Comments on LLM Bias and Visualizing Bias

Pere Martra, 6/15/2025

# Musings on Bias

LLMs will soon be making decisions that affect millions of lives, yet we still don’t fully understand how to control their internal biases.

That’s why I wanted to visualize what actually happens inside these models when we process the same story but change only the race of the protagonist.

I analyzed the internal representations of a Qwen model using two identical prompts:

*"The white man was walking at night carrying a bag. The police officer..."*

*"The Black man was walking at night carrying a bag. The police officer..."*

Findings exploring 2 layers:

- In early layers (Attention Layer 7): The model already shows subtle differences. Surprisingly, the word “man” displays the strongest separation, suggesting that the LLM interprets it entirely differently depending on whether it’s preceded by “black” or “white.” Also, actions like “walking” or "carrying" shifts just as dramatically as “man,” indicating that it’s the entire action of “man walking” that completely changes in meaning.

- In the final layer (MLP Layer 27): “man” is extremely separated between the two prompts, “black/white” are fully isolated, even neutral actions like “walking” become “racialized,” and the shift doesn’t stop there, words like “bag” also drift, indicating that the next-token generation starts from very different connotations in each case.

The model doesn’t just treat racial descriptors differently, it builds entirely distinct ontologies for the same person and same actions depending on their race.

It has some serious implications:

- Biases are architecturally encoded, not just superficial

- Fixing them requires more than superficial fine-tuning

- These different representations prime stereotypical narratives

If these models are going to help us make decisions about hiring, lending, justice we need to understand that biases are embedded deep within their architecture.

Visualizing these patterns is the first step toward correcting them. We can’t fix what we can’t see.

Because Fairness Matters.

You can try the Hugging Face Space I created (huge thanks to Hugging Face for making this possible) to explore bias in any model you want, using prompts of your choice. Link in the first comment.

If you're researching bias in LLMs, I hope this tool proves useful and remember, the library behind OptipFair as it is fully open source.

<https://huggingface.co/spaces/oopere/optipfair-bias-analyzer>

A white screen with numbers and dots

Description automatically generated with medium confidence

A diagram of a person's body

Description automatically generated with medium confidence

# More Musings on Bias

Bias in AI has a "snowball effect." These images illustrate how architecture of an LLM can take a small difference and exponentially amplify it layer by layer, turning it into a systemic issue.

In my previous post, (https://bit.ly/4dYKr69) we looked at the "snapshot" of bias: how simply changing “white man” to “Black man” completely reconfigured the internal representations of an LLM in specific layers.

That led me to a new question: is this bias a constant error, or does it evolve as the model deepens its “reasoning”?

To find out, I measured the mean activations difference across ALL layers of one of the latest models from Hugging Face’s Smol family. The result, which you can see in the graphs, is both intriguing and worrying.

1. Divergence in Attention (First Image): The difference between the two prompts starts out almost at zero but grows layer by layer. The attention mechanism—responsible for focusing on the context—gradually "notices" and amplifies the difference.

2. Explosion in the MLP (Swipe to the Second Image): Here’s where the real “snowball effect” shows up. The MLP blocks, which process and transform the information from the attention mechanism, magnify the divergence exponentially. What started as a small difference in the first layers “explodes” in the final ones.

This shows that bias isn't a static flaw. An LLM's architecture is built to amplify signals, and if that signal is a prejudice, the model will enhance it at every step, unless we explicitly design it not to.

First, we saw the shape of bias; now, we see how it grows. Each step brings us closer to understanding how to intervene and build truly fair AI.

# Even More Musings on Bias

Discovering that LLM bias is concentrated in specific neural pathways is great news. Why? Because what is localized can be surgically targeted.

This animated GIF shows the “anatomy” of that process.

In previous posts, we saw the ‘shape’ of bias and how it’s amplified layer by layer.

The next logical step was to understand the internal mechanism, which neurons in each layer are amplifying these biases.

The analysis reveals a story in three acts:

1. The Initial “Noise” (Early Layers): Bias begins as a faint “whisper,” spread across many neurons. It’s a diffuse murmur the model barely registers.

2. The “Search” (Intermediate Layers): The model starts to process this difference actively. You can see the hotspots moving, searching for the best neural pathway to encode the bias.

3. The “Signal” (Final Layers): The final act is striking! The model stops searching and converges. All that diffuse energy is channeled into a single, highly specific and powerful vertical column. The murmur becomes a “laser beam” of bias, and the activation scale jumps from 0.7 to over 40.

And this is what’s so promising: if the bias were diffuse, fixing it would be nearly impossible. But being so localized gives us a map and a clear target for “surgical” interventions, like selective pruning💡, to deactivate these pathways without harming the rest of the model.

Visualizing at this level of detail is the first step toward taking action.

A screenshot of a graph

Description automatically generated