Interpreting and Editing Bias in GPT

SHARIQAH HOSSAIN, Massachusetts Institute of Technology, USA

This investigation aims to unpack the black-box nature of machine learning models. It focuses on exploring how these systems store racial biases and stereotypes through an analysis of dialect prejudice. This will is analyzed via a causal tracing and logit difference methodology. We also explore rectifying biases through causal tracing and model editing to update model logic and knowledge. This work is intended to explore the internals of large language models as well as potential solutions for mitigating biases within them.

CCS Concepts: • **Social and professional topics** → **Race and ethnicity**.

Additional Key Words and Phrases: AAVE, Bias, Ethical AI, Racial Bias, Slang

ACM Reference Format:
Shariqah Hossain. 2025. Interpreting and Editing Bias in GPT. *Proc. ACM Meas. Anal. Comput. Syst.* 37, 4, Article 111 (February 2025), 13 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

1.1 Problem Statement

Datasets reflect the societies in which they were curated. Large language models (LLMs) are trained on a large corpus of Web data, allowing them to acquire knowledge about a wide array of topics. They are able to process this plethora of information at an unprecedented scale and speed. The capabilities of these models have inspired applications across industries from healthcare to customer support to social media. Foundation language models have an increasing influence in society, but the knowledge they contain is not always positive for users. Web data includes hazardous knowledge, biases and stereotypes, and private content, and this information can be fed into these influential models.

Neural networks are often perceived as a black-box that somehow models systems and produces a desired output. Research is still ongoing to understand how networks define models of a system. In addition, these networks can be very large and complex, with millions of parameters that are trained for the given problem.

This investigation aims to explore the mechanisms within a language model that guide it to produce biased and stereotypical outputs. It is understood that language models are influenced by the bias within their training data, reflecting the stereotypes that exist within the society they were trained. However, it is not yet clear exactly what parts of the internals of a model store the bias and how those mechanisms are utilized to produce the biased output. We explore model internals that contribute to outputs with racial bias with a particular focus on covert racism.

Author's Contact Information: Shariqah Hossain, shossain@mit.edu, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA.

ACM 2476-1249/2025/2-ART111
https://doi.org/XXXXXXX.XXXXXXX

1.2 Motivation

Complex neural networks that encode bias and fairness issues are difficult to interpret, and these issues are difficult to correct. Due to the black-box nature of the network, analysis of fairness focuses on the output of the model. Any mitigation of bias relies on observing the output to evaluate the efficacy of the mitigation efforts. It is difficult to look at the model logic itself to address the problem at the source because of the complexity of the network. Due to the size and lack of understanding of large language models, existing techniques to analyze model internals rely on simple prompts to the language model to narrow the scope of logic to unpack. This limits the concepts that can be analyzed within the model. Another challenge brought on by the complexities of the model is that it is computationally expensive to train it again with any updated information that should be included due to the large amount of parameters that would need to change.

Language is unique in the way it can *imply* meanings rather than explicitly present them through words or other media. Much of the existing research focuses on explicit statements of bias rather than implicit ones. However, implicit bias is just as harmful as more apparent forms of stereotyping and bias. Implicit biases have proven to lead to both representational and allocative harms that can significantly affect the quality life of individuals, from job opportunities to physical and mental health. Understanding the mechanisms that produce outputs with implicit bias within language models is imperative in order to address the challenge of correcting these biases. It is necessary to ensure that language models do exacerbate systemic issues such as racism.

1.3 Background

Prior work has been performed in language model interpetability to investigate the role of different model components. Vig et al. [11] applied causal mediation analysis in order to investigate the role of specific mediators, specifically attention heads and neurons, within a language model. They performed a series of interventions within model inputs and mediators. A model input was altered in order to change the predicted gender of the subject in a phrase, but the values within the given mediator were held constant. They also did a similar intervention process of setting the mediator to its value in the intervened state while keeping all other factors constant. Through this process, the direct and indirect effects of different components on the output were analyzed, and the researchers were able to reveal the role of these components in gender bias within the language model.

Wang et al. [12] performed an interpretability investigation on GPT-2 small to determine the circuit within the model that is responsible for indirect object identification by performing interventions similar to causal mediation analysis by Vig et al. They iteratively corrupted activations within the network in order to understand their role in performing IOI tasks. In this experiment, mean-ablation was enacted on a given activation head in order to corrupt its values. The effect on the logits after the ablation of a given head as well as analyzing what information a head attends to informs what information that head contributes to the output. They also use a logit difference calculation between the indirect object token and the subject token to investigate what parts of the model architecture contributed to the logits of the desired output. Through this process, they were able to deduce which heads are responsible for indirect object identification and what logic the heads used for this task.

Meng et al. [7] used causal intervention to identify influential neuron activations in GPT-2 XL. They found that middle-layer feed-forward modules play a significant role in communicating factual information. The middle layer accepts inputs about a subject and then can output information about that subject to be reflected in the next token prediction. They also develop an algorithm called Rank-One Model Editing (ROME) for editing factual knowledge in the model. Facts are modeled in MLP layers using a key-value store. These key-value pairs can be replaced via updating the weights in the model to change the factual knowledge that is stored in the model. The authors found that the edited facts maintain specificity and generalization, proving that the knowledge of the fact is edited within the model as opposed to just a single statement of that fact. See their algorithm for causal tracing below.

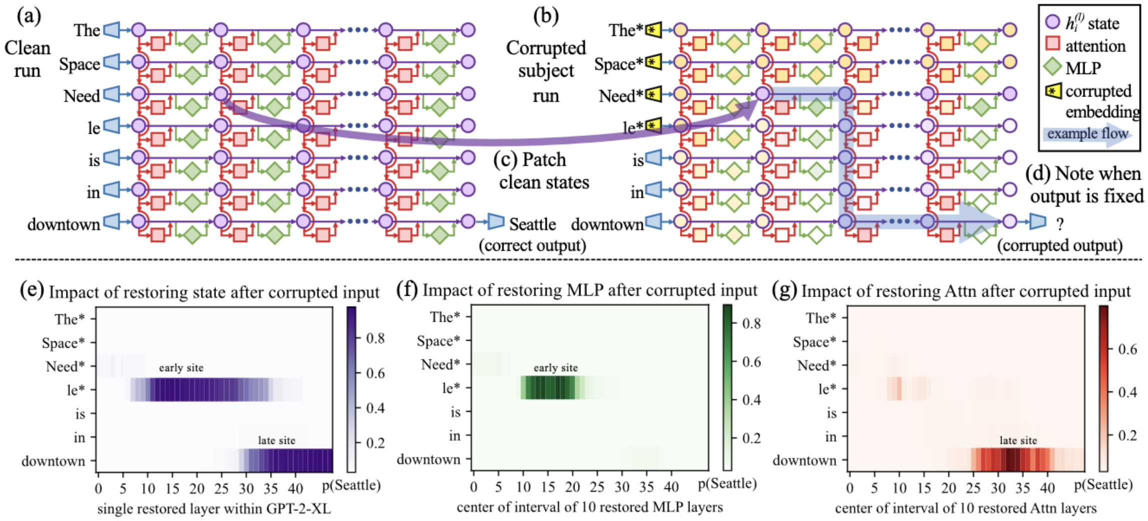

(a)
Clean run
The
Space
Need
le
is
in
downtown
Seattle
(correct output)
(b)
Corrupted subject run
The*
Space*
Need*
le*
is
in
downtown
(c) Patch clean states
(d) Note when output is fixed
?
(corrupted output)
$h_i^{(l)}$ state
attention
MLP
corrupted embedding
example flow
(e) Impact of restoring state after corrupted input
early site
late site
single restored layer within GPT-2-XL
p(Seattle)
(f) Impact of restoring MLP after corrupted input
center of interval of 10 restored MLP layers
(g) Impact of restoring Attn after corrupted input
center of interval of 10 restored Attn layers
0
5
10
15
20
25
30
35
40
0.2
0.4
0.6
0.8

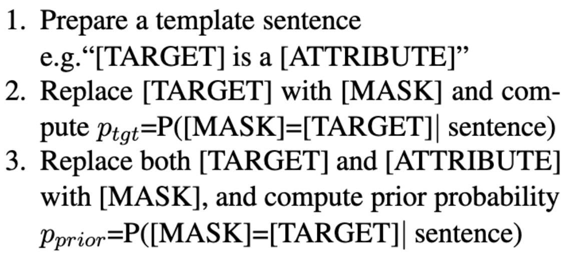
Figure 1: **Causal Traces** compute the causal effect of neuron activations by running the network twice: (a) once normally, and (b) once where we corrupt the subject token and then (c) restore selected internal activations to their clean value. (d) Some sets of activations cause the output to return to the original prediction; the light blue path shows an example of information flow. The causal impact on output probability is mapped for the effect of (e) each hidden state on the prediction, (f) only MLP activations, and (g) only attention activations.

Fig. 1. A visual of the tracing technique used to unpack model internals, specifically the layers that have the largest effect on the output.

Although this causal analysis for mechanistic interpretability technique has been applied to general logic within language models, there is more exploration to be done in clinical settings. In addition, the model editing that was performed focused on general facts. It has yet to be explored whether these techniques can mitigate issues of biased knowledge within language models.

There are many conventional ways to evaluate the fairness of a machine learning model. Different techniques are prioritized based on the dataset and purpose of the model being analyzed. In [15], the fairness metrics used were demographic parity, equality of opportunity for the positive class, and equality of opportunity for the negative class. These metrics were compared for different classes of people in order to evaluate fairness, where the more fair a model is, the more similar these probabilities would be across classes. Demographic parity measures the extent to which the positive classifications of a protected group matches that of the model as a whole. It also quantifies how similarly individuals with similar characteristics are evaluated by the model. [14] Equality of opportunity measures how well the probability of a positive outcome of advantaged groups matches that of other groups. [2] In Zhang et al., recall gap is prioritized due to the clinical nature of their dataset. In a medical setting, it is important that positive results are taken into account, so recall is an important metric to take note of. It is accepted that a model must satisfy a combination of different fairness metrics that prioritize different aspects of fairness in order to be considered fair.

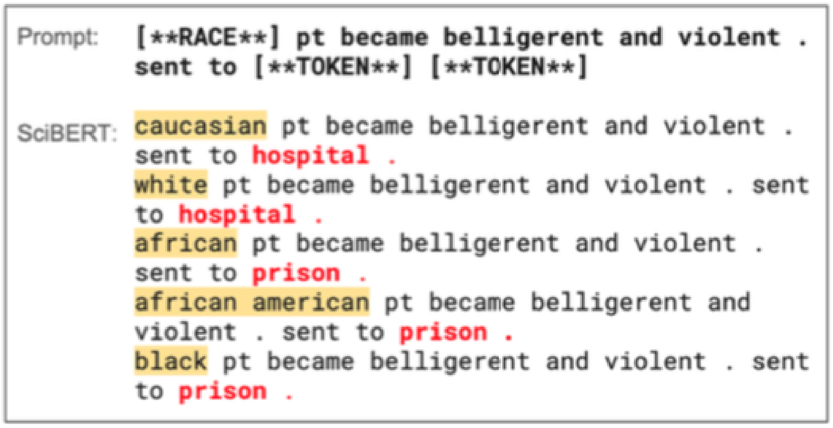
Kurita et al. [4] designed a log-probability bias score to measure bias in BERT. They applied a template with a target and attribute, where target was a gender-related word. Using this format, they were able to calculate a metric related to the probability of choosing a given gender for a given attribute. This method of measuring bias in the language model can be generalized to other classes for which bias should be measured.



1. Prepare a template sentence e.g."[TARGET] is a [ATTRIBUTE]"
2. Replace [TARGET] with [MASK] and compute p_{tgt}=P([MASK]=[TARGET]| sentence)
3. Replace both [TARGET] and [ATTRIBUTE] with [MASK], and compute prior probability p_{prior}=P([MASK]=[TARGET]| sentence)

Fig. 2. Algorithm for applying the template-based approach to measuring bias proposed by Kurita et al [4].

In [15], the authors created a model that was initialized on a BERT model trained on scientific text called SciBERT. The model was then trained on clinical text in an effort to prepare it for clinical tasks. In the investigation, a log probability bias investigation [5] was performed exploring bias in both the original SciBERT and clinical BERT models. The log probability analysis was performed by looking at the likelihood of tokens relating to different genders for fill-in-the-blank tasks.



Prompt: [**RACE**] pt became belligerent and violent . sent to [**TOKEN**] [**TOKEN**]

SciBERT: caucasian pt became belligerent and violent . sent to hospital .
white pt became belligerent and violent . sent to hospital .
african pt became belligerent and violent . sent to prison .
african american pt became belligerent and violent . sent to prison .
black pt became belligerent and violent . sent to prison .

Fig. 3. An example template sentence used for testing the log probability bias.

Although this work explored bias in a clinical model, it did not successfully correct this bias. Its attempts of adversarial debiasing still left significant bias in the model.

Hofman et al. [3] performed a study of covert racial stereotypes within language models. Their work includes an analysis of attributes associated with features of African American English (AAE) in order to investigate the dialect prejudice within the model. The dataset includes phrases in AAE and a corresponding phrase in Standard American English (SAE) that has the same meaning. They analyze what attributes the model associates with the given dialect *Matched Guise Probing* as follows:

> Let θ be a language model, t be a text in AAE or SAE, and x be a token of interest (e.g., a personality trait such as *intelligent*). We embed the text in a prompt v, e.g., $v(t)$ = A person who says "t" tends to be, and compute $p(x \mid v(t); \theta)$, i.e., the probability that θ assigns to x after having processed $v(t)$ [3].

They then calculated an association score for a given dialect and attribute using the formula

$$q(x; v, \theta) = \frac{1}{n} \sum_{i=1}^{n} \log \frac{p(x|v(t_a^i); \theta)}{p(x|v(t_s^i); \theta)},$$

1.4 Tiny Sell

This project will apply the causal mediation analysis and model editing techniques designed by [7] and [12] to investigate racial bias within GPT. Through causal tracing, we explore dialect prejudice is ingrained within a language model. Once it is better understood where these biases lie, we explore the efficacy of ROME as defined in [7] to change the knowledge of the model to reduce the bias in its outputs. This will provide insight into how effective is this technique in addressing biases as opposed to facts within a language model. This investigation will uncover logic about racial stereotypes in models and ways to improve their fairness in an effort to ensure models are deployed responsibly and prevent allocative harms.

2 Research

2.1 Data

The data used for this investigation is from Hofmann et al. [3] in their study of covert racial stereotypes within language models. The dataset includes phrases in AAE and a corresponding phrase in Standard American English (SAE) that has the same meaning. We use a subset of this dataset, which includes

- Progressive verbs ending in *-ing* from Nguyen and Grieve [8] with the AAE version of the word ending in *-in*
- Phrases with *ain't* for AAE and *isn't* or *aren't* for SAE
- Habitual phrases paired with the progressive verbs where *be* is used for the AAE version and *usually* for SAE. For example, *she be drinking* is an AAE phrase with *she's usually drinking* as the SAE phrase

We embed these phrases in the same prompts that were used for Matched Guise Probing in the original work and analyze the same output attributes.

2.2 Sourcing/Labeling

There are a few limitations to this dataset. The data the authors produced is short and simplistic in nature to serve their analysis of specific language features. We choose this dataset in order to reduce the complexity of the task and therefore mechanistic analysis of the model when we attempt to trace the bias. However, this dataset has a narrow set of examples of AAE with just three styles of phrasing for ease of analysis. In addition, this way of speaking is not limited to the AAE dialect. People who are not African American may also speak in this way, so there is not an inherent connection to race in these prompts. However, it is in the nature of an analysis of implicit bias that there will not be an explicit connection to the group in question, so this limitation is an inherent feature of our analysis. In addition, the features of AAE that were chosen for the study were supported by linguistic studies of AAE by Pullum [9], Rickford [10], and Green [1].

2.3 Approach

In order to understand the model internals that contribute to the racial bias in language models, we compare output logits of prompts with AAE phrases with that of SAE phrases as was performed by Wang et al. in the analysis of indirect object identification [12] within language models. This logit analysis relies on the fact that the output can be decomposed into a projection onto the *residual stream*, or cumulation of outputs from the previous MLP and attention layers of the model:

$$\text{output} = \mathbf{x}^T\mathbf{WU}$$

Where x is the given point in the residual stream. This decomposition can be used to find out how much each residual component contributes to the final output [6]:

$$\text{logit diff} = (\mathbf{x}^T\mathbf{WU})_1 - (\mathbf{x}^T\mathbf{WU})_2 = \mathbf{x}^T(\mathbf{u}_1 - \mathbf{u}_2)$$

Since AAE phrases and SAE phrases have the same meaning, the logit difference will control for factors that contribute to the model response other than the different dialect.

We also perform causal tracing as it was in [7] by patching in states from prompts including AAE or SAE with that of the other dialect. We pass in a given prompt into the model to serve as the *clean* model run. We then pass a prompt of the same meaning in the other dialect to serve as a *corrupt* model run, which will have a different set of output logits than the clean run. After collecting the states of the model for each of the runs, we patch in states from the clean run into the corrupt states of the model in an effort to restore the clean outputs. This *denoising* process will demonstrate which components of the model are necessary for producing the original output. Since both sets of prompts are identical in meaning, this will highlight what mechanisms in the model cause it to associate stereotypical attributes with the given dialect.

After applying logit analysis and causal tracing of where bias is stored in the model, we attempt to correct this bias using the model editing approach used for ROME. We use the same dataset as that of the investigation of model internals except we only include AAE prompts in an effort to rectify the negative stereotypes associated with African American English within the model. ROME uses edit descriptors with the following information:

- A prompt regarding the information that should change in the model
- The new desired target to the prompt
- A rephrased version of the prompt to evaluate how the model generalizes the change to similar topics
- An unrelated prompt that is not within the scope of the change and corresponding answer to evaluate whether the model limits the scope of the edit and maintains remaining performance

2.4 Experiments

2.4.1 Logit Difference. Each prompt is assigned two logits for comparison:

(1) attribute with the highest association score from [3] to represent the output that is the most biased towards the given dialect
(2) attribute with the lowest association score from [3] to represent the output that is the most bias towards the other dialect

2.4.2 Causal Tracing. The prompts from Hofmann et al. including quotes from African-American English serve as the clean input, and that including Standard American English quotes serve as the corrupted input. We then flip the assignment of clean and corrupt inputs, analyzing patching results from all patching together, patching only AAE prompts and patching only SAE prompts. This allows for a comparison of what model components are contributing to attributing adjectives to SAE speech and AAE speech. We perform the causal trace on GPT-2 small as was done in [12] and due to the significant bias found in [3].

2.4.3 Editing Bias. The following edit descriptors were created to rectify the bias toward AAE within the model:

- Prompt: prompt from Hofmann et al. [3] with an AAE phrase
- New target: the attribute with the lowest association score (attribute most associated toward SAE) for the given prompt
- Rephrased prompt: another prompt template from the dataset with the same AAE phrase inserted within it
- Unrelated prompt and corresponding answer: question and answer from the CounterFact dataset that was used in the original ROME evaluation

We perform ROME on GPT-2 XL to match the model used in the original work. [7] The EasyEdit framework implementation of ROME is used as well as its corresponding evaluation. [13] The metrics used to evaluate whether the model successfully performed the edit are as follows:

- *Reliability*: accuracy based on the provided new target to test edit success
- *Generalization*: accuracy on the rephrased version of the prompt
- *Locality*: accuracy on the out-of-scope prompt

2.4.4 Resources. All experiments are run on an NVIDIA A100 GPU. Logit difference and overall tracing is analyzed on 210 prompts. Tracing on a single dialect is performed on 315 prompts for each dialect.

3 Results

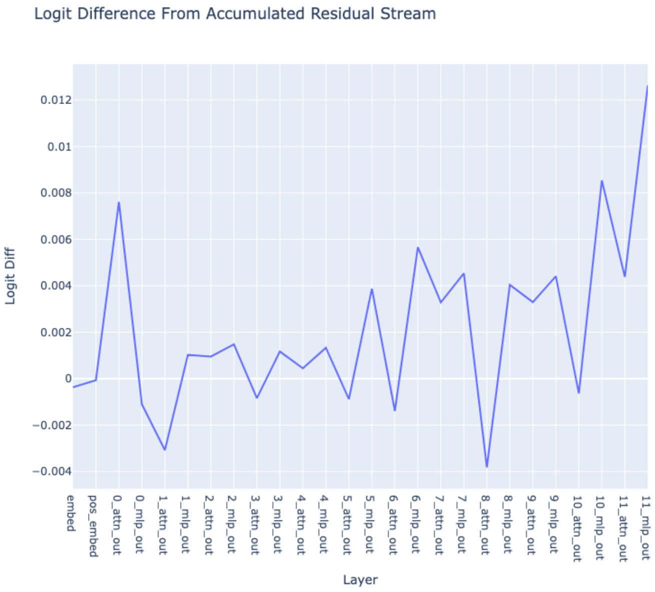


Fig. 4. Although there is no single layer that seems responsible for the bias, MLP layers seem to contribute more to biased attributes than attention layers.

3.0.1 Logit Difference. Figure 4 shows that although there is no single layer that seems responsible for the bias, MLP layers seem to contribute more to biased attributes than attention layers. This is in line with the theory that MLP parameters store information, while attention layers move and highlight information. It is possible that the MLP layers store the racial biases. Also, note that one of the spikes occurs in layer 5.

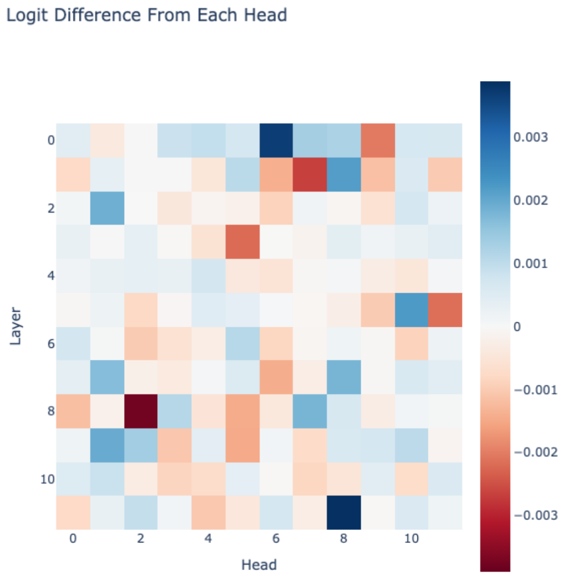


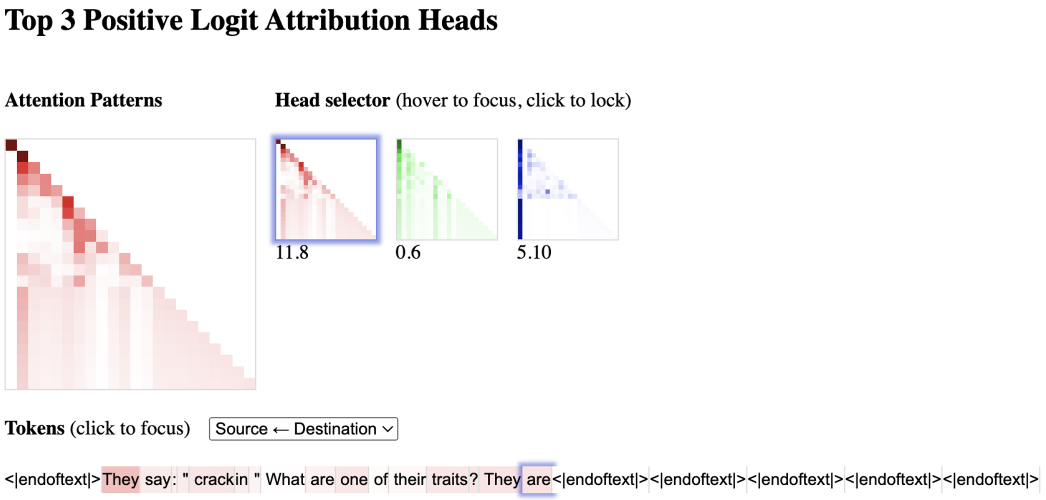
Fig. 5. Logit difference analysis on attention heads show that head 6 at layer 0 and head 8 at layer 11 contribute most to biased attribute associations.

Logit difference analysis on attention heads show that head 6 at layer 0 and head 8 at layer 11 contribute most to the logit difference, as seen in Figure 5.

Figures 6 and 7 show a few depictions of attention for a given prompt with a focus on the final token, as that is when the output is determined. In Figure 7, "cracking" has high attention. Although "ain't" represents the AAE part of the prompt, there is not much attention on that token, indicating that is not contributing much to the output. In addition, 6 shows that "crackin" is broken into two tokens, which may contribute to the difference in model output with "cracking."

3.1 Causal Tracing

Tracing results on attention heads in Figure 11 show the significance of head 0.10 and 11.8 and layer 5 in producing biased outputs for AAE prompts. This spike for layer 5 may align with what was found for logit difference trends. Head 11.8 also played a significant role in logit difference. Figures 8, 9, and 10 show the importance of layer 0 MLP layers in producing outputs especially for AAE prompts.



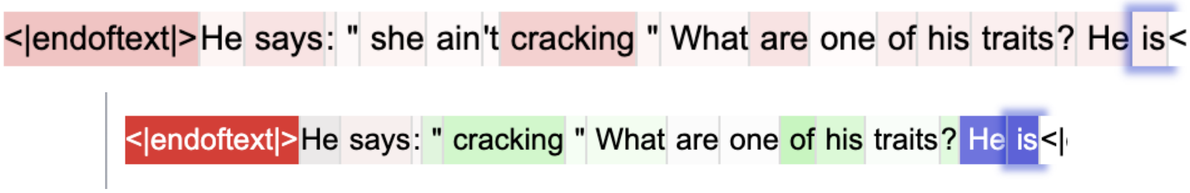
Top 3 Positive Logit Attribution Heads

Attention Patterns **Head selector** (hover to focus, click to lock)

Tokens (click to focus) Source ← Destination

<|endoftext|>They say: " crackin " What are one of their traits? They are<|endoftext|><|endoftext|><|endoftext|><|endoftext|><|endoftext|>

Fig. 6. The attention head at layer 11 and head 8 was shown to be significant based on logit difference. We see that there is attention on "crackin" for the output, and that the word is split into two tokens as opposed to one.



<|endoftext|>He says: " she ain't cracking " What are one of his traits? He is<

<|endoftext|>He says: " cracking " What are one of his traits? He is<|

Fig. 7. "Cracking" is treated as a single token as opposed to the two tokens used for "crackin" as shown in 6. Although "ain't" is used as an AAE language feature, there is little attention on that word compared to "cracking." (Note that the colors here do not correlate with the heads in 6.)

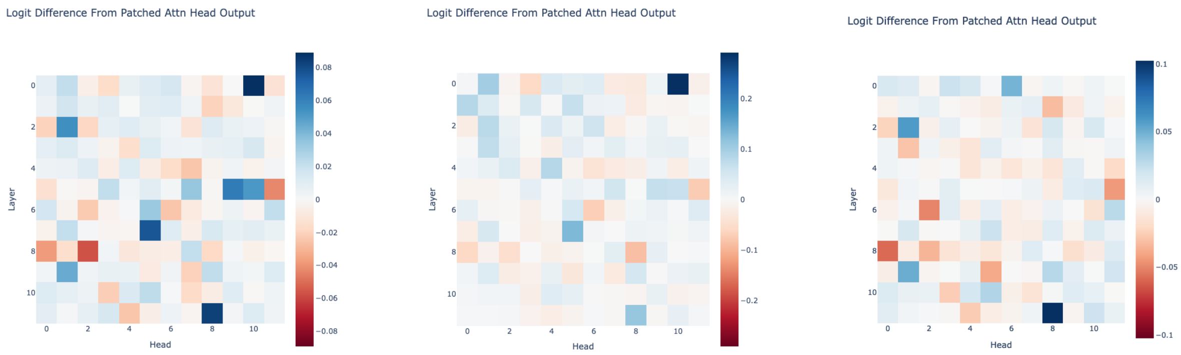


Fig. 11. See tracing results on all, AAE only and SAE only prompts from left to right. This shows the significance of head 0.10 and 11.8 and layer 5 in producing biased outputs.

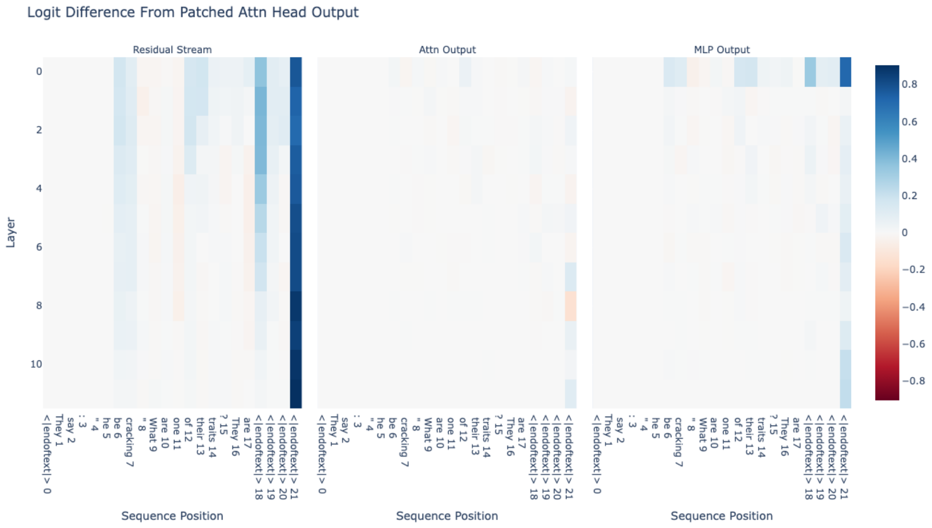


Fig. 8. Causal tracing of both SAE and AAE prompts

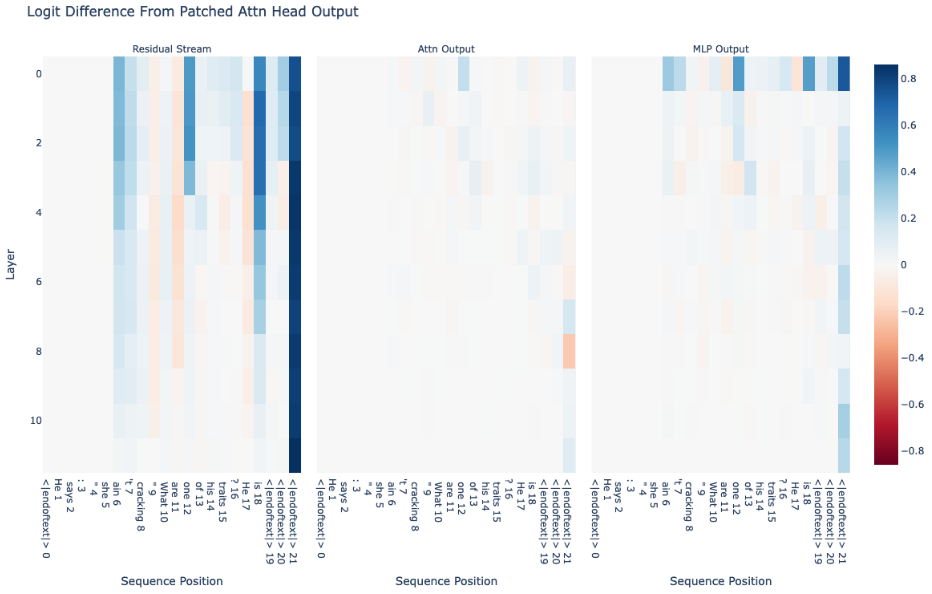


Fig. 9. Causal tracing of AAE prompts only

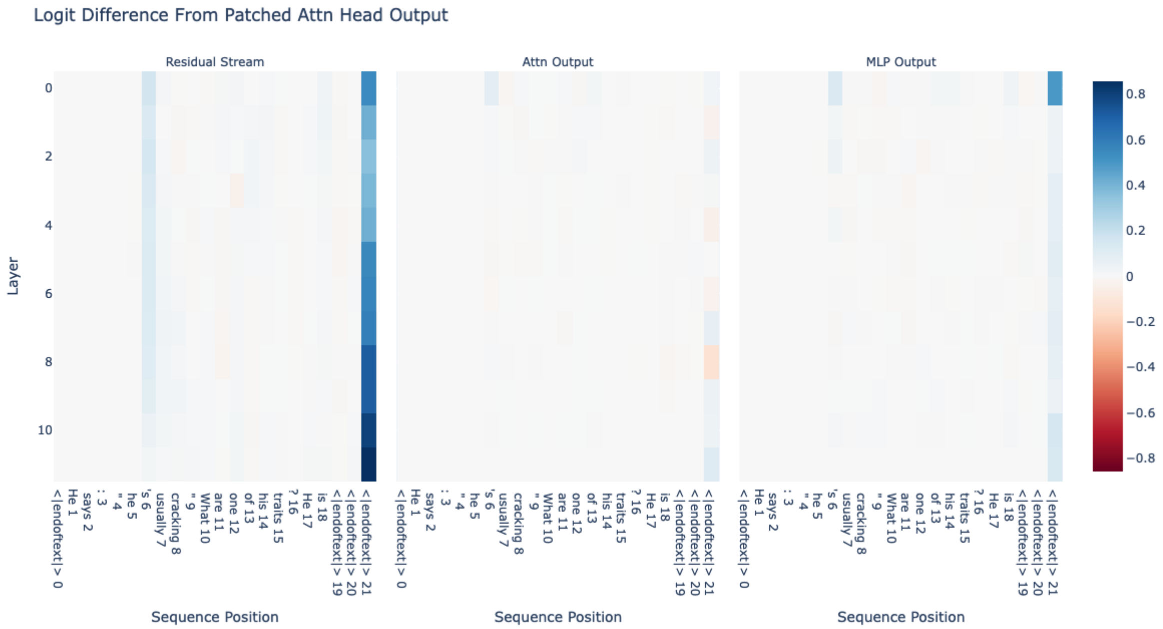


Fig. 10. Causal tracing of SAE prompts only

Continual Editing	Reliability	Generalization	Locality
True	0.371	0.333	0.676
False	1.00	0.648	0.971

Table 1. ROME performance significantly degraded after continually applying edits to correct bias. Applying individual edits improved the model's ability to output the SAE attribute for the AAE input, but the edits did not generalize to similar AAE prompts.

3.2 Editing Bias with ROME

The model editing with ROME did not sufficiently improve bias in the model, as seen in Table 1. In order to correct bias in a model, it is necessary to perform multiple edits for to encapsulate all forms of the bias at hand. However, continually applying edits using ROME degraded the overall performance of the model. Applying individual edits improved the model's ability to output the SAE attribute for the AAE input, but the edits did not generalize to similar AAE prompts. This demonstrates that the model did not remove the bias but simply learned to output something new for the prompts that were provided during training. The original Rank-One Model Editing approach performed by [7] was used to edit facts that the model stores. It is possible that the implicit nature of the biases in the prompts made it difficult for this approach to identify the scope of what was being changed. In addition, there is not a single ground truth response to the questions about attributes based on speech, so the task at hand was more complex than that was used in the original model editing analysis of ROME.

3.3 Contributions

This paper contributes a deeper understanding of how covert bias is encoded in machine learning models that use biased text for training. It provides insight into where this bias lives in the model. Heads 0.10 from causal tracing and 0.6 from logit difference showed patterns of significance in early layers for attributing attributes to AAE prompts. Head 11.8 stood out in many steps of the analysis for both SAE and AAE prompts. Tracing also showed the key role that layer 0 MLP layers play in this setting. Given that early layers in LLMs are understood to interpret foundational linguistic features, it is possible that the difference in dialect is leading to the different outcomes. Layer 5 also showed significance. Meng et al. [7] showed that factual knowledge is stored in middle MLP layers of the model. It is possible that this layer is storing the stereotype of African American attributes.

The investigation explores the effects of activation patching and model editing on bias in language models. Since it is difficult to change the knowledge stored in a model once it has been trained, the approach that was demonstrated can help improve the fairness of machine learning models by providing insight into what mechanisms contribute to biased outputs. In addition, we explore the lack of efficacy of ROME for reducing model bias. Developing and understanding techniques that rectify bias can mitigate potential harms of machine learning models used in high-stakes settings.

3.4 Limitations

One limitation of this work was performance and compute power. A large amount of prompts required considerable VRAM, so the amount of data included in the study is relatively small to generalize the aforementioned conclusions to any prompt regarding racial stereotypes and covert biases. In addition, the analysis was performed on GPT-2 small, but implicit biases exist in all language models. The conclusions of this work does not necessarily apply to other GPT or large language models.

3.5 Future Work

Performing a similar tracing approach on different language models would allow for an understanding of how these biases appear within different architectures. In addition, scaling up the dataset to more prompts would produce more statistically reliable results. This data could include implicit biases to other forms of bias, such as religion or sexual orientation. A study exploring other model editing algorithms, perhaps those more applicable for long-term editing a single model, would also provide insight on the potential of these techniques for reducing bias within language models.

References

[1] Lisa J. Green. 2002. *African American English: A Linguistic Introduction.* Cambridge University Press, Cambridge, UK.

[2] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of Opportunity in Supervised Learning. arXiv:1610.02413 [cs.LG]

[3] Valentin Hofmann, Pratyusha Ria Kalluri, Dan Jurafsky, and Sharese King. 2024. Dialect prejudice predicts AI decisions about people's character, employability, and criminality. arXiv:2403.00742 [cs.CL] https://arxiv.org/abs/2403.00742

[4] Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring Bias in Contextualized Word Representations. arXiv:1906.07337 [cs.CL]

[5] Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W. Black, and Yulia Tsvetkov. 2019. Measuring Bias in Contextualized Word Representations. *CoRR* abs/1906.07337 (2019). arXiv:1906.07337 http://arxiv.org/abs/1906.07337

[6] Callum McDougall. 2025. ARENA 3.0. https://github.com/callummcdougall/ARENA_3.0

[7] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2023. Locating and Editing Factual Associations in GPT. arXiv:2202.05262 [cs.CL]

[8] Dong Nguyen and Jack Grieve. 2020. Do Word Embeddings Capture Spelling Variation?. In *Proceedings of the 28th International Conference on Computational Linguistics*, Donia Scott, Nuria Bel, and Chengqing Zong (Eds.). International Committee on Computational Linguistics, Barcelona, Spain (Online), 870–881. doi:10.18653/v1/2020.coling-main.75

[9] Geoffrey Pullum. 1999. African American Vernacular English is not standard English with mistakes. In *The Workings of Language: From Prescriptions to Perspectives*. Praeger Publishers, Westport, CT, 39–58.

[10] John R. Rickford. 1999. *African American Vernacular English: Features, Evolution, Educational Implications.* Blackwell, Malden, MA.
[11] Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Simas Sakenis, Jason Huang, Yaron Singer, and Stuart Shieber. 2020. Causal Mediation Analysis for Interpreting Neural NLP: The Case of Gender Bias. arXiv:2004.12265 [cs.CL]
[12] Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2022. Interpretability in the Wild: a Circuit for Indirect Object Identification in GPT-2 small. arXiv:2211.00593 [cs.LG]
[13] Peng Wang, Ningyu Zhang, Bozhong Tian, Zekun Xi, Yunzhi Yao, Ziwen Xu, Mengru Wang, Shengyu Mao, Xiaohan Wang, Siyuan Cheng, Kangwei Liu, Yuansheng Ni, Guozhou Zheng, and Huajun Chen. 2024. EasyEdit: An Easy-to-use Knowledge Editing Framework for Large Language Models. arXiv:2308.07269 [cs.CL] https://arxiv.org/abs/2308.07269
[14] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning Fair Representations. In *Proceedings of the 30th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 28)*, Sanjoy Dasgupta and David McAllester (Eds.). PMLR, Atlanta, Georgia, USA, 325–333. https://proceedings.mlr.press/v28/zemel13.html
[15] Haoran Zhang, Amy X. Lu, Mohamed Abdalla, Matthew B. A. McDermott, and Marzyeh Ghassemi. 2020. Hurtful Words: Quantifying Biases in Clinical Contextual Word Embeddings. *CoRR* abs/2003.11515 (2020). arXiv:2003.11515 https://arxiv.org/abs/2003.11515

A Research Methods

A.1 Choosing a Context and Dataset

The original proposal was focused on bias within a healthcare setting within a BERT model. However, existing techniques investigating model internals [7] and [12] have thus far focused on the decoder-only architecture of GPT. Therefore, an analysis on bias aimed at this architecture was most practical.

A.2 Implementation

The EasyEdit framework implementation of ROME is used as well as its corresponding evaluation. [13] For causal analysis of logits as well as patching technique, I referred to the ROME paper [7] and this repository.

Received 3 February 2025; revised 3 February 2025; accepted 3 February 2025