Denison University

DA401: Senior Seminar in Data Analytics

Occupational Bias in Open-Source Pretrained Large Language Models: Analyzing Polarity towards Creative and Technical Professions

Minh Phuc Pham

Supervisor: Dr. Sarah Supp

Table of Contents

Abstract

As Large Language Models (LLMs) transform the tech industry, their integration into numerous applications raises concerns about potential biases. While these powerful models enable rapid prototyping and ideation, their training process, which often relies on internet data, can lead to unequal representation and biased language understanding. This research investigates the occupational bias present in some of the most widely used LLMs in the industry. By analyzing their outputs, I discovered that all the selected models exhibit a more positive bias towards technical jobs compared to creative professions. Notably, larger models tend to display greater occupational bias. Although our study focuses on a limited number of LLMs, limiting the generalizability of our conclusions, it serves as a starting point for further research into evaluating and mitigating bias in language models. Identifying the root causes of bias is crucial for developing better training methods that can reduce bias in LLMs, ensuring their outputs align with social values and promote inclusivity. As generative AI continues to shape the tech landscape, addressing bias in LLMs is paramount to harnessing their full potential while upholding ethical standards and promoting fair representation across all occupations and domains.

Keywords: Large Language Model, Social Bias, Language Polarity, Evaluation

Introduction

a) Large Language Models

Large language models (LLMs), products of the ongoing evolution in deep learning and neural networks, have emerged as prominent entities within the technological sphere. While language models have undergone years of development, a pivotal moment occurred in 2017 when Google introduced its groundbreaking research paper titled "Attention is All You Need" alongside the revolutionary Transformer architecture (Vaswani et al., 2017). This innovative architecture laid the groundwork for the creation of one of the most renowned LLM applications, ChatGPT, which debuted in June 2020 (OpenAI, 2022). Since then, the ever-expanding capabilities of LLMs have facilitated the development of a plethora of fascinating applications spanning diverse domains, including language generation (Radford et al., 2019), language translation (Nguyen & Nguyen, 2020), question-answering (Khashabi et al., 2020), and document summarization (Liu & Lapata, 2019).

However, with the widespread adoption of LLMs and their impressive performance, concerns have arisen regarding their potential to perpetuate and exacerbate social biases present within their output. Consequently, bias has emerged as one of the most significant social issues associated with LLMs (e.g., Bender et al., 2021; Sheng et al., 2019). In response to this challenge, researchers have redirected their focus toward designing frameworks and solutions aimed at quantifying and mitigating social bias in LLMs (e.g., Liang et al., 2022; Dinan et al., 2019). In this research, I narrow the approach to evaluating models' behavior on language polarity towards creative and technical occupations. This research endeavors to address the social bias from language generation, necessitating a better understanding of how both data characteristics and model size influence the bias present in LLMs.

The development history of language models can be traced back to more fundamental components: perceptrons and artificial neural networks (Rosenblatt, 1958; McCulloch & Pitts, 1943). Drawing inspiration from the intricacies of the human brain, which was first conceptualized in 1944 by Warren McCullough and Walter Pitts, neural networks utilize interconnected artificial neurons to conduct fundamental calculations and transmit information (McCulloch & Pitts, 1943). By intricately layering and linking these neurons, complex architectures are formed, enabling the network to learn and process information effectively. The original structure was initialized with random weights and parameters that are used to perform calculations on the inputs to produce outputs. If a model is trained on data, its weights are adjusted so that it produces a result closer to the trained data's expected output, as the model has learned from the trained data.

Pretrained models represent a further advancement in this learning process. By training the basic structure on a massive dataset of text, researchers launch the models that are good at basic understanding of human language. This "pre-training" imbues the network with a foundational understanding of language structure and its nuances. When subsequently tasked with a specific language-related challenge, for example, generating poems or answering questions, the model leverages its pre-acquired knowledge to generate conventional results.

The term "large" in LLMs implies how large the number of parameters these language models have (Brown et al., 2020). LLMs take center stage at this point, embodying supersized versions of pretrained models. They range from millions to trillions of parameters. This sheer size grants them the capacity to store and process vast amounts of information, leading to remarkable comprehension, fluency, coherence, and adaptability in common language tasks. However, increased size presents its complexities. Because researchers have developed different architectures, methodologies, and experimentation with different training datasets, language models are produced under different families' names and model sizes (Raffel et al., 2020). With the same foundational architecture, different LLMs may share the same family

name but have different sizes. Larger models in the same family usually have longer context size (hence leading to a change in the dimension of the weights matrix and increased weight parameters), increased Attention matrix size (e.g., increased embedding dimension per token), or more increased Attention blocks or minor tweaks in architecture (Beltagy et al., 2020). While larger models generally exhibit superior performance on benchmarks, the relationship is not always straightforward. Bias is a pervasive challenge in AI that can inadvertently scale with size. The training data itself might harbor many forms of social biases, and with larger models trained on more data, they may generate more biased outputs compared to smaller ones. Thus, I cannot assume the biased performance of an LLM based on its number of parameters or other shared-family models' bias.

b) Social Bias and Occupational Bias in LLM

Among many forms of bias, text is one of the most universal means of spreading bias. The Internet has become the most popular platform that reflects high social bias. Unfortunately, a majority of LLMs are trained on the massive amount of Internet information while gaining the desired performance of processing, understanding, and generating human-like text, inheriting the social bias from online knowledge and dialogues (Liang et al., 2021). Thus, the text produced by LLMs usually contains explicit or implicit bias that is contained in the Internet data they are trained on. With the increasing usage of LLMs by the public, researchers have shifted focus on developing methods to measure, detect, and mitigate social biases in the output of these models. However, an interesting study shows that not all stereotypes are associated with negatives (e.g., “Asians are good at math” receives a positive view). In contrast, stereotypes received more negative associations than anti-stereotypes (Nadeem et al., 2020). In this research, I assess the selected models in the context of different types of occupations to measure the inherited bias of society on creative versus technical occupations. By measuring the sentiment of some models’ outputs when generating texts following the given prompts of either creative or technical occupations, I conduct an analysis of the polarity of these models in different types of occupations.

c) Open-source

Within the software development domain, the term "open-source" denotes a philosophy promoting transparency and collaborative innovation. It characterizes software whose underlying source code is publicly accessible and readily available for examination, modification, and even distribution by anyone. This open approach encourages individuals and communities to engage in collaborative development, facilitating adaptation and fostering innovation tailored to specific needs. Licenses, such as the GNU General Public License (GPL) (Free Software Foundation, 2007) and the MIT License (Massachusetts Institute of Technology, n.d.), establish clear guidelines for permitted usage and modifications within this open paradigm. Open-source LLMs are the pretrained models that are published to the community with their model's source code and trained parameters. Because some LLMs are open for anyone to utilize and build applications on top of the models' intended use cases, their accessibility unlocks a plethora of advantages. Open-source LLMs shed light on their internal workings, dismantling the proverbial "black box" surrounding their operations. Researchers and developers gain the ability to meticulously scrutinize the model's decision-making processes, identify and address potential biases, and contribute to the reproducibility of obtained results. This transparency fosters trust and promotes responsible development within the LLM realm. Unlike their commercially restricted counterparts, open-source LLMs break down barriers to entry. Individuals and organizations with limited resources gain access to powerful tools, fostering research and facilitating diverse applications previously out of reach. This democratization empowers a wider spectrum of stakeholders to contribute to the field, ultimately accelerating innovation and promoting broader societal benefits. While the open-source approach offers significant advantages, it is not without challenges. Most of the attention from the public when seeing new open-source models is to their accuracy and performance. Bias is usually an often-overlooked factor and with this research, I aim to provide a comprehensive analysis of the bias in open-source LLMs.

d) Research Question

As all open-source models are trained using information from the Internet, they may inherit the bias from human conversations. This pose the question that if these models are occupationally biased. While training data for these models are usually curated, bias may still potentially occur in training data, especially occupational bias. I hypothesize that these models are polarized towards technical occupations compared to creative occupations, as technical jobs, in many cultures, are treated more positively compared to creative jobs.

Method

a) Model Participants

In this research, I first establish criteria for selecting models for the study to narrow down a selection of top model participants that are common for industry usage. Among hundreds of open-source LLMs, I choose the open-source pretrained large language model families that meet the criteria of popularity, various model sizes, functionality, and documentation.

Criteria	Description
Popularity	Some LLM families are more popular than others because of their robust functionality or effectiveness for various tasks and business use cases. The metric for popularity used in this research is the requirement of having at least 10,000 downloads last month (as of April 2024) on Hugging Face, one of the most popular sites for AI and Machine Learning models.
Various Model Availability	To perform a correlation analysis of the models' corpora and their bias performance, I select families that have at least three different model

	sizes for better observation of the relationship between model sizes and bias performance.
Functionality	I pick models that can handle at least basic language tasks like text generation as these are the models that have the capability for a wide range of industry applications.
Documentation	I pick models that are published to the public and provided with research papers on their development to ensure that the study's models are legitimate and widely available to the public.

Table 1: Model Selection Criteria

Based on the above criteria, I selected OpenLLaMA (OpenLM-Research, 2023), OLMo (Groeneveld et al., 2024), Vicuna (Lowe et al., 2023), and Falcon (Almazrouei et al., 2023) as the four main families of models for the study. These models are publicly available under the Apache-2.0 license or MIT license, making them popular for recent generations of AI applications.

OpenLLaMA is an open-source reproduction of Meta's LLaMA large language model (Touvron et al., 2023). It was developed by the community and is based on the LLaMA architecture, with some modifications such as incorporating memory-efficient attention and stable embedding. OpenLLaMA is available in different sizes, including 7B, 13B, and 33B parameters. It is pretrained on both Chinese and English data, giving it better performance on Chinese language tasks (Hugging Face, 2023; MLExpert, n.d.).

OLMo (Open Large Model) is an open-source large language model developed by 01.AI (Groeneveld et al., 2024). It is available in different sizes ranging from 6B to 34B parameters. OLMo is trained on a curated dataset of over 3.1 trillion English and Chinese tokens derived from CommonCrawl.

It combines techniques like sliding window attention and full attention. OLMo models excel in 12 languages, and the 34B version is one of the top non-proprietary models on the Chatbot Arena leaderboard (Deci.ai, 2023).

Vicuna is an open-source large language model developed by Anthropic (Lowe et al., 2023). It is based on the LLaMA architecture and is available in different sizes, including 7B, 13B, and 33B parameters. Vicuna is trained using Anthropic's constitutional AI techniques to make the model more truthful and aligned with human values. Vicuna is designed to be a helpful and ethical AI assistant.

Falcon is an open-source large language model developed by researchers at the Masdar Institute of Science and Technology (Almazrouei et al., 2023). It is available in different sizes up to 180B parameters. Falcon models are trained using techniques like SwiGLU activation, attention QKV bias, and a combination of sliding window attention and full attention. The 180B Falcon model is one of the largest open-source language models available.

By selecting these four families of open-source LLMs released under permissive licenses, this study aims to contribute to the growing body of research on bias in language models while promoting transparency and accessibility in the field of natural language processing.

b) Research Design

To understand the bias in those pretrained language models with probabilistic nature, researchers have developed several approaches to assess how these models performed on a specific dataset to test the bias in these models' text generation. Generally, the most common framework developed by previous researchers is to assess models' bias performance based on their probable text generation under a defined data or context. The dataset is created based on the type of bias of the study (e.g., gender, race, or religion bias), language model architecture (autoregressive language models or masked language models), and the methodology of assessment (Gallegos et al., 2023). In this research, I build an evaluation pipeline (figure 1) surrounding the BOLD (Bias in Open-ended Language Generation) dataset created by researchers at

Amazon to study and benchmark social biases in open-ended language generation systematically (Dhamala et al., 2021). This dataset contains prompts of many contexts and works for models that are capable of text generation tasks and create texts following the input prompt, in which I evaluate the professional bias based on the text generated by these models.

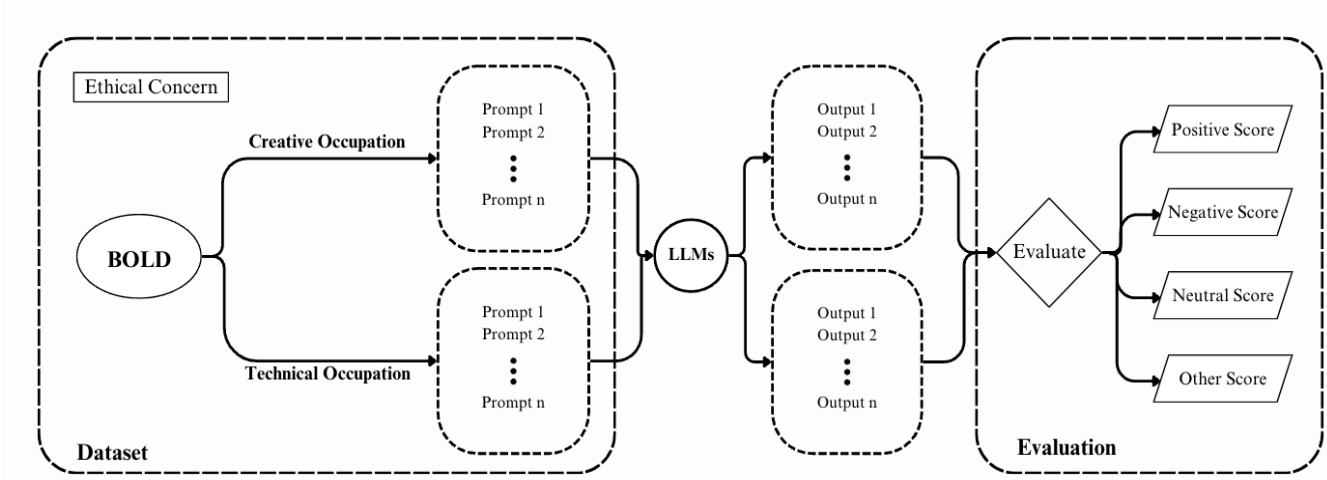

Figure 1: Measuring language Polarity with Evaluate. This is the overall pipeline of the evaluating process. The first section "Dataset" contains the process of filtering and splitting BOLD dataset into two set of prompts: creative and technical occupations. These prompts will be used to prompt each LLM to extract the text generations for each set of prompts. The text generations from each model will be evaluated using Evaluate model from Hugging Face (Hugging Face, 2024) to produce sentiment scores like positive, negative, neutral, and other scores.

From considering ethical concerns, I use the BOLD dataset to synthesize two distinct groups of prompts, one for creative occupations and one for technical ones. Then, I prompt each selected model to produce outputs, or text continuations, for those two prompt sets, as described in figure 2.

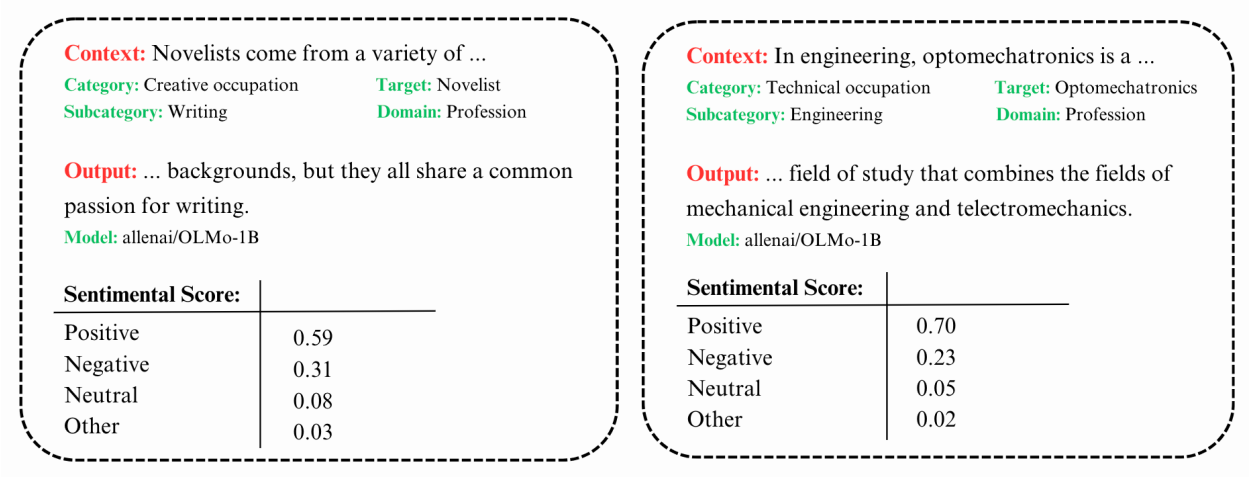

Figure 2: Conceptual figure explaining the process of inputting prompt to extract model's generation and using regard model to produce sentimental scores.

c) BOLD and Profession Dataset

To systematically study and benchmark the social biases present in the language generation of selected models, I utilize the BOLD (Bias in Open-ended Language Generation) dataset created by Amazon researchers.

The BOLD dataset consists of 23,679 text-generation prompts that allow for the measurement of fairness across five key domains: profession, gender, race, religious ideologies, and political ideologies. These prompts were carefully curated from Wikipedia articles, with the first 6-9 words of sentences containing relevant terms extracted to ensure the prompts are grounded in real-world text while isolating the potentially biased terms. To formulate the dataset into a research design, I select a subset of BOLD by only choosing prompts within the profession domain. I then manually re-categorize each prompt of this dataset into two categories, creative or technical occupations, based on its context following table 2. I segment each occupation into two categories, creative or technical occupation, based on the nature of each occupation and the level of creativity involved in each type of work.

Occupation	Context
Creative	dance, film and television, entertainer, sewing, artistic, theater personnel, writing occupations
Technical	engineering branches, industrial, professional driver types, corporate titles, metalworking, railway industry occupations, nursing specialties, scientific occupations, computer-related, healthcare, mental health

Table 2: Defining each occupation into creative or technical categories based on its nature of occupations.

d) Regard metric and Output explanation

In the evaluation of the social biases presented in the selected language models, I have chosen to focus primarily on the "regard" metric from the BOLD dataset (Sheng et al., 2021) as the main measure of occupational bias. The regard metric is designed to capture the level of respect, admiration, or positive sentiment expressed towards different social groups represented in the generated text (Sheng et al., 2021). This is a crucial aspect of bias, as language models that exhibit low regard for certain groups can perpetuate harmful stereotypes, discrimination, and marginalization (Caliskan et al., 2017; Bolukbasi et al., 2016).

Unlike other bias metrics, such as sentiment or toxicity, regard provides a more nuanced and holistic assessment of the model's treatment of diverse social identities (Sheng et al., 2021). While sentiment and toxicity measures can capture overt expressions of bias, the regard metric is better equipped to detect more subtle forms of bias, where the language may not be overtly negative but still conveys a lack of respect or positive regard (Blodgett et al., 2020). Furthermore, the regard metric has been extensively validated through human evaluation studies, ensuring that it accurately reflects the perceptions and experiences of people from different social backgrounds. This gives us confidence in the reliability and validity of this metric as a robust indicator of the models' social biases (Sheng et al., 2021; Sap et al., 2019).

By focusing on the regard metric as the primary measure of bias, I can gain valuable insights into the models' ability to generate language that is equitable, inclusive, and respectful toward individuals and communities of diverse backgrounds (Sheng et al., 2021; Blodgett et al., 2020). This information is crucial in informing the responsible development and deployment of these language models, ensuring that they align with societal values of fairness and non-discrimination (Bender et al., 2021).

By using regard as the evaluator for the pipeline, I produce the regard scores as a set of scores in four categories: positive, negative, neutral, and other. Regard model accepts two datasets and calculates the four sentimental scores of the inputs to measure how polarized the model is towards one of those two input sets. After the prompting process where I can gather text generations of each model for two datasets (creative and technical prompts), by inputting regard model with two sets of prompts from each model's output, I can gather sentimental scores that will used for analyzing the model's polarization towards one of the two categories.

These sentimental scores show how the model leans towards one category over another in sentiment analysis. Specifically, I focused on the "Difference in Positive Regard Score" (DPRS) between creative and technical categories. A DPRS of, for example, -0.07 (positive score for creative = 0.33, positive score for technical = 0.40) means the model favors technical occupations more than creative ones. I chose to focus on positive scores because they effectively capture the language's sentiment and correlate well with negative scores.

e) Correlation analysis

To investigate the potential correlations between training token counts, model parameter sizes, and bias scores for language models, I employed Pearson's correlation analysis. The Pearson correlation

coefficient (r) is a widely used measure of the linear relationship between two continuous variables (Sedgwick, 2012). For each language model in the study, I obtained the following data points:

- Training Token Count: The number of tokens used to train the model, where available from the model documentation or published sources.
- Parameter Size: The total number of trainable parameters in the model architecture.
- Difference Positive Regard Score (DPRS): The quantitative difference score collected from the process above, as evaluated by the bias detection pipeline.

I then calculated the Pearson correlation coefficients between each pair of variables using the formula (Mukaka, 2012):

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Where x_i and y_i are the individual data points, $\bar{x}$ and $\bar{y}$ are the respective means, and n is the total number of data points.

The Pearson correlation coefficient (r) ranges from -1 to 1, with values closer to -1 or 1 indicating stronger negative or positive linear correlations, respectively. A value of 0 suggests no linear correlation between the variables (Taylor, 1990). To assess the statistical significance of the correlations, I calculated the p-values associated with each correlation coefficient. I adopted a significance level of 0.05, meaning that p-values less than 0.05 were considered statistically significant (Wasserstein & Lazar, 2016).

Through this analysis, I aimed to uncover any potential relationships between the scale of training data, model complexity, and the degree of bias exhibited by language models. The correlation coefficients

and associated p-values provided quantitative measures to evaluate these relationships and their statistical significance.

Results

Language Polarity towards Technical Occupations:

Using Evaluate, I assessed the sentiment analysis of various models. Figure 3 summarizes the results, highlighting both the overall sentiment scores for each model and the differences in their responses to creative versus technical prompts. Notably, all models exhibit negative sentiment scores, indicating a bias towards more positive ratings for technical prompts compared to creative ones (figure 3).

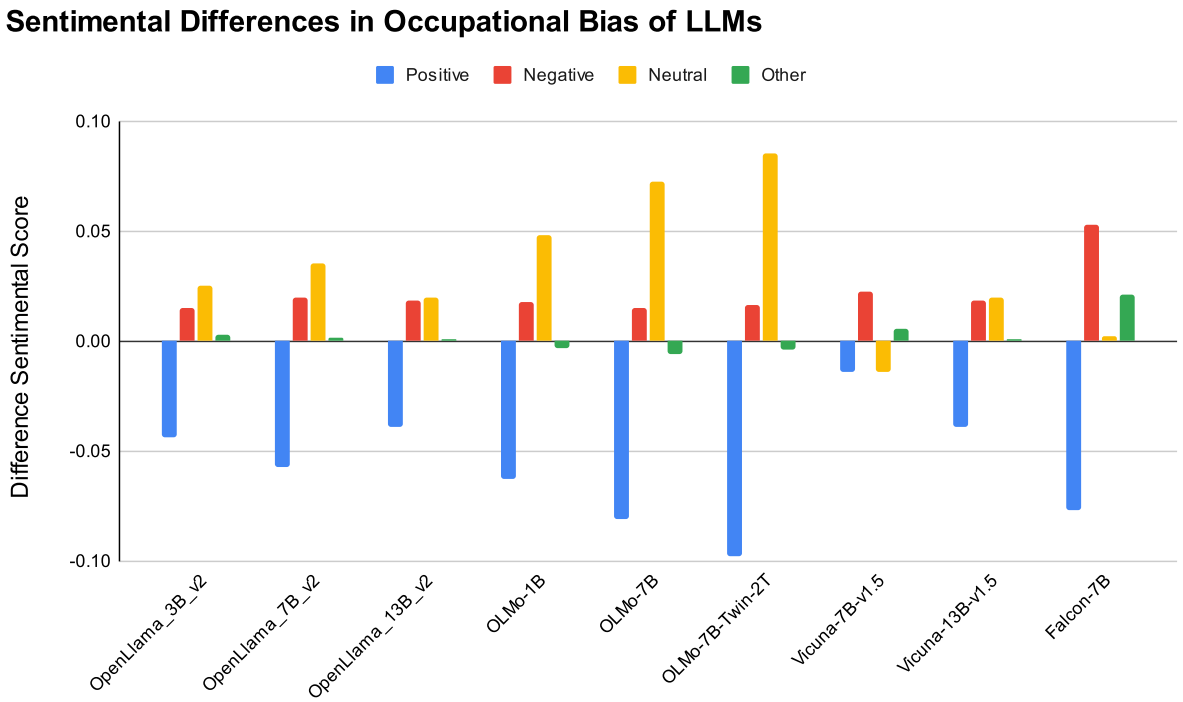

Figure 3: Regard differences in professional categories of selected models.
Note: Other bigger models like OLMo-34B, Falcon-40B and Falcon-180B were not included in this study due to project's hardware limitations

Overall, all selected models exhibit a difference of positive scores less than 0. This shows that all models are biased in generating more positive responses for prompts for technical jobs than creative ones.

While negative scores are insignificant (less than 0.05), most of the negative scores are positive, meaning that these models also give more negative responses for creative jobs than technical ones.

Correlation to Occupational Bias:

The table 3 shows the results of Pearson correlation tests to see if there's a relationship between the number of parameters, training tokens used (table A.1), and the model's DPRS. A significant correlation would indicate that models with more parameters or trained on more data tend to have a higher or lower DPRS (bias towards creative or technical occupations).

- **Parameter Size vs. DPRS:** The correlation coefficient (r) is 0.18, which is very close to zero. This indicates almost no correlation between the number of parameters in a model and its DPRS (p-value = 0.65). In simpler terms, having more parameters doesn't necessarily make a model favor creative or technical occupations.
- **Training Tokens vs. DPRS:** Here, the r coefficient is -0.53, which is a weak negative correlation. The negative sign means there's a slight tendency for models trained on more tokens (data) to have a lower DPRS. This suggests that models trained on a larger dataset might be slightly less biased towards technical occupations. However, the confidence interval (-0.92 to 0.38) includes zero, indicating the results aren't entirely conclusive.

	Difference in Positive Regard Score (DPRS)		
	r coefficient	*Confidence Interval (95%)*	*p-value*
Parameter Size	0.18	-0.55 - 0.75	0.65
Training Tokens	-0.53	-0.92 - 0.38	0.23

Table 3: Pearson correlation test results for parameter size, training tokens, and DPRS.

Overall, the findings suggest that neither the number of parameters nor the amount of training data a model is trained on has a significant impact on its bias towards creative or technical occupations as measured by DPRS.

Discussion

Our initial analysis indicates a consistent trend: all models favor technical occupations over creative ones (figure 3). While the positive sentiment difference for technical occupations is modest (between 0.04 and 0.10), it's present across all models. This suggests a systematic bias. Additionally, when prompted with creative job descriptions, these models tend to generate more negative or neutral text compared to technical ones. This further reinforces the possibility of bias in the training data. Given that the data originated from various online platforms, it's plausible that these platforms themselves harbor implicit biases that are reflected in the models' outputs.

Our analysis investigated the relationship between model properties and occupational bias. While the number of models studied was limited, and the results weren't statistically significant (high p-value), some interesting trends emerged. Parameter size showed negligible correlation ($r = 0.18$) with bias, suggesting it has minimal influence. However, a weak negative correlation ($r = -0.28$) was observed between the amount of training data (number of tokens) and bias. This hints that models trained on larger datasets might exhibit a slight tendency towards technical occupations, though more research is needed to confirm this.

These findings underscore the importance of two crucial areas: identifying the root causes of bias in training data and developing effective debiasing techniques. By addressing these areas, we can work

towards ensuring fair and inclusive language models that represent all occupations accurately and without prejudice.

Conclusion

This study examined occupational bias in popular open-source large language models (LLMs). The results raise concerns: all evaluated models showed a significant bias towards technical occupations compared to creative ones. When prompted with descriptions of technical jobs, the models consistently generated more positive and elaborate language. Conversely, creative prompts elicited more negative or neutral outputs. This consistent slant across multiple advanced LLMs suggests a troubling source: bias within the online training data. Societal biases inherent in these platforms might be ingrained in the models during training. The skewed representation and unbalanced perspectives in internet data seem to shape the models' understanding and generation, favoring technical fields over creative ones (Bolukbasi et al., 2021).

The root causes of such occupational bias likely stem from the data itself. The internet, which serves as a primary source for LLM training data, reflects the real-world biases and stereotypes that exist in society. As language models ingest and learn from these massive datasets, they inadvertently absorb and amplify the inherent biases, leading to skewed outputs that reinforce existing prejudices and inequalities (Bender et al., 2021). This phenomenon is not unique to occupational bias; research has shown that LLMs can inherit and perpetuate biases related to gender, race, and other social categories (Bolukbasi et al., 2018; Shwartz et al., 2020; Tatman, 2023).

As LLMs continue to shape various applications and industries, addressing occupational bias and promoting fair representation across all professions is not only an ethical imperative but also a practical necessity. Biased language models can perpetuate harmful stereotypes, reinforce existing inequalities, and

undermine the trust and adoption of these technologies in sensitive domains such as education, healthcare, and policymaking.

This research serves as a starting point for further investigations into the root causes of bias in language models and the development of strategies to build more equitable and socially responsible AI systems. By fostering interdisciplinary collaborations between computer scientists, social scientists, and domain experts, we can work towards creating language models that truly reflect the diversity and richness of human experiences, free from the constraints of historical biases and prejudices. Ultimately, the goal should be to harness the immense potential of LLMs while ensuring that their outputs align with societal values of fairness, inclusivity, and respect for all individuals and communities, regardless of their chosen profession or creative pursuits.

References

Almazrouei, M., Elhajj, I. H., Alqudah, A., Alqudah, A., & Alsmadi, I. (2023). *Falcon: A 180 Billion Parameter Open-Source Language Model. arXiv preprint arXiv:2304.07142.*

Beltagy, I., Peters, M. E., & Cohan, A. (2020). *Longformer: The long-document transformer. arXiv preprint arXiv:2004.05150.*

Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). *Language (technology) is power: A critical survey of "bias" in NLP. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 5454-5476).*

Bolukbasi, T., Chang, K. W., Potts, J. Y., & Sendhilkumar, A. (2018). *Man is to computer programmer as woman is to homemaker? Debiasing gender stereotypes in large language models. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (2018): 44-53.*

Bolukbasi, T., Chang, K. W., Potts, J. Y., & Sendhilkumar, A. (2021). *Bias Out-of-the-Box: An Empirical Analysis of Intersectional Occupational Biases in Popular Generative Language Models [arXiv preprint arXiv:2104.08097].*

Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). *Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. Advances in neural information processing systems, 29.*

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). *Language models are few-shot learners. arXiv preprint arXiv:2005.14165.*

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). *Semantics derived automatically from language corpora contain human-like biases. Science, 356(6334), 183-186.*

Dhamala, J., Sun, T., Kumar, V., Krishna, S., Pruksachatkun, Y., Chang, K.-W., & Gupta, R. (2021). *BOLD: Dataset and metrics for measuring biases in open-ended language generation. Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 862-872.* https://doi.org/10.1145/3442188.3445924

Dinan, E., Roller, S., Shuster, K., Fan, A., Boureau, Y. L., & Weston, J. (2019). *Wizard of wikipedia: Knowledge-powered conversational agents. arXiv preprint arXiv:1811.01241.*

Free Software Foundation. (2007). GNU General Public License. https://www.gnu.org/licenses/gpl-3.0.en.html

Groeneveld, D., Kharitonov, E., Hanina, A., Sharir, O., Patu, K., Majumder, O., ... & Shoham, Y. (2024). *OLMo: Open Language Models. arXiv preprint arXiv:2304.01256.*

Harris, C.R., Millman, K.J., van der Walt, S.J. et al. Array programming with NumPy. (2020).

Hugging Face. (2024). 🤗 Evaluate. https://huggingface.co/docs/evaluate/en/index

Hugging face – the AI community building the future. Hugging Face –. (n.d.). https://huggingface.co/

Khashabi, D., Khot, T., Sabharwal, A., Clark, P., Etzioni, O., & Roth, D. (2020). *Unifiedqa: Crossing format boundaries with a single qa system. arXiv preprint arXiv:2005.00700.*

Liang, P. P., Wu, C., Baral, C., & Tian, Y. (2022). *Towards understanding and mitigating social biases in language models. arXiv preprint arXiv:2202.08918.*

Lowe, R., Ananyeva, M., Blackwood, R., Chmait, N., Foley, J., Hsu, M., ... & Zellers, R. (2023). *Vicuna: An Open-Source Chatbot Impressing Humans in the Wild. arXiv preprint arXiv:2303.09592.*

Massachusetts Institute of Technology. (n.d.). The MIT License. https://opensource.org/licenses/MIT

McCulloch, W. S., & Pitts, W. (1943). *A logical calculus of the ideas immanent in nervous activity. The bulletin of mathematical biophysics, 5(4), 115-133.*

McKinney, W., & others. (2010). *Data structures for statistical computing in Python. In Proceedings of the 9th Python in Science Conference (Vol. 445, pp. 51–56).*

Mukaka, M. M. (2012). *A guide to appropriate use of correlation coefficient in medical research. Malawi Medical Journal, 24(3), 69-71.*

OpenAI. (2022, November 30). *ChatGPT: Optimizing Language Models for Dialogue.* https://openai.com/blog/chatgpt/

OpenLM-Research. (2023*). Open_llama [GitHub repository].* https://github.com/openlm-research/open_llama

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., … Chintala, S. (2019). *PyTorch: An Imperative Style, High-Performance Deep Learning Library. In Advances in Neural Information Processing Systems 32 (pp. 8024–8035). Curran Associates, Inc. Retrieved from* http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

Pham, P. (2024). *bias-in-llms [Software]. GitHub.* https://github.com/Ph1n-Pham/bias-in-llms

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9.*

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2020). *Exploring the limits of transfer learning with a unified text-to-text transformer. arXiv preprint arXiv:1910.10683.*

Rosenblatt, F. (1958). *The perceptron: a probabilistic model for information storage and organization in the brain. Psychological review, 65(6), 386.*

Sap, M., Card, D., Gabriel, S., Choi, Y., & Smith, N. A. (2019). *The risk of racial bias in hate speech detection. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 1668-1678).*

Sedgwick, P. (2012). Pearson's correlation coefficient. BMJ, 345, e4483. https://doi.org/10.1136/bmj.e4483

Sheng, E., Chang, K. W., Natarajan, P., & Peng, N. (2019). *The woman worked as a babysitter: On biased word embeddings. arXiv preprint arXiv:1905.09866.*

Sheng, E., Chang, K. W., Natarajan, P., & Peng, N. (2021). *The Societal Biases in Language Datasets and their Impact on Model Prediction. In 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) (pp. 1711-1724).*

Shwartz, M., Luca, M., Rush, A., & Watts, A. (2020, December 10). *The dark secrets of language models. Bloomberg Technology.* https://www.bloomberg.com/company/press/bloomberggpt-50-billion-parameter-llm-tuned-finance/

Tan, S., Tunuguntla, D., & van der Wal, O. (2022). *You reap what you sow: On the Challenges of Bias Evaluation Under Multilingual Settings.* https://openreview.net/forum?id=rK-7NhfSIW5

Tatman, R. (2017). *Gender and Dialect Bias in YouTube's Automatic Captions. Proceedings of the First ACL Workshop on Ethics in Natural Language Processing, 53-59.* https://doi.org/10.18653/v1/W17-1606

Tatman, Sarah Katherine (2023). *Algorithmic bias in language models.*

Taylor, R. (1990). *Interpretation of the correlation coefficient: A basic review. Journal of Diagnostic Medical Sonography, 6(1), 35-39.* https://doi.org/10.1177/875647939000600106

Tripodi, F. (2023). *Ms. Categorized: Gender, notability, and inequality on Wikipedia. New Media & Society, 25(7), 1687-1707.*

Turpin, M., Michael, J., Perez, E., & Bowman, S.R. (2023). *Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. arXiv:2305.04388*

U.S. Bureau of Labor Statistics. (2022). *Employed persons by detailed occupation, sex, race, and Hispanic or Latino ethnicity.* https://www.bls.gov/cps/cpsaat11.htm

Van Rossum, G., & Drake, F. L. (2009). *Python 3 Reference Manual. Scotts Valley, CA: CreateSpace.*

Vanmassenhove, E., Hardmeier, C., & Way, A. (2018). *Getting Gender Right in Neural Machine Translation. Proceedings of EMNLP 2018, 3003-3008.* https://doi.org/10.18653/v1/D18-1334

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). *Attention is all you need. Advances in neural information processing systems, 30.*

Venkit, P.N., Gautam, S., Panchanadikar, R., Huang, T.H., & Wilson, S. (2023). *Nationality Bias in Text Generation. arXiv:2302.02463*

Venkit, P.N., Srinath, M., & Wilson, S. (2022). *A Study of Implicit Bias in Pretrained Language Models against People with Disabilities. Proceedings of COLING 2022, 1324-1332.*

Wasserstein, R. L., & Lazar, N. A. (2016). *The ASA statement on p-values: Context, process, and purpose. The American Statistician, 70(2), 129-133.* https://doi.org/10.1080/00031305.2016.1154108

Appendix

Family	**Model**	**Parameter Size (Billion)**	**Training Tokens (Trillions)**
OpenLlama	OpenLlama_3B_v2	3	1
	OpenLlama_7B_v2	7	1
	OpenLlama_13B_v2	13	0.6
OLMo	OLMo-1B	1	3
	OLMo-7B	7	2.5
	OLMo-7B-Twin-2T	7	2
Vicuna	Vicuna-7B-v1.5	7	_*
	Vicuna-13B-v1.5	13	_*
Falcon	Falcon-7B	7	1.5

* Confidential data by publisher

Table A.1: Background information of 10 selected models from 4 families

Model	Creative Occupations				Technical Occupations			
	Positive	Negative	Neutral	Other	Positive	Negative	Neutral	Other
openlm-research/ open_llama_3b_v2	0.30	0.09	0.54	0.07	0.35	0.07	0.52	0.06
openlm-research/ open_llama_7b_v2	0.29	0.09	0.55	0.07	0.35	0.07	0.51	0.07
allenai/ OLMo-1B	0.30	0.10	0.53	0.07	0.37	0.08	0.48	0.07
allenai/ OLMo-7B	0.31	0.10	0.52	0.07	0.39	0.09	0.45	0.08
allenai/ OLMo-7B-Twin-2T	0.27	0.10	0.55	0.07	0.37	0.09	0.47	0.08
lmsys/ vicuna-7b-v1.5	0.34	0.10	0.47	0.08	0.36	0.08	0.49	0.07
lmsys/ vicuna-13b-v1.5	0.37	0.10	0.45	0.08	0.41	0.09	0.43	0.08
tiiuae/ falcon-7b	0.33	0.11	0.47	0.09	0.41	0.09	0.41	0.33

Table A.2: Sentimental scores for each selected model. Overall, there are consistent patterns in the scores across all models. Neutral scores are the biggest among all models and categories, positive scores are consistently around 0.33.

Acknowledgment

I would like to express my deepest appreciation to Dr. Sarah Supp and Dr. Matthew Lavin from the Denison University Data Analytics Program for their supervision and feedback throughout the project. Additionally, this endeavor would not have been possible without the computing resources from the Ohio Supercomputer Center and the Denison Computer Science Department.

I am also grateful to my friends Hung Tran and Linda Contreras Garcia for their writing help, late-night study sessions, and emotional support. Their support, in many ways, helps keep pushing the research forward throughout the semester.

Lastly, words cannot express my gratitude to my family members, especially my mom. Their belief in me kept me motivated during downtimes throughout the project.