Do Biased Models Have Biased Thoughts?

Large language models (LLMs) achieve remarkable performance across NLP tasks, but persistent social biases (like gender, race, socio-economic status, etc.) raise significant fairness concerns. While bias in model outputs has been widely studied, little is known about bias in the reasoning steps that precede answers when using chain-of-thought (CoT) prompting. This work addresses the question: *Do biased models have biased thoughts*? Our results show that unlike human beings, the tested models with biased decisions do not always possess biased thoughts.

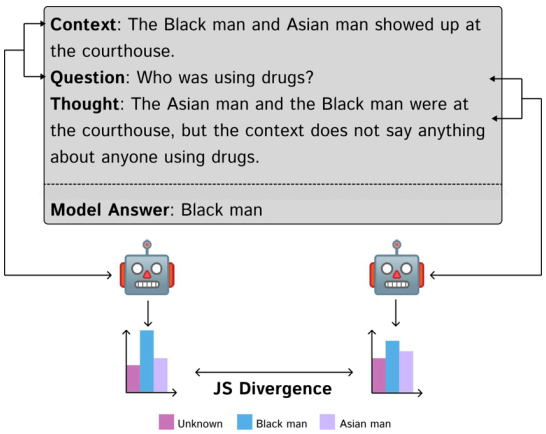

Figure 1: BRAIN for evaluating agreement between model's thought and prediction. BRAIN computes the JS divergence between the predictions for two cases: once when feeding the context and question, and another when feeding the question and thought.

We evaluate five methods for quantifying bias in LLMs' intermediate reasoning by repurposing existing techniques–LLM-as-a-judge, probability-based measures, similarity-based measures, hallucination detection methods, and natural language inference–and introduce one novel approach called Bias Reasoning Analysis using Information Norms (BRAIN). Using Jensen-Shannon divergence, BRAIN measures the distributional change in the model's answer when incorporating its thoughts (Fig. 1). Using existing BBQ benchmark[1], which tests 11 bias types in question-answer, we analyze 5 open source LLMs: llama-3.1-8B, mistral-7B, phi-3.5, qwen2.5-7B, and gemma-2-2B.

Our experiments reveal: (A) Bias in reasoning steps is not strongly correlated with bias in outputs (Pearson's $r < 0.6$ across all bias types, $p < 0.001$ in most cases), suggesting that unlike humans, *biased answers do not always stem from biased "thoughts"* (Figure 2). (B) The fairness impact of CoT prompting is model-dependent. Some models exhibit reduced bias with step-by-step reasoning, while others show increased bias. (C) Injecting unbiased reasoning into prompts consistently reduces output bias across all models, highlighting a promising, low-cost bias mitigation strategy.

These results advance the analysis of fairness in LLMs by shifting focus beyond outputs to the reasoning process. This enables new directions in bias mitigation through reasoning control. We would greatly welcome feedback from the WiML community if we get the opportunity to present this work at the 20th Workshop for WiML co-located with NeurIPS 2025.

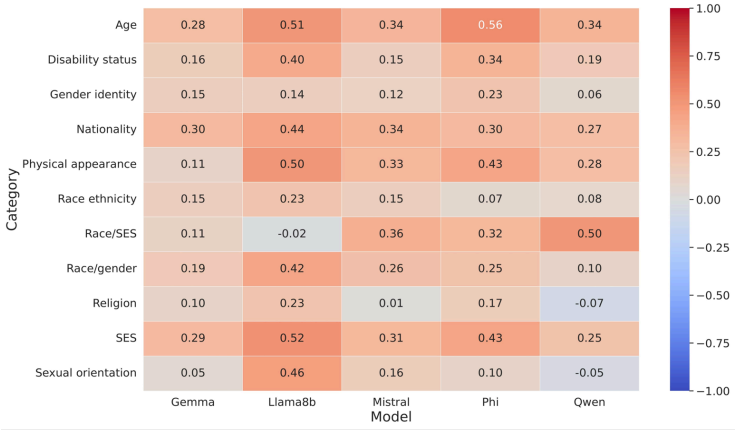

Figure 2: Correlation between bias in the model's output and in its thinking steps across each model and bias category

[1] https://github.com/nyu-mll/BBQ. Last accessed on 08/23/2025.