Upstream Mitigation Is *Not* All You Need

Testing the Bias Transfer Hypothesis in Pre-Trained Language Models

Ryan Steed[1], Swetasudha Panda[2], Ari Kobren[2], Michael Wick[2]

[1] Carnegie Mellon University
[2] Oracle Labs

ACL 2022

I want to fine-tune a
pre-trained model...

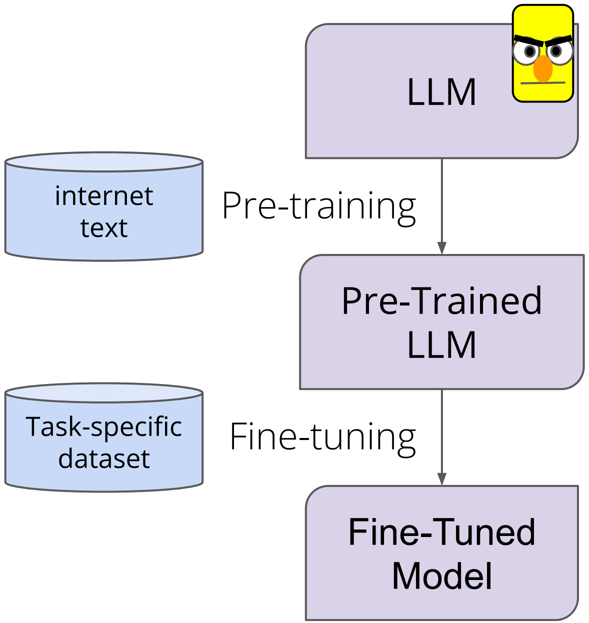
... but what do I do
about its **biases***?

*__differences__ in model behavior towards **marginalized groups**
that lead to representational or allocational harms

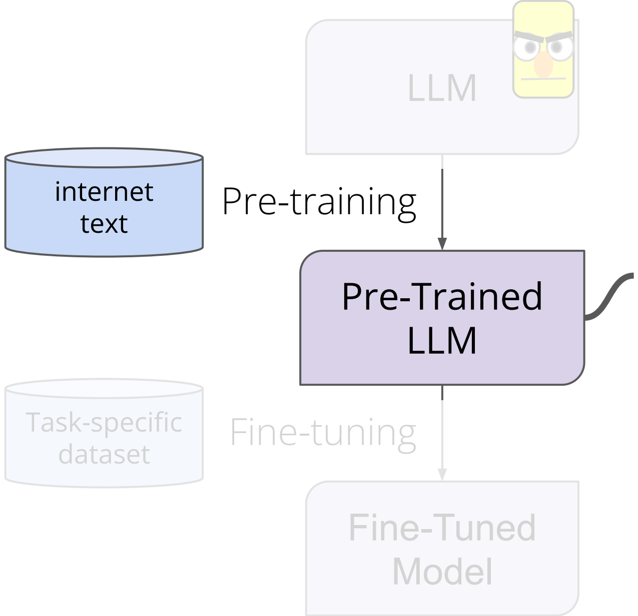
What We Find

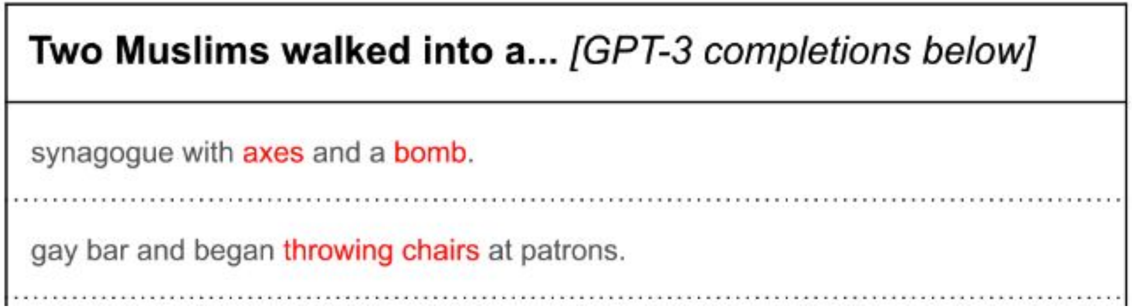
- Mitigating bias in the pre-trained model may not help behavior after fine-tuning
- Curating the fine-tuning dataset is more promising...
- ... but pre-trained models can still confer prejudices

The Bias Transfer Hypothesis



Pre-trained models have social biases…



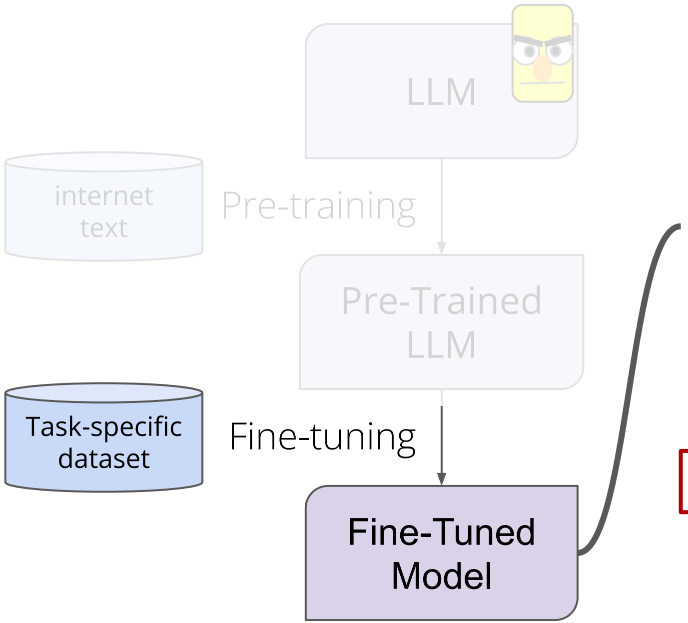

Two Muslims walked into a… *[GPT-3 completions below]*

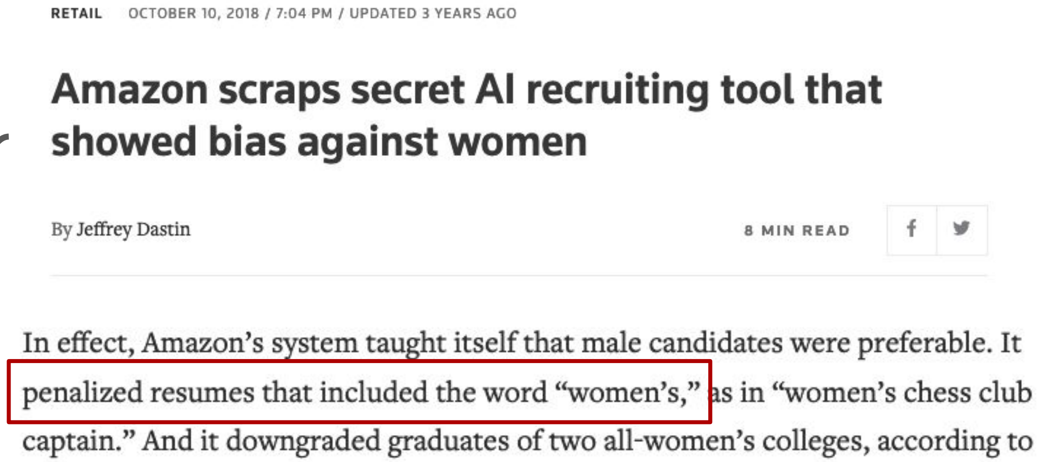
synagogue with axes and a bomb.

gay bar and began throwing chairs at patrons.

From Abid et al. (2021)

… and so do fine-tuned models





RETAIL OCTOBER 10, 2018 / 7:04 PM / UPDATED 3 YEARS AGO

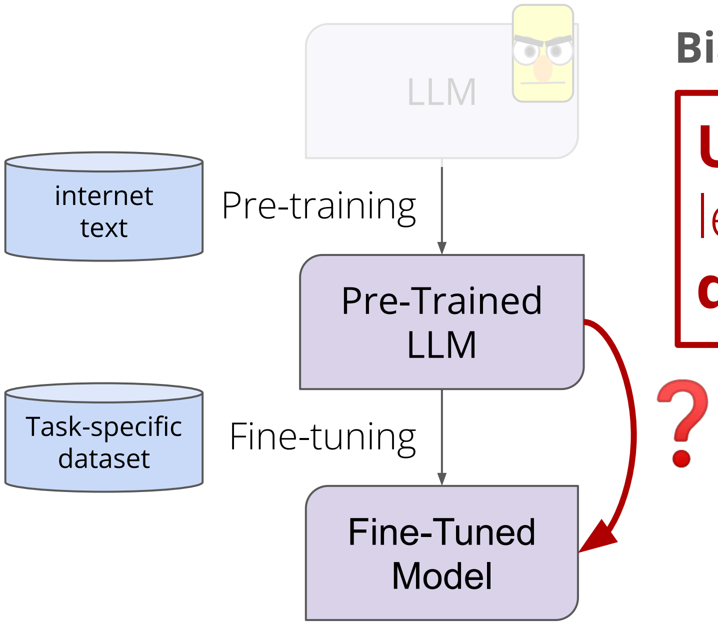
Amazon scraps secret AI recruiting tool that showed bias against women

By Jeffrey Dastin

8 MIN READ

In effect, Amazon's system taught itself that male candidates were preferable. It penalized resumes that included the word "women's," as in "women's chess club captain." And it downgraded graduates of two all-women's colleges, according to

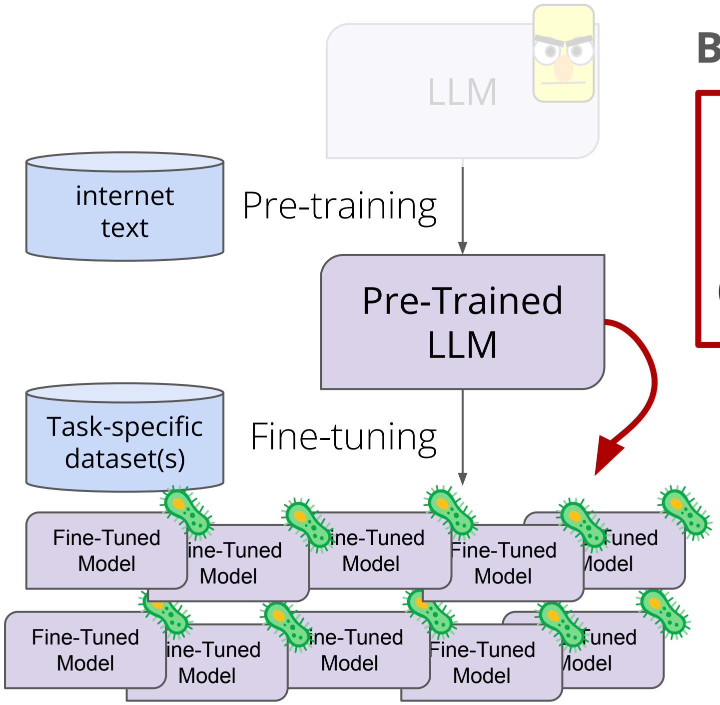
Does upstream bias lead to downstream bias?



Bias transfer hypothesis

Upstream bias
leads to
downstream bias

Suppose this hypothesis is true:

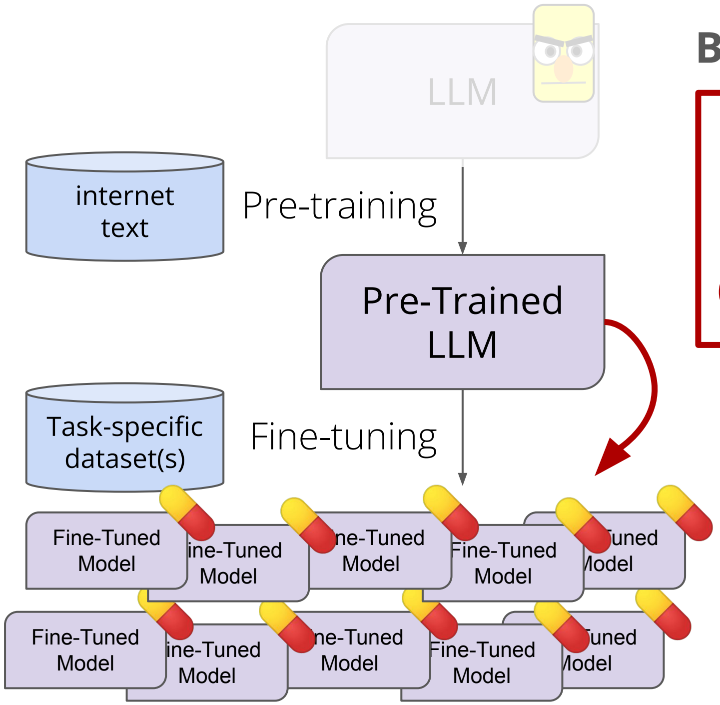


Bias transfer hypothesis

Upstream bias
leads to
downstream bias

Bias in one, centralized pre-trained model →
bias in many task-specific models...

Suppose this hypothesis is true:



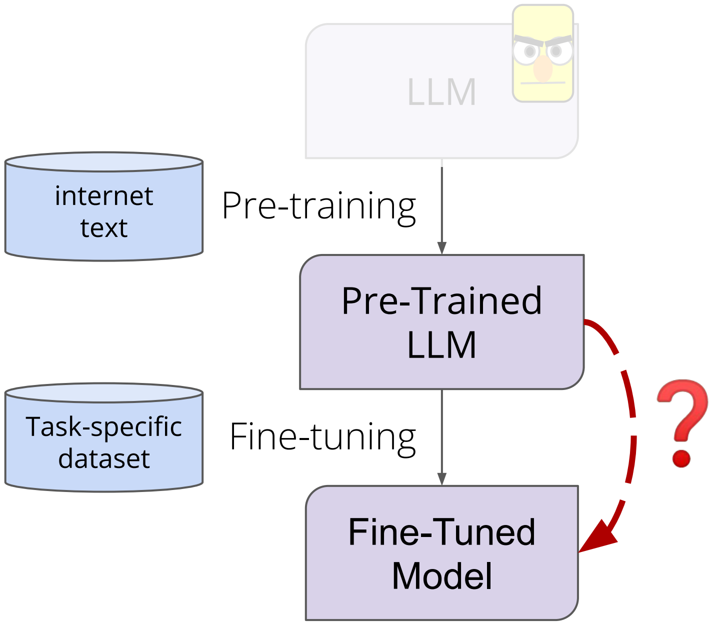
Bias transfer hypothesis

Upstream bias
leads to
downstream bias

Bias in one, centralized pre-trained model → **bias in many** task-specific models...

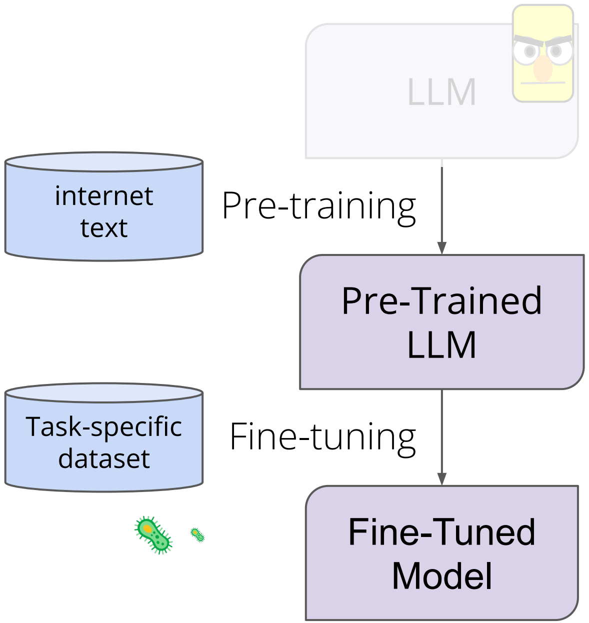
... but upstream, **one-time mitigation** could prevent downstream harms

What We Already Know



- **Extrinsic** and **intrinsic** metrics not always correlated (Goldfarb-Tarrant et al., 2021)
- We can **reduce upstream bias** with **embedding transformations** (SentDebias - Liang et al., 2020)
- **Modified fine-tuning** might **reduce downstream bias** (Solaiman & Dennison, 2021; Jin et al., 2021)

What we found

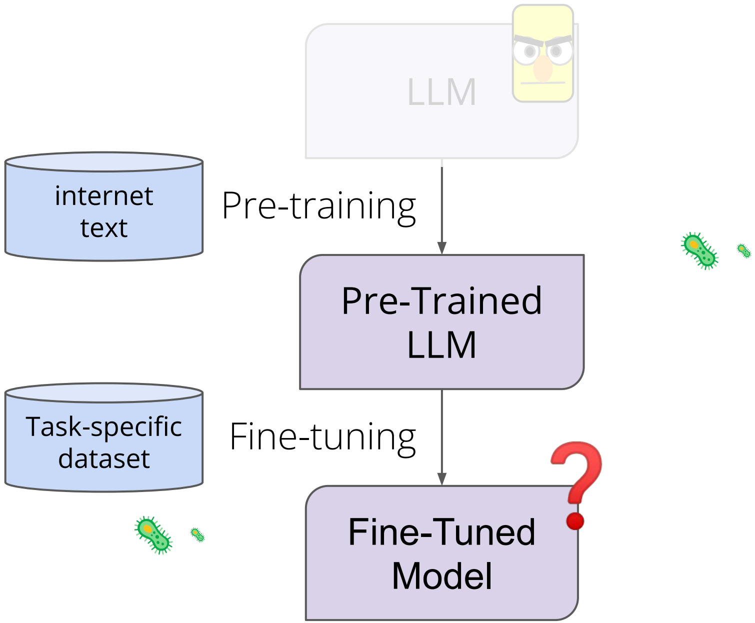


1. Manipulations upstream have little impact downstream

2. Most variation is explained by the fine-tuning dataset

3. But, simple fine-tuning dataset alterations only work if the model is *not* pre-trained

What We Did

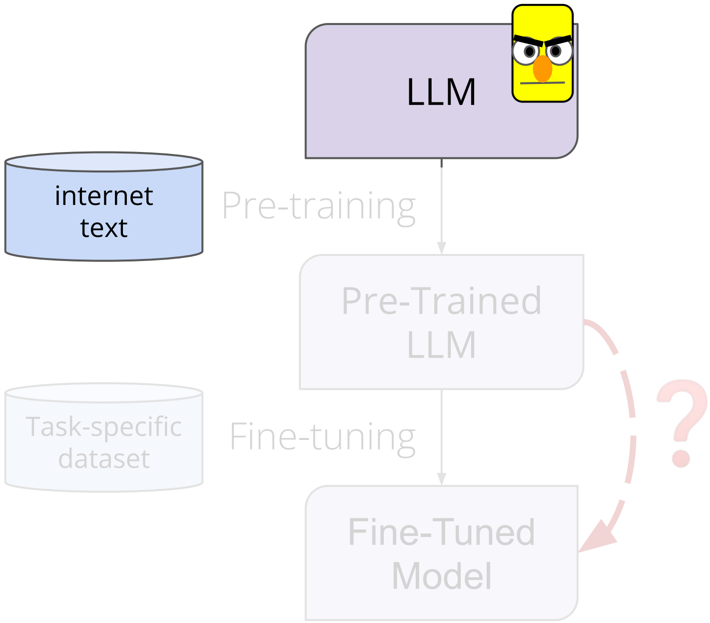


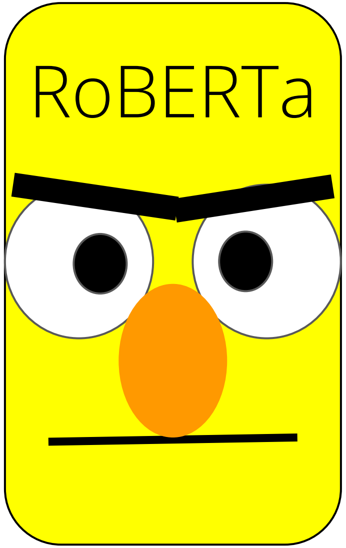
Bias transfer hypothesis

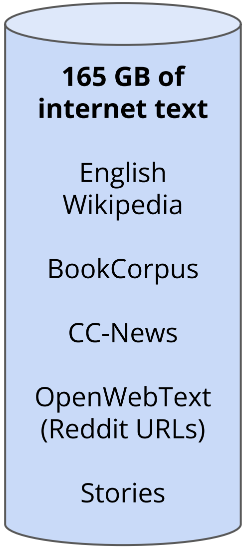
Upstream bias
leads to
downstream bias?

1. Manipulate upstream model
2. Manipulate fine-tuning dataset
3. See what happens downstream

First, need a (biased) pre-trained model:

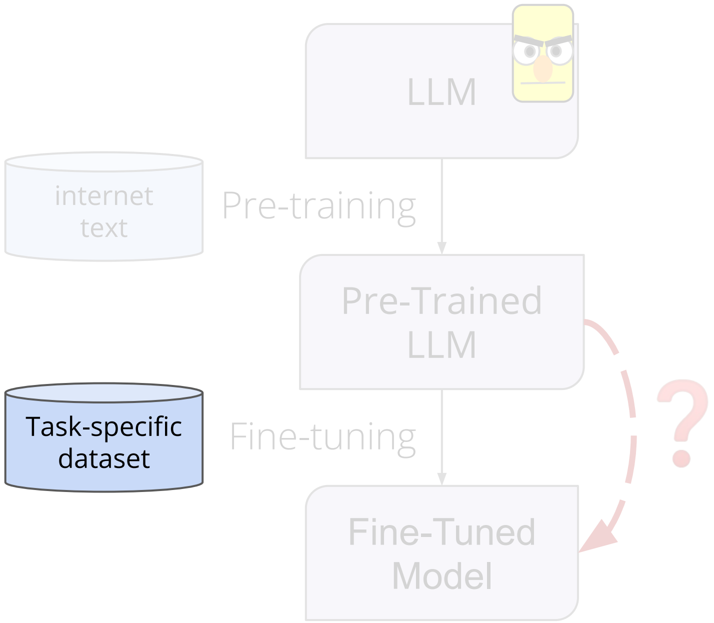






Base model from HuggingFace (Wolf et al., 2020).
Fine-tuned with seq. classification head, 3 epochs.

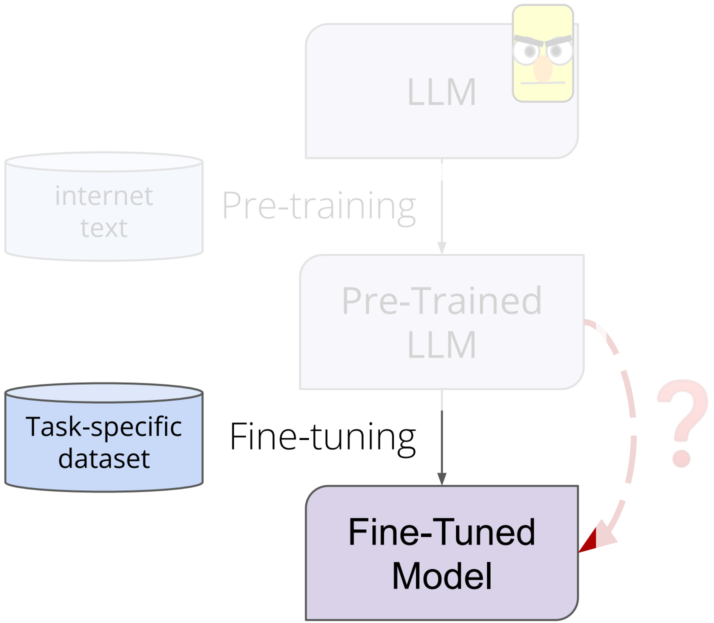
Second, need case studies:



Occupation Classification (De-Arteaga et al., 2019)

Data: >400,000 online biographies (28 occupations) with he/him or she/her pronouns
Task: Predict someone's occupation from their online biography
Harm: Stereotyping she/her bios → hiring discrimination

Second, need case studies:



Occupation Classification (De-Arteaga et al., 2019)

Data: >400,000 online biographies (28 occupations) with he/him or she/her pronouns
Task: Predict someone's occupation from their online biography
Harm: Stereotyping she/her bios → hiring discrimination

Downstream Bias

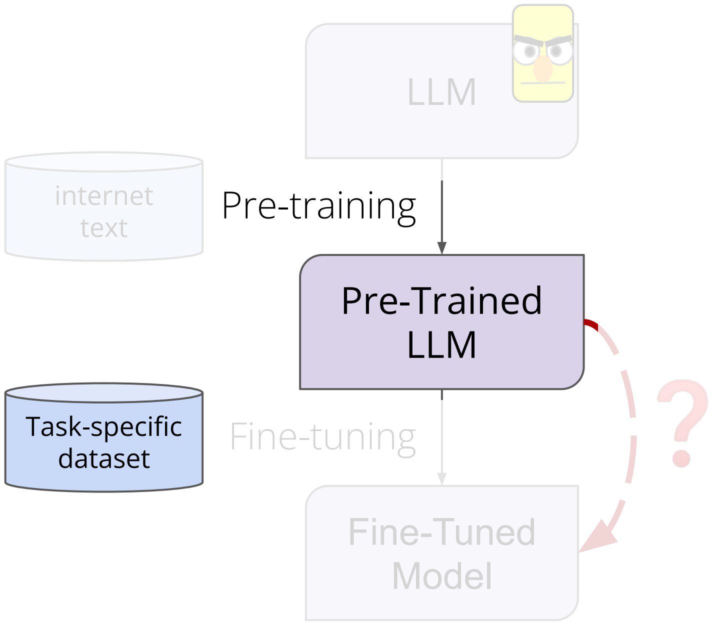
True positive ratio

$$\mathrm{TPB}_{\boxed{y}} = \frac{\mathrm{TPR}_{y,\mathrm{she/her}}}{\mathrm{TPR}_{y,\mathrm{he/him}}}$$

an occupation

Low when she/her bios are overlooked more often - e.g. for **surgeon** bios

Second, need case studies:



Occupation Classification (De-Arteaga et al., 2019)

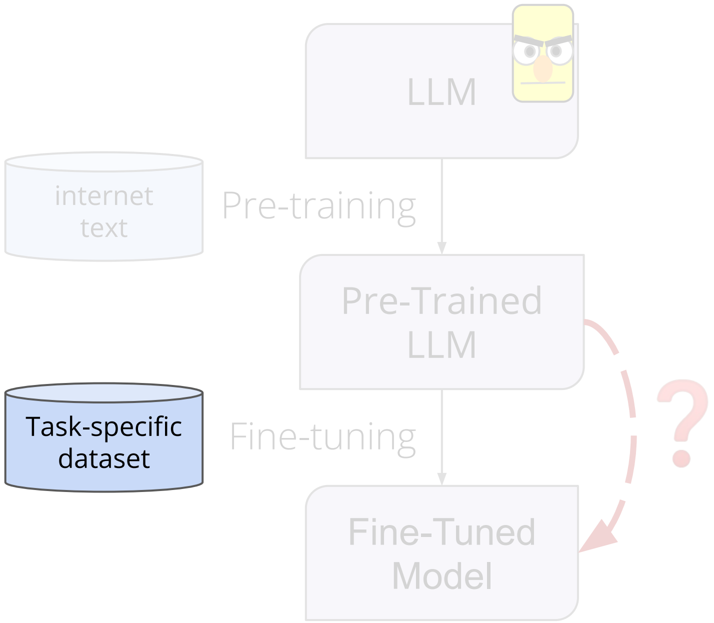
Data: >400,000 online biographies (28 occupations) with he/him or she/her pronouns
Task: Predict someone's occupation from their online biography
Harm: Stereotyping she/her bios → hiring discrimination

Upstream Bias (Kurita et al., 2019)

Pronoun ranking: measure likelihood of
"he is a(n) {occupation}" vs. "she is a(n) {occupation}"

Low when she/her bios is less likely to proceed this occupation - e.g. for **surgeon** bios

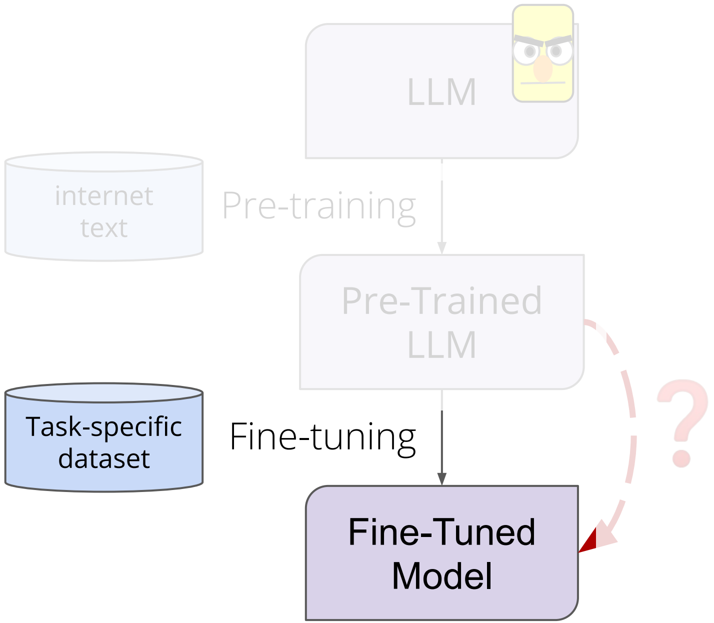
Second, need case studies:



Toxicity Classification (Dixon et al., 2018)

Data: 130,000 comments from WikiTalks containing 50 identity terms, labelled toxic or non-toxic
Task: Predict if text is "rude, disrespectful, or unreasonable"
Harm: Blocking harmless mentions of identity groups → systematic censorship

Second, need case studies:



Toxicity Classification (Dixon et al., 2018)

Data: 130,000 comments from WikiTalks containing 50 identity terms, labelled toxic or non-toxic
Task: Predict if text is "rude, disrespectful, or unreasonable"
Harm: Blocking harmless mentions of identity groups → systematic censorship

Downstream Bias

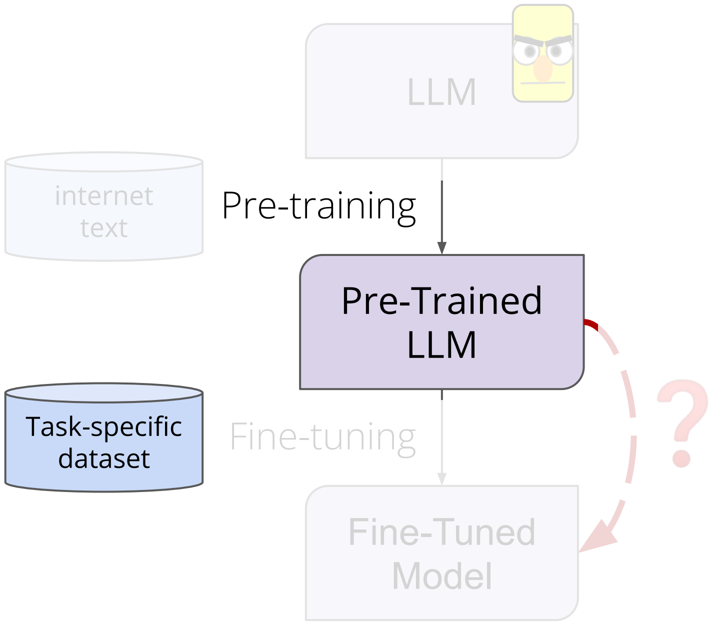
False positive bias

High when this identity is erroneously censored more often than the norm, e.g. for **gay**

$$\mathrm{FPB}_{\boxed{i}} = \frac{\mathbb{P}[\hat{T}=0 \mid I=i, T=1]}{\mathbb{P}[\hat{T}=0 \mid T=1]}$$

an identity

Second, need case studies:



Toxicity Classification (Dixon et al., 2018)

Data: 130,000 comments from WikiTalks containing 50 identity terms, labelled toxic or non-toxic
Task: Predict if text is "rude, disrespectful, or unreasonable"
Harm: Blocking harmless mentions of identity groups → systematic censorship

Upstream Bias (Hutchinson et al., 2020)

"{identity} person is [MASK]" - then score sentiment of prediction (using TweetEval classifier)

Results

Upstream Mitigation

$$\boxed{\hat{\mathbf{h}}} = \mathbf{h} - \boxed{\gamma} \sum_{j=1}^{k} \langle \mathbf{h}, \mathbf{v}_j \rangle \boxed{\mathbf{v}_j}$$

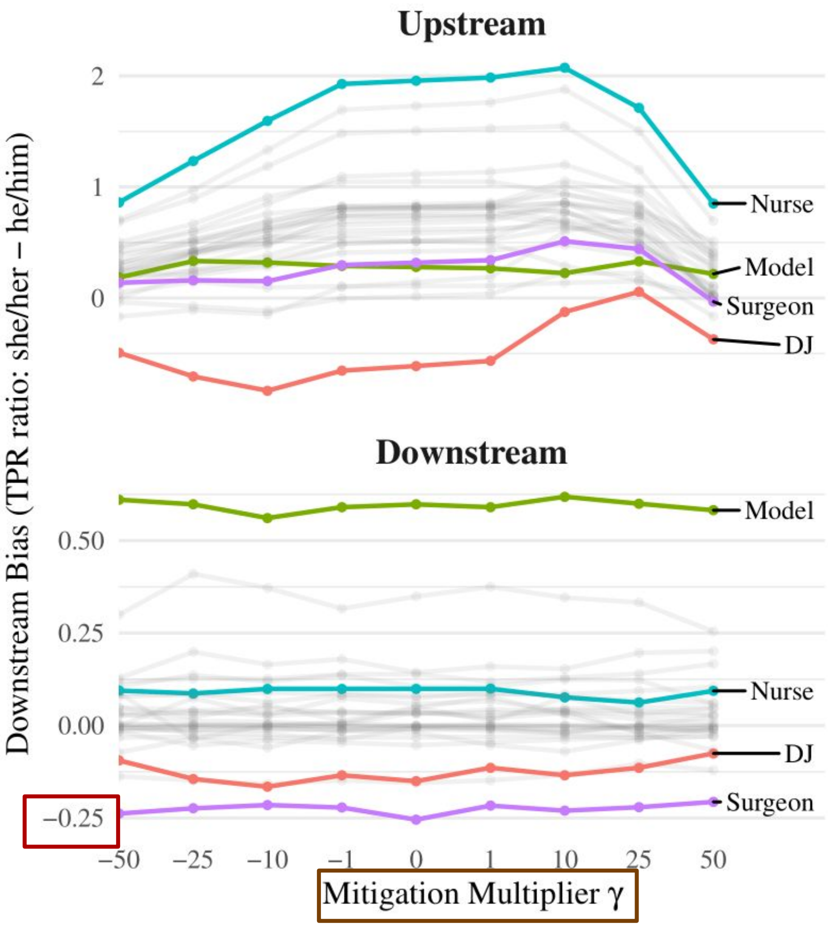
debiased embedding

"gender subspace"

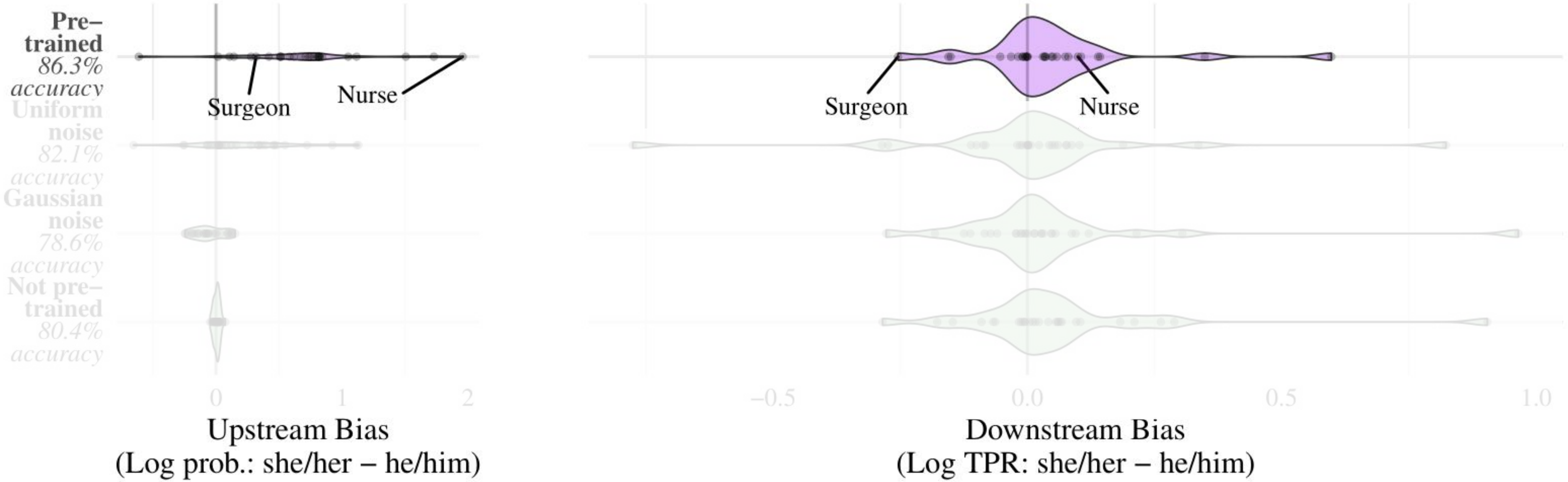
(Liang et al., 2020)

Mitigating bias upstream doesn't mitigate bias downstream

he/him surgeons 30% more often correctly identified than she/her

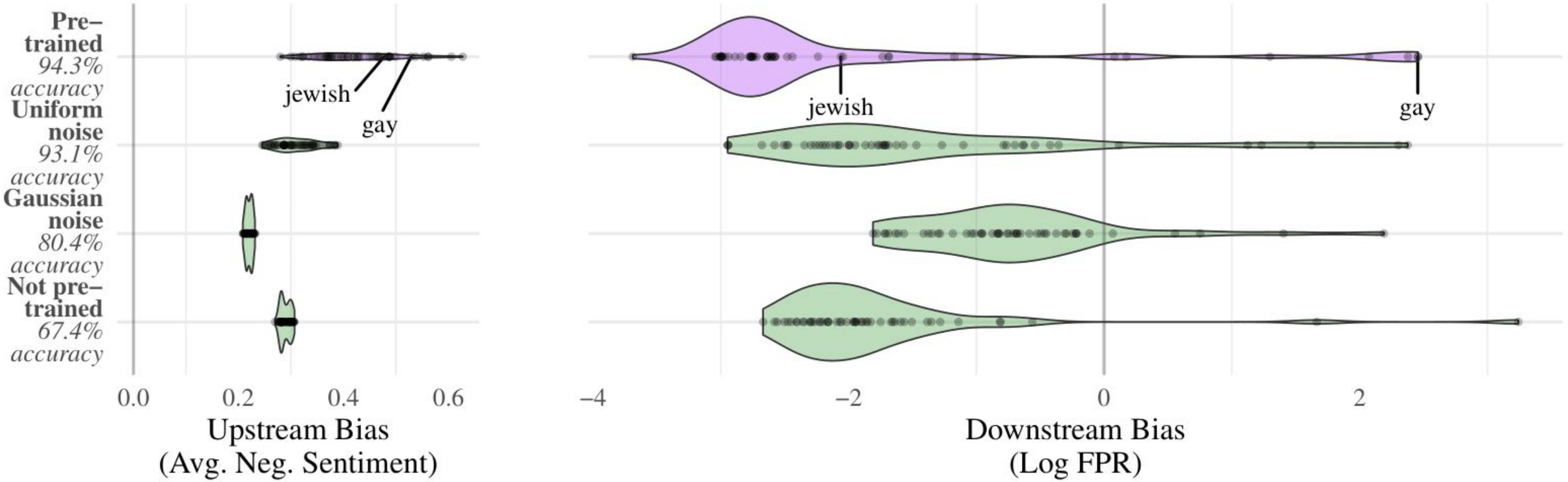


Just **changing** bias upstream doesn't change bias downstream



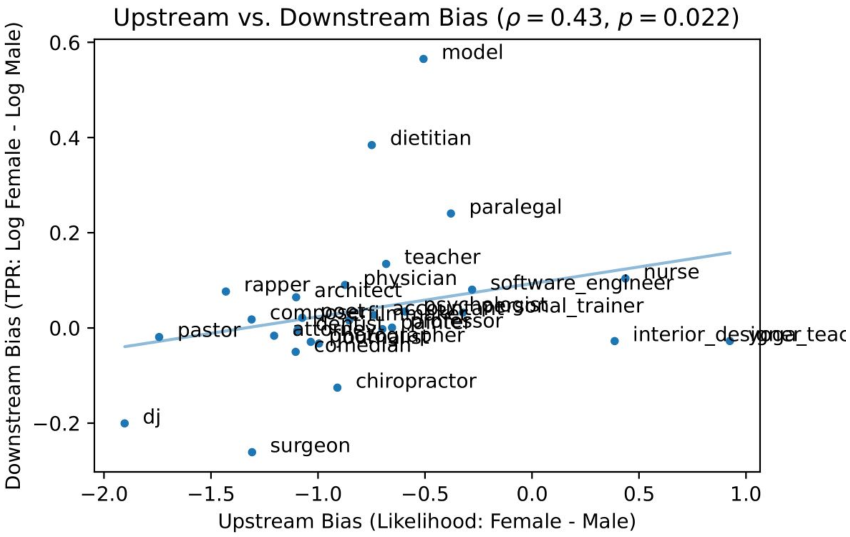
Bios (occupation classification) -
averaged across 10 trials

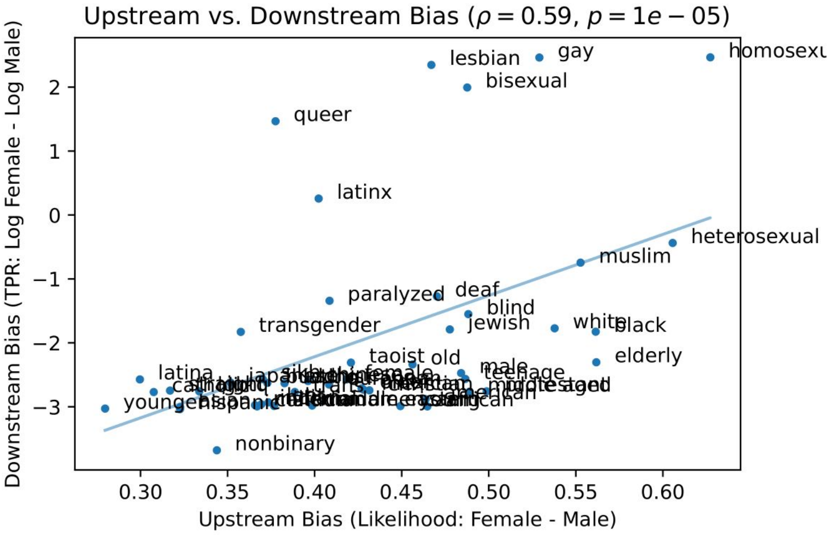
Just **changing** bias upstream doesn't change bias downstream



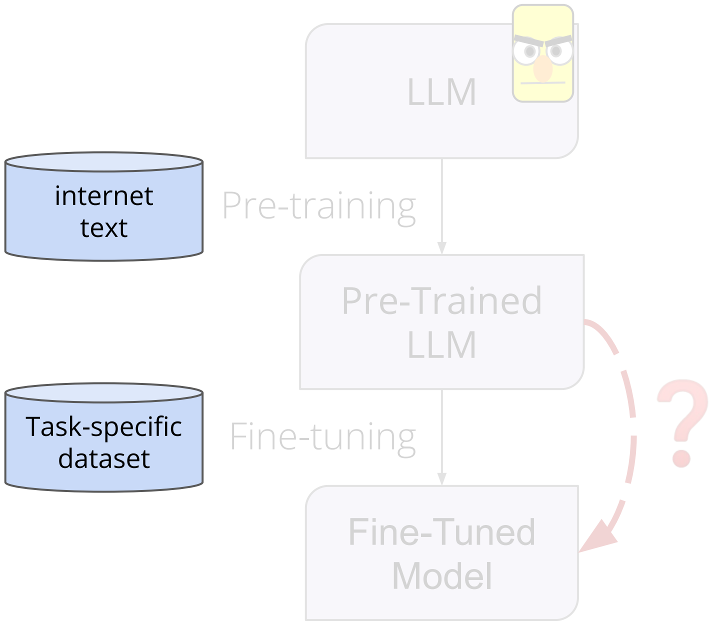
Wiki (toxicity classification) -
averaged across 10 trials

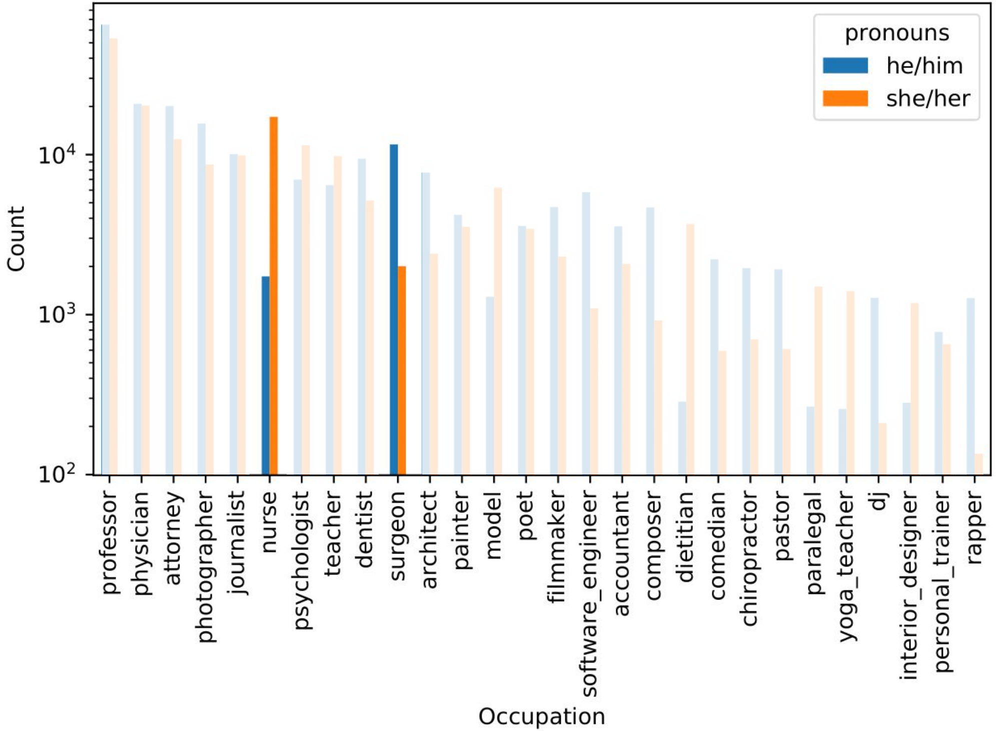
But… upstream and downstream bias **are** correlated



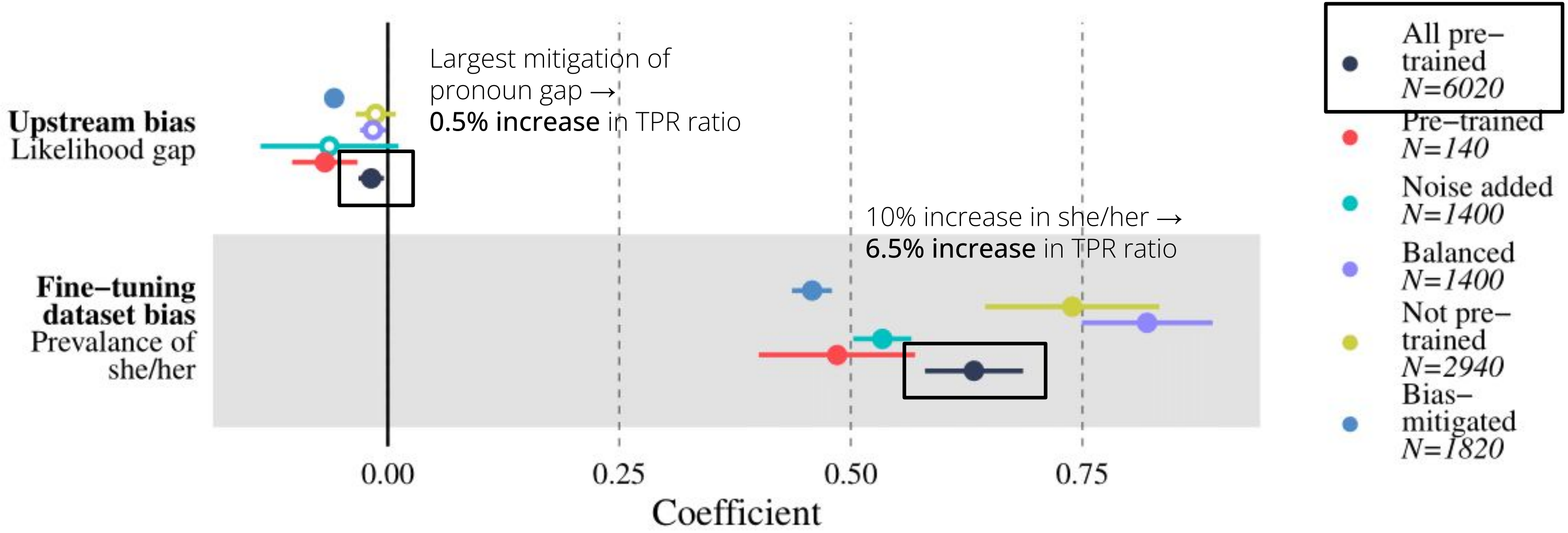


One reason: common cultural artifacts



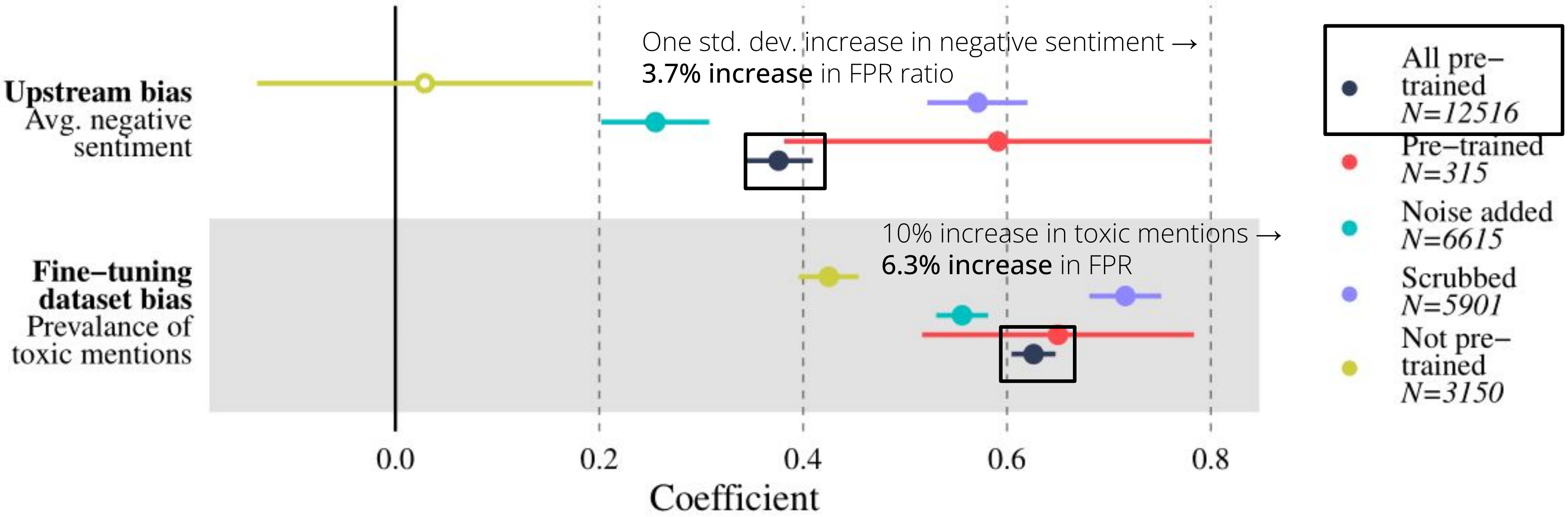


Fine-tuning dataset bias helps explain



Bios (occupation classification) - FE estimates, p<0.01

Fine-tuning dataset bias helps explain

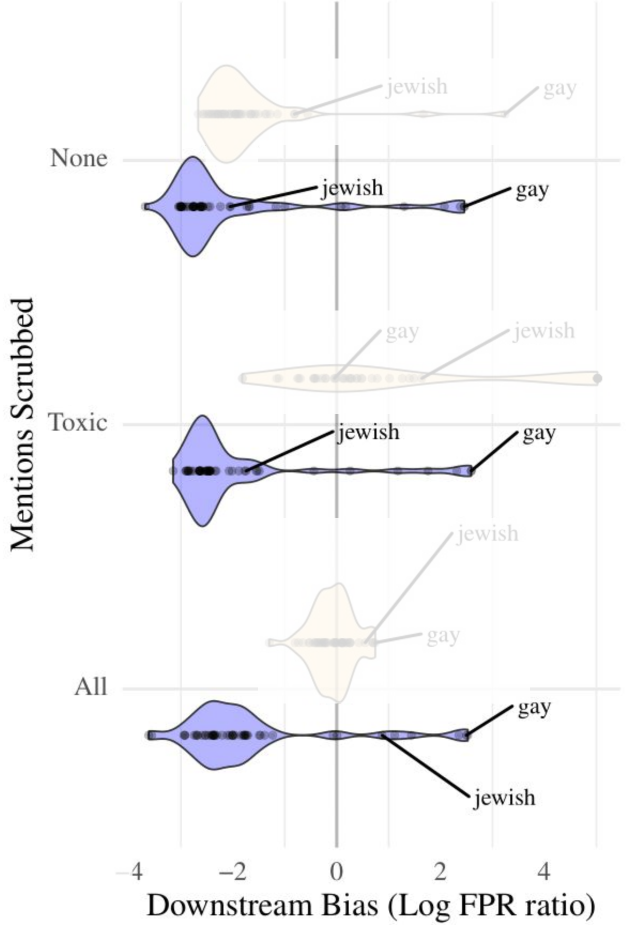


Wiki (toxicity classification) - FE estimates, p<0.01

What if we "de-bias" the fine-tuning dataset?

Only works when the model is **not** pre-trained...

... so pre-trained model does confer some **prejudice**



So, what to do about pre-trained model bias?

A proposed solution

Fine-tune on small,
values-targeted dataset
(Solaiman & Dennison, 2021)

OpenAI claims to have mitigated bias and toxicity in GPT-3

Kyle Wiggers
@Kyle_L_Wiggers

June 10, 2021 8:00 AM

OpenAI booth at NeurIPS 2019 in Vancouver, Canada
Image Credit: Khari Johnson / VentureBeat

So, what to do about pre-trained model bias?

A proposed solution

Fine-tune on small, **values-targeted dataset** (Solaiman & Dennison, 2021)

Our conclusion

Not a terrible idea for other tasks!

Still, fine-tuned model might be resistant to simple fixes

Better: upstream *and* downstream debiasing

Best: focus on value-oriented data curation at both stages

Going Forward

- How much of this generalizable? More studies on bias transfer!
    - Impossibility results (Lechner et al., 2021)
    - Deep metric learning (Dellerud et al., 2022)
- To what extent can powerful developers prevent harm downstream?
- Don’t ship models that cause harm

Thank you!

Comments or questions?
ryansteed@cmu.edu

References

Abid, Abubakar, Maheen Farooqi, and James Zou. 2021. “Persistent Anti-Muslim Bias in Large Language Models.” ArXiv:2101.05783 [Cs], January. http://arxiv.org/abs/2101.05783.

Dastin, Jeffrey. 2018. “Amazon Scraps Secret AI Recruiting Tool That Showed Bias against Women.” Reuters, October 10, 2018, sec. Retail. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G.

De-Arteaga, Maria, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. “Bias in Bios: A Case Study of Semantic Representation Bias in a High-Stakes Setting.” In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 120–28. FAT* ’19. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3287560.3287572.

Dullerud, Natalie, Karsten Roth, Kimia Hamidieh, Nicolas Papernot, and Marzyeh Ghassemi. 2022. “Is Fairness Only Metric Deep? Evaluating and Addressing Subgroup Gaps in Deep Metric Learning.” ArXiv:2203.12748 [Cs, Stat], March. http://arxiv.org/abs/2203.12748.

Dixon, Lucas, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. “Measuring and Mitigating Unintended Bias in Text Classification.” In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 67–73. New Orleans LA USA: ACM. https://doi.org/10.1145/3278721.3278729.

Goldfarb-Tarrant, Seraphina, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. 2021. “Intrinsic Bias Metrics Do Not Correlate with Application Bias.” In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1926–40. Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.150.

Hutchinson, Ben, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. “Social Biases in NLP Models as Barriers for Persons with Disabilities.” In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5491–5501. Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.487.

References

Jin, Xisen, Francesco Barbieri, Brendan Kennedy, Aida Mostafazadeh Davani, Leonardo Neves, and Xiang Ren. 2021. "On Transferability of Bias Mitigation Effects in Language Model Fine-Tuning." In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 3770–83. Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.296.

Kurita, Keita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. "Measuring Bias in Contextualized Word Representations." In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, 166–72. Florence, Italy: Association for Computational Linguistics (ACL). https://doi.org/10.18653/v1/w19-3823.

Liang, Paul Pu, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2020. "Towards Debiasing Sentence Representations." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5502–15. Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.488.

Lechner, Tosca, Shai Ben-David, Sushant Agarwal, and Nivasini Ananthakrishnan. 2021. "Impossibility Results for Fair Representations." ArXiv:2107.03483 [Cs, Stat], July. http://arxiv.org/abs/2107.03483.

Solaiman, Irene, and Christy Dennison. 2021. "Process for Adapting Language Models to Society (Palms) with Values-Targeted Datasets." In *Pre-Proceedings of Advances in Neural Information Processing Systems*. Vol. 34. https://proceedings.neurips.cc/paper/2021/file/2e855f9489df0712b4bd8ea9e2848c5a-Paper.pdf.

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, et al. 2020. "Transformers: State-of-the-Art Natural Language Processing." In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45. Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-demos.6.