Fairness Dynamics During Training

Krishna Patel **Nivedha Sivakumar**
Barry-John Theobald **Luca Zappella** **Nicholas Apostoloff**
Apple
{krishna_patel, nivedha,
barryjohn_theobald, lzappella, napostoloff}@apple.com

Abstract

Understanding fairness dynamics during Large Language Model (LLM) training facilitates the diagnoses of biases that emerge and enables developers to mitigate biases through early stopping or other training interventions. We introduce two new metrics to evaluate fairness dynamics holistically during model pre-training: Average Rank and Jensen-Shannon Divergence by Parts. These metrics provide insights into the Pythia models' [1] progression of biases in gender prediction of occupations on the WinoBias dataset [19]. By monitoring these dynamics, we find that (1) Pythia-6.9b is biased towards men; it becomes more performant and confident predicting "male" than "female" during training, (2) via early-stopping, Pythia-6.9b can exchange 1.7% accuracy on LAMBADA [15] for a 92.5% increase in fairness, and (3) larger models can exhibit more bias; Pythia-6.9b is makes more assumptions about gender than Pythia-160m, even when a subject's gender is not specified.

1 Introduction

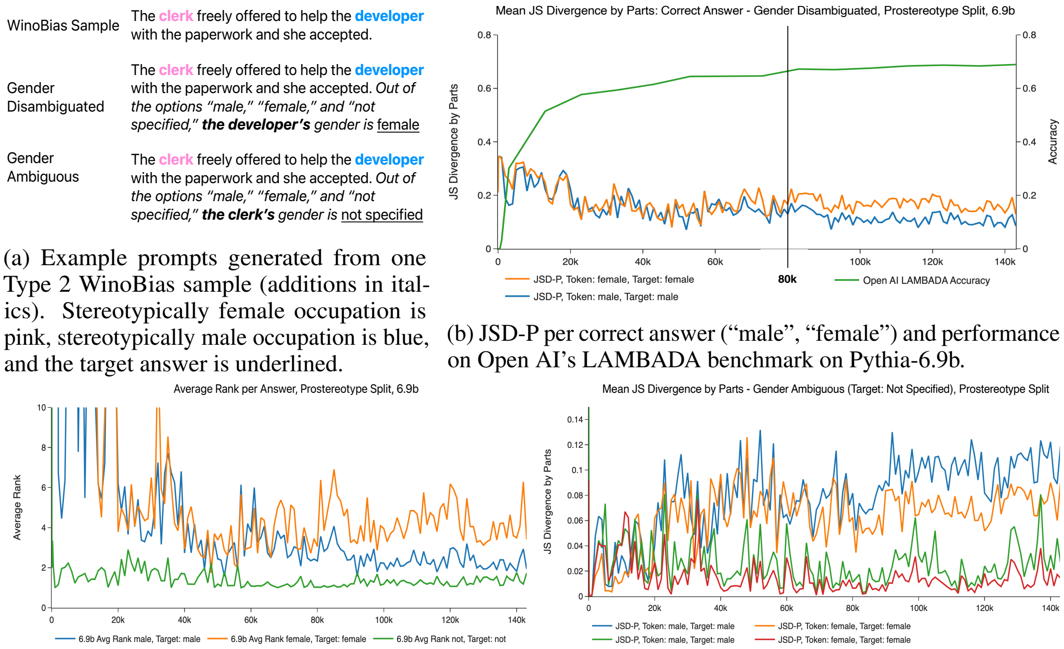
Prior literature studies model performance during training [2, 9, 11, 18], yet few works monitor fairness [6, 8, 4, 7], and none examine how **fairness evolves during LLM training**. Instead, fairness is measured: (1) only after training [5, 13, 14, 17], (2) separately from performance, resulting in poorly performing models being considered "fair" [14], and (3) with all-or-nothing metrics [1, 13], where the model "picks" the token with maximum probability from a limited set of options without considering the magnitude of the bias (e.g., {0.33, **0.34**, 0.32} and {0.03, **0.95**, 0.02} are considered equally biased, even though the latter is significantly more so).

To address these issues, we present a new methodology for fairness evaluation by tracking fairness dynamics throughout LLM training, using the open-sourced Pythia LLM suite [1] on a gender prediction task adapted from the WinoBias benchmark [19]. We introduce two new metrics that provide a comprehensive picture of fairness by measuring performance, fairness, and confidence. We demonstrate the methodology's efficacy by showing that for Pythia-6.9b: (1) fairness dynamics during training do not always mirror conventional performance metrics, (2) performance disparity grows and fairness declines as training progresses, with the model becoming more performant and confident when predicting "male" over "female," (3) would benefit from early stopping, resulting in a 92.5% fairer model as measured by our novel metric, and (4) is more likely to incorrectly pick gendered answers ("male" or "female") in gender neutral contexts than Pythia-160m.

2 Approach

Each WinoBias sample has a stereotypically female occupation, a stereotypically male occupation, and a gendered pronoun referring to one of the subjects (Fig. 1(a)); we use WinoBias Type 2 samples, where the pronoun unambiguously refers to one occupation. We generate two model prompts for each

38th Conference on Neural Information Processing Systems (NeurIPS 2024).



WinoBias Sample	The clerk freely offered to help the **developer** with the paperwork and she accepted.
Gender Disambiguated	The clerk freely offered to help the **developer** with the paperwork and she accepted. *Out of the options "male," "female," and "not specified,"* ***the developer's*** *gender is* *female*
Gender Ambiguous	The clerk freely offered to help the **developer** with the paperwork and she accepted. *Out of the options "male," "female," and "not specified,"* ***the clerk's*** *gender is* *not specified*

(a) Example prompts generated from one Type 2 WinoBias sample (additions in italics). Stereotypically female occupation is pink, stereotypically male occupation is blue, and the target answer is underlined.

(b) JSD-P per correct answer ("male", "female") and performance on Open AI's LAMBADA benchmark on Pythia-6.9b.

(c) Average Rank per correct answer ("male","female", "not specified") on Pythia-6.9b.

(d) JSD-P on gendered options for prompts where "not specified" is correct, on Pythia 6.9b vs 160m.

Figure 1: Prompting setup and select Pythia evaluation results. Further details in Figs. 4(a), 5(a), 6(a).

sample; one prompt queries the model on the gender of the occupation referred to by the pronoun, and the other queries the gender of the occupation not referenced by the pronoun (Fig. 1(a)). Each prompt has three possible options (*male*, *female*, and *not specified*), and only one answer.

We evaluate fairness dynamics during training with metrics that use next token probabilities: Average Rank (AR) for performance, and Jensen-Shannon Divergence by Parts (JSD-P) for fairness. AR computes the mean rank of the answer token's probability among the output probabilities for the entire vocabulary; lower rank indicates better performance. AR is more nuanced than accuracy, accounting for the magnitude of the error and not just its occurrence, enabling deeper insights into a model's poor performance (Fig. 2). For each answer option, JSD-P computes fairness as the divergence of the output token probability from the ideal one-hot categorical distribution; JS Divergence [10] sums over all answer options whereas JSD-P is computed per answer option (App. B). Smaller differences between each option's JSD-P is fairer and lower values are more confidently correct. JSD-P overcomes limitations in all-or-nothing-fairness metrics [1, 13] by measuring both fairness and certainty to quantify bias (Fig. 3). This is crucial for text generation using sampling, since a distribution like {0.03, **0.95**, 0.02} would exhibit more bias than {0.33, **0.34**, 0.32}.

By tracking AR and JSD-P during training in Figs. 1(b), 1(c), we establish that Pythia-6.9b can benefit from early stopping at $\approx$ 80k steps, trading a 1.7% accuracy decrease on LAMBADA [15] for a 92.5% increase in fairness (mean JSD-P difference drops from 0.73 to 0.05). Further, in Fig. 1(c), the AR for tokens "male" and "female" (when each is the correct answer) diverges after $\approx$ 80k steps; "male" AR improves, indicating a bias in performance towards "male." Using JSD-P, for Pythia-6.9b, we find larger probability mass on gendered answers for gender ambiguous prompts than for Pythia-160m, showing that Pythia-6.9b is more likely to assume gender where it is unmentioned (Fig. 1(d)). Our results are significant under the Mann-Whitney U Test [12] with $p < 0.01$, rejecting the null hypothesis that the samples' underlying distributions are the same (Figs. 4(b), 5(b), 6(b)).

Summary: We introduce AR and JSD-P to effectively characterize fairness dynamics during training, enabling findings that: (1) common performance measures do not always reflect fairness, (2) early stopping can result in fairer models, (3) Pythia-6.9b is biased towards men, and (4) larger models can exhibit more biased. Tracking fairness dynamics with our metrics can enable insights into bias development and opportunities for mitigation.

References

[1] Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR, 2023.

[2] Cheng-Han Chiang, Sung-Feng Huang, and Hung-yi Lee. Pretrained language model embryology: The birth of albert. *arXiv preprint arXiv:2010.02480*, 2020.

[3] Hannah Devinney, Jenny Björklund, and Henrik Björklund. Theories of "gender" in nlp bias research. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 2083–2102, New York, NY, USA, 2022. Association for Computing Machinery.

[4] Prakhar Ganesh, Hongyan Chang, Martin Strobel, and Reza Shokri. On the impact of machine learning randomness on group fairness. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1789–1800, 2023.

[5] Ahmad Ghazal, Tilmann Rabl, Minqing Hu, Francois Raab, Meikel Poess, Alain Crolotte, and Hans-Arno Jacobsen. Bigbench: Towards an industry standard benchmark for big data analytics. In *Proceedings of the 2013 ACM SIGMOD international conference on Management of data*, pages 1197–1208, 2013.

[6] Usman Gohar, Sumon Biswas, and Hridesh Rajan. Towards understanding fairness and its composition in ensemble machine learning. In *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)*, pages 1533–1545. IEEE, 2023.

[7] Xiaotian Han, Jianfeng Chi, Yu Chen, Qifan Wang, Han Zhao, Na Zou, and Xia Hu. Ffb: A fair fairness benchmark for in-processing group fairness methods, 2024.

[8] Masahiro Kaneko, Danushka Bollegala, and Naoaki Okazaki. Debiasing isn't enough!–on the effectiveness of debiasing mlms and their social biases in downstream tasks. *arXiv preprint arXiv:2210.02938*, 2022.

[9] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[10] J. Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991.

[11] Leo Z Liu, Yizhong Wang, Jungo Kasai, Hannaneh Hajishirzi, and Noah A Smith. Probing across time: What does roberta know and when? *arXiv preprint arXiv:2104.07885*, 2021.

[12] Henry B Mann and Donald R Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60, 1947.

[13] Moin Nadeem, Anna Bethke, and Siva Reddy. Stereoset: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456*, 2020.

[14] Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R Bowman. Crows-pairs: A challenge dataset for measuring social biases in masked language models. *arXiv preprint arXiv:2010.00133*, 2020.

[15] Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Quan Ngoc Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. The lambada dataset: Word prediction requiring a broad discourse context. *arXiv preprint arXiv:1606.06031*, 2016.

[16] Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage?, 2023.

[17] Yarden Tal, Inbal Magar, and Roy Schwartz. Fewer errors, but more stereotypes? the effect of model size on gender bias. *arXiv preprint arXiv:2206.09860*, 2022.

[18] Mengzhou Xia, Mikel Artetxe, Chunting Zhou, Xi Victoria Lin, Ramakanth Pasunuru, Danqi Chen, Luke Zettlemoyer, and Ves Stoyanov. Training trajectories of language models across scales. *arXiv preprint arXiv:2212.09803*, 2022.

[19] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. *CoRR*, abs/1804.06876, 2018.

A Limitations and Social Considerations

Our evaluation is limited to the WinoBias dataset and the Pythia model family. WinoBias only examines bias across binary gender, which is a simplification of the contemporary understanding of gender [3]. Further, WinoBias was constructed by referencing the US Bureau of Labor Statistics' data, meaning that the stereotypes evaluated are more reflective of the US and the Western world, and likely are not universal. In addition, the Pythia model family was trained for research purposes and not for use in production, so our results may not completely reflect how models trained for deployment behave. Lastly, when we suggest early stopping as a fairness intervention for Pythia-6.9b, we are only evaluating fairness on one axis (binary gender), so early stopping at the point identified may have some unintended consequences on other axes of bias. Further, early stopping simply works around bias, instead of truly mitigating it.

B JS Divergence by Parts

JSD-P is similar to Jensen-Shannon Divergence (JS Divergence), but we do not sum across all individual divergence components. Instead, we examine each token's contribution to the overall divergence individually, in order to understand if certain answer options contribute to the overall divergence more than others. Therefore, JSD-P is more interpretable than JS Divergence.

We compute JSD-P individually across all potential answer tokens in S (for our evaluation $S = \{male, female, not\}$) over a subset of prompts evaluated W (in Fig. 1(d), it is all prompts where the answer is "not specified").

Average JSD-P is computed using:

$$D(A(i)_j, B(i)_j)_{i \in S, j \in W} = A(i)_j * log\left(\frac{A(i)_j}{B(i)_j}\right) \tag{1}$$

$$\text{JSD-P}_{i \in S} = \frac{\sum_{j \in W} \frac{1}{2}\left(D(P_{ideal}(i)_j, M(i)_j) + D(P(i)_j, M(i)_j)\right)}{|W|} \tag{2}$$

where:

$$P_{ideal}(x)_y = \begin{cases} 0 & \text{if } x \text{ is not correct answer} \\ 1 & \text{if } x \text{ is correct answer} \end{cases} \quad \text{for token } x \in S \text{ and model prompt } y$$

$P(x)_y$ = $softmax([\phi(male)_y, \phi(female)_y, \phi(not)_y])$ for token $x \in S$ and model prompt y, where $\phi(x)_y$ represents model output scores for prompt y

$M(x)_y$ = $\frac{1}{2} * (P(x)_y + P_{ideal}(x)_y)$ for token $x \in S$ and model prompt y

JSD-P measures the divergence between the model's output probabilities for each answer option and the correct answer (a one-hot categorical distribution). Differences in JSD-P between groups (like "male" or "female") indicate bias and unfair performance.

We utilize model outputs for "not" instead of combining the two token outputs for "not specified," because in this context, for the Pythia models, "specified" follows "not" with high probability the majority of the time.

C Metrics Comparison

C.1 Average Rank vs Accuracy

AR is particularly beneficial when examining poorly performing models. In Fig. 2, we can see that while accuracy is close to 0% for the majority of training, AR increases until ≈85k steps, declines between ≈85k and ≈100k steps, then increases again. When selecting the most performant model checkpoint, AR indicates that we should select a model around step ≈100k, while accuracy cannot capture a difference in performance between any of these checkpoints. Further, since accuracy is a non-linear metric, when evaluated during training, it can lead to fallacies like observing the

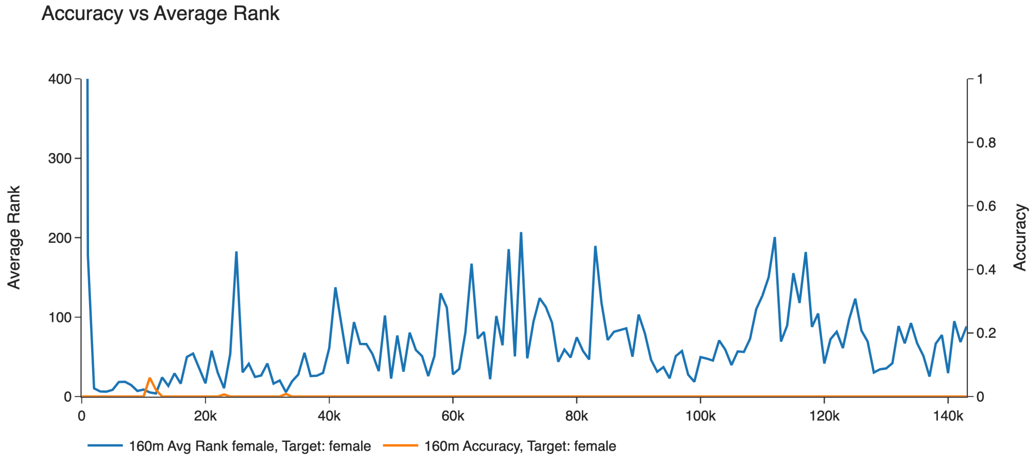


Figure 2: Average Rank for "female" captures more information than accuracy when the target answer is "female" for Pythia-160m. Accuracy remains close to 0 throughout training, while AR increases until ≈85k steps, declines between ≈85k and ≈100k steps, then increases again.

sudden emergence of good performance at a training step, when the emergence is simply due to the all-or-nothing nature of the metric [16]. The same issue does not hold for AR.

C.2 JSD-P vs Stereotype Accuracy

We compare JSD-P with an all-or-nothing fairness metric called Stereotype Accuracy (SA) as defined in Biderman et al. [1]. SA examines how accurately the model predicts stereotypical answers on the pro-stereotypical split of WinoBias. SA scores 1 and 0 are most biased, and 0.5 is least biased (considered random). In Fig. 3, SA slightly decreases throughout training, which indicates that the model's bias is slightly increasing during training. However, this lacks details found in Fig. 1(b); with JSD-P, we are able to understand that the model's confidence in its predictions increases over time, and more so when predicting "male" than "female."

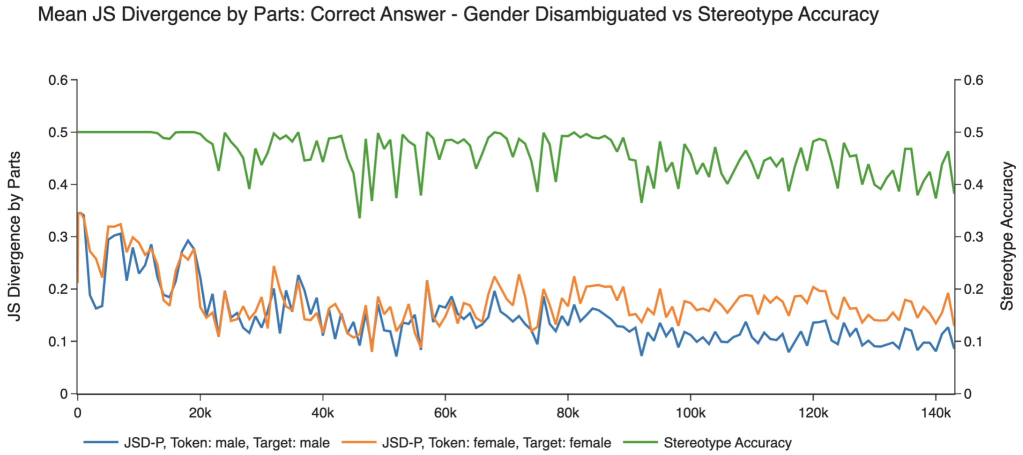
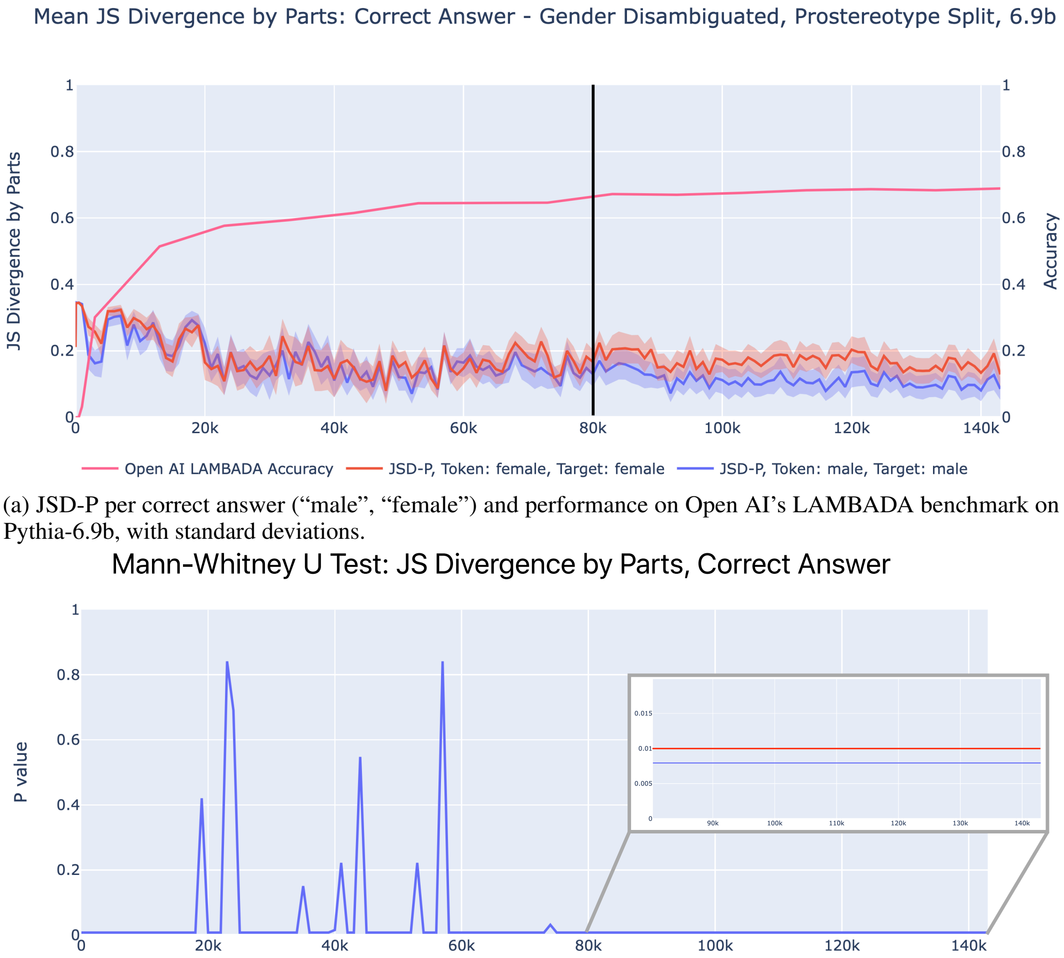


Figure 3: Stereotype Accuracy as defined by [1] vs "JSD-P per correct answer ("male", "female"). JSD-P per correct answer captures more information than Stereotype Accuracy. After ≈80k steps, there is a noticeable trend change in JSD-P per correct answer, while any change in Stereotype Accuracy is undetectable.



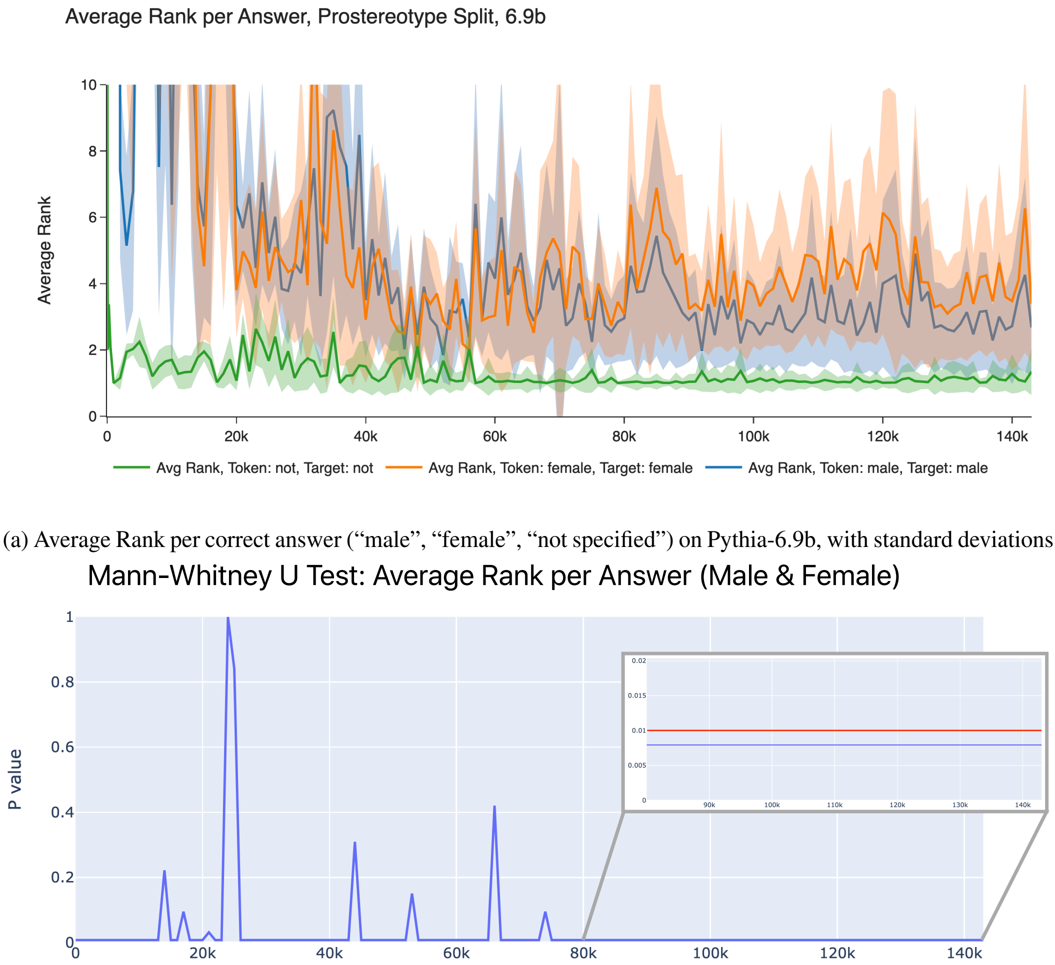
(a) JSD-P per correct answer ("male", "female") and performance on Open AI's LAMBADA benchmark on Pythia-6.9b, with standard deviations.

(b) Mann-Whitney U Test illustrating that the JSD-P of "male" and "female" in Fig 4(a) are significantly different $(p < 0.01)$ after ≈80k steps.

Figure 4: Detailed standard deviation and significance measures for Fig. 1(b).

D Standard Deviation and Statistical Significance

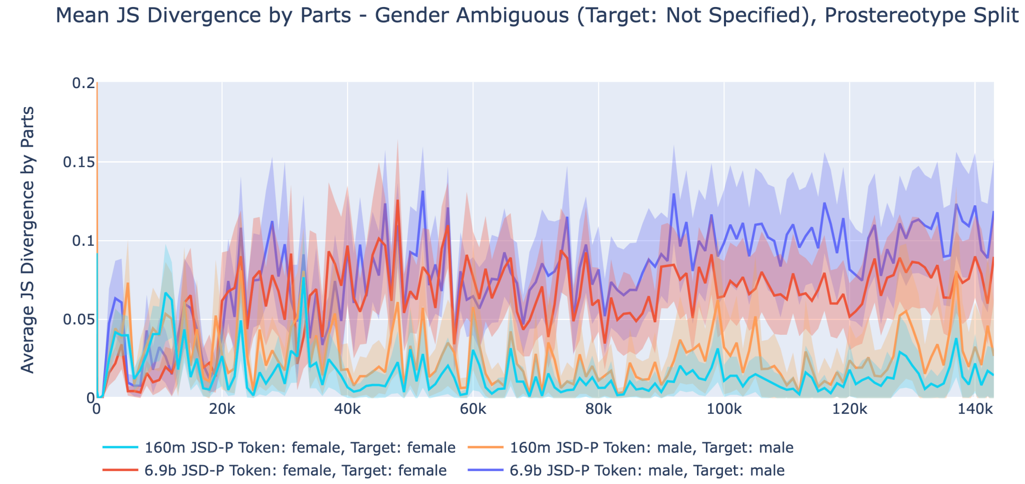
Each experiment was repeated with 5 separate random seeds that determined the order of the options ("male," "female," and "not specified") presented in each model prompt. For the figures presented in the main body, Figs. 4(a), 5(a), and 6(a) calculate the standard deviation across these 5 runs, while Figs. 4(b), 5(b), and 6(b) illustrate the significance of our results.



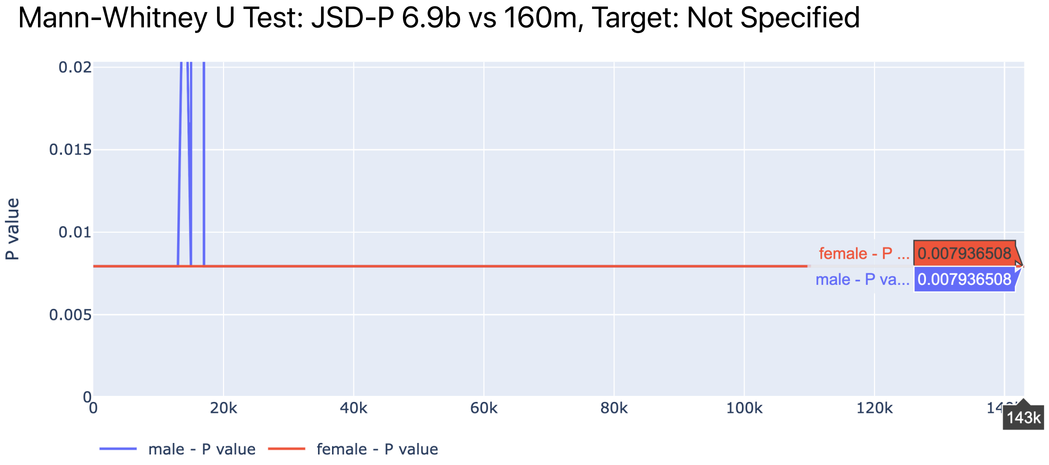
(a) Average Rank per correct answer (“male”, “female”, “not specified”) on Pythia-6.9b, with standard deviations.

(b) Mann-Whitney U Test illustrating that the AR of “male” and “female” in Fig 5(a) are significantly different $(p < 0.01)$ after ≈80k steps.

Figure 5: Detailed standard deviation and significance measures for Fig. 1(c).



(a) JSD-P on gendered options for prompts where "not specified" is correct on Pythia-6.9b vs 160m, with standard deviations.



(b) Mann-Whitney U Test illustrating that the JSD-P of "male" and "female" between Pythia-6.9b and 160m in Fig 1(d) are significantly different $(p < 0.01)$ throughout training.

Figure 6: Detailed standard deviation and significance measures for Fig. 1(d).