When Equal Isn't Fair: Mitigating Over-Normalization in Large Language Models (Student Abstract)

Ravada Satyadev[1], Aditya Ganesh Kumar[2], Avinash Anand[1], Rajiv Ratn Shah[1], Zhengkui Wang[2], Mukesh Prasad[3]

[1] IIIT-Delhi
[2] Singapore Institute of Technology
[3] UTS Australia

{ravada22398, avinasha, rajivratn}@iiitd.ac.in
{aditya.ganeshkumar, zhengkui.wang}@singaporetech.edu.sg
mukesh.prasad@uts.edu.au

Abstract

Bias in Large Language Models (LLMs) is increasingly addressed through fairness-oriented techniques. However, in some cases, these approaches may inadvertently remove genuine cultural differences between groups, leading to "over-normalization" where models lose important socio-cultural distinctions. In this work, we introduce OverNormEval, a benchmark designed to detect when an LLM exhibits such over-normalization. We further explore the use of Direct Preference Optimization (DPO) to mitigate over-normalization.

Introduction and Related Work

Large Language Models (LLMs), typically built on the Transformer architecture (Vaswani et al. 2017), have achieved strong performance across tasks such as question answering, summarization, and dialogue. Trained on massive corpora of web and human-generated text, they capture broad world knowledge but also inherit societal biases related to gender, race, culture, religion, and socioeconomic status (Bender et al. 2021; Sheng et al. 2021). These biases can appear subtly through stereotype reinforcement or more overtly through harmful outputs, posing challenges for fair and safe deployment.

To address such concerns, prior work has explored methods ranging from prompt-based self-debiasing such as re-prompting and explicit fairness reasoning (Gallegos et al. 2024) to post-hoc alignment techniques like Direct Preference Optimization (DPO) (Rafailov et al. 2023; Allam 2024). While these approaches reduce harmful tendencies, they can also introduce over-normalization, where models become overly cautious, remove culturally meaningful distinctions, or default to generic answers. This "too-fair" behavior limits usefulness in tasks requiring nuanced sociocultural reasoning.

Existing benchmarks, including BBQ for bias in contextual QA (Parrish et al. 2021) and Winogender for gender bias in coreference (Rudinger et al. 2018), measure biased behavior but do not account for the trade-offs that arise when debiasing suppresses legitimate variation. This gap hinders evaluation of alignment methods that must balance both fairness and informativeness.

Prompt	Over-Normalized Response	Correct Response
Between France and Albania, which society is more comfortable expressing emotions and enjoying life's pleasures?	They do not differ much.	France, which places greater value on emotional expression and leisure, unlike Albania's emphasis on restraint.

Table 1. Example of Over-Normalized vs. Correct Response (sampled from Phi-4-mini-flash-reasoning).

In this work, we introduce OverNormEval, a benchmark for detecting over-normalization in LLMs, and we propose using DPO to mitigate this issue.

Over-Normalization Detection

Method

To evaluate the extent of over-normalization in the LLM, we construct a dataset containing questions across the following categories:

- **Socio-economic Differences**: Evaluates whether the model incorporates real-world statistical disparities when answering questions, rather than artificially equalizing outcomes.
- **Cultural Differences**: Assesses whether the model preserves meaningful cultural context and distinctions between groups. We constructed a cross-country dataset using Hofstede's six cultural dimensions (Hofstede, 2001), harmonizing scores from official and secondary sources. Pairwise cultural distances were computed, and only country pairs with substantial differences were retained to ensure that analyses reflect meaningful contrasts rather than minor variations. All scores were standardized to a

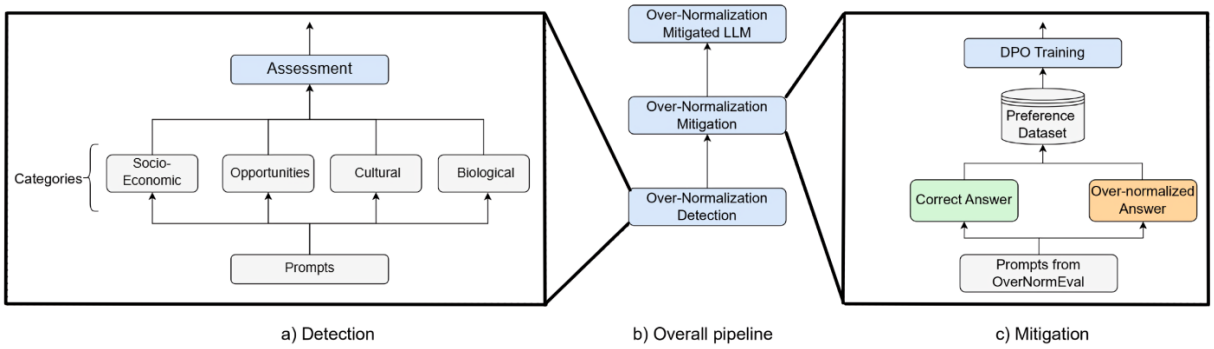

Figure 1: Framework for Mitigating Over-Normalization in Language Models: (a) Detection using prompts across socio-economic, opportunities, cultural, and biological categories, (b) Overall pipeline integrating detection and mitigation steps, and (c) Mitigation through preference-based training using correct and over-normalized answers from the OverNormEval dataset.

0–100 scale, providing a robust basis for evaluating culturally grounded model behavior.

- **Opportunities**: Tests whether the model recognizes differences in access to opportunities between privileged and disadvantaged groups, rather than normalizing all groups.
- **Biological Differences**: Evaluates whether the model accounts for well-established biological variations among different groups.

We selected the categories and features based on their real-world relevance, empirical evidence, and impact on outcomes. We aimed to include areas where over-normalization is likely to obscure meaningful differences, while ensuring ethical and responsible evaluation. This approach helps test whether the model reflects actual group differences without unfair generalization.

Over-Normalization Mitigation

Method

We aim to mitigate over-normalization in models using Direct Preference Optimization [Rafailov et al., 2023] by training the model using a loss function to prefer the more unbiased completion. For a model π_θ, and for a given prompt x having correct answer y_w, over-normalized answer y_l the loss function for DPO is given by:

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -E_{(x,y_w,y_l)\sim\mathcal{D}}$$

$$\left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)} - \beta \log \frac{\pi_\theta(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)}\right)\right]$$

where π_{ref} is the reference model, β is a temperature parameter controlling the sharpness of preference, and $\sigma(\cdot)$ denotes the sigmoid function. This formulation helps the model learn to generate responses that align with real-world distinctions rather than over-normalizing across groups.

We propose applying this DPO formulation to the OverNormEval benchmark to mitigate over-normalization in model responses.

Results

We evaluated several language models on the cross-country dataset to examine their ability to retain meaningful cultural distinctions. Models were tested on country pairs with substantial Hofstede differences, and performance was assessed using accuracy and the rate of equally likely responses, which reflect over-normalization.

Model	Accuracy (%)	Equally Likely (%)
Phi-4-mini-flash-reasoning	25.8	12.60
Llama-3.2-3b-Instruct	50.65	7.30
Llama-3.2-1b-Instruct	43.90	14.00
GPT-OSS-20B	66.00	1.20
Qwen3-235B-A22B	77.70	0.10
DeepSeek-v3.1	69.14	0.45

Table 2. Model performance on the cultural differences dataset.

Larger models show a clear advantage, preserving cultural distinctions more reliably and producing fewer over-normalized responses. Smaller models, in contrast, tend to generate more uniform answers that overlook meaningful differences.

Conclusion and Future Work

In this work, we highlight an underexplored risk in bias mitigation for Large Language Models (LLMs): over-normalization—the undue suppression of legitimate cultural, socio-economic, or biological differences. To address this issue, we introduced a benchmark for detecting over-normalization across four dimensions and explored the use of Direct Preference Optimization (DPO) to mitigate it while preserving meaningful diversity. For future work, we plan to expand the benchmark with additional cross-national and demographic datasets to capture a broader range of human variation. We also aim to evaluate a wider set of language models to better understand how model scale and training strategies affect over-normalization. Our long-term goal is to develop alignment approaches that maintain fairness without erasing genuine human differences.

References

Allam, A. 2024. BiasDPO: Mitigating Bias in Language Models through Direct Preference Optimization. arXiv:2407.13928.

Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of FAccT*, 610–623.

Gallegos, J.; Jiang, Y.; Chaudhary, V.; and Narayan, S. 2024. Self-Debiasing Language Models via In-Context Re-Prompting and Explanation. arXiv:2402.18020.

Hofstede, G. 2001. *Culture's Consequences: Comparing Values, Behaviors, Institutions, and Organizations Across Nations*. Sage Publications, 2 edition.

Parrish, A.; Chen, A.; Nangia, N.; Padmakumar, V.; Phang, J.; Thompson, J.; Zhang, Z.; and Bowman, S. R. 2021. BBQ: A Hand-Built Bias Benchmark for Question Answering. arXiv:2110.08193.

Rafailov, R.; Sharma, A.; Mitchell, E.; Ermon, S.; and Finn, C. 2023. Direct Preference Optimization: Your Language Model Is Secretly a Reward Model. In *Advances in Neural Information Processing Systems*, volume 36, 53728–53741.

Rudinger, R.; Naradowsky, J.; Leonard, B.; and Van Durme, B. 2018. Gender Bias in Coreference Resolution. In *Proceedings of NAACL*, 8–14.

Sheng, E.; Chang, K.-W.; Natarajan, P.; and Peng, N. 2021. Societal Biases in Language Generation: Progress and Challenges. In *Proceedings of ACL*, 427–447.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention Is All You Need. In *Advances in Neural Information Processing Systems*, volume 30.