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

Bias Transfer Hypothesis

- Homogenous large language models (LLMs) undergird many machine learning systems.
- LLMs exhibit social biases (e.g., stereotypes) before *and* after fine-tuning.

Do biases internalized by LLMs during pre-training transfer into harmful behavior after fine-tuning?

Our Findings

- In these tasks, reducing downstream bias via upstream interventions is mostly futile.
- The fine-tuning dataset plays a larger role than upstream bias in determining downstream harms.
- But, a pre-trained model learns biases more easily.
- Practitioners should focus on task-specific harms.

Results (Toxicity)

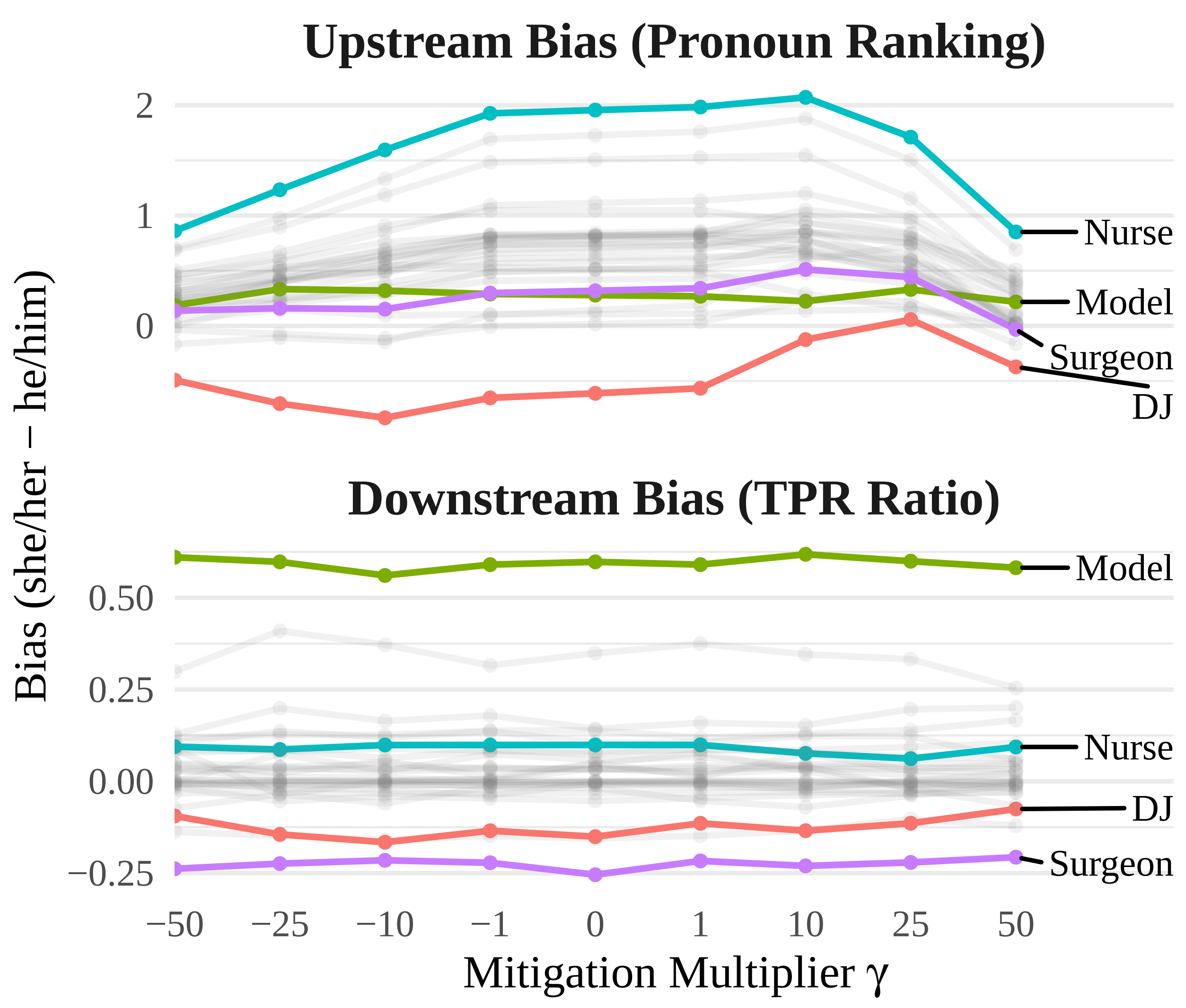

Figure 1: **(Upstream Intervention)** Even when upstream bias is mitigated [5] in BIOS, the distribution of downstream bias remains mostly the same.

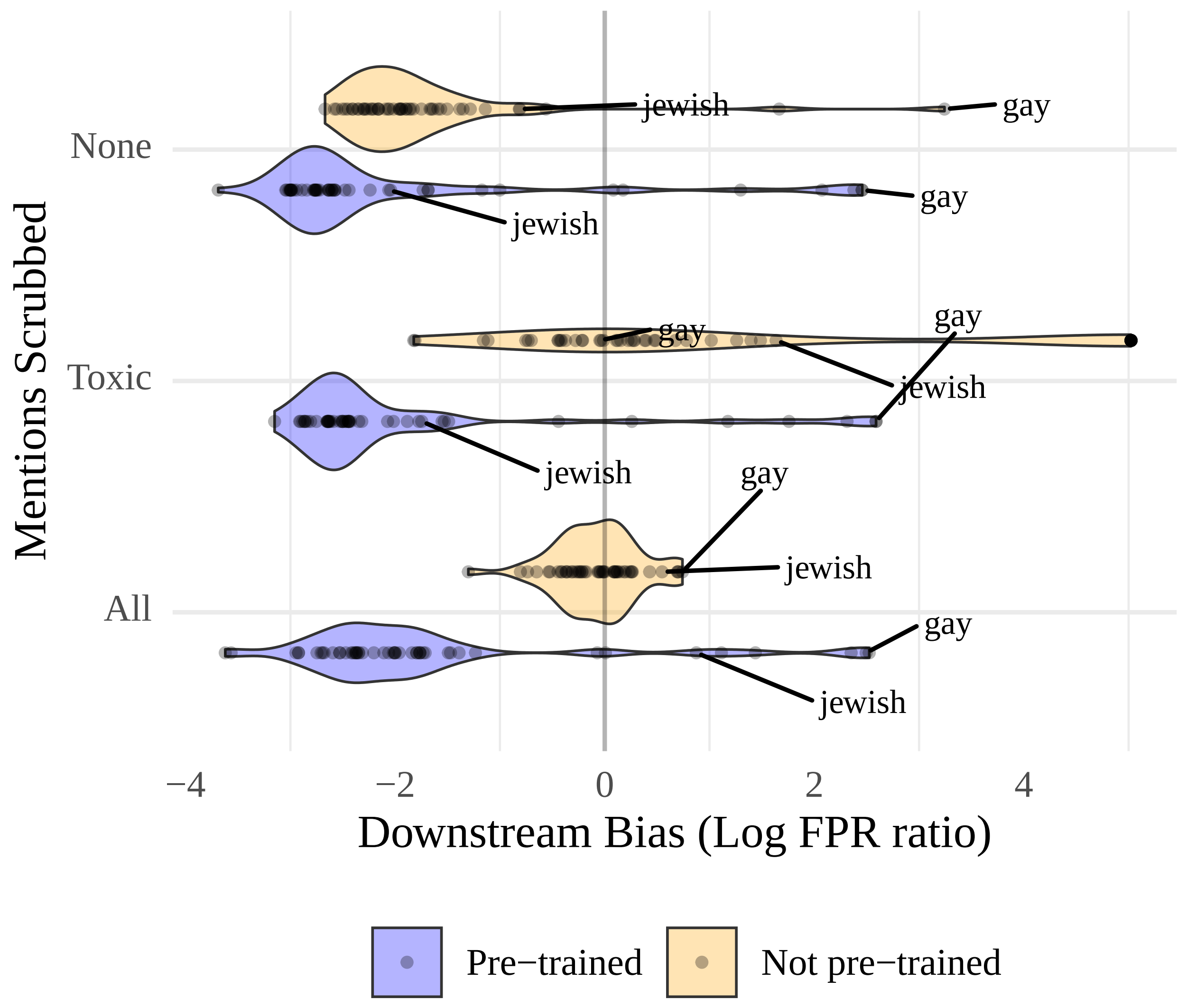

Figure 2: **(Downstream Intervention)** Scrubbing toxic mentions from the fine-tuning dataset reduces downstream bias *only when* the model is not pre-trained (yellow).

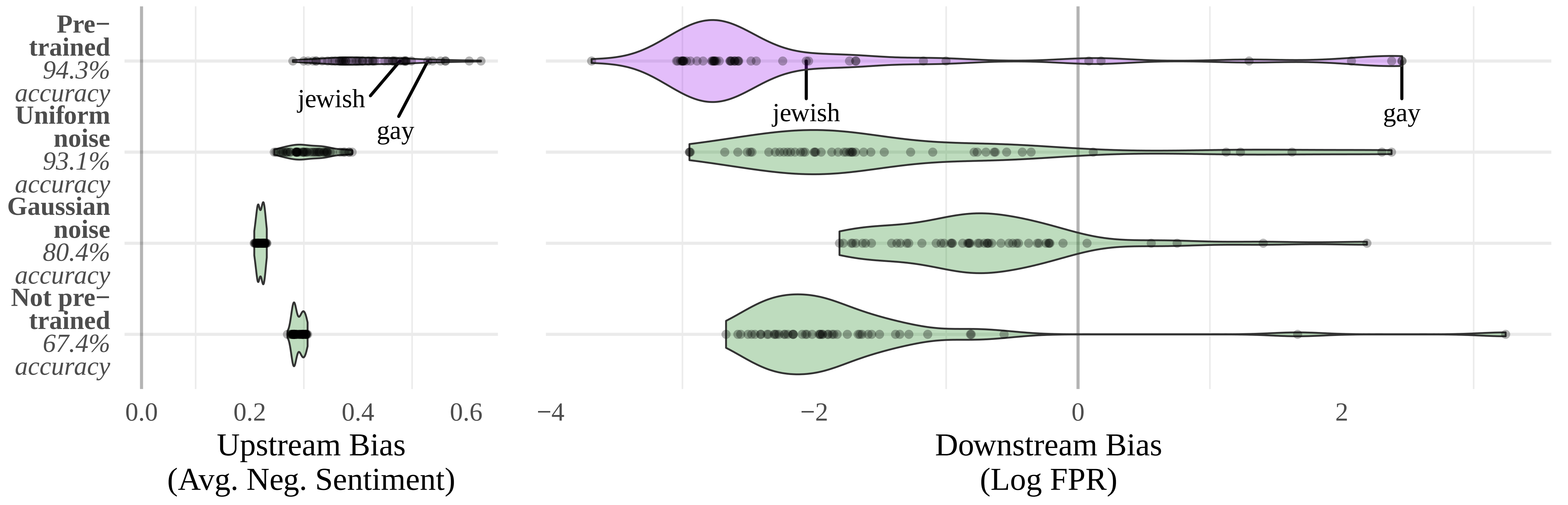

Figure 3: **(Upstream Intervention)** Despite drastic random changes to upstream bias (left), downstream bias (right) per identity remains roughly stable. RoBERTa [6] learns bias even without pre-training. (Averaged over 10 trials.)

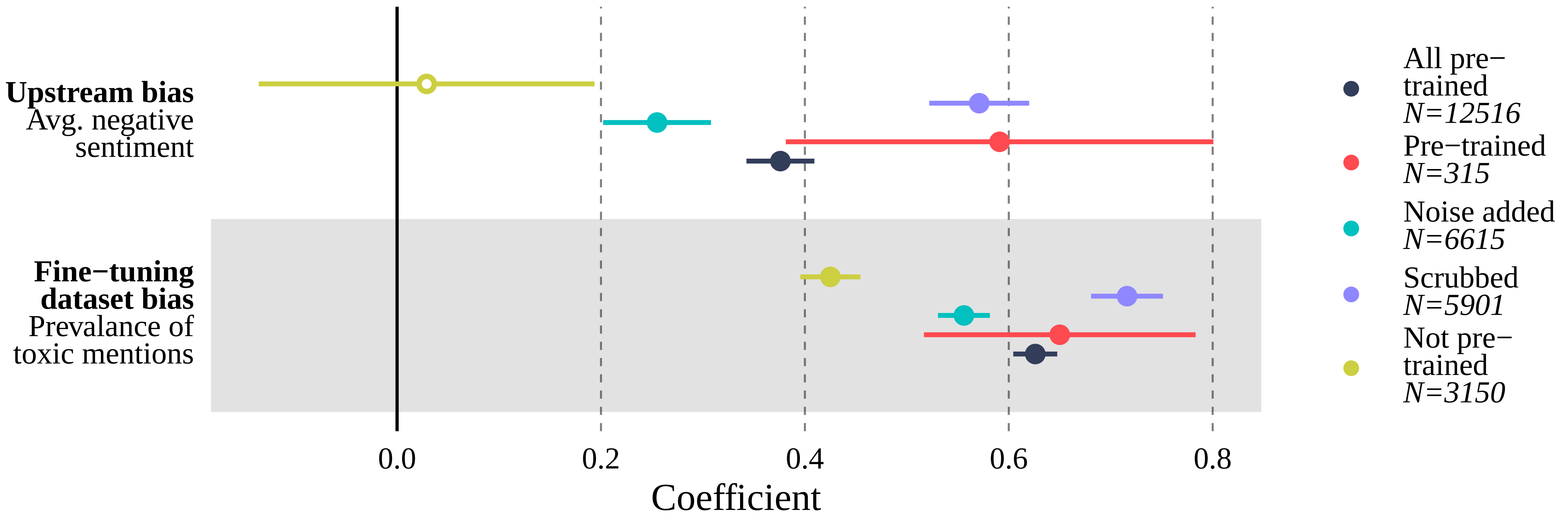

Figure 4: We regress upstream bias and fine-tuning dataset bias (proxied by co-occurrence rates) on downstream bias, controlling for template effects. Bars depict standard errors. In BIOS, upstream bias has an even **smaller** impact.

- **Large** 0.1 SD increase in negative sentiment (upstream bias) → **modest** 3.7% increase in FPR (downstream bias).
- **Modest** 10% increase in toxic mentions of an identity term → **even larger** 6.3% increase in FPR.

Experiments

For two tasks, we measured upstream and downstream bias after several interventions.

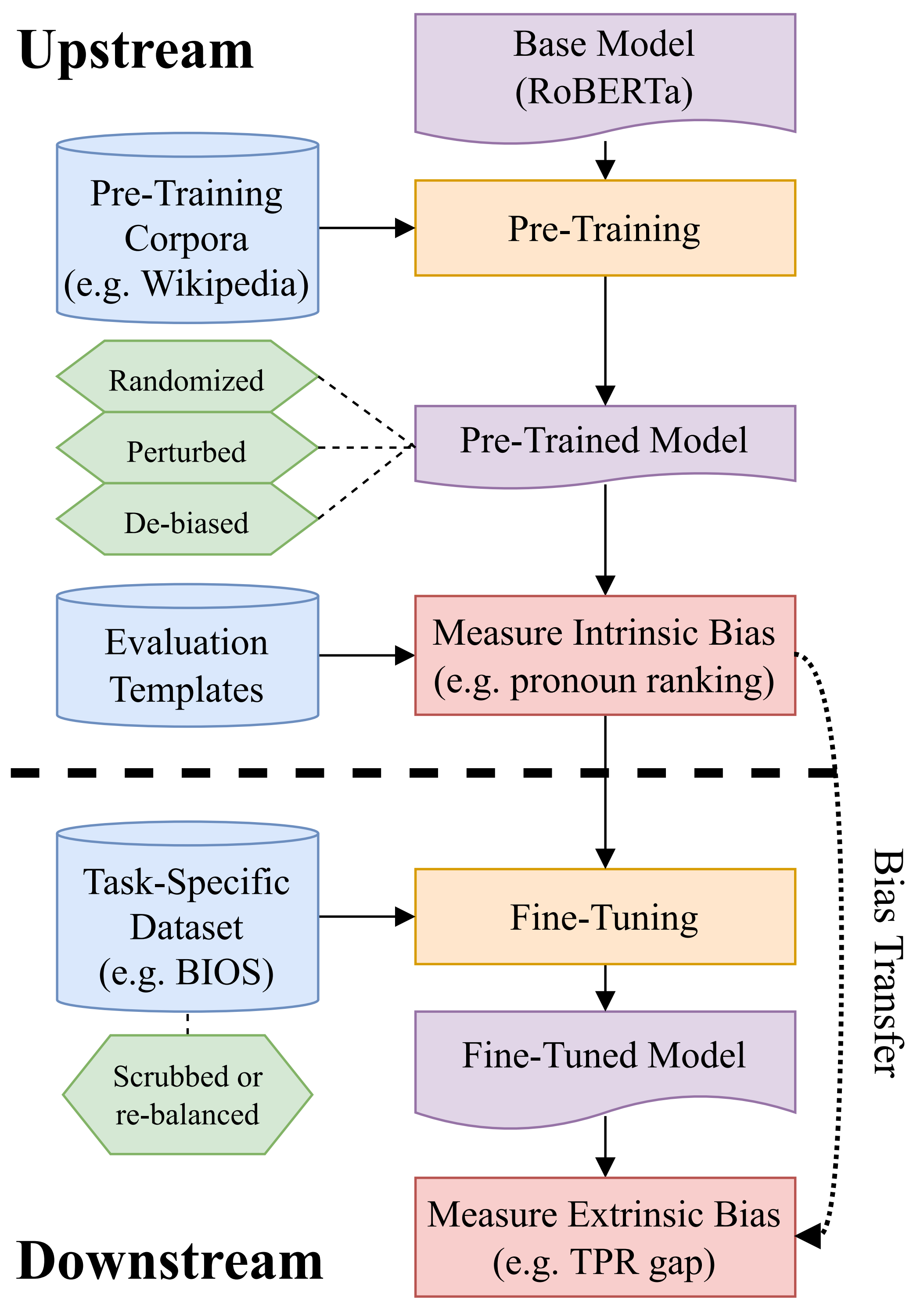

Figure 5: Fine-tuning pipeline, with experimental interventions (hexagons) to test bias transfer.

Toxicity Classification (WIKI) [2]

Predict 28 occupations from 400k online bios.

Harm: Stereotyping she/her bios
→ hiring discrimination.
Downstream Bias: True positive ratio
Upstream Bias: Pronoun ranking [4]

Biography Classification (BIOS) [1]

Predict toxicity in 130k posts about 50 identities.

Harm: Blocking innocuous mentions
→ systematic censorship.
Downstream Bias: False positive ratio
Upstream Bias: Negative sentiment [3]

[1] M. De-Arteaga, A. Romanov, H. Wallach, J. Chayes, C. Borgs, A. Chouldechova, S. Geyik, K. Kenthapadi, and A. T. Kalai. 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*, FAT* '19, pages 120–128, New York, NY, USA, Jan. 2019. Association for Computing Machinery.

[2] L. Dixon, J. Li, J. Sorensen, N. Thain, and L. Vasserman. Measuring and Mitigating Unintended Bias in Text Classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73, New Orleans LA USA, Dec. 2018. ACM.

[3] B. Hutchinson, V. Prabhakaran, E. Denton, K. Webster, Y. Zhong, and S. Denuyl. Social Biases in NLP Models as Barriers for Persons with Disabilities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5491–5501, Online, July 2020. Association for Computational Linguistics.

[4] K. Kurita, N. Vyas, A. Pareek, A. W. Black, and Y. Tsvetkov. Measuring Bias in Contextualized Word Representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172, Florence, Italy, 2019. Association for Computational Linguistics (ACL).

[5] P. P. Liang, I. M. Li, E. Zheng, Y. C. Lim, R. Salakhutdinov, and L.-P. Morency. Towards Debiasing Sentence Representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5502–5515, Online, July 2020. Association for Computational Linguistics.

[6] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. Rush. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, 2020. Association for Computational Linguistics.