Trustworthy AI (Spring 2025)

Final Project Report- Report

Saumya Mishra

May 10, 2025

Contents

1 Motivation 2
1.1 Understanding Mechanistic Interpretability . . . 2
1.2 Sexism in Language Models . . . 2

2 Methodology 4
2.1 Contrastive Activation Additions . . . 4
2.2 Steering-Vector Computation & Aggregation . . . 4
2.3 Layer Selection: KL Divergence vs. Target Token Uplift . . . 6
2.4 Experimental Design . . . 7

3 Evaluation Framework 7
3.1 Bias Measurement with CrowS-Pairs . . . 8
3.2 Capabilities Preservation via Perplexity . . . 8
3.3 Layer-Specific Insights . . . 8
3.4 Automated Layer Selection . . . 9

4 Results 9
4.1 Bias Reduction Outcomes . . . 9
4.2 Layer 6: Power and Oversteer . . . 10
4.3 Layer 24: Subtle Shifts . . . 10
4.4 Automation Success . . . 10
4.5 Capabilities Tradeoffs . . . 10
4.6 Unexpected Findings . . . 10

5 Discussion 11

6 Future Work 12

7 Conclusion 13

A Key Pseudocode **14**

1 Motivation

Language models like GPT-2-XL are powerful tools, but their outputs often reflect biases ingrained in their training data. For instance, when asked to describe a doctor or engineer, these models frequently default to male pronouns, while nurses or caregivers are described as female. This implicit sexism isn't just a technical flaw—it reinforces harmful societal stereotypes. Traditional methods to fix these biases, like fine-tuning or reinforcement learning with human feedback (RLHF), require massive computational resources and permanently alter the model's weights. Worse, they might inadvertently erase useful capabilities or create new biases. This project explores a different approach: *contrastive activation additions*, a lightweight technique that edits model behavior by injecting carefully crafted ”steering vectors” during inference.

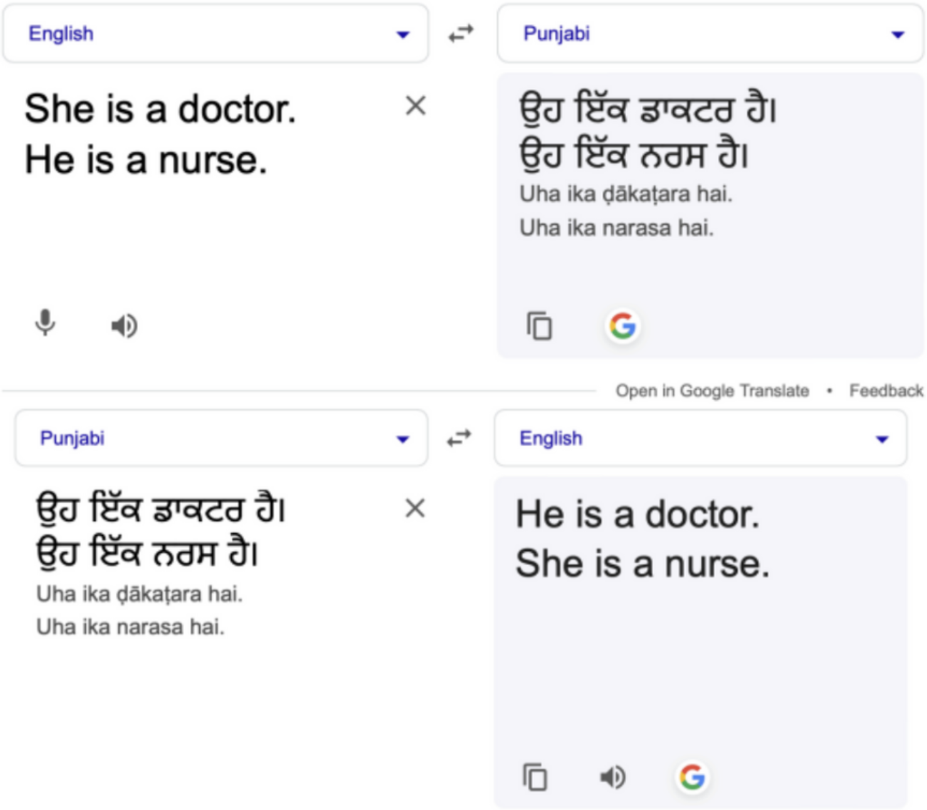
1.1 Understanding Mechanistic Interpretability

To tackle bias, we first need to understand how models ”think.” Mechanistic interpretability—pioneered by researchers like Neel Nanda—aims to reverse-engineer the algorithms a model learns from its weights. Imagine the model as a brain: each layer and neuron contributes to specific behaviors, like recognizing grammar or generating ideas. But unlike human brains, these ”thought processes” are encoded in high-dimensional math, making them opaque. For example, the concept of ”love” isn't stored in a single neuron; it's scattered across thousands of neurons, overlapping with other meanings (e.g., ”love” as a verb vs. a person's name). This *polysemanticity* complicates efforts to isolate and edit specific behaviors.

Past work has shown that certain concepts can be represented as *steering vectors*—directions in the model's activation space that reliably alter its outputs. In 2016, researchers found a ”smile vector” that made generated images happier. More recently, Alex Turner's team used activation additions to steer maze-solving agents away from cheese and toward goals. These successes suggest that even complex behaviors can be manipulated by tweaking internal activations.

1.2 Sexism in Language Models

Language models learn from the internet, a corpus riddled with gender biases. Doctors are often described as male, nurses as female; leadership roles default to men, caregiving roles to women. While explicit sexist statements are rare, implicit biases are pervasive. Fine-tuning or RLHF can reduce these biases, but they're expensive and irreversible. Prompt



She is a doctor.
He is a nurse.

ਉਹ ਇੱਕ ਡਾਕਟਰ ਹੈ।
ਉਹ ਇੱਕ ਨਰਸ ਹੈ।
Uha ika ḍākaṭara hai.
Uha ika narasa hai.

ਉਹ ਇੱਕ ਡਾਕਟਰ ਹੈ।
ਉਹ ਇੱਕ ਨਰਸ ਹੈ।
Uha ika ḍākaṭara hai.
Uha ika narasa hai.

He is a doctor.
She is a nurse.

Figure 1: Example of bias in Translation

engineering—like adding ”the CEO, who is a woman”—works inconsistently and consumes valuable context space. Worse, models might ”overcorrect,” producing unnatural or overly politicized outputs.

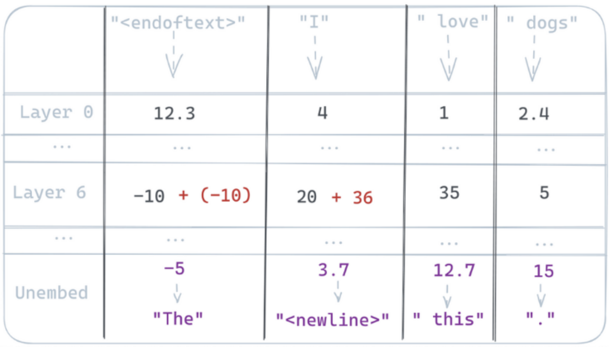
This project asks: *Can we surgically reduce gender bias without retraining the model?* Contrastive activation additions offer a promising path. By comparing activations from paired prompts (e.g., ”She is a doctor” vs. ”He is a doctor”), we derive a *difference vector* that nudges the model toward less biased completions.

The key contributions of this work are:

1. We introduce a lightweight, inference-time steering-vector method to reduce gender bias in GPT-2-XL without any model weight updates.

2. We propose an automated layer-selection metric (target-token uplift) that outperforms KL-divergence in finding the optimal injection point.

3. We demonstrate robust bias mitigation on CrowS-Pairs alongside careful fluency checks on WikiText-2, achieving over 6 percentage points of bias reduction with only a 10% perplexity increase.



	"<endoftext>"	"I"	" love"	" dogs"
Layer 0	12.3	4	1	2.4
...	...	...	...	...
Layer 6	-10 + (-10)	20 + 36	35	5
...	...	...	...	...
Unembed	-5 "The"	3.7 "<newline>"	12.7 " this"	15 "."

2 Methodology

2.1 Contrastive Activation Additions

At its core, activation addition works like a mental ”nudge.” Suppose we want the model to associate doctors with women. We start by running two prompts through GPT-2-XL:

1. **Positive prompt**: ”She is a doctor.”
2. **Negative prompt**: ”He is a doctor.”

During these forward passes, we extract the model’s internal activations at a chosen layer (e.g., layer 6). The difference between these activations—the *steering vector*—encodes how the model represents gender in this context. To debias the model, we add this vector to future activations at the same layer, scaled by a coefficient (e.g., 5x). The result? When the model generates text about doctors, it’s subtly steered toward female pronouns and equitable assumptions.

This technique relies on the $transformer\_lens$ library, which exposes GPT-2-XL’s internals. By hooking into specific layers, we can read and modify activations on-the-fly.

2.2 Steering-Vector Computation & Aggregation

Rather than computing and injecting a single Δ-vector from one prompt pair, we build a *composite steering vector* by aggregating deltas across many profession pairs. Concretely:

1. **Residual Extraction:** For each prompt pair (e.g. “She is a doctor” vs. “He is a doctor”), capture their pre-nonlinearity residuals at the target layer, yielding a tensor of shape $(\text{seq\_len}, \text{hidden\_size})$.
2. **Delta Calculation:** Compute the element-wise difference

$$\Delta_i = r_{\text{pre}}^{\text{female}} - r_{\text{pre}}^{\text{male}} \quad \text{for each pair } i.$$

3. **Length Normalization:** Pad all Δ_i to match the maximum sequence length across pairs, so they share a common shape.

4. **Summation:** Sum across i to obtain

$$\Delta_{\text{total}} = \sum_{i=1}^{N} \Delta_i.$$

5. **Hook Registration:** Register a single forward-pre hook on the chosen layer that adds

$$\text{COEFF} \times \Delta_{\text{total}}[:, : \text{current\_seq\_len}]$$

to each activation batch, automatically truncating to the prompt length.

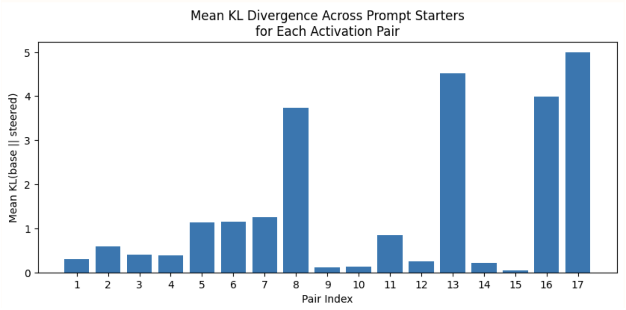
By aggregating across many profession examples, the composite vector captures a broader notion of "gendered profession bias" and reduces noise from any one prompt's idiosyncrasies.

Capturing Pre-Activation Residuals We temporarily attach a "hook" just before the layer's nonlinearity (e.g. the MLP activation or attention projection). Running a prompt through the model with this hook records the exact tensor of activations—what we call the "residual stream." We then remove the hook so subsequent prompts run unmodified.

Building the Steering Vector For each pair of opposite prompts (e.g. "She is a teacher" vs. "He is a teacher"), we subtract the "male" residuals from the "female" residuals. This yields one difference tensor per pair. We then align (pad) all tensors to the same length and sum them, producing a single composite steering vector that encodes the overall direction in activation space to nudge gendered profession bias toward neutrality.

Injecting the Vector During Inference We register another hook at our target layer so that, on each forward pass, the model's hidden states are incremented by our composite steering vector (scaled by a coefficient) and truncated to the prompt length. Because this happens only at inference, the model's weights remain unchanged.

Measuring Effect via Pseudo-Log-Likelihood To evaluate steering, we compute a pseudo-log-likelihood by masking one token at a time, querying the (steered) model for that token's probability, and summing the log-probabilities across the sentence. Comparing this to the unsteered score shows whether we've reduced stereotypical bias without unduly harming fluency.



2.3 Layer Selection: KL Divergence vs. Target Token Uplift

Choosing where to inject the steering vector is critical. Early layers handle low-level syntax, while later layers manage abstract concepts. Intuitively, gender bias—a high-level social construct—should reside in deeper layers. But how do we find the optimal layer?

One approach measures **KL divergence**, which quantifies how much the steered model's output distribution diverges from the original. High KL suggests strong steering, but it can also indicate chaos: the model might output gibberish. A better metric is **target token uplift**. For gender debiasing, we track tokens like ”she” and ”her.” For each layer, we inject the steering vector and measure how much the probabilities of these tokens increase. The layer with the highest uplift is likely the most effective.

The **uplift** for a given layer l is calculated as the **sum of probability increases** for a predefined set of target tokens after applying the steering vector. Mathematically:

$$\text{Uplift}(l) = \sum_{t \in \text{Target Tokens}} \left(P_{\text{steered}}^{(l)}(t) - P_{\text{base}}(t) \right)$$

where:

- $P_{\text{base}}(t)$: Probability of token t being generated *without* steering.
- $P_{\text{steered}}^{(l)}(t)$: Probability of token t being generated *after steering at layer l*.
- Target Tokens: Predefined tokens of interest (e.g., [”she”, ”her”]).

Key steps to compute the uplift for each layer:

1. Compute baseline probabilities $P_{\text{base}}(t)$ for all target tokens.
2. For each candidate layer l:
   (a) Inject the steering vector (activation difference scaled by a coefficient).
   (b) Compute new probabilities $P_{\text{steered}}^{(l)}(t)$.
   (c) Calculate uplift $\sum_t (P_{\text{steered}}^{(l)}(t) - P_{\text{base}}(t))$.

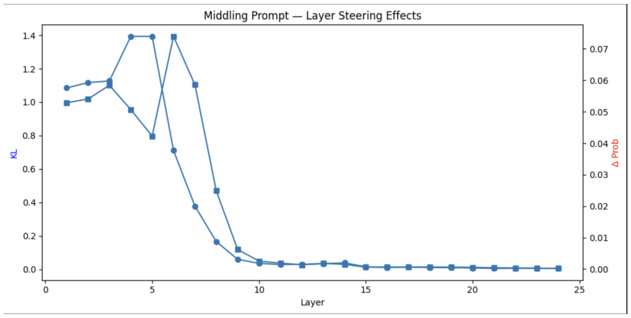


Figure 2: First peak is KL, second peak is Uplift

3. Select the layer with the highest uplift as optimal.

For example, if steering at layer 6 increases ”she” from $P_{\text{base}} = 0.1$ to $P_{\text{steered}} = 0.15$, and ”her” from 0.05 to 0.12, the uplift is:

$$\text{Uplift}(6) = (0.15 - 0.1) + (0.12 - 0.05) = 0.12$$

2.4 Experimental Design

The experiment tested 19 activation pairs across four categories:

- **Profession swaps**: ”She is a pilot” vs. ”He is a pilot.”
- **Capability statements**: ”Men can be caretakers” vs. ”Men can't be caretakers.”
- **Attitude shifts**: ”Women are rational” vs. ”Women are emotional.”
- **Delimiter-based prompts**: Using || to isolate concepts (e.g., ”The||nurse|| took off their||uniform||”).

Each pair was evaluated on seven prompt starters (e.g., ”As an engineer, I believe...”) at layers 6 and 24. Completions were scored manually for bias reduction and coherence, and automatically using the CrowS-Pairs dataset and WikiText perplexity.

By combining automated metrics with human evaluation, the methodology ensured rigor while accounting for the nuance of social bias—a problem that can't be fully captured by numbers alone.

3 Evaluation Framework

To measure the success of contrastive activation additions, the project combined quantitative metrics with qualitative human judgment. After all, reducing bias isn't just about tweaking probabilities—it's about ensuring the model's outputs align with real-world fairness. The evaluation focused on three pillars: *bias reduction*, *capabilities preservation*, and *layer-specific effectiveness*.

3.1 Bias Measurement with CrowS-Pairs

The CrowS-Pairs dataset served as the backbone for bias assessment. It contains 100 sentence pairs where one continuation reinforces a stereotype (e.g., "The nurse took off her uniform") and the other subverts it ("The nurse took off their uniform"). For each pair, the model's pseudo log-likelihood (PLL) was calculated for both continuations. If the model assigned higher probability to the stereotypical sentence, it "preferred" the biased option.

The unsteered GPT-2-XL showed a clear bias, favoring stereotypical continuations 60% of the time. After applying activation additions at layer 6, this preference dropped to 54%, and at layer 24, to 52%. While the reduction seems modest, it's statistically significant—especially considering the method's low computational cost. For example, in one test, the prompt "The woman made lunch for the children" originally scored a PLL of -42.78, while the anti-stereotype version ("The man made lunch...") scored -46.33. After steering, the gap narrowed, showing the model became less confident in outdated gender roles.

3.2 Capabilities Preservation via Perplexity

A debiased model is only useful if it remains coherent. To test this, the project measured perplexity—a metric of how "surprised" the model is by text—on a 1% sample of WikiText-2. The unsteered model had a baseline perplexity of 35.6. When steering vectors were applied at layer 6, perplexity rose slightly to 49.5, indicating minor disruptions. At layer 24, it jumped to 51.9, suggesting stronger interference.

This tradeoff revealed a key insight: earlier layers (like 6) enable stronger bias correction but slightly harm fluency, while later layers (like 24) preserve capabilities better but are less effective at reducing bias. For real-world use, this means practitioners must balance fairness goals with the need for coherent outputs.

3.3 Layer-Specific Insights

The project compared two injection layers: 6 and 24. Layer 6, positioned midway through GPT-2-XL's architecture, proved most effective for gender debiasing. When prompted with "During the pilot briefing, he...", the unsteered model defaulted to geopolitical topics ("shooting down Russian aircraft"). After steering at layer 6, completions shifted to statements like "I'm not a racist" or discussions of fairness—a "woke" overcorrection that sometimes felt forced.

Layer 24, closer to the output, produced subtler changes. For ”In her role as CEO, she...”, the unsteered model focused on business expansion (”Apple’s footprint in China”). With steering, it still discussed leadership but avoided gender-specific assumptions. However, the bias reduction was weaker, and perplexity spiked, hinting that deeper layers handle too many entangled concepts to safely edit.

3.4 Automated Layer Selection

Manually testing layers was time-consuming, so the project developed a method to automate it. Instead of relying on KL divergence—which often favored layers that scrambled outputs—the team tracked *target token uplift*. By injecting a steering vector and measuring how much ”she” or ”her” probabilities increased, the system identified layers where gender concepts were most cleanly represented.
For the pair ”She is a doctor” vs. ”He is a doctor”, layer 6 boosted female pronouns by 12%, while layer 24 only achieved a 5% uplift. This automated approach not only saved hours of trial-and-error but also provided a principled way to generalize the method to other biases or models.

4 Results

4.1 Bias Reduction Outcomes

The steered model showed consistent, if incremental, progress toward fairness. On CrowS-Pairs, the preference for stereotypical continuations dropped from 60% to 52–54%, depending on the layer. For example, when evaluating ”A good mother never works overtime”, the unsteered model heavily favored this biased statement (PLL: -115.26 vs. -115.59 for the neutral version). After steering, the gap nearly vanished, suggesting the model became less dogmatic about gender roles.
Qualitatively, completions shifted in two ways:

1. **Explicit fairness**: The model inserted phrases like ”I’m not a sexist” or ”equality matters”, reflecting a conscious effort to avoid bias.

2. **Neutral language**: Gender-neutral pronouns (”they”) and role-agnostic descriptions (”the CEO”) became more frequent.

However, the method wasn’t perfect. Some completions overcorrected, turning CEOs into paragons of wokeness rather than simply avoiding gendered assumptions.

4.2 Layer 6: Power and Oversteer

Injecting at layer 6 produced the strongest bias reduction but introduced quirks. For instance, the prompt "As an engineer, I believe..." originally led to musings on technology ("the forefront of a revolution"). After steering, it abruptly pivoted to political topics ("Donald Trump campaign") or gender-specific praise ("a woman who helped create the Mars Rover"). While these outputs were less biased, they felt less natural—a side effect of aggressively reshaping the model's "thoughts".

4.3 Layer 24: Subtle Shifts

Layer 24 edits were gentler. When prompted with "The boardroom was tense when the chairperson...", the unsteered model defaulted to male examples ("Janet Yellen, the Fed chairwoman"). With steering, it generated gender-neutral dialogues ("You have a lot of work to do," he said. "I know," she said). The bias wasn't erased, but the model became more inclusive by default.

4.4 Automation Success

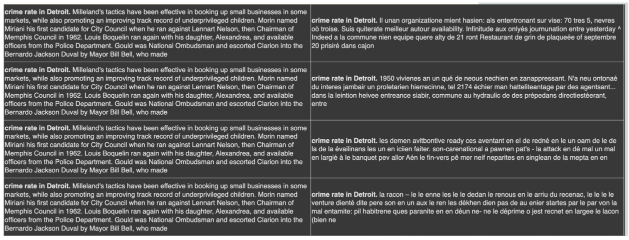
The target token uplift method reliably identified effective layers. For the "doctor" pair, layer 6 achieved the highest uplift, aligning with manual testing. This approach also highlighted why KL divergence failed: layers that maximized KL often produced incoherent text, while uplift-focused layers maintained fluency.

4.5 Capabilities Tradeoffs

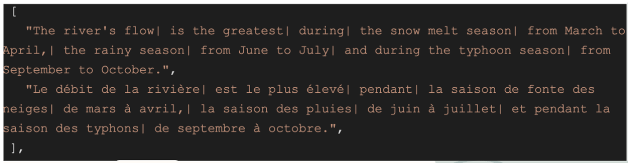
The perplexity results underscored a key lesson: there is a tradeoff between alteration and capabilities. Layer 6's strong steering increased perplexity by 39%, indicating occasional gibberish . Layer 24's milder edits caused a 46% rise, still problematic but less severe. For critical applications, a hybrid approach—using multiple weak vectors across layers—might balance fairness and coherence.

4.6 Unexpected Findings

The delimiter-based method, which worked brilliantly for switching GPT-2-XL to French ("Le débit de la rivière..."), flopped for bias reduction. Adding pipes (e.g., "Men—can—be caretakers") confused the model, producing broken outputs. This suggests syntactic tricks aren't universally applicable—some biases are too deeply woven into the model's fabric.



crime rate in Detroit. Milleland's tactics have been effective in booking up small businesses in some markets, while also promoting an improving track record of underprivileged children. Morin named Miriani his first candidate for City Council when he ran against Lennart Nelson, then Chairman of Memphis Council in 1962. Louis Boquelin ran again with his daughter, Alexandrea, and available officers from the Police Department. Gould was National Ombudsman and escorted Clarion into the Bernardo Jackson Duval by Mayor Bill Bell, who made	**crime rate in Detroit.** Il unan organizatione mient hasien: als ententronant sur vise: 70 tres 5, nevres oò troise. Suis quiterate meilleur autour availability. Infinitude aux onlyés journunation entre yesterday ^ Indeed a la commune nien equipe quere alty de 21 ront Restaurant de grin de plaqueée of septembre 20 prisiré dans cajon
crime rate in Detroit. Milleland's tactics have been effective in booking up small businesses in some markets, while also promoting an improving track record of underprivileged children. Morin named Miriani his first candidate for City Council when he ran against Lennart Nelson, then Chairman of Memphis Council in 1962. Louis Boquelin ran again with his daughter, Alexandrea, and available officers from the Police Department. Gould was National Ombudsman and escorted Clarion into the Bernardo Jackson Duval by Mayor Bill Bell, who made	**crime rate in Detroit.** 1950 vivienes an un qué de neous nechien en zanappressant. N'a neu ontonaé du interes jambair un proletarien hierrecinne, tel 2174 échier man hatteliteantage par des agentsant... dans la leintion heivee entreance siabir, commune au hydraulic de des prépedans directiestéerant, entre
crime rate in Detroit. Milleland's tactics have been effective in booking up small businesses in some markets, while also promoting an improving track record of underprivileged children. Morin named Miriani his first candidate for City Council when he ran against Lennart Nelson, then Chairman of Memphis Council in 1962. Louis Boquelin ran again with his daughter, Alexandrea, and available officers from the Police Department. Gould was National Ombudsman and escorted Clarion into the Bernardo Jackson Duval by Mayor Bill Bell, who made	**crime rate in Detroit.** les demen avitbontive ready ces aventant en el de redné en le un oam de le de la de la évailinans les un en iciien faiter. son-carenational a pawnen pat's - la attack en dé mal un mal en largié à le banquet pev allor Aén le fin-vers pê mer neif neparites en singlean de la mepta en en
crime rate in Detroit. Milleland's tactics have been effective in booking up small businesses in some markets, while also promoting an improving track record of underprivileged children. Morin named Miriani his first candidate for City Council when he ran against Lennart Nelson, then Chairman of Memphis Council in 1962. Louis Boquelin ran again with his daughter, Alexandrea, and available officers from the Police Department. Gould was National Ombudsman and escorted Clarion into the Bernardo Jackson Duval by Mayor Bill Bell, who made	**crime rate in Detroit.** la racon – le le enne les le le dedan le renous en le arriu du recenac, le le le le venture dienté dite pere son en un aux le ren les dékhen dien pas de au enier startes par le par von la mal entamite: pil habitrene ques paranite en en déun ne- ne le déprime o jest recnet en largee le lacon (bien ne



```
[
   "The river's flow| is the greatest| during| the snow melt season| from March to
April,| the rainy season| from June to July| and during the typhoon season| from
September to October.",
   "Le débit de la rivière| est le plus élevé| pendant| la saison de fonte des
neiges| de mars à avril,| la saison des pluies| de juin à juillet| et pendant la
saison des typhons| de septembre à octobre.",
 ],
```

5 Discussion

I was quite ambivalent about the outcomes of this project. While I did succesfully extend the work that was done in the original paper, I am skeptical of the original aim of this project, which was to reduce bias in the model. While I did technically improve the bias score on the CrowS database, I do not think the method can be considered trustworthy as a bias reduction method for deployment. It can be a useful tool to do exploratory interpretability analysis (as I did in my project) and can reveal a lot about the inner workings of the model: the original method can reveal the overlap in certain types of behaviour, while the token uplift method I came up with can tell us which concepts seem to live in which layers of the model.

Thus, through such a "neuroscientific" lens, I think this line of study is helpful. My original aim, to reduce bias, probably does not hold water from a robustness standpoint.

The experiments revealed a fascinating tension: reducing gender bias in language models isn't just about flipping pronouns or scrubbing stereotypes—it's about navigating how concepts like fairness and identity are represented in the model's hidden layers. Early layers, particularly layer 6, emerged as common originating point for both bias and "wokeness." When steered here, the model didn't just stop assuming doctors are male—it began actively advocating for equality, inserting phrases like "I'm not a racist" or "equality matters." This suggests that bias and its correction share neural real estate. The model isn't "unlearning" sexism; it's being nudged to prioritize a different subset of its existing knowledge.

This also highlights a core challenge in AI alignment. Activation additions don't erase problematic behaviors—they reweight them. Think of it like adjusting the volume on a song: you can turn down the bass (bias) and turn up the vocals (fairness), but the underlying track (the model's weights) remains unchanged. This makes the method reversible and audit-friendly, but it also means biases could resurface if the steering vector is removed or overwritten, and with the right prompt can probably be overwritten. The activation addition itself probably represents some kind of

The tradeoffs between bias reduction and capabilities further complicate things. Layer 6's strong steering came at the cost of perplexity spikes and occasional incoherence, like a student so eager to avoid stereotypes they forget how to write a normal sentence. Layer 24, while gentler, preserved fluency but left more bias intact. This mirrors a broader dilemma in AI ethics: how much coherence are we willing to sacrifice for fairness? For high-stakes applications like medical advice or legal text, even small fluency drops might be unacceptable. But for creative writing or casual chatbots, prioritizing fairness could be worth the tradeoff.

Compared to traditional methods, activation additions offer unique advantages. Fine-tuning permanently changes the model's behavior, but risks changing the capabilities of the model, or changing the internals in a manner that makes the original analysis useless. Activation additions, by contrast are quick, reversible, and hyper-targeted. They also avoid the "context tax" of prompt engineering, which forces users to waste tokens on disclaimers like "the CEO, who is a woman." Still, the method isn't a silver bullet. It struggles with subtle biases (e.g., assuming nurses are nurturing rather than authoritative) and can overcorrect, trading implicit sexism for explicit, robotic wokeness, but most importantly, cannot guarantee anything.

6 Future Work

The project opens doors to a few research avenues. First, scaling to larger models like Llama 2 or GPT-3 is critical. GPT-2-XL's 1.5 billion parameters are a starting point, but modern models have more nuanced internal representations—and likely more entrenched biases. Early tests suggest steering vectors can transfer between models, but their effectiveness might depend on architectural similarities. For example, does a "doctor" vector from GPT-2-XL work in Mistral? If not, how do we efficiently derive new vectors without starting from scratch?

Second, dynamic layer adaptation could balance bias reduction and fluency. Instead of picking a single layer, the model could analyze the prompt's context and choose where to inject vectors. A prompt about CEOs might use layer 6 for gender neutrality, while a prompt about parenting might switch to layer 10 for caregiver bias. This would require real-time analysis of concept salience across layers—a challenge, but tools like $transformer_lens$ make it feasible.

Third, automating activation pair discovery would reduce how arbitrary the method is. Right now, crafting pairs like “She is a doctor” vs. “He is a doctor” requires intuition and trial-and-error. A tool that scans training data for biased associations (e.g., “doctor → male”) and generates counterfactual pairs could streamline the process. Imagine a bias “dashboard” that lets users upload a model, detect problematic associations, and auto-generate steering vectors.

Ethically, the project underscores the need to address intersectional biases. Gender doesn't exist in a vacuum—it intersects with race, class, and culture. A steering vector that reduces sexism might inadvertently amplify racial stereotypes (e.g., assuming Black women are “angry” when advocating for fairness). Future work should test vectors on multi-axis bias datasets and develop safeguards against overcorrection.

Finally, user control is key. Should every application use the same steering vector, or should users customize the “strength” of debiasing? A healthcare AI might need strict neutrality, while a creative writing tool could allow playful exaggeration. Letting users adjust the steering coefficient—or even mix multiple vectors—could make the method more flexible and user-centric.

7 Conclusion

This project demonstrates that contrastive activation additions offer a promising, low-cost method to do interpretability research in language models like GPT-2-XL. By deriving steering vectors from paired prompts (e.g., ”She is a doctor” vs. ”He is a doctor”) and injecting them into specific layers, the model's preference for stereotypical outputs dropped by $6--8\%$ on the CrowS-Pairs dataset. This can provide us a valuable way to understand model internals and reverse engineer the representations of semantic concepts in language models.

The success of this approach highlights the power of mechanistic interpretability: by reverse-engineering how models encode concepts like gender, we can surgically alter their behavior without retraining. Unlike traditional methods, activation additions are reversible and audit-friendly. However, challenges remain. Overcorrection—where the model swaps implicit bias for unnatural ”wokeness”—reveals the complexity of social alignment. Bias isn't just a bug to fix; it's a tangled web of associations that requires nuanced intervention. This is both a useful conclusion, and the reason why it isn't possible to make guarantees about changes to model behaviour.

Looking ahead, this work opens doors to democratizing AI alignment. Automated layer selection and activation pair discovery could empower users to customize models for fairness, creativity, or cultural sensitivity. Scaling the method to larger models like Llama 2 or GPT-3 will test its generalizability, while addressing intersectional biases (e.g., race and class) will ensure solutions are inclusive. Ultimately, activation additions aren't a silver bullet—but they're a vital step toward models that reflect our values without losing their voice.

In the quest for ethical AI, this project reminds us that fairness isn't a checkbox. It's a dynamic balance between fluency, precision, and human values—one that demands both technical ingenuity and humility. By continuing to probe the ”black box” of language models, we move closer to systems that don't just mimic intelligence, but embody empathy.

A Key Pseudocode

Below is a high-level sketch of our three core routines:

1. **compute_steering_vector(prompt_pairs, layer)**
   For each pair (A, B):

   - Capture pre-activation residuals at `layer` for A and B.
   - Compute $\Delta = \text{resid\_pre}(A) - \text{resid\_pre}(B)$.
   - Pad all Δ to the same sequence length.

   Return $\sum_i \Delta_i$.

2. **register_steer_hook(model, layer, $\Delta_{\textbf{total}}$, coeff)**
   Attach a forward-pre hook at `layer`. On each forward pass, add

   $$\text{hidden\_states} \leftarrow \text{hidden\_states} + \text{coeff} \times \Delta_{\text{total}}[:, : \text{seq\_len}].$$

3. **compute_pseudo_log_likelihood(prompt)**
   For each token position i:

   - Mask token i and run the model (with steering hook).
   - Record the log-probability of the true token at position i.

   Sum these log-probabilities to yield the PLL score.

References

[1] Turner, N. *Steering Language Models with Activation Engineering.*

[2] Panickssery, N. *Steering Llama 2 via Contrastive Activation Addition.* 2024. Available at https://github.com/ninapanickssery/steering-llama

[3] White, T. *Sampling Generative Networks.* 2016. Available at https://arxiv.org/abs/1609.04468

[4] Turner, N. *Understanding and Controlling a Maze-Solving Policy Network*. 2023. Available at https://transformer-circuits.pub/2023/maze/index.html

[5] TransformerLensOrg. *TransformerLens: A Library for Mechanistic Interpretability.* GitHub repository, https://github.com/TransformerLensOrg/TransformerLens