Following Stereotypes in LLMs' Learning Trajectories

Michele Dusi[a,b,*]

[a]University of Brescia
[b]Sapienza University of Rome
ORCID (Michele Dusi): https://orcid.org/0000-0002-6564-3096

Abstract. This paper outlines my PhD research in Artificial Intelligence, which investigates the presence and origins of bias in language models. In the early stages of the project I developed methods to visualize, measure, and statistically validate stereotypes in model embeddings. More recently, my focus has shifted to the training process itself, with the goal of tracing how biases propagate from data to model behavior. Through controlled experiments on models fine-tuned with ideologically distinct corpora, I analyze learning trajectories and the emergence of distorted knowledge. Adopting a white-box perspective, this work aims to make large language models more transparent, interpretable, and ultimately more trustworthy over time.

1 Introduction and Research Vision

This document presents the current stage of my PhD research in Artificial Intelligence, with a specific focus on the critical analysis of language model behavior — particularly regarding the presence of stereotypes, prejudices, and, to use a now-ubiquitous but still meaningful term, **bias**.

The paper is structured into two main parts. The first (§2) offers an overview of the work conducted so far, highlighting the logical thread that connects each step and its contribution to my scientific development. The second (§3) focuses on ongoing research directions, outlining the final phase of my doctoral project.

During the first two years, my research aimed to understand how and when language models exhibit **preferential or discriminatory attitudes** toward specific social groups. I began by designing a visualization protocol [9] to make the presence of bias in model embeddings more explicit. This was followed by the development of a quantitative method [9], designed to be both flexible and agnostic to the model architecture and the type of social groups under analysis.

Currently, my work investigates the **origins of bias**, with particular attention to the role that training data — both from the pretraining and fine-tuning phases — plays in shaping model behavior. I am exploring the possibility of a causal relationship between the composition of training datasets and the model's responses on socially sensitive or controversial topics. To do this, I analyze the behavior of models trained on politically oriented corpora, evaluating their outputs in relation to historically marginalized groups.

* Corresponding Author. Email: michele.dusi@unibs.it.

2 Research Achievements

2.1 Bias Visualization

In our initial work [9], my coauthors and I introduced a procedure to visualize distortions in how language models represent concepts related to socially stereotyped categories. The goal was to highlight how, at a statistical level, words associated with certain protected groups are systematically represented differently from others — not by chance, but as a reflection of recurring linguistic patterns in the training data, which themselves mirror pre-existing social stereotypes.

The method is based on analyzing the distribution of semantic vectors corresponding to different categories and draws inspiration from previous works, such as [4], while introducing key innovations. Our approach is applicable to contextual models (like BERT [7]) and requires only a minimal set of terms to produce interpretable results. This makes it suitable not only for technical analysis but also for more accessible communication, for example, towards non-expert users interested in understanding model behavior.

2.2 Bias Quantification

In the follow-up work [10], we expanded and deepened the analysis initiated with bias visualization by proposing a procedure to quantify distortions present in language model embeddings. The goal was to transform qualitative observations into a robust, statistically validated measure capable of objectively capturing the presence of stereotypes in the semantic representation of certain social categories.

The underlying hypothesis is that latent components within word vector representations unintentionally — but systematically — **encode sensitive information** such as gender, ethnicity, or religion. To detect these components, we employed a Support Vector Machine (SVM) trained to distinguish the protected attribute within embeddings. Next, we tested the classifier on a second set of words associated with common stereotypes linked to the protected attribute. If the SVM can correctly separate these words as well, there is a **correlation** between the protected attribute and other stereotypically associated traits (Figure 1) To quantify this correlation, we used Cramér's V metric [6] and we tested the method on several language models known to exhibit gender, ethnicity, and religion biases, demonstrating its effectiveness in detecting stereotypical distortions within embeddings.

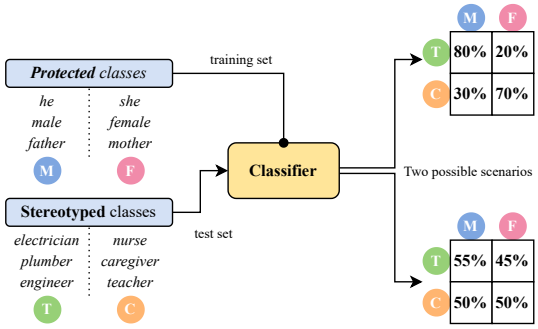

Figure 1. Schema of the **bias quantification procedure**. The classifier is trained on the protected words, and then tested on the stereotyped words. The more the outcome is unbalanced w.r.t. the ideal perfectly balanced outcome, the more correlation we observe between the attributes.

3 Current and Future Work

In the final phase of my doctoral research, I am focusing on investigating how biases transfer from data to models. In this section, I will briefly outline the research framework, starting from the key questions I intend to explore and moving towards a preliminary review of the relevant literature and the current state of the field.

3.1 Research Questions

Previous studies have treated language models as "frozen" artifacts: biases were visualized or measured only *ex post*, without investigating how these biases were introduced into the model. The final phase of my PhD reverses this perspective: I aim to observe the dynamic transfer between data and internal representations, with the goal of preventing — not just diagnosing — undesirable distortions.

The guiding questions of this project are:

1. What is the **current state of the art** regarding studies that link training phases to the final behavior of models?
2. To what extent can existing techniques be adapted to **trace the emergence of stereotypes** (such as gender, ethnicity, religion, etc.) throughout training?
3. What practical modifications are needed to transform these techniques into **effective auditing tools during model development**?
4. What impact would bias tracing have on the **safety and user experience** of those interacting daily with large-scale language models?

3.2 Related Work and State of the Art

A preliminary survey reveals that the literature on training dynamics of large language models (LLMs) remains fragmented. Some relevant strands include:

- **Bias tracking from data to model** — studies whose purpose is to approach bias in LLMs with a causal perspective [18, 11, 3, 16].
- **Studies on memorization and knowledge emergence** — studies that research how training influences the model's knowledge [12, 17, 14, 5].
- **Machine unlearning** — methods to "forget" portions of datasets and measure their impact on model behavior [8, 20, 2, 13, 19].

While these works provide useful building blocks, there is still a lack of a unified framework that causally connects dataset composition, to bias formation, to model output.

3.3 Bias Tracing (Work in Progress)

The experimental strategy proceeds along two complementary paths:

1. **Training Trajectory Analysis**.
   - *Starting model:* Pythia GPT-NeoX [1] (160 million parameters).
   - *Dataset:* BIGNEWS [15], two parallel corpora of US news articles labeled by political orientation ("left" / "right").
   - *Objective:* to track, checkpoint by checkpoint, shifts in semantic representations and world knowledge induced by exposure to ideologically distinct content.
2. **Targeted Fine-tuning and Sensitivity Testing**.
   - Starting from a pre-trained LLM, we apply controlled fine-tuning on "biased" subsets.
   - For each model version, we measure the emergence of bias using the SVM + Cramér's V framework developed in prior work [10].

In the coming months, tests will be extended to other sensitive attributes (gender, religion) and larger models, aiming to produce dataset-aware training guidelines and early-warning mechanisms for bias.

4 Conclusions and Perspectives

Throughout my doctoral journey, I have tackled the issue of stereotypes in language models through a systematic, *white-box* approach. From the start, I chose to investigate models not as black boxes to be queried, but as transparent systems whose **internal representations** can and must be analyzed, understood, and ultimately improved.

The guiding thread of my work is a critical yet constructive vision of the current Natural Language Processing landscape. Language models are extraordinarily powerful tools, but **inherently non-neutral**: they incorporate and amplify biases, hallucinations, and distorted reasoning patterns, often directly reflecting the data on which they are trained.

I do not believe the answer to these challenges lies in rejecting the technology or seeking purely technical fixes. Rather, it is essential to make these models more accessible, measurable, and interpretable. Recognizing, quantifying, and transparently communicating their behavioral distortions — as well as tracing their origins back to the training data — is a crucial step toward a fairer, more aware, and more reliable NLP.

Acknowledgements

This work is carried out while the author, Michele Dusi, is enrolled in the *Italian National Doctorate on Artificial Intelligence* run by Sapienza University of Rome in collaboration with the University of Brescia.

References

[1] A. Andonian, Q. Anthony, S. Biderman, S. Black, P. Gali, L. Gao, E. Hallahan, J. Levy-Kramer, C. Leahy, L. Nestler, K. Parker, M. Pieler, J. Phang, S. Purohit, H. Schoelkopf, D. Stander, T. Songz, C. Tigges, B. Thérien, P. Wang, and S. Weinbach. GPT-NeoX: Large Scale Autoregressive Language Modeling in PyTorch, 9 2023. URL https://www.github.com/eleutherai/gpt-neox.

[2] A. Blanco-Justicia, N. Jebreel, B. Manzanares-Salor, D. Sánchez, J. Domingo-Ferrer, G. Collell, and K. Eeik Tan. Digital forgetting in large language models: a survey of unlearning methods. *Artificial Intelligence Review*, 58(3), Jan. 2025. ISSN 1573-7462. doi: 10.1007/s10462-024-11078-6. URL http://dx.doi.org/10.1007/s10462-024-11078-6.

[3] S. L. Blodgett, S. Barocas, H. Daumé III, and H. Wallach. Language (technology) is power: A critical survey of "bias" in NLP. In D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.485. URL https://aclanthology.org/2020.acl-main.485/.

[4] T. Bolukbasi, K. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In D. D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 4349–4357, July 2016.

[5] H. Chang, J. Park, S. Ye, S. Yang, Y. Seo, D.-S. Chang, and M. Seo. How do large language models acquire factual knowledge during pretraining? In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=TYdzj1EvBP.

[6] H. Cramér. *Mathematical methods of statistics*, page 575. Princeton: Princeton University Press, 1946. ISBN 0-691-08004-6.

[7] J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, and T. Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, June 2019.

[8] O. Dige, D. Arneja, T. F. Yau, Q. Zhang, M. Bolandraftar, X. Zhu, and F. K. Khattak. Can machine unlearning reduce social bias in language models? In F. Dernoncourt, D. Preoţiuc-Pietro, and A. Shimorina, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 954–969, Miami, Florida, US, Nov. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-industry.71. URL https://aclanthology.org/2024.emnlp-industry.71/.

[9] M. Dusi, N. Arici, A. E. Gerevini, L. Putelli, and I. Serina. Graphical identification of gender bias in bert with a weakly supervised approach. In *NL4AI 2022: Sixth Workshop on Natural Language for Artificial Intelligence*. CEUR-WS, Nov. 2022. URL http://sag.art.uniroma2.it/NL4AI/wp-content/uploads/2022/11/paper16.pdf.

[10] M. Dusi, N. Arici, A. Emilio Gerevini, L. Putelli, and I. Serina. Discrimination bias detection through categorical association in pre-trained language models. *IEEE Access*, 12:162651–162667, 2024. doi: 10.1109/ACCESS.2024.3482010.

[11] S. Fulay, W. Brannon, S. Mohanty, C. Overney, E. Poole-Dayan, D. Roy, and J. Kabbara. On the relationship between truth and political bias in language models. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9018, Miami, Florida, USA, Nov. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.508. URL https://aclanthology.org/2024.emnlp-main.508/.

[12] E. Gjinika, N. Arici, A. E. Gerevini, L. Putelli, and I. Serina. An analysis on how pre-trained language models learn different aspects. In *5th Italian Workshop on Explainable Artificial Intelligence, XAI.it 2024*, volume 3839, page 28 – 41, 2024. URL https://iris.unibs.it/handle/11379/619127.

[13] L. Han, H. Huang, D. Scheinost, M.-A. Hartley, and M. R. Martínez. Unlearning information bottleneck: Machine unlearning of systematic patterns and biases, 2024. URL https://arxiv.org/abs/2405.14020.

[14] D. D. Leybzon and C. Kervadec. Learning, forgetting, remembering: Insights from tracking LLM memorization during training. In *The 7th BlackboxNLP Workshop*, 2024. URL https://openreview.net/forum?id=gvwo8txXeO.

[15] Y. Liu, X. F. Zhang, D. Wegsman, N. Beauchamp, and L. Wang. POLITICS: Pretraining with same-story article comparison for ideology prediction and stance detection. In M. Carpuat, M.-C. de Marneffe, and I. V. Meza Ruiz, editors, *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1354–1374, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-naacl.101. URL https://aclanthology.org/2022.findings-naacl.101/.

[16] M. Thaler, A. Köksal, A. Leidinger, A. Korhonen, and H. Schütze. How far can bias go? – tracing bias from pretraining data to alignment, 2024. URL https://arxiv.org/abs/2411.19240.

[17] K. Tirumala, A. H. Markosyan, L. Zettlemoyer, and A. Aghajanyan. Memorization without overfitting: analyzing the training dynamics of large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

[18] F. Wang, W. Mo, Y. Wang, W. Zhou, and M. Chen. A causal view of entity bias in (large) language models. In H. Bouamor, J. Pino, and K. Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15173–15184, Singapore, Dec. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.1013. URL https://aclanthology.org/2023.findings-emnlp.1013/.

[19] Y. Xu. Machine unlearning for traditional models and large language models: A short survey, 2024. URL https://arxiv.org/abs/2404.01206.

[20] C. Yu, S. Jeoung, A. Kasi, P. Yu, and H. Ji. Unlearning bias in language models by partitioning gradients. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6032–6048, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.375. URL https://aclanthology.org/2023.findings-acl.375/.