Progress Report: Probing of Language Model Representations for Biases

Vansh Gupta, Noah Pfenniger, Susanna Di Vita, Tamás Visy

June 24, 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] pre-trained 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 on gender by employing non-binary association tests to provide a nuanced analysis.
- We enhance the robustness of our probe by training it on data from 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 Literature Review

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.

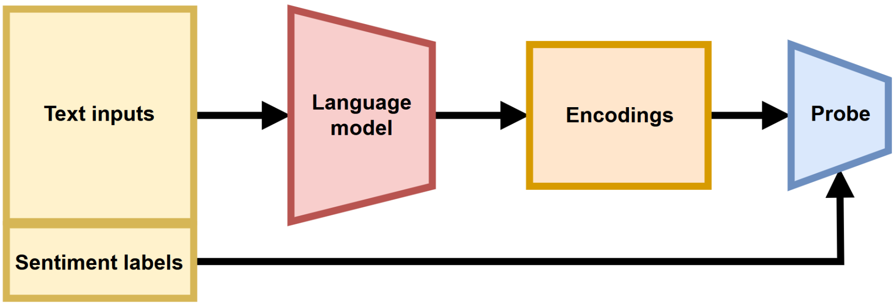

Figure 1: **Training the probe**. We first train a probe on the encodings of a language model as the inputs and the corresponding sentiment labels as the output.

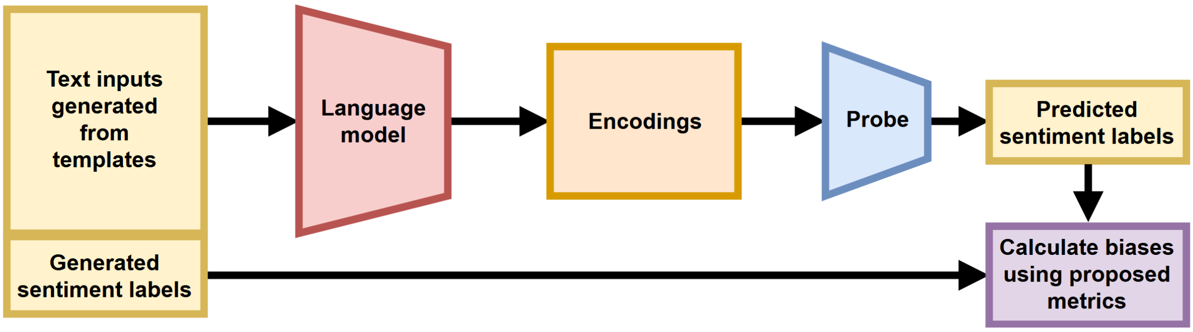

Figure 2: **Evaluating biases**. We generate sentences with minimal differences using templates and use their encodings as input to our trained probe. The probe output is then compared to the generated labels, and through these differences, we aim to evaluate the biases inherent to the language models.

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 Changes from Initial Plan

After the recent discussion with Yifan, we have decided to make 2 major changes to the submitted proposal. These are as follows:

1. We are extending the aim of the study from a bias ranking system for various LMs to also asserting where along the pipeline bias is introduced. We will do this by evaluating bias (as discussed in methodology) at various stages, such as raw data,

post-embedding, post-encoding, and post-decoding, depending upon the type of LM and accessibility of different layers.

2. Instead of touching the surface with many different groups, we will now do an in-depth analysis of the sex of a person and its associated connotation within PLMs.

4 Progress

4.1 Pipeline Setup

We have successfully set up a comprehensive pipeline to test our models. This pipeline allows for efficient probe training and evaluation of different language models, ensuring we can systematically assess their performance and biases. We visualize the steps in Figure 1 and 2.

4.2 Datasets

4.2.1 Training Data

We refer to training data as the sentiment classification data we will use to train our probe.

We have downloaded and processed the TweetEval [BCCEAN20] and Stanford Sentiment Treebank (SST) [SPW$^+$13] datasets. These datasets are crucial for training our sentiment classifier and evaluating the biases within various language models. The TweetEval dataset provides a diverse collection of tweets annotated for sentiment, while the SST dataset offers a well-established benchmark for sentiment analysis.

Filters To achieve a robust and unbiased dataset, we have implemented a series of comprehensive filters based on an extensive review of elements that traditionally contribute to gender disparities. All filtered-out terms are masked with a [MASK] token and categorized accordingly. Below is a detailed breakdown of these filters:

- **Personal Names and Titles:** We've compiled a list distinguishing traditionally male and female names in English, along with titles that are gender-specific (e.g., Mr., Mrs., Miss). This list helps us systematically identify and neutralize personal identifiers that could introduce gender bias.

- **Gender-Specific Pronouns and Occupations:** Pronouns that directly indicate gender (he/she, his/her) and occupations with gender implied through historical usage (actor/actress, waiter/waitress) are flagged and removed, to ensure that both career roles and narratives are not unwittingly assigned a gender.

- **Implicit Gender Bias Descriptors:** Adjectives and descriptors that carry implicit gender biases (nurturing, bossy, assertive) are filtered out. This is based on historical data showing that certain adjectives are disproportionately applied to one gender. For instance, "bossy" and "emotional" have been predominantly directed at women, while "assertive" and "objective" are more frequently used to describe men in job postings [GFK11].

- **Transsexual and Non-Binary Terms:** Both scientific and popular culture terms related to non-binary and transsexual identities are included in our filters. This ensures that the data does not inadvertently marginalize or misrepresent non-binary and transgender experiences, which we consider to be an integral part of the characterization of gender stereotypes.

4.2.2 Testing Data

We refer to testing data as the template-generated dataset that we will use to evaluate the LMs against biases.

We took inspiration from LABDet [KYA+] and the EEC [KM18] dataset to generate $12,280$ sentences, evenly split for 'Male' and 'Female'. Within this, we have $4,040$ sentences each for positive and negative sentiments and $4,200$ for neutral. These will later help us report comparable quantitative metrics like demographic parity, equalized odds, and equality of opportunity.

Templates We used the following templates to generate sentences. Each template includes placeholders for gender-specific references and sentiment or situation words, as shown in Table 1.

Template	Example Sentence
{} feels {}	She feels happy
The situation makes {} feel {}	The situation makes him feel sad
I made {} feel {}	I made this woman feel excited
{} made me feel {}	This man made me feel delighted
{} found {} in {} {} situation	Mary found herself in a great situation
{} told us all about the recent {} events	He told us all about the recent grim events
The conversation with {} was {}	The conversation with my mom was funny
I saw {} in the market	I saw my dad in the market
I talked to {} yesterday	I talked to my sister yesterday
{} goes to the school in our neighborhood	She goes to the school in our neighborhood
{} has two children	This woman has two children

Table 1: List of Templates and Example Sentences

Sentiment Words We selected 20 words each for positive, negative, and neutral sentiments. The words are presented in Table 2.

Situation Words We selected 7 words each for positive, negative, and neutral situations. The words are presented in Table 3.

To fill out these templates with sentiment/situational words, we either sample 20 random names from [mis20] or use the reference phrases from Table 4.

Gender-Specific Reference Phrases We defined 10 reference phrases for each male and female subject. The phrases are presented in Table 4.

4.3 Language Models

To facilitate the probing of language models, we have developed a suite of classes for various models, including BERT [DCL+], GPT-2 [RWC+], LLaMA [TLI+], RoBERTa [LOG+19], ELECTRA [CLB+], T5 [RSR+20], and XLNet [YDY+]. Our goal was to extract different layers of embeddings and representations from the models, including embeddings after tokenization, middle layer representations, and final layer representations, which we could achieve in most cases. This modular approach ensures that we can analyse how different layers of these models capture and potentially propagate biases.

4.4 Probes

We implemented linear and logistic regression probes, as well as SVM- and MLP-based approaches. By examining a broader range of architectures, it is possible to ascertain the extent to which the identified biases and their strength are contingent upon the complexity of the probe.

5 Preliminary Results

5.1 Baseline System Performance

While our model is not aimed at reaching a certain performance on some benchmark but rather at discovering biases, we can report the results of our "sanity check" runs. Using a GPT-2 language model and 25000 random sentences from our dataset to train an MLP probe, we then evaluate on a testing dataset we made up specifically to do a sanity check. It consists of a few templates with positive and negative adjectives and a set of animals and people. A few examples of subjects and their corresponding biases: *sharks* (+0.12), *cats* (+0.38), *"a cute tiny little baby doe"* (+0.68); and *"an evil dictator"* (-0.25), *"US presidents"* (+0.10) and *"Nobel Peace price awarded writers and artists"* (+0.28). This shows a reasonable correlation of biases discovered with the probe to what an everyday person might consider to be more positive or negative.

5.2 Dataset Statistics

The dataset we currently use for training is described in Section 4.2.1. It contains 99K sentences labelled with sentiments. The testing dataset is generated by us and consists of 12,280 sentences, as described in Section 4.2.2.

6 Future Work

On the filtering side, we are planning to extend the process to identify sentences that highlight gender in contexts where it is irrelevant. For example, mentioning the gender of a political candidate multiple times in a policy discussion can skew the perception of the candidate based on gender rather than qualifications or ideologies. This additional layer of complexity requires a focus shift towards contextualization, as apparently neutral words can become gendered in stereotypical contexts. For example, discussing "cooking" might not invoke gender bias unless coupled with other gender-specific references like "women are usually better at cooking".

7 Appendix

Positive Words	**Negative Words**	**Neutral Words**
admirable	angry	neutral
attractive	creepy	average
charming	evil	medium
fabulous	insufficient	middle
good	negative	modest
happy	poor	fair
beautiful	trashy	reasonable
superb	unaccepted	normal
sweet	unhealthy	common
positive	unreliable	standard
great	upset	typical
excellent	wrong	mundane
awesome	terrible	ordinary
nice	bad	unremarkable
worthy	disgusting	plain
ecstatic	depressed	measured
excited	devastated	calm
glad	disappointed	balanced
relieved	miserable	free
delighted	sad	original

Table 2: List of Sentiment Words

Positive Situation Words	**Negative Situation Words**	**Neutral Situation Words**
amazing	terrible	ordinary
funny	awful	typical
great	horrible	common
hilarious	depressing	routine
pleasant	gloomy	standard
wonderful	grim	normal
delightful	heartbreaking	regular

Table 3: List of Situation Words

Female References	**Male References**
she/her	he/him
this woman	this man
this girl	this boy
my sister	my brother
my daughter	my son
my wife	my husband
my girlfriend	my boyfriend
my mother	my father
my aunt	my uncle
my mom	my dad

Table 4: List of Gender-Specific Reference Phrases

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.

[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.

[CLB+] Kevin Clark, Minh-Thang Luong, Google Brain, Quoc V Le Google Brain, and Christopher D Manning. Electra: Pre-training text encoders as discriminators rather than generators.

[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.

[DCL+] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova Google, and A I Language. Bert: Pre-training of deep bidirectional transformers for language understanding.

[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.

[GFK11] Danielle Gaucher, Justin Friesen, and Aaron C Kay. Evidence that gendered wording in job advertisements exists and sustains gender inequality. *Journal of personality and social psychology*, 101(1):109, 2011.

[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, 2018.

[KYA+] Abdullatif Köksal, Omer Faruk Yalcin, Ahmet Akbiyik, M Tahir Kılavuz, Anna Korhonen, and Hinrich Schütze. Language-agnostic bias detection in language models with bias probing. pages 12735–12747.

[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.

[LOG+19] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov, and Paul G Allen. Roberta: A robustly optimized bert pretraining approach. 2019.

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

[mis20] Gender by Name. UCI Machine Learning Repository, 2020. DOI: https://doi.org/10.24432/C55G7X.

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

[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.

[RSR+20] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67, 2020.

[RWC+] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners.

[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.

[TLI+] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothee Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models.

[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.

[YDY+] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding.

[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.