# Probing of Language Model Representations for Biases


Vansh Gupta — guptav
Noah Pfenniger — pfnoah
Susanna Di Vita — sdivita
Tamás Visy — tavisy


May 3, 2024

## 1 Introduction

As the field of Natural Language Processing (NLP) has advanced in recent years, the status quo has shifted from recurrent models such as the LSTM [HS97] to transformer-based [VSP+23] pretrained language models (PLMs) such as BERT [DCLT19] to finally the large language models (LLMs) such as GPT [OAA+24]. Although these models offer high efficiency and usefulness, their limitations, including the learning and perpetuating harmful biases, must not be ignored and require attention. Often, these biases are not only embedded in the representations of language models but also carry over into downstream tasks, resulting in disparate treatment of various socio-demographic groups [BLO+21, SA21, SSZ19, VSW22]. This work investigates such biases in language model representations through probing. Our contributions are as follows:

- We identify biases in the representations of LMs across various socio-demographic groups, namely, religion, race, and gender, employing non-binary association tests to provide a nuanced analysis.
- We enhance the interpretability and robustness of our probe by training it on diverse datasets that closely mirror the characteristics of downstream applications while increasing its utility for the second step of our methodology.
- We introduce a novel bias ranking system for various LMs, utilizing previously unexplored evaluation metrics for group fairness.

## 2 Related Works

### 2.1 Biases in Language Models and Group Fairness

As language models become increasingly integrated into everyday applications, their potential to propagate/amplify existing biases has prompted significant scientific attention [GVWP, DARW+19, WWT+20, NBR20, GAK+21, NV23, AO21, NVBB20]. This concern is addressed through the lens of group fairness, where researchers aim to understand and mitigate biases against specific demographic groups within the models' outputs.

Our evaluation of such biases follows directly from a very recent work [MSGA24] on social bias probing. Here, the authors argue that the binary association tests on small datasets predicated on a single "ground truth" regarding stereotypical statements have constrained the depth of analysis and oversimplified the intricate nature of social identities and their linked stereotypes.

### 2.2 Probing Techniques in Machine Learning Models

Probing techniques have become a cornerstone in the interpretability of machine learning models, particularly in understanding how deep neural networks encode information. These techniques involve using auxiliary classifiers, *probes*, to extract and analyze representations learned by models during training [AB18, AKB+17]. The primary goal is to determine if specific types of information are captured in the models' representations. Previous works have extensively focused on linguistic, semantic and syntactic properties [CFX+21, ZB18, NRS+18, CKL+18, BG17, HM19, TXC+19, PNZY18, LRF20, Ett20, JSS19, HL19, VPL+20], while more recent works have shifted towards evaluating the models' world knowledge and comprehensive capacities [PRL+19, DDH+21, JXAN20, ZFC21, BCL+23, LMZ+23].

## 3 Methodology

Our approach combines LABDet's [KYA+23] probing technique with aspects of the SoFa [MSGA24] methodology to detect socio-demographic biases for religious, racial and gender groups effectively.

We adopt LABDet's two-step training process, balancing simplicity and effectiveness in probe design. Using the encodings of a fixed PLM, we first train a sentiment classifier (probe). This classifier is trained on a sentiment dataset specifically curated to exclude any references to the demographic groups under study, ensuring the focus remains purely on sentiment analysis. To enhance the robustness of the classifier, we utilize diverse datasets that prepare it to handle out-of-distribution data, which is critical for the evalua-

tion phase.
In the second step, we apply what we term *minimal groups* of sentences corresponding to different socio-demographic groups to examine how these groups elicit varying sentiments. This approach allows us to detect subtle biases in sentiment associations that might emerge when the classifier is applied to specific group-related contexts. This structured yet flexible approach to training and evaluation provides a rigorous assessment of how different demographic groups are represented and potentially biased within language models.
To compare these biases among different groups and LMs, we will use the following evaluation metrics:

- **Demographic Parity**, to ensure that the model predictions are equally accurate across the analyzed demographic groups
- **Equalized Odds**, to ensure the model's true positive rates and false positive rates are equal across different demographic groups

# 4 Dataset

The project requires two different kinds of datasets to facilitate the processes outlined in Section 3: First, a sentiment analysis dataset devoid of socio-demographic identifiers is required to train the probe. Subsequently, a second dataset involving these groups will be utilized to evaluate the LMs for biases.

## 4.1 Probe Training

We use existing, real-world sentiment analysis datasets like the Stanford Sentiment Treebank [SPW+13], the TweetEval dataset [BCCEAN20] and the Multi-Domain Sentiment Dataset [BDP07] to train the probe. LABDet deliberately avoids the existing corpora, instead favouring a smaller, synthetically generated dataset crafted from templates to circumvent inherent biases. Nonetheless, we assert that training on real-world data will better align the probe with downstream applications. To address potential biases inherent in this data, cleaning measures such as filtering via string matching or Named Entity Recognition will be necessary. This approach ensures that the probe is rigorously tested in the context where the LM is utilized, thereby enabling a comprehensive assessment of biases.

## 4.2 Minimal groups

This dataset is constructed using templates that generate sentences with various subjects and adjectives inserted. We plan to utilize resources such as the Equity Evaluation Corpus (EEC) [KM18] or HONEST [NBH21] to explore positive and negative adjectives in sentences, in addition to neutral statements. Specifically, we will draw neutral statement templates from LABDet and complement them with positive and negative templates sourced from EEC or HONEST. This strategy ensures a comprehensive coverage of sentiment variations for thorough evaluation.

## 4.3 Tools

We employ a classifier like a Support Vector Machine (SVM) as our probe. We are also exploring the option of employing more complex architectures, such as a Deep Neural Network, to compare the effectiveness of different probes.
The language models under evaluation include Glove, BERT, Llama-2 7b, GPT-2, etc. We will select diverse models from these to represent a spectrum of architectures and capabilities, allowing for a comprehensive performance assessment across various tasks and contexts.

# 5 Resources

To generate training data for the probe, it is essential to first process a sentiment analysis dataset through each language model (LM) under evaluation, capturing the resulting encodings. Our estimate for this task is 5 (LMs) $\times$ 10 GPU hours. Subsequently, the probe itself will be trained using these encodings as inputs and the original sentiment labels as outputs. While the duration of each training session is expected to be relatively short, the need to iterate through this process multiple times for optimal performance suggests planning for 10 runs at 5 GPU hours each. Finally, for the probing of LMs we would need to run them again on the test dataset. This will likely again require 5 $\times$ 10 GPU hours. Considering all tasks, the computation time commitment is projected to be roughly 150 hours.

# 6 Expected Results

We aim to effectively identify and measure biases within language models across various datasets and linguistic contexts, including gender, race, and religion. By assembling datasets that contain biased training data reflecting societal stereotypes, we aim to devise practical strategies to detect these biases and support the development of robust models.

# References


[AB18] Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes, 2018.

[AKB+17] Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks, 2017.

[AO21] Jaimeen Ahn and Alice Oh. Mitigating language-dependent ethnic bias in bert. *arXiv preprint arXiv:2109.05704*, 2021.

[BCCEAN20] Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In Trevor Cohn, Yulan He, and Yang Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online, November 2020. Association for Computational Linguistics.

[BCL+23] Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*, 2023.

[BDP07] John Blitzer, Mark Dredze, and Fernando Pereira. Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In Annie Zaenen and Antal van den Bosch, editors, *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 440–447, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[BG17] Yonatan Belinkov and James Glass. Analyzing hidden representations in end-to-end automatic speech recognition systems. *Advances in Neural Information Processing Systems*, 30, 2017.

[BLO+21] Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *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)*, pages 1004–1015, Online, August 2021. Association for Computational Linguistics.

[CFX+21] Boli Chen, Yao Fu, Guangwei Xu, Pengjun Xie, Chuanqi Tan, Mosha Chen, and Liping Jing. Probing bert in hyperbolic spaces. *arXiv preprint arXiv:2104.03869*, 2021.

[CKL+18] Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. What you can cram into a single vector: Probing sentence embeddings for linguistic properties. *arXiv preprint arXiv:1805.01070*, 2018.

[DARW+19] Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman 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*, pages 120–128, 2019.

[DCLT19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[DDH+21] Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. Knowledge neurons in pretrained transformers. *arXiv preprint arXiv:2104.08696*, 2021.

[Ett20] Allyson Ettinger. What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48, 2020.

[GAK+21] Aparna Garimella, Akhash Amarnath, Kiran Kumar, Akash Pramod Yalla, N Anandhavelu, Niyati Chhaya, and Balaji Vasan Srinivasan. He is very intelligent, she is very beautiful? on mitigating social biases in language modelling and generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4534–4545, 2021.

[GVWP] Vipul Gupta, Pranav Narayanan Venkit, Shomir Wilson, and Rebecca J Passonneau. Sociodemographic bias in language models: A survey and forward path.

[HL19] John Hewitt and Percy Liang. Designing and interpreting probes with control tasks, 2019.

[HM19] John Hewitt and Christopher D Manning. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, 2019.

[HS97] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, nov 1997.

[JSS19] Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. What does bert learn about the structure of language? In *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*, 2019.

[JXAN20] Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438, 2020.

[KM18] Svetlana Kiritchenko and Saif M. Mohammad. Examining gender and race bias in two hundred sentiment analysis systems. *CoRR*, 2018.

[KYA+23] Abdullatif Köksal, Omer Yalcin, Ahmet Akbiyik, M Kilavuz, Anna Korhonen, and Hinrich Schuetze. Language-agnostic bias detection in language models with bias probing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12735–12747, 2023.

[LMZ+23] Tianle Li, Xueguang Ma, Alex Zhuang, Yu Gu, Yu Su, and Wenhu Chen. Few-shot in-context learning for knowledge base question answering. *arXiv preprint arXiv:2305.01750*, 2023.

[LRF20] Jindřich Libovickỳ, Rudolf Rosa, and Alexander Fraser. On the language neutrality of pre-trained multilingual representations. *arXiv preprint arXiv:2004.05160*, 2020.

[MSGA24] Marta Marchiori Manerba, Karolina Stańczak, Riccardo Guidotti, and Isabelle Augenstein. Social bias probing: Fairness benchmarking for language models, 2024.

[NBH21] Debora Nozza, Federico Bianchi, and Dirk Hovy. "HONEST: Measuring hurtful sentence completion in language models". In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2398–2406, Online, June 2021. Association for Computational Linguistics.

[NBR20] Moin Nadeem, Anna Bethke, and Siva Reddy. Stereoset: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456*, 2020.

[NRS+18] Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. Stress test evaluation for natural language inference. *arXiv preprint arXiv:1806.00692*, 2018.

[NV23] Pranav Narayanan Venkit. Towards a holistic approach: Understanding sociodemographic biases in nlp models using an interdisciplinary lens. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 1004–1005, 2023.

[NVBB20] 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.

[OAA+24] OpenAI, Josh Achiam, Steven

Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas

Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024.

[PNZY18] Matthew E Peters, Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih. Dissecting contextual word embeddings: Architecture and representation. *arXiv preprint arXiv:1808.08949*, 2018.

[PRL+19] Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*, 2019.

[SA21] Karolina Stanczak and Isabelle Augenstein. A survey on gender bias in natural language processing, 2021.

[SPW+13] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In David Yarowsky, Timothy Baldwin, Anna Korhonen, Karen Livescu, and Steven Bethard, editors, *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA, October 2013. Association for Computational Linguistics.

[SSZ19] Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. Evaluating gender bias in machine translation. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy, July 2019. Association for Computational Linguistics.

[TXC+19] Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, et al. What do you learn from context? probing for sentence structure in contextualized word representations. *arXiv preprint arXiv:1905.06316*, 2019.

[VPL+20] Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. Probing pretrained language models for lexical semantics. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240, Online, November 2020. Association for Computational Linguistics.

[VSP+23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

[VSW22] Pranav Narayanan Venkit, Mukund Srinath, and Shomir Wilson. A study of implicit bias in pretrained language models against people with disabilities. In Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na, editors, *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1324–1332, Gyeongju, Republic of

Korea, October 2022. International Committee on Computational Linguistics.

[WWT+20] Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, Ed Chi, and Slav Petrov. Measuring and reducing gendered correlations in pre-trained models. *arXiv preprint arXiv:2010.06032*, 2020.

[ZB18] Kelly W Zhang and Samuel R Bowman. Language modeling teaches you more syntax than translation does: Lessons learned through auxiliary task analysis. *arXiv preprint arXiv:1809.10040*, 2018.

[ZFC21] Zexuan Zhong, Dan Friedman, and Danqi Chen. Factual probing is [mask]: Learning vs. learning to recall. *arXiv preprint arXiv:2104.05240*, 2021.