By now, we all know that House Intelligence Committee ...
5	F	...	Donald Trump won't tell you this on his Twitter feed. The ...
2	F	...	His accusation that President Obama wiretapped him has ...
...	...	...	...
4	T	...	The rainbow flag has become a symbol for LGBT rights ...

X1	X2	...	Y
1	T	...	By now, we all know that House Intelligence Committee ...
5	F	...	Donald Trump won't tell you this on his Twitter feed. The ...
2	F	...	His accusation that President Obama wiretapped him has ...
...	...	...	...
4	T	...	The rainbow flag has become a symbol for LGBT rights ...

Annotate by hand

X1	X2	...	Y
1	T	...	By now, we all know that House Intelligence Committee ...
5	F	...	Donald Trump won't tell you this on his Twitter feed. The ...
2	F	...	His accusation that President Obama wiretapped him has ...
...	...	...	...
4	T	...	The rainbow flag has become a symbol for LGBT rights ...

Annotate by hand

Generate annotations with an LLM

X1	X2	…	Y
1	T	…	By now, we all know that House Intelligence Committee …
5	F	…	Donald Trump won't tell you this on his Twitter feed. The …
2	F	…	His accusation that President Obama wiretapped him has …
…	…	…	…
4	T	…	The rainbow flag has become a symbol for LGBT rights …

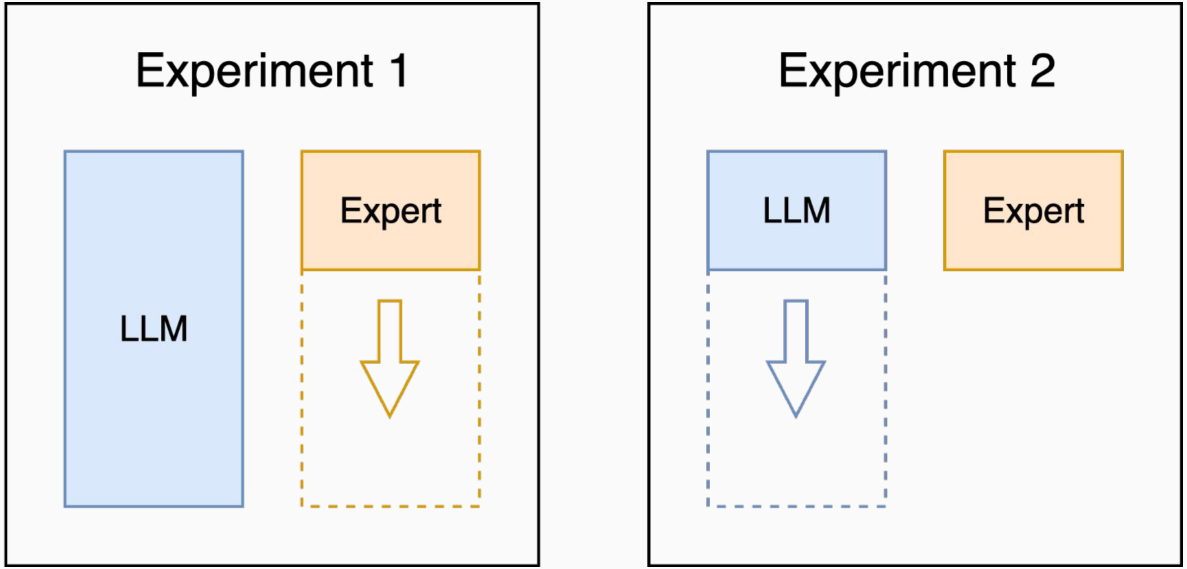
Debiasing methods let you use both!

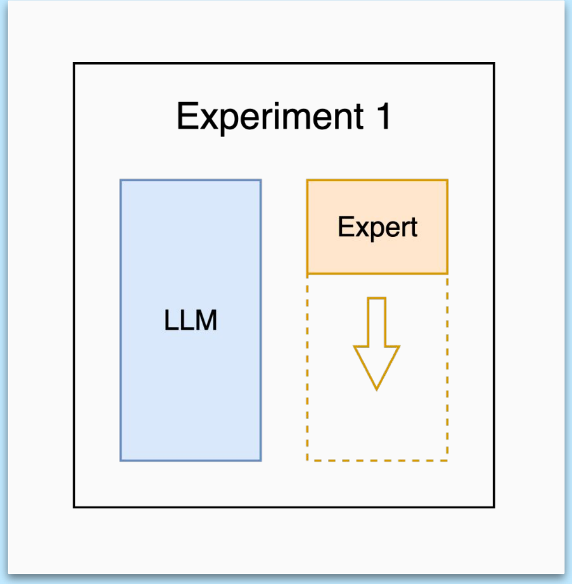
Annotate by hand

Generate annotations with an LLM

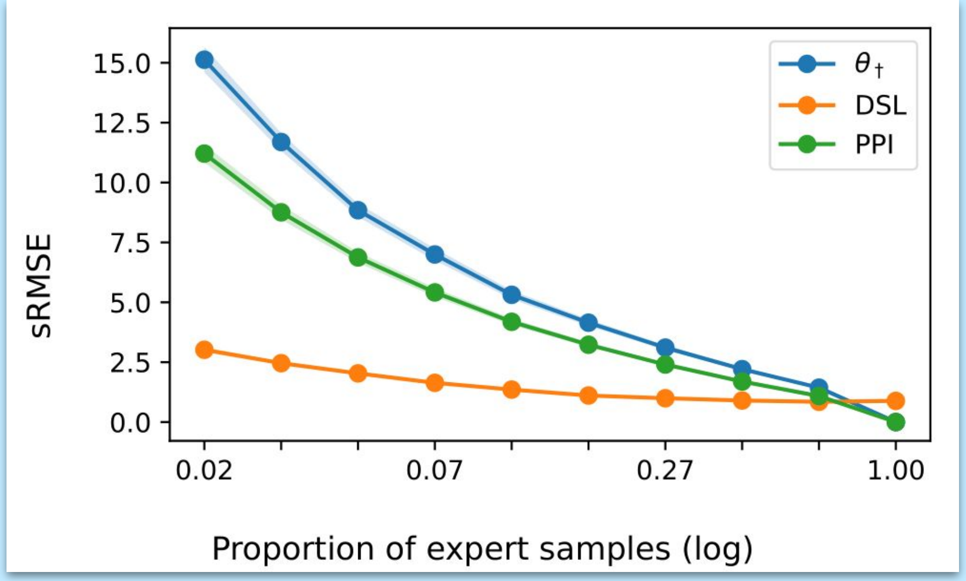
RQ1: When is it preferable to use debiasing methods over just the expert annotations?

RQ2: What are the performance differences between debiasing methods?

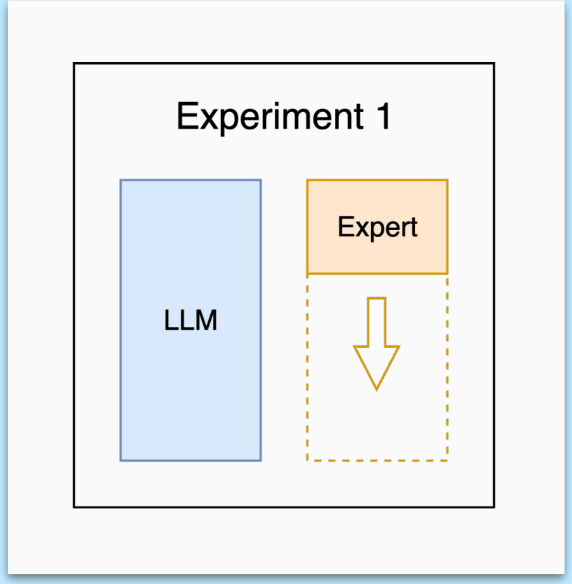

Experiment 1
LLM
Expert
Experiment 2
LLM
Expert

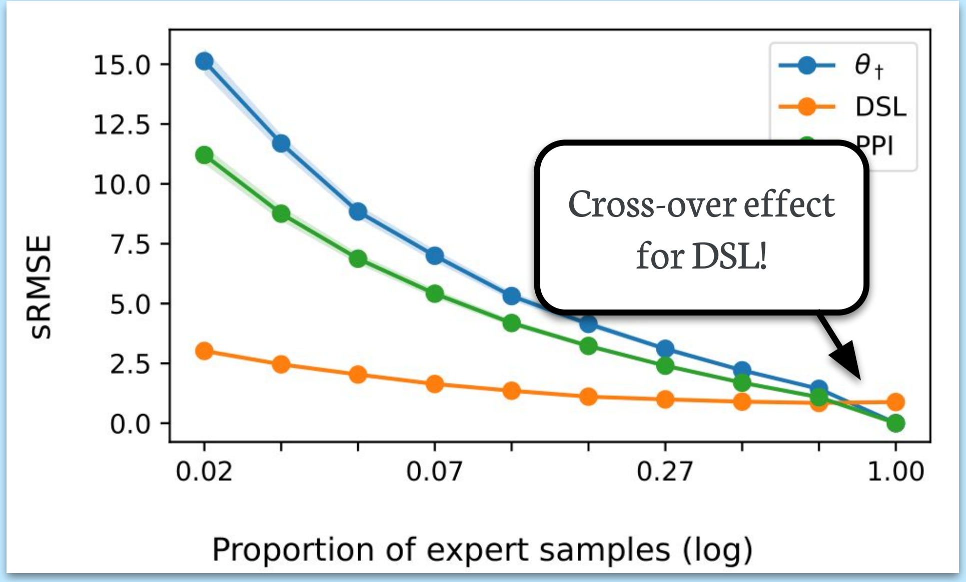


θ†: only expert annotations



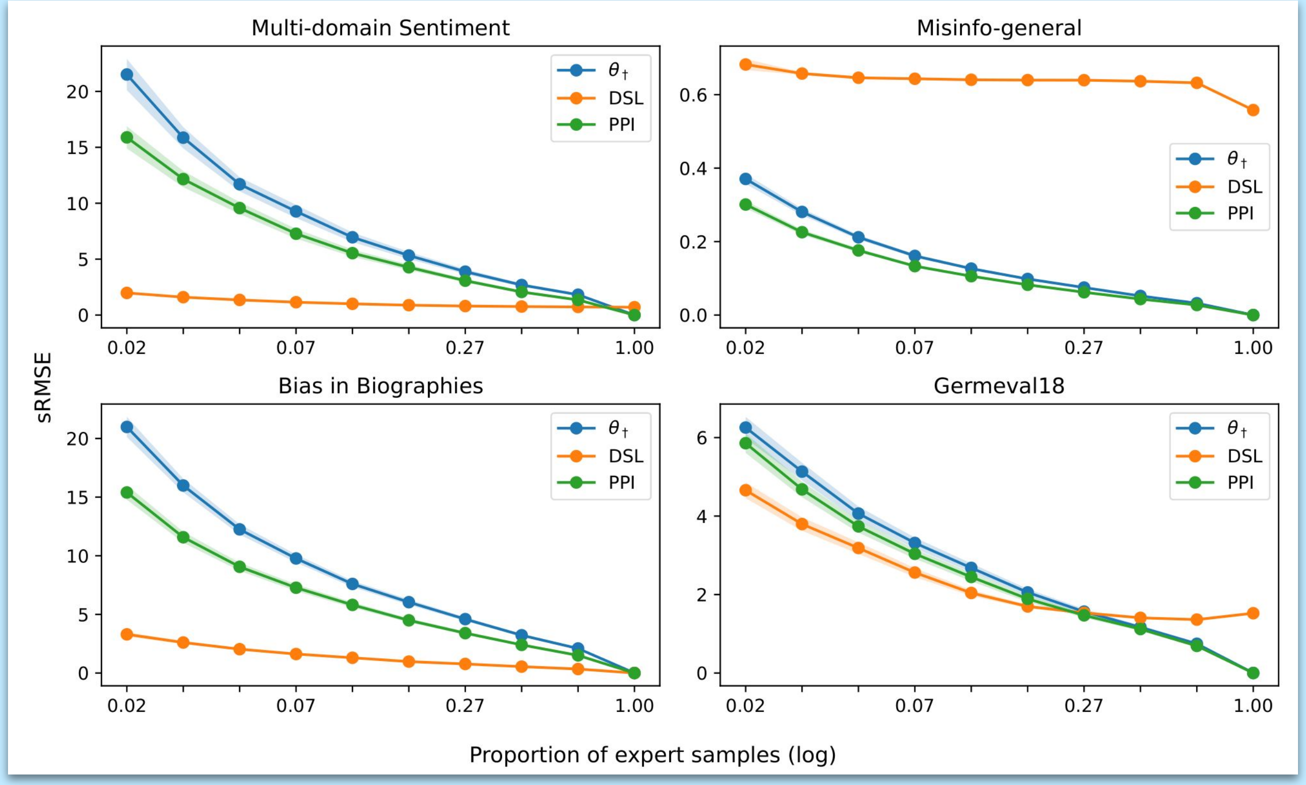
$\theta_\dagger$: only expert annotations

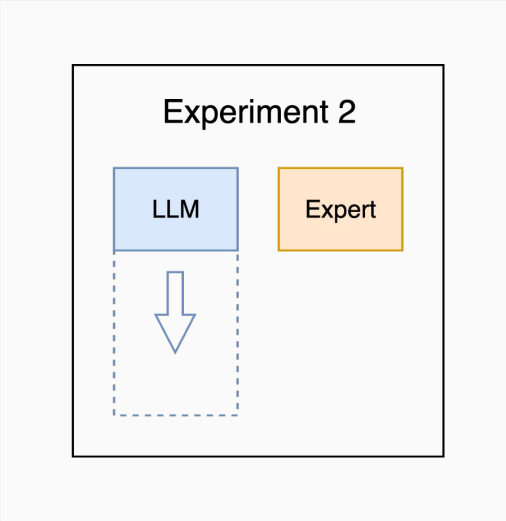




Cross-over effect for DSL!

$\theta_\dagger$: only expert annotations



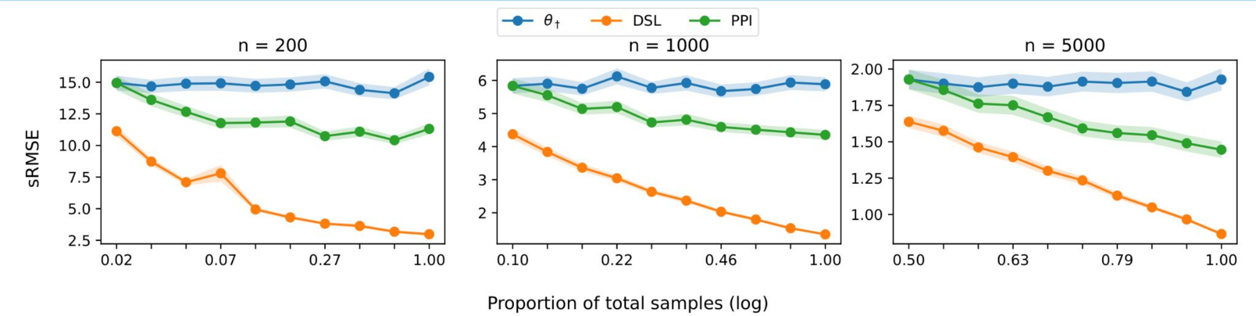


Experiment 2

LLM

Expert

$\theta_\dagger$: only expert annotations



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

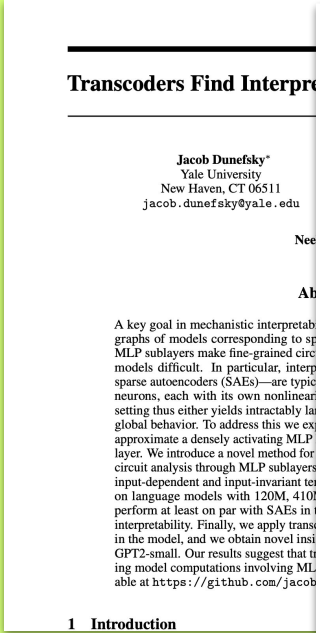
Both DSL or PPI is more efficient than using only expert annotations.

RQ2: What are the performance differences between debiasing methods?

DSL tends to outperform PPI, but performance is more dataset-dependant.

Representation Blinding

Advances in mechanistic interpretability …



Transcoders Find Interpre

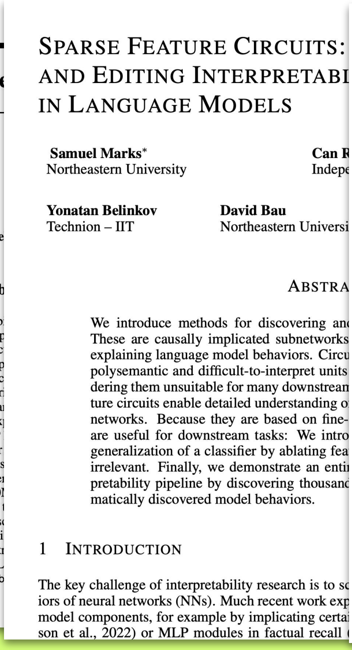
Jacob Dunefsky*
Yale University
New Haven, CT 06511
jacob.dunefsky@yale.edu

Nee

Ab

A key goal in mechanistic interpretab
graphs of models corresponding to sp
MLP sublayers make fine-grained circ
models difficult. In particular, interp
sparse autoencoders (SAEs)—are typic
neurons, each with its own nonlinear
setting thus either yields intractably la
global behavior. To address this we ex
approximate a densely activating MLP
layer. We introduce a novel method for
circuit analysis through MLP sublayers
input-dependent and input-invariant te
on language models with 120M, 410M
perform at least on par with SAEs in t
interpretability. Finally, we apply trans
in the model, and we obtain novel insi
GPT2-small. Our results suggest that tr
ing model computations involving ML
able at https://github.com/jacob

1 Introduction



SPARSE FEATURE CIRCUITS:
AND EDITING INTERPRETABI
IN LANGUAGE MODELS

Samuel Marks*
Northeastern University

Can R
Indepe

Yonatan Belinkov
Technion – IIT

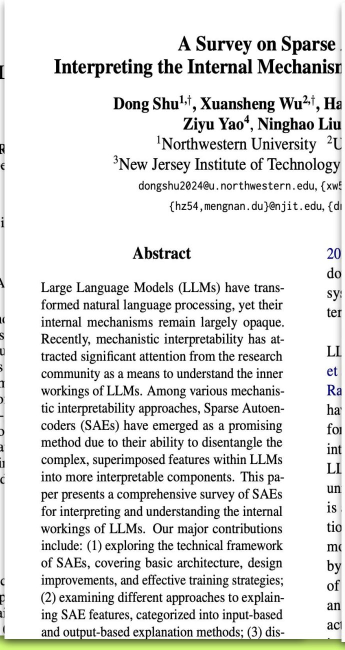
David Bau
Northeastern Universi

ABSTRA

We introduce methods for discovering and
These are causally implicated subnetworks
explaining language model behaviors. Circu
polysemantic and difficult-to-interpret units
dering them unsuitable for many downstream
ture circuits enable detailed understanding o
networks. Because they are based on fine-
are useful for downstream tasks: We intro
generalization of a classifier by ablating fea
irrelevant. Finally, we demonstrate an enti
pretability pipeline by discovering thousan
matically discovered model behaviors.

1 INTRODUCTION

The key challenge of interpretability research is to sc
iors of neural networks (NNs). Much recent work exp
model components, for example by implicating certa
son et al., 2022) or MLP modules in factual recall (



A Survey on Sparse
Interpreting the Internal Mechanisn

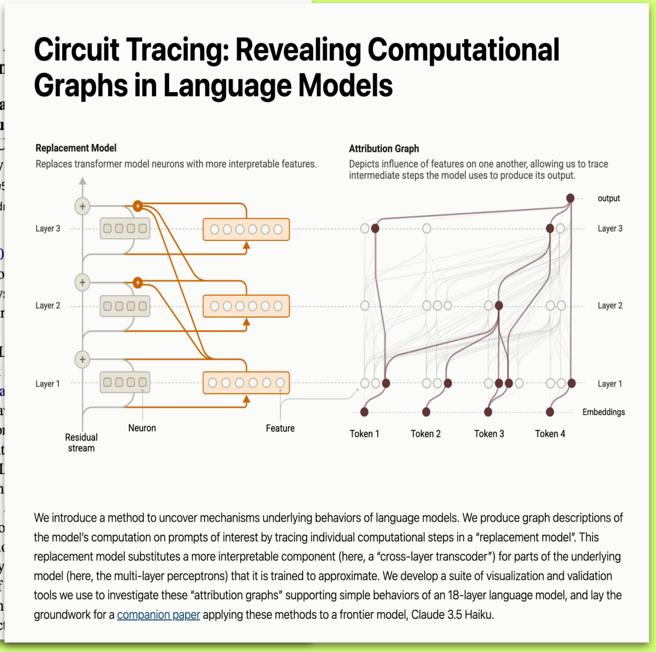
Dong Shu[1,†], Xuansheng Wu[2,†], Ha
Ziyu Yao[4], Ninghao Liu
[1]Northwestern University [2]U
[3]New Jersey Institute of Technology
dongshu2024@u.northwestern.edu, {xw5
{hz54,mengnan.du}@njit.edu, {d

Abstract

Large Language Models (LLMs) have transformed natural language processing, yet their internal mechanisms remain largely opaque. Recently, mechanistic interpretability has attracted significant attention from the research community as a means to understand the inner workings of LLMs. Among various mechanistic interpretability approaches, Sparse Autoencoders (SAEs) have emerged as a promising method due to their ability to disentangle the complex, superimposed features within LLMs into more interpretable components. This paper presents a comprehensive survey of SAEs for interpreting and understanding the internal workings of LLMs. Our major contributions include: (1) exploring the technical framework of SAEs, covering basic architecture, design improvements, and effective training strategies; (2) examining different approaches to explaining SAE features, categorized into input-based and output-based explanation methods; (3) dis-

20
do
sy
ter

LL
et
Ra
ha
fo
int
LL
un
is
tio
mo
by
of
an
act



Circuit Tracing: Revealing Computational Graphs in Language Models

Replacement Model
Replaces transformer model neurons with more interpretable features.

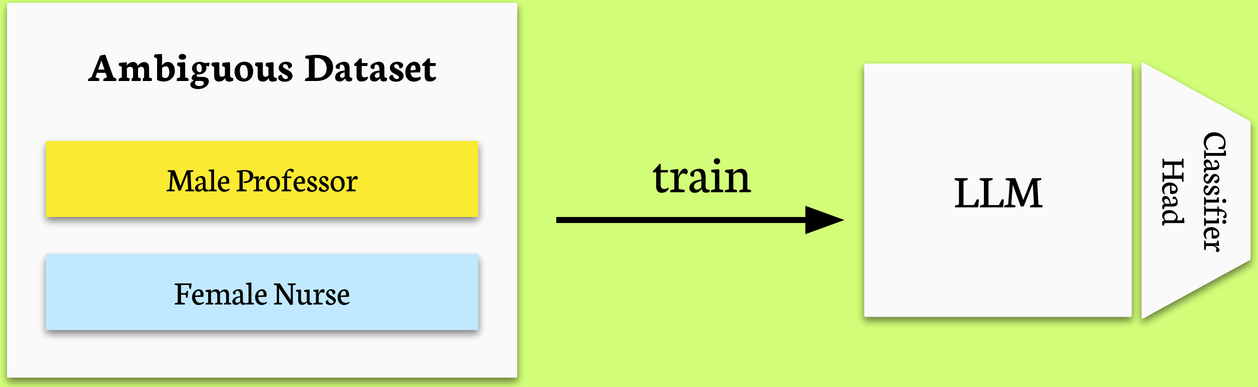
Attribution Graph
Depicts influence of features on one another, allowing us to trace intermediate steps the model uses to produce its output.

We introduce a method to uncover mechanisms underlying behaviors of language models. We produce graph descriptions of the model's computation on prompts of interest by tracing individual computational steps in a "replacement model". This replacement model substitutes a more interpretable component (here, a "cross-layer transcoder") for parts of the underlying model (here, the multi-layer perceptrons) that it is trained to approximate. We develop a suite of visualization and validation tools we use to investigate these "attribution graphs" supporting simple behaviors of an 18-layer language model, and lay the groundwork for a companion paper applying these methods to a frontier model, Claude 3.5 Haiku.

Ambiguous Dataset

Male Professor

Female Nurse



Ambiguous Dataset

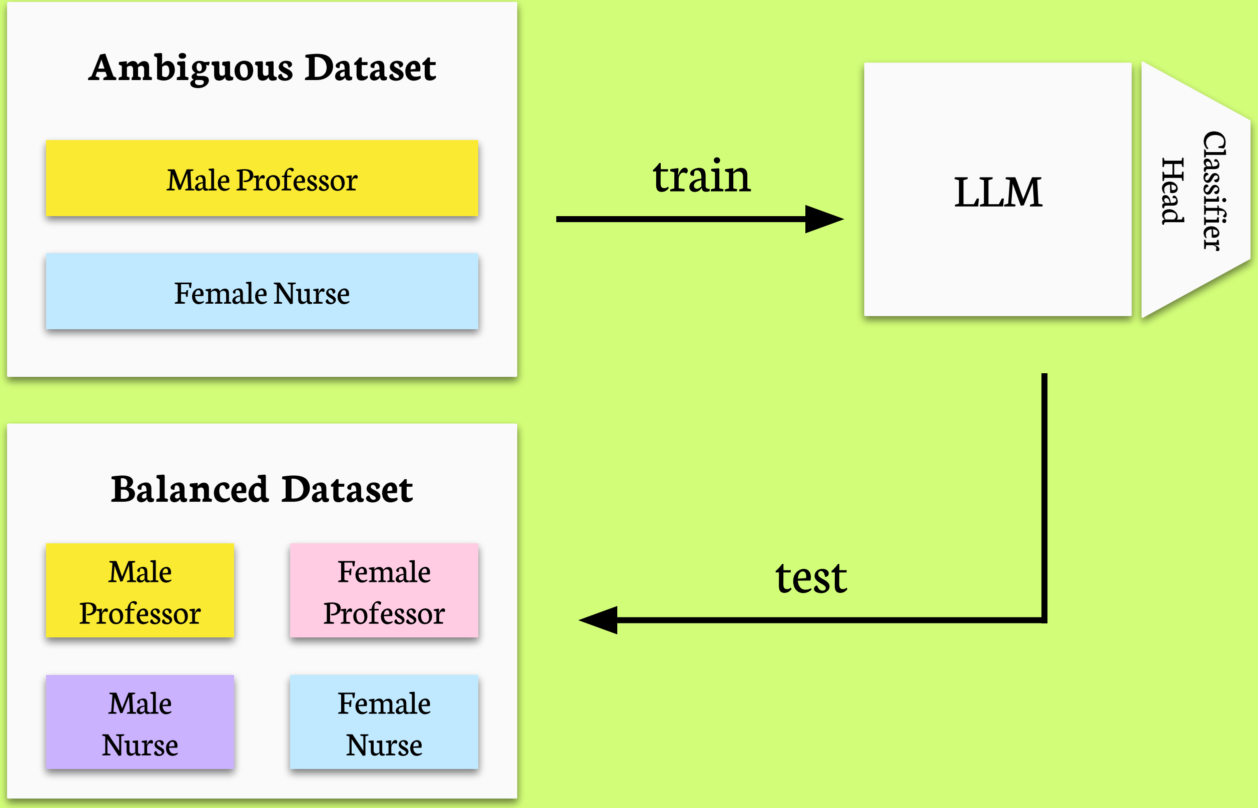
Male Professor

Female Nurse

train

LLM

Classifier Head



Ambiguous Dataset

Male Professor

Female Nurse

train

LLM

Classifier Head

test

Balanced Dataset

Male Professor

Female Professor

Male Nurse

Female Nurse

Can we remove male/female information from the text representation?

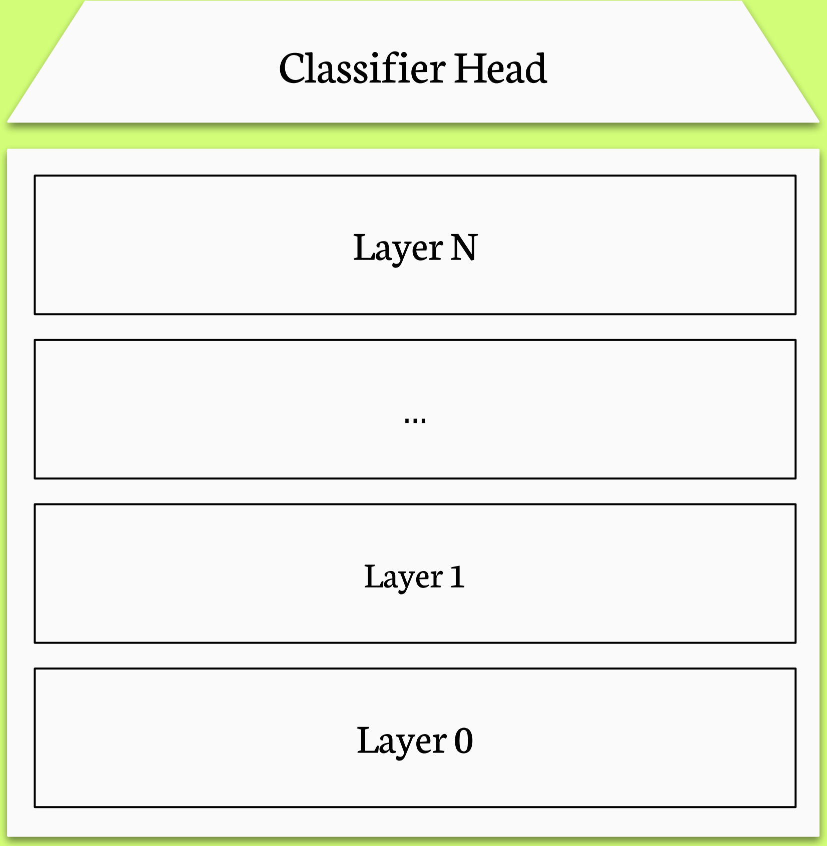
Can we remove male/female information from the text representation?

SHIFT method:

Can we remove male/female information from the text representation?

SHIFT method:

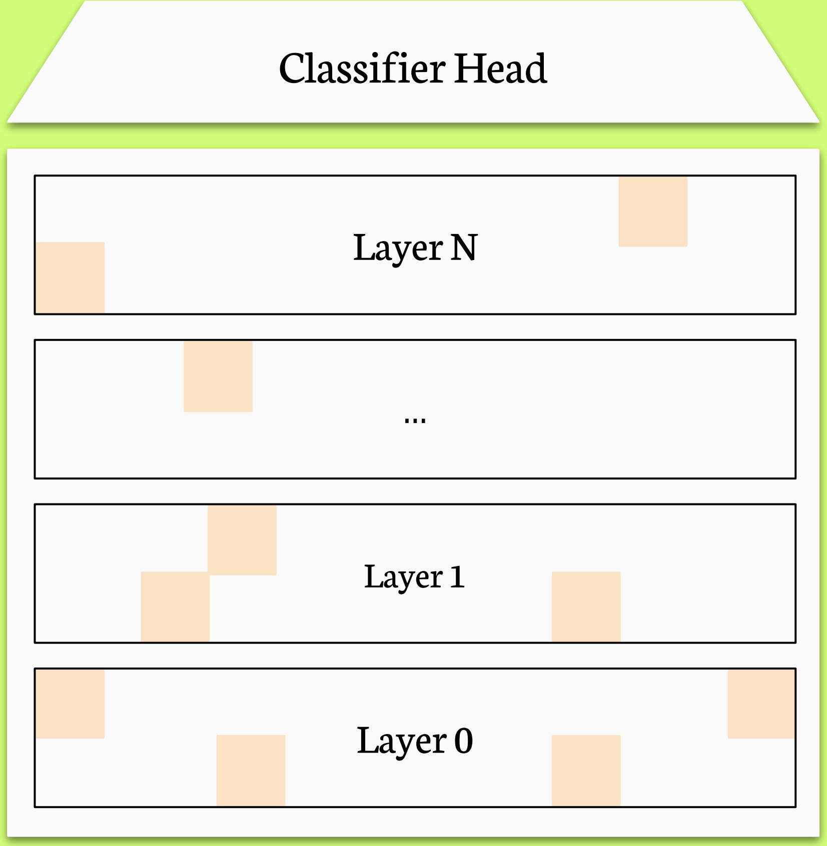
1) Train the classifier



Can we remove male/female information from the text representation?

SHIFT method:

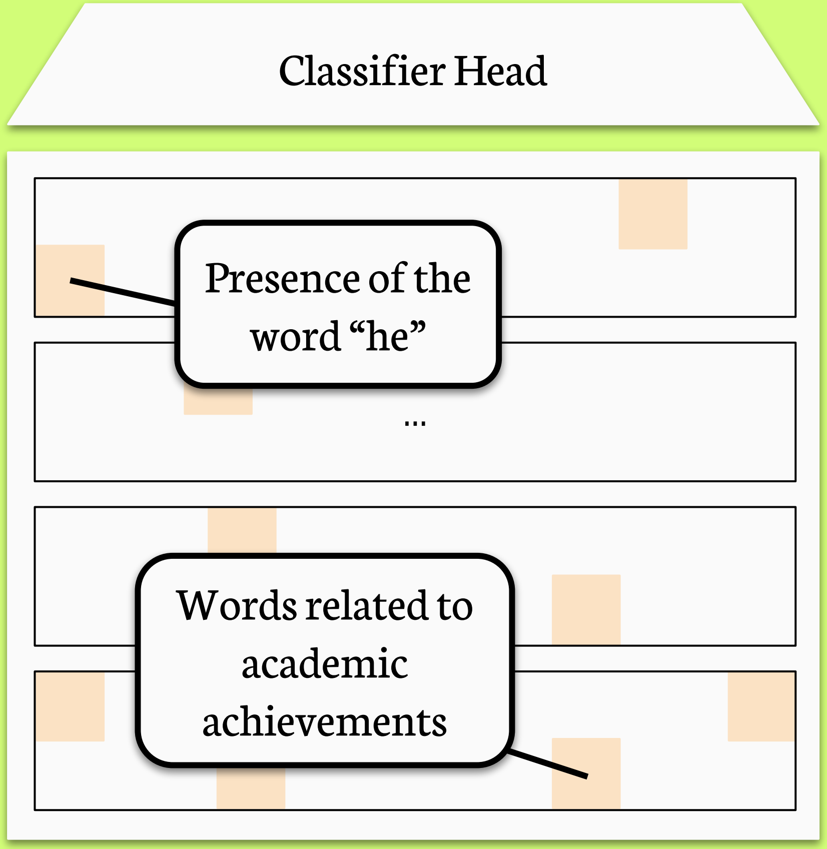
1) Train the classifier
2) Find relevant features



Can we remove male/female information from the text representation?

SHIFT method:

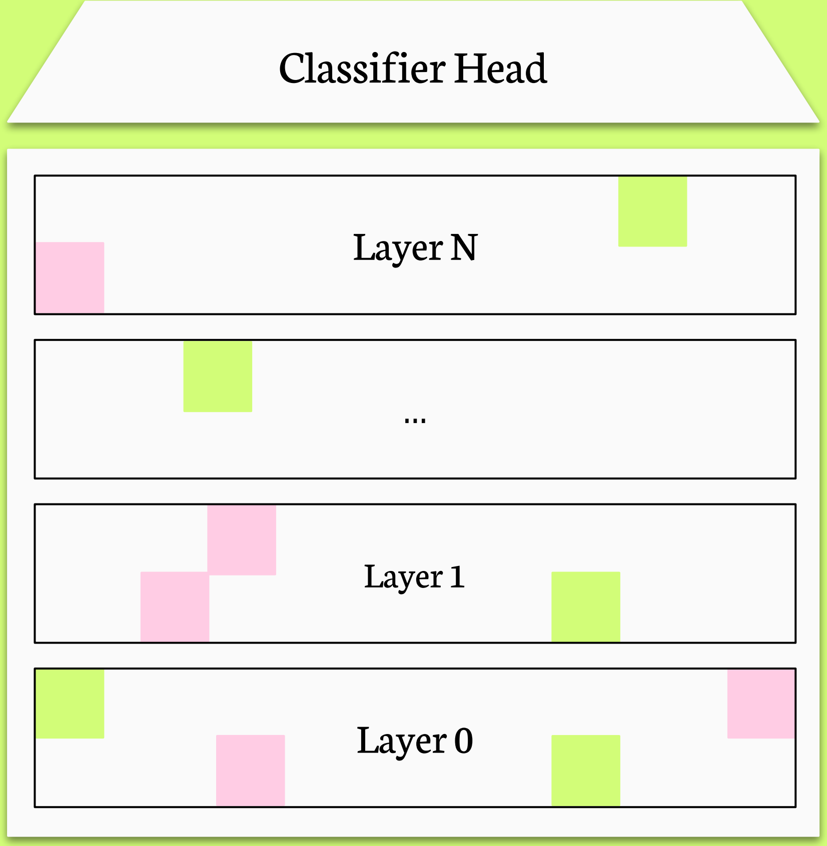
1) Train the classifier
2) Find relevant features
3) Interpret features


Classifier Head
Presence of the word “he”
...
Words related to academic achievements

Can we remove male/female information from the text representation?

SHIFT method:

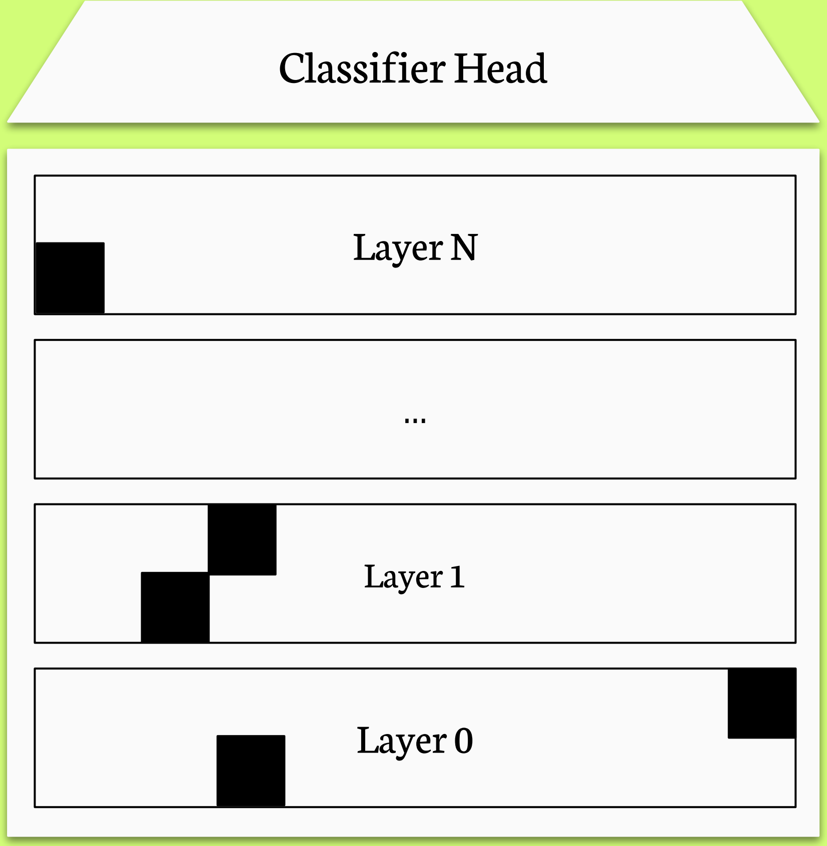
1) Train the classifier
2) Find relevant features
3) Interpret features
4) Select features

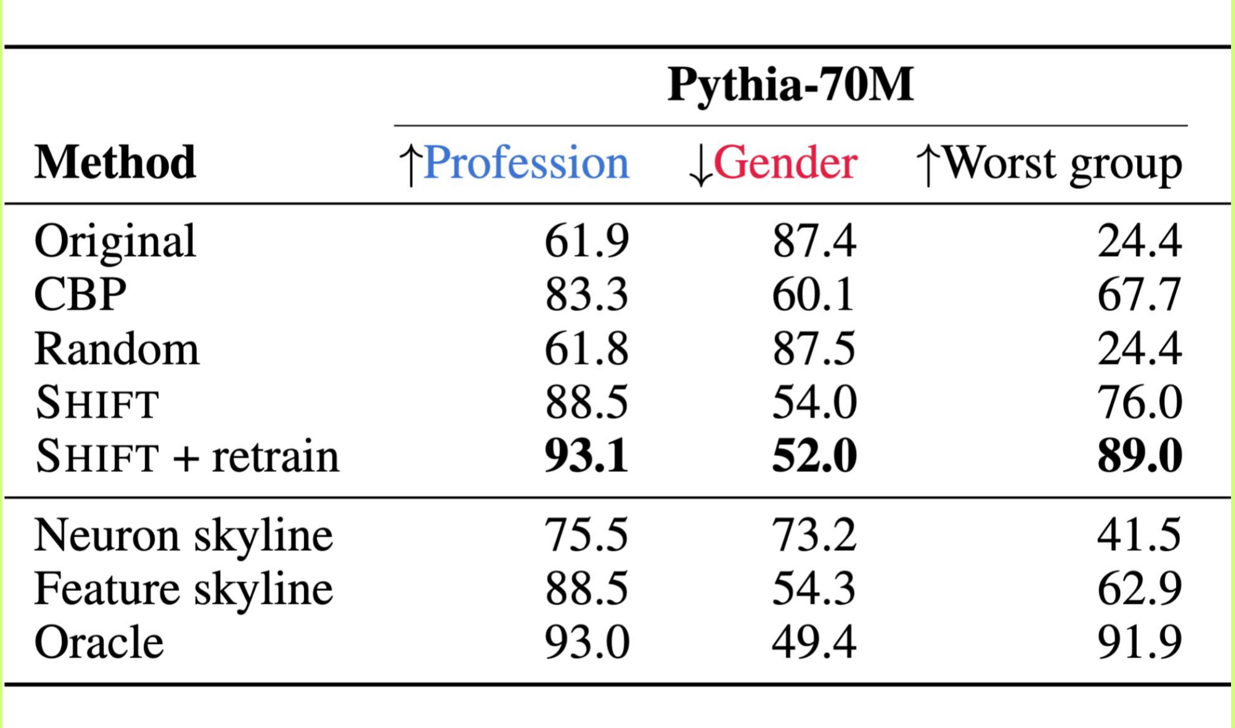


Can we remove male/female information from the text representation?

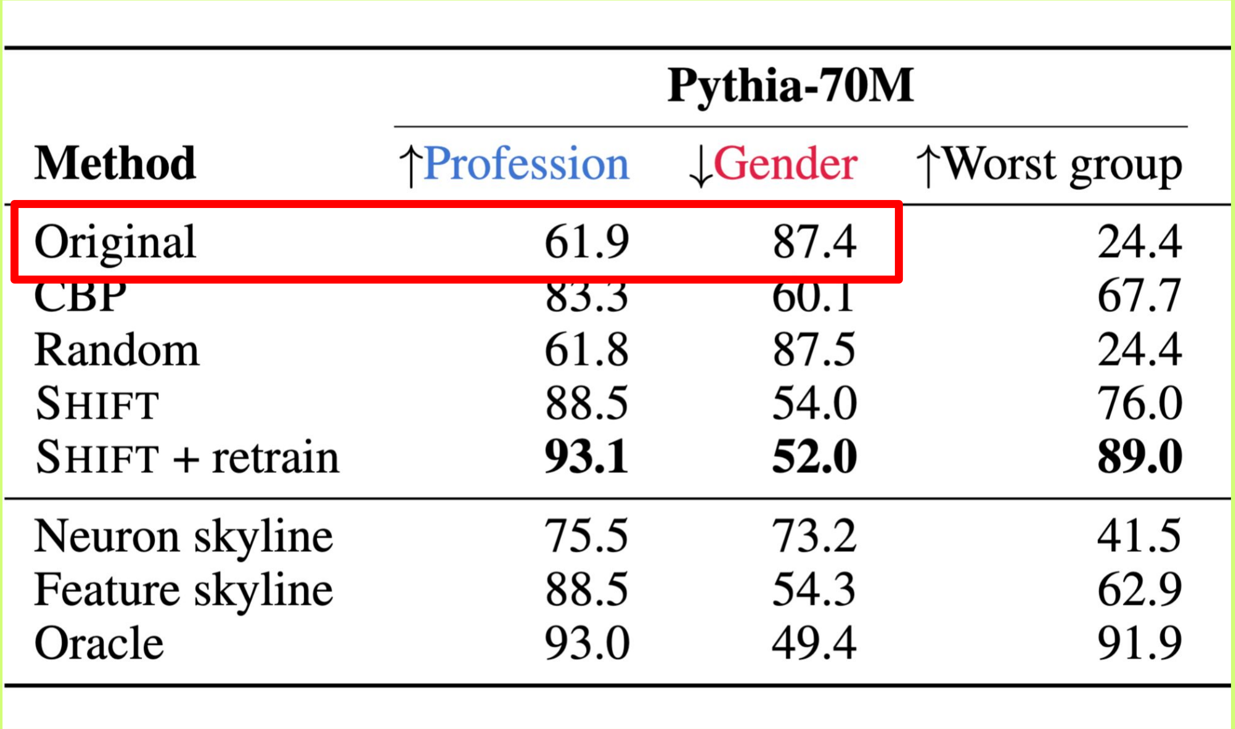
SHIFT method:

1) Train the classifier
2) Find relevant features
3) Interpret features
4) Select features
5) Remove and re-train

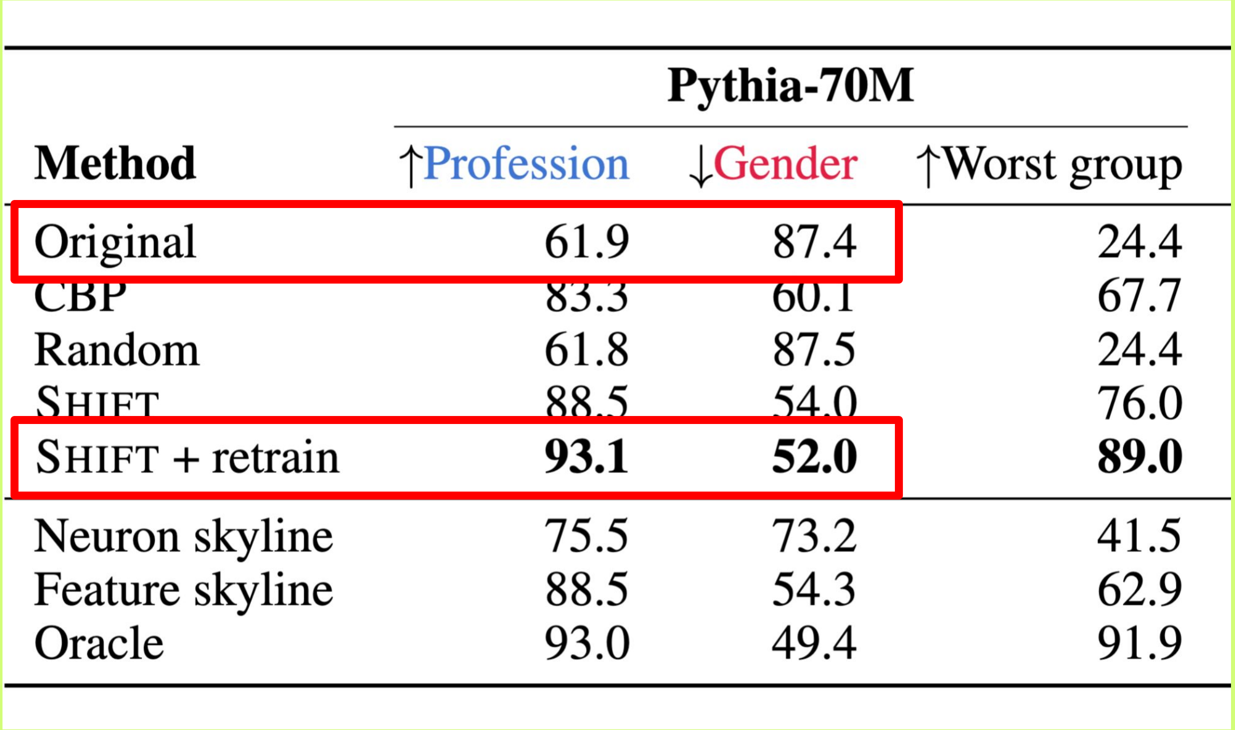




	Pythia-70M		
Method	↑Profession	↓Gender	↑Worst group
Original	61.9	87.4	24.4
CBP	83.3	60.1	67.7
Random	61.8	87.5	24.4
SHIFT	88.5	54.0	76.0
SHIFT + retrain	**93.1**	**52.0**	**89.0**
Neuron skyline	75.5	73.2	41.5
Feature skyline	88.5	54.3	62.9
Oracle	93.0	49.4	91.9

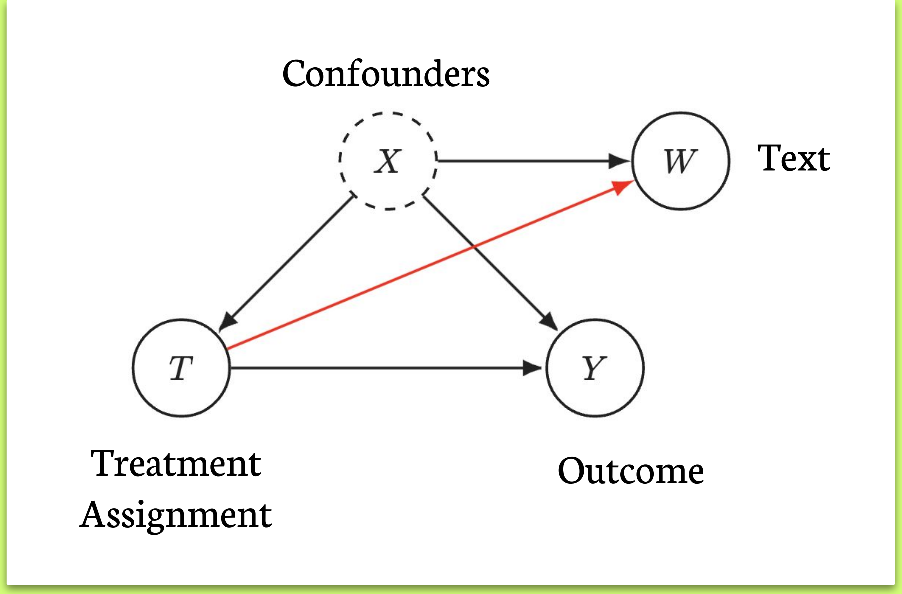


	Pythia-70M		
Method	↑Profession	↓Gender	↑Worst group
Original	61.9	87.4	24.4
CBP	83.3	60.1	67.7
Random	61.8	87.5	24.4
SHIFT	88.5	54.0	76.0
SHIFT + retrain	**93.1**	**52.0**	**89.0**
Neuron skyline	75.5	73.2	41.5
Feature skyline	88.5	54.3	62.9
Oracle	93.0	49.4	91.9

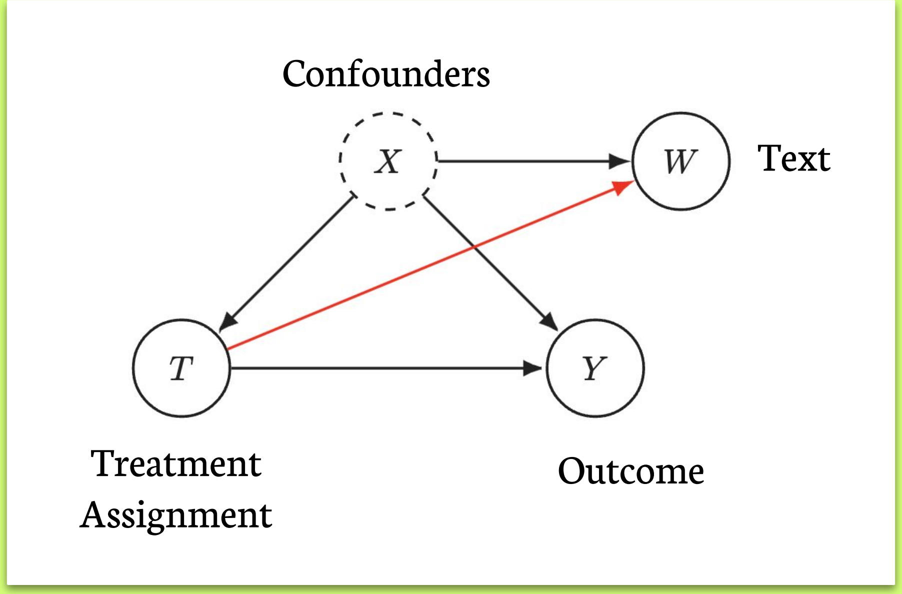


	Pythia-70M		
Method	↑Profession	↓Gender	↑Worst group
Original	61.9	87.4	24.4
CBP	83.3	60.1	67.7
Random	61.8	87.5	24.4
Shift	88.5	54.0	76.0
Shift + retrain	**93.1**	**52.0**	**89.0**
Neuron skyline	75.5	73.2	41.5
Feature skyline	88.5	54.3	62.9
Oracle	93.0	49.4	91.9

Can we apply this to treatment leakage?



Can we apply this to treatment leakage?



… we'll see!

Right now the results look encouraging but are not there yet.

I’d like to work together!

If you found any of it useful or interesting, come have a chat :)

Thank you for listening!

References

1. Feder et al; Causal Inference in Natural Language Processing: Estimation, Prediction, Interpretation and Beyond. TACL 2022

2. Daoud et al; Conceptualizing Treatment Leakage in Text-based Causal Inference. NAACL 2022

3. Egami et al; Using Imperfect Surrogates for Downstream Inference: Design-based Supervised Learning for Social Science Applications of Large Language Models. NeurIPS 2023

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

5. Marks et al; Sparse Feature Circuits: Discovering and Editing Interpretable Causal Graphs in Language Models. ICLR 2025