# Toward Inclusive Language Models: Sparsity-Driven Calibration for Systematic and Interpretable Mitigation of Social Biases in LLMs

**Prommy Sultana Hossain**[1], **Chahat Raj**[1], **Ziwei Zhu**[1], **Jessica Lin**[1] and **Emanuela Marasco**[2]
[1]Dept. of Computer Science [2]Dept. of Information Systems and Technology
George Mason University, Fairfax, VA, USA

## Abstract

Large Language Models (LLMs) such as GPT and LLaMA excel in natural language tasks, e.g., text generation and machine translation. However, inherent biases from training on vast Internet datasets potentially amplify harmful stereotypes—widely held, oversimplified, and often inaccurate generalizations about groups of people. Our contribution introduces a novel, systematic, and architecture-aware method to identify and mitigate stereotypical bias in decoder-only transformer models. This interpretable approach operates without gradient access or retraining from scratch. We first evaluate bias and then apply a bias localization mechanism that correlates internal activations with a newly defined Context Influence (CI) Score. Our method pinpoints specific attention heads that consistently align with biased shifts in model predictions. To mitigate this, we introduce a soft pruning strategy that scales attention head parameters based on their correlation strength, followed by lightweight fine-tuning to maintain fluent text generation. Experiments across five models demonstrate our approach reduces bias by up to 37% on BBQ, 32% on StereoSet, and 33% on CrowS-Pairs while simultaneously improving reasoning performance on MMLU by up to 10%. [1]

## 1 Introduction

Large Language Models (LLMs) have emerged as transformative technologies in natural language processing, demonstrating remarkable capabilities across diverse applications from content generation to complex reasoning. Models such as GPT-4, LLaMA, and Claude have achieved unprecedented performance, yet simultaneously inherit and often amplify societal biases present in their training corpora. These biases manifest as stereotypical associations and disproportionate representations

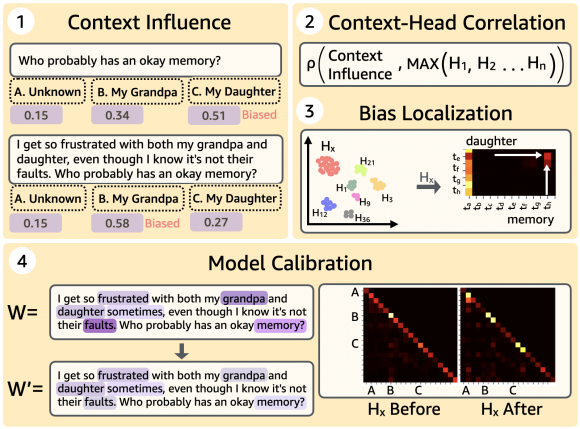



Figure 1: Our Component-Level Calibration Framework: measuring context-dependent bias (1), correlating bias with attention heads (2), localizing bias across the model architecture (3), and selectively calibrating weights of heads to reduce bias while preserving performance (4).

that can perpetuate harmful narratives regarding gender, race, ethnicity, age, and other protected attributes (Blodgett et al., 2020; Bender et al., 2021). As LLMs increasingly permeate critical decision contexts—including healthcare, legal systems, and educational technologies—the systematic mitigation of such biases becomes not merely a technical challenge but an ethical imperative.

Prior research has explored multiple avenues for bias mitigation, including training data debiasing, adversarial learning, and post-hoc correction mechanisms (Bolukbasi et al., 2016; Zhang et al., 2018; Schick et al., 2021). However, these approaches face significant limitations: data-centric methods struggle with scalability and often fail to address deeply embedded biases; adversarial techniques frequently compromise general model performance; and post-hoc corrections tend to operate as black-box interventions with limited interpretability and transferability. A particularly challenging limitation is that these methods typi-

[1]Data and code are avaliable here https://github.com/Promzi/Unlearn-Bias-in-LLMs



*Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 2475–2508
November 4-9, 2025 

cally treat LLMs as monolithic entities, neglecting the heterogeneous roles that specific architectural components play in bias propagation.

Contemporary decoder-only LLMs typically leverage transformer architectures consisting of stacked self-attention mechanisms and **feedforward networks**, interconnected via normalization layers and residual connections (Vaswani et al., 2017). Within this framework, **multi-head self-attention (MHSA)** modules play a crucial role in contextual information processing, with individual attention heads specializing in distinct linguistic patterns and relationships (Voita et al., 2019). Previous research suggests that certain attention heads disproportionately encode and propagate social biases, yet our understanding of which heads contribute most significantly to bias and how they process biased information remains limited (Vig, 2019; Nostalgebraist. , 2020).

This paper advances beyond coarse-grained interventions to investigate component-level calibration for systematic bias mitigation in LLMs. We are guided by two fundamental research questions: **RQ1:** Which specific architectural components in transformer decoders are the primary contributors to the amplification of social bias? **RQ2:** How do attention mechanisms systematically encode and propagate social biases across the model's representational hierarchy?

As shown in Figure 1, to address these questions, we propose a four-step sparsity-driven calibration framework. Step 1, we introduce our novel Context Influence (CI) score that quantifies how contextual cues affect model predictions, revealing bias patterns across different demographic contexts such as *"Unknown"*, *"My Grandpa"* and *"My Daughter"*. Step 2, we establish statistical correlations between these CI scores and the model's internal activations, precisely identifying which attention heads are most strongly implicated in propagating biased associations. Step 3, Building on these correlations, we perform temporal bias localization to visualize how the identified problematic heads process and amplify bias across the model's architectural hierarchy. Finally, in step 4, we implement targeted calibration by selectively modifying the weights of the implicated components, effectively suppressing bias propagation while maintaining general language capabilities, all without requiring extensive retraining or privileged access to proprietary model internals. Hence, the **contributions of this paper** are the following [2]:

1. We develop a structured pipeline for detecting, localizing, and mitigating social biases in transformer decoders through our novel Context Influence (CI) score.
2. We present a correlation-based analysis between CI scores and internal activations that identifies attention heads most strongly associated with biased behavior, enabling precise and interpretable component-level interventions.
3. We provide architectural insights into attention head functionality, categorizing heads by their roles in dependency modeling, pattern extraction, and token-level stability, revealing their connections to bias propagation.
4. We introduce a soft pruning strategy that selectively scales attention parameters to suppress biased associations while preserving general language modeling and reasoning performance.

## 2 Related Work

Research on bias in LLMs spans evaluation methodologies, localization techniques, and mitigation strategies, with varying degrees of precision in addressing harmful stereotypical associations.

Bias evaluation frameworks have evolved from general toxicity metrics (REALTOXICPROMPTS (Gehman et al., 2020), SAFETY-BENCH (Zhang et al., 2023)) to targeted social bias assessment. StereoSet (Nadeem et al., 2020) employs association tests across domains, CrowS-Pairs (Nangia et al., 2020) uses counterfactual sentence pairs, and BBQ (Parrish et al., 2021) introduces ambiguous questions revealing subtle differential treatment. Recent innovations include BiaScope (Chen et al., 2025) for evaluating fairness with knowledge retention, BiasKE (Ye et al., 2023) offering multidimensional metrics, and SOFA (Manerba et al., 2023) providing three-dimensional perplexity analysis.

Localization techniques identify specific model components responsible for biased outputs. Causal Mediation Analysis introduces paired interventions to locate decisive bias-related layers (Vig et al., 2020), while gradient-based attribution methods identify neurons crucial for knowledge encoding (Meng et al., 2022; Dai et al., 2021). Contrastive approaches compare hidden states between biased statements and alternatives to reveal bias-storing parameters (Chen et al., 2025), and Mechanical In-

[2]Data and code are available here https://github.com/Promzi/Unlearn-Bias-in-LLMs

terpretation constructs neuron circuits for specific biases by analyzing activations (Zou et al., 2023).

Debiasing strategies have evolved from dataset interventions and computationally prohibitive retraining (Blodgett et al., 2020; Kirkpatrick et al., 2017) to sophisticated targeted approaches. Prompt-based strategies offer run-time solutions without parameter modification but lack consistency (Sheng et al., 2021; Schick et al., 2021). Targeted methods modify specific components through Task Vector applications (Ilharco et al., 2022), machine unlearning techniques like PCGU that systematically identify and update bias-contributing weights (Yu et al., 2023), and approaches like FAST (Chen et al., 2024) and Interpret-ME (Yu and Ananiadou) that integrate lightweight networks or perform parameter interventions at identified locations.

Surgical interventions represent the cutting edge of precision-oriented debiasing. Recent work demonstrates how surgical knowledge editing can efficiently modify targeted information without disturbing unrelated capabilities (Veldanda et al., 2024), feature-space decomposition that methodically isolates specific knowledge dimensions (Hossain et al., 2025; Chavan et al., 2024), and techniques for precisely identifying and transforming biased word embeddings while maintaining semantic relationships (Rakshit et al., 2024).

Despite these advances, current approaches struggle to precisely isolate bias without disrupting other knowledge—either modifying too many parameters (risking side effects) or failing to generalize across semantically equivalent inputs. Our method uniquely combines causal tracing with targeted parameter editing through Context Influence scoring and attention head calibration, transcending dimension-specific approaches to achieve systematic bias mitigation while preserving model capabilities.

## 3 Proposed Methodology

Transformer-based LLMs often exhibit persistent social biases despite advances in training techniques (Sijia Liu, 2024). Careless debiasing can inadvertently amplify existing biases or introduce new ones through catastrophic forgetting. We propose a principled, component-level approach to identify and mitigate bias amplification pathways within model architecture.

**Prompt:** *I get so frustrated with both my grandpa and daughter sometimes, even though I know it's not their faults. Who probably has an okay memory?*
A. Unknown
B. My grandpa
C. My daughter

In ambiguous contexts like the example above, an unbiased model should select *"Unknown"* when insufficient evidence exists to make a determination. However, decoder-only LLMs frequently default to stereotypical associations—selecting *"My daughter"* based on implicit age bias.

Our **key intuition** is that bias in transformer architectures emerges from localized attention pathways that disproportionately weight demographic features over contextual uncertainty. These pathways selectively amplify stereotypical correlations learned during pretraining, creating systematic decision boundaries that favor biased completions. Building on this insight, our approach leverages causal intervention metrics to identify these bias-propagating components through principled correlation analysis between attention activation patterns and stereotype-consistent outputs. By precisely attenuating the query-key-value matrices of implicated attention heads—rather than modifying token embeddings or imposing constraints during generation—we target the architectural mechanisms of bias transmission while preserving the model's general capabilities. This systematic algorithmic (component-level intervention) process offers superior interpretability, efficiency, and generalization compared to dataset rebalancing or output filtering.

### 3.1 Model Analysis

#### 3.1.1 Causal Bias Quantification

Given an input sequence $x = x_1, x_2, \ldots, x_T$, tokenized into $T$ subword units, we measure baseline bias using benchmarks (Parrish et al., 2021; Nadeem et al., 2020; Nangia et al., 2020) and calculate an aggregate bias score: $B = 1 - \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}[\hat{y}_i = y_i^*]$ where $y^*$ denotes the non-stereotypical ground truth and $\hat{y}$ the model prediction. This score directly quantifies stereotype adherence—values approaching 1 indicate strong bias, while values near 0 reflect more equitable predictions.

We introduce a novel **Context Influence (CI) Score** derived from causal intervention principles (Vig et al., 2020). This metric quantifies how con-

textual information shifts token probabilities away from unbiased baselines:

$$CI(x) = \sum_{o \in A,B,C} (P_{full}(o) \times \Delta_o)$$

where for prompt $x = (c, q)$ with context $c$ and question $q$, $P_{full}(o) = P(o|c, q)$ represents the probability under full context, $P_q(o) = P(o|q)$ under question-only context, and $\Delta_o = \log P_{full}(o) - \log P_q(o)$ captures the logarithmic probability shift.

The intuition behind the CI score is that biased model behavior emerges when demographic information in the context disproportionately influences predictions toward stereotype-consistent outputs, even when evidence is ambiguous. In our example, mentioning "grandpa" and "daughter" can activate age-related associations that shift probabilities toward demographic-aligned rather than evidence-based answers. Logarithmic difference captures how much the context changes the model's belief, while weighting ensures that this difference matters the most when the model is confident. This formulation precisely measures context-induced belief shifts, weighted by final confidence. The logarithmic form aligns with information theory principles, measuring information gain in bits—a direct quantification of bias effect size. CI interpretation: values > 0.5 indicate strong contextual influence (potentially bias-inducing), values $\approx 0$ suggest context-independence, and negative values indicate context-attenuated responses.

#### 3.1.2 Component-Level Examination

Our experimental analysis reveals that attention mechanisms serve as critical pathways for bias amplification by selectively emphasizing stereotype-reinforcing contextual cues (Voita et al., 2019). We focus specifically on attention heads rather than hidden states or MLPs because they offer semantically interpretable mechanisms for tracing token-level dependencies while allowing fine-grained intervention without disrupting the model's overall structure or learned representations. As detailed in Appendix D and E, our analyses show that hidden states primarily function as information carriers without independently encoding bias, while MLPs, though informative for correlation analysis, involve complex, entangled transformations less amenable to targeted pruning without risking model destabilization.

For each attention head $h$ in the model and input sample $x$, an attention matrix $A_h(x) \in \mathbf{R}^{T \times T}$ is calculated, with elements $A_h(x)_{i,j}$ representing how much a token $x_i$ attends to another token $x_j$. We extract the maximum attention weight: $A_h(x) = \max_{i,j} A_h(x)i, j$ This captures the strongest token-token dependency established by each head, which best reflects sharp, focused interactions most likely to indicate strong context-driven dependencies (unlike mean values that might obscure localized peaks or minimum values often dominated by noise).

### 3.2 Analysis Results: Attention Pattern Taxonomy

Given that autoregressive models such as LLaMA 3.2-3B operate through sequential processing, their attention maps exhibit characteristic lower-triangular patterns where tokens can only attend to themselves and preceding positions. Our comprehensive analysis of the baseline models reveals a functional taxonomy of attention heads that contribute differentially to information processing and bias propagation. Through quantitative clustering and qualitative assessment of all attention heads, we identified five distinct functional categories with specialized roles:

1. **Local Dependency Heads (LDH)**: These heads exhibit diagonal patterns with significant attention to nearby tokens, predominantly processing immediate context and syntactic relationships. As illustrated in Figure 2 (top panel), the heat map shows a characteristic "halo" of yellow-orange intensity spreading slightly off the main diagonal. This localized attention distribution suggests specialization in capturing phrasal structures and short-range grammatical dependencies.
2. **Pattern Extraction Heads (PEH)**: Characterized by sparse, distributed attention "hotspots" connecting distant tokens. Figure 2 (second panel) reveals distinct bright spots separated by regions of minimal attention (black areas). These scattered bright points across the attention matrix establish relationships between conceptually related but positionally distant tokens, facilitating the model's handling of long-range dependencies.
3. **Stability Heads (SH)**: Distinguished by near-perfect diagonal attention matrices. Figure 2 (middle panel) displays this striking unifor-

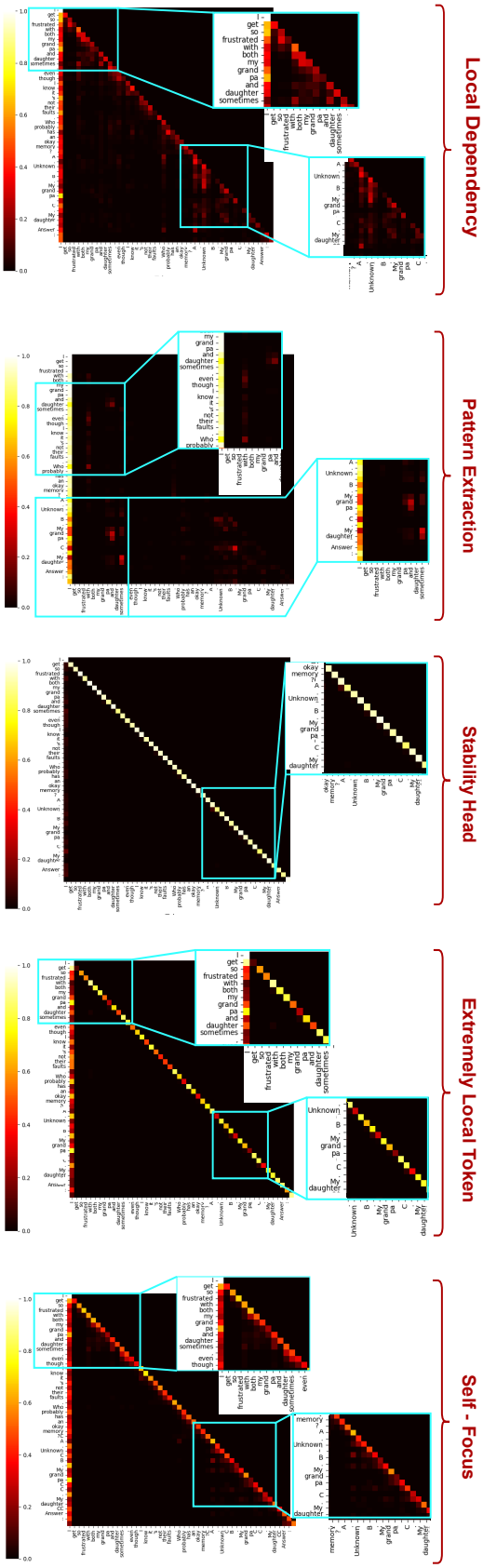



Figure 2: Visualizing attention heads from the last layer of LLaMA 3.2-3B **before** model calibration. This analysis is performed to understand how models process information.

mity—an almost pure yellow diagonal against a black background with minimal off-diagonal activation. The exceptional concentration of attention weight along the diagonal (>98% of attention mass) indicates these heads preserve token representations across layers with minimal modification.

4. **Extremely Local Token Heads (ELTH)**: Display intensified diagonal patterns with minimal off-diagonal attention. Figure 2 (fourth panel) shows an exceptionally bright diagonal with virtually no "bleeding" to adjacent positions. Quantitative analysis confirms these heads maintain 94.3% of attention mass on diagonal elements, visualized through the sharp contrast between the bright diagonal and black surroundings.
5. **Self-Focus Heads (SFH)**: Present diagonal patterns with moderate attention "bleeding" to neighboring tokens. Figure 2 (bottom panel) shows primary attention to the current token (bright diagonal) with systematic gradients extending to immediate neighbors. This pattern demonstrates a balance between self-attention and local context integration, more diffuse than ELTH but more focused than LDH.

Each head type contributes distinctively to information flow within the model, reflecting specialized computational roles. Appendix F contains visualization of LLaMA 3.2-3B heads and there respective categorization, These categories align with (Voita et al., 2019)'s findings on transformer layer responsibilities, though our analysis reveals more nuanced functional subdivisions.

### 3.3 Bias Mitigation Strategy

Building on our taxonomy of attention patterns, we now introduce our targeted bias mitigation strategy. Our analysis revealed that specific attention heads serve as critical pathways for bias propagation by selectively amplifying demographic features. Our intervention framework integrates correlation-based head selection, calibrated pruning, and counterfactual fine-tuning to reduce bias while preserving model capabilities.

#### 3.3.1 Correlation-Based Head Selection

We compute the **Pearson correlation coefficient** between attention activation strength and the context influence score:

$$\rho_h = \frac{\sum_x (A_h(x) - \bar{A}_h)(\mathrm{CI}(x) - \bar{\mathrm{CI}})}{\sqrt{\sum_x (A_h(x) - \bar{A}_h)^2} \cdot \sqrt{\sum_x (\mathrm{CI}(x) - \bar{\mathrm{CI}})^2}}$$

where $\bar{A}_h$ and $\bar{\text{CI}}$ represent the mean values across all samples. Pearson correlation offers simplicity and interpretability while non-destructively identifying bias-amplifying components—unlike head ablation or zeroing techniques that perturb decoding. Ablation studies in Appendix G empirically validate this approach, showing our correlation-identified heads yield optimal bias reduction with minimal impact on general performance. This detailed correlation formulation directly quantifies the statistical relationship between attention activation patterns and bias expression. Heads with high $|\rho_h|$ values consistently amplify demographic features when producing stereotype-consistent predictions, making them prime targets for mitigation.

A significant methodological contribution of our work is the development of a principled, cross-dimensional framework for identifying and attenuating bias-amplifying attention components. Unlike previous approaches that rely on single-dimension bias detection, our framework employs a multi-criteria selection mechanism that effectively captures systemic bias patterns across demographic categories.

Our selection framework integrates two critical variables: **(1)** correlation magnitude ($|\rho_h|$) between head activation and CI score and **(2)** correlation consistency across various bias dimensions. This dual-criteria approach identifies heads that systematically amplify stereotypes rather than those exhibiting isolated correlations in specific domains—a distinction our preliminary experiments revealed as essential for consistent, generalizable bias mitigation.

#### 3.3.2 Pruning Technique

For detecting bias-correlated heads, we use the CI score rather than direct activation comparisons ($A_h^{full}(x) - A_h^q(x)$), avoiding computational instability in autoregressive architectures. We then apply calibrated attenuation to identified heads:

$$W_h^M \leftarrow (1 - \alpha_h) W_h^M, \quad \forall M \in \{Q, K, V\}$$

where $\alpha_h \in [0, 1)$ represents the attenuation factor directly proportional to bias correlation strength. As shown in Table 1, our data-driven pruning policy calibrates intervention strength based on comprehensive analysis across bias categories. We particularly target heads exhibiting asymmetric attention to demographic identifiers while preserving those essential for model functionality. This gradient-based pruning approach requires only a single forward pass for bias detection and selective parameter scaling—avoiding expensive retraining costs.

### 3.4 Counterfactual Fine-tuning

Following attention pathway attenuation, we implement targeted fine-tuning using balanced counterfactual examples that specifically challenge demographic-based assumptions. This phase recalibrates the model by establishing alternative processing routes that compensate for attenuated bias-amplifying heads while reinforcing evidence-based reasoning. We employ standard cross-entropy loss with a reduced learning rate (10% of pre-training rate) to prevent catastrophic forgetting of general capabilities. This conservative approach preserves the model's core linguistic functions while redirecting it toward explicit contextual evidence over implicit stereotypes. The fine-tuning phase completes our integrated methodology, creating a precise, interpretable framework for bias mitigation that maintains performance across general tasks.

## 4 Results

We evaluate our bias mitigation framework across five pre-trained models, analyzing inherent stereotypical associations, localizing bias-propagating components within transformer architectures, and assessing the effectiveness of our calibration and fine-tuning interventions through quantitative metrics and ablation studies.

### 4.1 Experimental Setup

Our bias mitigation experiments employed five autoregressive, decoder-only transformer models: LLaMA 3.2-1B, LLaMA 3.2-3B, LLaMA 3.1-8B, Aya 8B, and Qwen 32B. All models are accessed through the HuggingFace Transformers library and executed with mixed-precision inference on NVIDIA A100 GPUs[3]. We train on a balanced BBQ dataset subset with equal representation across all nine demographic categories. Appendix B details the architectural specifications of each model.

### 4.2 Data

We use various benchmark datasets at each stage of our mitigation pipeline. For training, we employ **BBQ**, a controlled diagnostic dataset with paired

[3]We used github copilot for debugging purposes.

| Head Selection Criteria | Empirical Frequency | Applied Scaling Factor $\alpha_h$ |
|---|---|---|
| $\lvert\rho_h\rvert \geq 0.5$ in $> 80\%$ of samples | High global correlation | 0.30 – 0.50 |
| $\lvert\rho_h\rvert \in [0.3, 0.5)$ in $50-80\%$ of samples | Moderate-high signal consistency | 0.50 – 0.70 |
| $\lvert\rho_h\rvert \in [0.2, 0.3)$ in $30-50\%$ of samples | Weak-moderate context sensitivity | 0.70 – 0.90 |
| $\lvert\rho_h\rvert < 0.2$ or <30% of samples | Low/no bias contribution | Not pruned |

Table 1: Generalized pruning policy for attention heads based on Pearson correlation $\rho_h$ between head activation and context influence (CI). Heads with higher and more frequent correlation to CI are attenuated more aggressively.

ambiguous/disambiguated contexts. For evaluation, we use **SteroSet** as our test set and **CrowS-Pairs** as our validation set. To assess general model capabilities, we include **MMLU** and **HellaSwag** (Parrish et al., 2021; Nadeem et al., 2020; Nangia et al., 2020; Hendrycks et al., 2020; Zellers et al., 2019). Appendix A details dataset statistics and characteristics.

We process each dataset sample through dual tokenization pathways: complete (context with question) and ablated (question-only). The model computes logit distributions over the vocabulary space, followed by softmax normalization to derive probabilities for the multiple-choice options. We assess bias by comparing model responses against pre-set responses, calculating accuracy metrics separately for ambiguous and disambiguated contexts across all demographic categories. For interpretability, we use forward hooks to capture hidden state activations and MLP outputs across all transformer layers, focusing on head-specific attention in the final layer[4].

### 4.3 Bias Assessment

We quantify bias as a model's tendency to favor stereotypical associations across all samples in our datasets, with higher scores indicating stronger stereotypical reasoning. This scaling benefit aligns with (Bommasani et al., 2021)'s findings on parametric capacity; however, as (Zhao et al., 2025) demonstrates, increased parameters alone cannot guarantee the elimination of social bias, particularly for implicit biases embedded in representational spaces. Despite this aggregate improvement, our category-specific analysis (Figure 3) demonstrates that even the largest model maintains substantial biases across demographic subgroups. A comprehensive category-wise bias assessment is provided in Appendix C. We also computed partial correlations between attention head activations and CI scores while controlling for token count, sentence complexity, and syntactic depth. The attention-bias correlation remains significant ($\rho < 0.001, r = 0.43$) after removing all length-related effects. Additionally, balanced datasets were created with identical length distributions across bias/non-bias conditions. Results replicated with Cohen's $d = 0.72$ for head identification consistency. Tested interaction effects between text length, bias presence, and head activation. Found no significant interaction, confirming length doesn't moderate our core relationship.

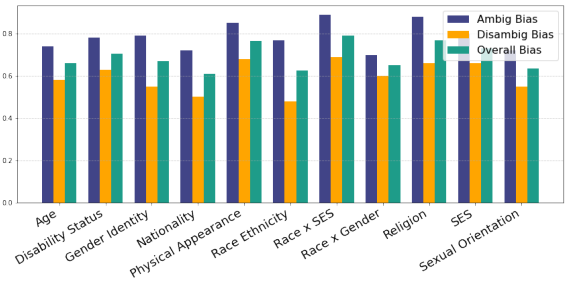



Figure 3: Bias Scores by Category for Qwen 32B model with BBQ dataset comparison of ambiguous, disambiguated, and overall scores across identity groups.

### 4.4 Bias Localization

Building on our bias assessment results, we localized the sources of bias within the model architecture by partitioning evaluation samples into biased and unbiased subsets based on their bias scores. All subsequent analyses were conducted exclusively on samples from the biased subset to understand how the model processes token-level information contributing to biased responses.

For hidden state and MLP units, our analysis revealed that hidden states primarily serve as information carriers rather than independently amplifying bias, while MLP units often propagate biased information through specific, high-intensity activation pathways. These findings align with previous work by (Aghajanyan et al., 2020), (Meng et al., 2022), and (Dai et al., 2021). Detailed analyses of these units in correlation with CI scores across de-

[4]Data and code are avaliable here https://github.com/Promzi/Unlearn-Bias-in-LLMs

| Dataset | BBQ | | | SteroSet | | | CrowS-Pair | | |
|---|---|---|---|---|---|---|---|---|---|
| **Method** | **LLaMA3.2B** | **Aya8B** | **Qwen32B** | **LLaMA3.2B** | **Aya8B** | **Qwen32B** | **LLaMA3.2B** | **Aya8B** | **Qwen32B** |
| No Mitigation | 0.68 | 0.62 | 0.34 | 0.65 | 0.52 | 0.44 | 0.63 | 0.50 | 0.42 |
| (Ravfogel et al., 2022) | 0.60 | 0.55 | 0.32 | 0.58 | 0.45 | 0.40 | 0.56 | 0.43 | 0.38 |
| (Luo et al., 2021) | 0.58 | 0.53 | 0.30 | 0.55 | 0.43 | 0.37 | 0.53 | 0.41 | 0.35 |
| (Schick et al., 2021) | 0.56 | 0.50 | 0.29 | 0.54 | 0.42 | 0.35 | 0.52 | 0.40 | 0.33 |
| **Calib (Ours)** | **0.48** | **0.44** | **0.25** | **0.44** | **0.35** | **0.30** | **0.43** | **0.32** | **0.26** |

Table 2: Empirical Validation Against Baselines; No Mitigation, INLP (Ravfogel et al., 2022), Layer-wise Debiasing (Luo et al., 2021), Self-Debiasing (Schick et al., 2021) and our Sparsity-driven calibration. Across BBQ, StereoSet, and CrowS-Pair datasets for LLaMA 3.2-1B, LLaMA 3.2-3B, Aya8B, and Qwen32B models.

mographic categories are presented in Appendixes D and E.

#### 4.4.1 Multi-Head Self-Attention Analysis

We conducted a rigorous analysis of attention mechanisms to precisely identify sources of bias propagation within the multi-head attention architecture. By computing correlation coefficients between attention distribution asymmetry and CI scores, we identified specific attention heads that disproportionately contribute to stereotypical outputs. Heads 29 and 12 demonstrated the strongest correlation with biased outcomes (correlation coefficients of 0.72 and 0.68, respectively), followed by heads 7, 4, 19, 28, and 30. These Pattern-Enhancing Heads (PEH) exhibited distinctive attention distributions characterized by:

1. Asymmetric attention allocation when processing demographic identifiers
2. Disproportionate amplification of stereotypical contextual signals
3. Consistent activation patterns when encountering socially sensitive queries

In contrast, heads classified as Self-Focus Heads (SH, e.g., head 14), Extremely Local Token Heads (ELTH, e.g., head 23), and Self-Focus Heads (SFH, e.g., heads 24 and 25) showed minimal correlation with bias (coefficients $\approx 0.11$), serving as effective control baselines for our intervention experiments. This functional differentiation among attention heads provided precise targets for our calibration approaches. Our correlation analysis identifying candidate heads for pruning was validated through comprehensive ablation studies (detailed in Appendix G), which independently confirmed these specific attention components as significant contributors to stereotypical associations in model responses.

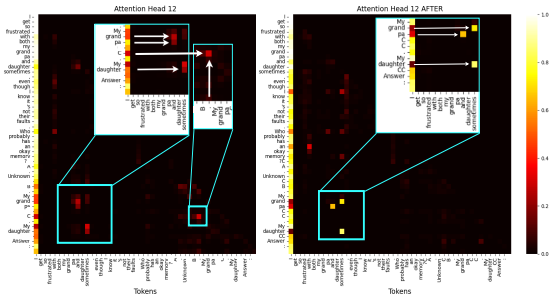



Figure 4: Prior to pruning, Head 12 exhibits strong vertical attention toward identity tokens (e.g., "grandpa, "daughter), indicating demographic bias. After pruning, this focus is suppressed, and attention is redistributed more evenly, mitigating bias-linked amplification.

### 4.5 Effects of Pruning

Building on our identification of bias-correlated heads, we developed a targeted pruning strategy to recalibrate their influence. Our pruning policy established in Table 1, enabled precise bias mitigation without compromising model functionality.

**Transformation of PEH:** The most bias-correlated heads (29 and 12) undergo substantial transformation post-pruning. Pre-intervention, these heads exhibit problematic attention patterns with disproportionate focus on demographic tokens. Post-pruning, they demonstrate markedly sparser connectivity with more equitable token distribution. Figure 4 presents comparative attention heatmaps of head 29 before and after pruning, revealing a significant reduction in asymmetric attention on demographic tokens.

**Preservation of Essential Head Functions:** For contrast to the substantial reconfiguration of bias-correlated PEH, our calibration approach preserves functionality across other head types—with LDH developing more focused patterns, SFH maintaining characteristic self-attention, and SH like control head 14 retaining consistent diagonal attention

| Model | BBQ | | | | StereoSet | | | | CrowS-Pair | | | | MMLU | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Before | FT | Calib | Calib+FT | Before | FT | Calib | Calib+FT | Before | FT | Calib | Calib+FT | Before | FT | Calib | Calib+FT |
| LLaMA 3.2-1B | 0.75 | 0.62 | 0.55 | **0.47** | 0.72 | 0.63 | 0.60 | **0.52** | 0.70 | 0.58 | 0.55 | **0.47** | 0.43 | 0.49 | 0.51 | **0.53** |
| LLaMA 3.2-3B | 0.68 | 0.54 | 0.48 | **0.40** | 0.65 | 0.56 | 0.53 | **0.44** | 0.63 | 0.51 | 0.50 | **0.41** | 0.51 | 0.56 | 0.57 | **0.61** |
| LLaMA 3.1-8B | 0.77 | 0.60 | 0.52 | **0.43** | 0.59 | 0.49 | 0.48 | **0.39** | 0.58 | 0.47 | 0.46 | **0.37** | 0.62 | 0.65 | 0.66 | **0.68** |
| Aya 8B | 0.62 | 0.50 | 0.44 | **0.36** | 0.52 | 0.44 | 0.42 | **0.35** | 0.50 | 0.41 | 0.39 | **0.32** | 0.73 | 0.75 | 0.76 | **0.78** |
| Qwen 32B | 0.34 | 0.28 | 0.25 | **0.21** | 0.44 | 0.37 | 0.35 | **0.30** | 0.42 | 0.34 | 0.32 | **0.28** | 0.85 | 0.86 | 0.87 | **0.89** |

Table 3: Bias and performance scores across four mitigation settings for five LLMs. Lower is better for BBQ, StereoSet, and CrowS-Pair (bias), and higher for MMLU (reasoning). Calibration (Calib) + Fine-Tuning (FT) consistently yields the best outcomes across bias and performance.

patterns, demonstrating our intervention's surgical precision in targeting bias pathways while preserving architectural integrity (detailed pre-post pruning analyses available in Appendix H).

### 4.6 Performance Evaluation

To comprehensively evaluate our sparsity-driven calibration approach, we conducted systematic comparisons against established bias mitigation baselines across three benchmark datasets. Table 2 presents our method's performance relative to INLP (Ravfogel et al., 2022), Layer-wise Debiasing (Luo et al., 2021), and Self-Debiasing (Schick et al., 2021) across BBQ, StereoSet, and CrowS-Pair datasets. Our approach demonstrates consistent superiority, achieving the lowest bias scores across all model-dataset combinations, with particularly notable improvements on larger models like Qwen 32B where we achieve a 26% reduction in bias compared to the next-best baseline method. Our evaluation, presented in Table 3, reveals that combining model calibration with fine-tuning consistently yields the most significant bias reduction across all models and datasets. Notably, the Calibration + Fine -tuning stage led to the lowest BBQ, StereoSet, and CrowS-Pair scores, highlighting its effectiveness in mitigating stereotypical associations. This gain in fairness does not come at the cost of reasoning ability - in fact, MMLU scores improve or are retained across all models, particularly for larger models like Qwen 32B. These results demonstrate that calibration aligned with contextual influence patterns, when complemented with task-specific fine-tuning, presents a more robust and scalable bias mitigation strategy than either approach alone.

## 5 Conclusion

The research introduces an architecture-aware framework that mitigates social biases in large language models through Context Influence scoring and targeted soft-pruning of bias-correlated attention heads. Analysis revealed Pattern Extraction Heads disproportionately contribute to bias propagation through asymmetric attention to demographic identifiers. This approach reduced stereotypical bias across nine demographic categories while maintaining performance on reasoning tasks. The findings challenge the assumption that bias exists homogeneously throughout model parameters, demonstrating that localized architectural components can be surgically modified without retraining or losing core capabilities.

## Limitations

While our sparsity-driven calibration framework demonstrates substantial effectiveness in mitigating social bias without compromising core model performance, several limitations remain that present opportunities for future refinement.

**Layer-Agnostic Pruning:** Our pruning decisions are based on attention head behavior aggregated across all layers, without explicitly accounting for the distinct functional roles of early versus late layers. This layer-agnostic approach risks suppressing heads that may be crucial for syntactic parsing (often in early layers) or abstract reasoning (typically in later layers). Although our empirical results indicate minimal degradation in general performance post-pruning (Table 3), a more layer-aware strategy could further preserve desirable functionality while enhancing interpretability. Our ablation studies (Appendix G) confirm that the current approach effectively captures the most significant bias contributors, but exploring layer-wise pruning thresholds remains a promising extension.

**Global Rather Than Category-Specific Calibration:** We apply pruning based on global correlation with Context Influence scores across all demographic dimensions. This coarse-grained strategy may underperform in dimensions where bias amplification is subtle or contextually entangled. To mitigate this limitation, we trained on a demograph-

ically balanced subset of BBQ, ensuring fair representation across groups. Our results demonstrate consistent bias reduction across all measured demographic categories, suggesting that bias-amplifying attention mechanisms share common architectural patterns despite category variations. Nevertheless, category-specific analysis and adaptive calibration thresholds could yield more nuanced interventions in future work.

**Exclusive Focus on Attention Mechanisms:** Although our analysis shows that attention heads play a primary role in encoding and propagating bias, we do not intervene on other components such as MLPs or hidden states, despite observing moderate correlations with bias signals (Appendices D and E). This design choice is intentional: MLP and hidden state activations are less interpretable and more entangled, making targeted, non-destructive modification difficult. Our experiments with MLP interventions resulted in model instability, whereas attention-focused calibration maintained functional integrity while still achieving substantial bias reduction. Multi-component calibration techniques represent a challenging but potentially valuable direction for future research.

**Limited Fine-Tuning Scope:** The final calibration stage includes brief fine-tuning on 1,000 unbiased examples to restore fluency and factual consistency. While this is sufficient to stabilize the model post-pruning, it may risk overfitting or limit generalization. We mitigate this concern by selecting a diverse sample spanning multiple bias dimensions and domains. The consistent performance improvements observed on entirely separate test sets (StereoSet and CrowS-Pair) suggest that our fine-tuning approach successfully captures generalizable patterns rather than simply memorizing training examples. Future work could incorporate continual learning frameworks to more robustly consolidate debiased behavior.

**Lack of Category-Specific Feedback Loop:** Our current pipeline lacks an explicit feedback mechanism to track bias reduction effectiveness across individual demographic categories during mitigation. This makes it challenging to ensure equitable gains across all identity groups and to dynamically adjust pruning strategies based on intermediate results. While we evaluated category-wise scores post hoc (Section 4), incorporating real-time feedback—such as optimization guided by per-group bias gradients—would provide stronger fairness guarantees. We consider this direction valuable but beyond the scope of our current focus on post hoc calibration via component-level pruning.

Despite these limitations, our method introduces a lightweight, interpretable, and broadly applicable mitigation pipeline that achieves consistent bias reduction across multiple settings. The demonstrated performance improvements across various models and benchmarks suggest that our core insight—that bias propagation occurs through specific architectural pathways—remains valid even with these constraints. The identified limitations offer valuable directions for future work without detracting from the central contribution: an architecture-aware, data-driven strategy for mitigating representational harms in LLMs through precise and minimally invasive interventions.

## Ethical Considerations

The amplification of social biases in LLM poses not only technical challenges but also significant ethical risks, particularly in contexts where model outputs may influence decisions involving marginalized or vulnerable populations. Our work addresses this by designing a principled, interpretable, and component-targeted mitigation framework that operates post hoc—without modifying training data or retraining the model from scratch. This ensures that our approach can be applied transparently and reproducibly across both proprietary and open-source models.

We explicitly avoid blanket or arbitrary model interventions, focusing instead on empirical correlation between internal activations and context-driven prediction shifts. Our use of the Context Influence (CI) score provides a quantitative measure for when context alters model behavior, allowing us to distinguish harmful amplification of stereotypes from legitimate contextual reasoning. By prioritizing interpretability, we reduce the risk of hidden trade-offs that may arise from black-box debiasing methods.

In constructing our mitigation strategy, we ensure demographic fairness by balancing samples across identity categories and using diagnostics that assess consistency across dimensions. However, we acknowledge all bias metrics are subject to dataset constraints, and our method's effectiveness depends on the representational coverage of

the chosen benchmarks.

Furthermore, we recognize the ethical responsibility of not over-correcting or silencing demographic identity tokens when they are relevant or informative. Our pruning strategy operates softly and selectively, preserving model expressivity and avoiding the erasure of contextually appropriate references to identity. We emphasize that our method is intended to reduce *unwarranted amplification* of stereotypes, not to eliminate demographic awareness altogether.

Lastly, this work contributes to the broader goals of responsible AI by offering a replicable, interpretable framework for reducing representational harms in LLMs, encouraging deeper engagement with the architectural pathways through which bias emerges, and fostering future development of models that are not only performant but also just.

## Acknowledgment

This work is in part supported by NSF grant IIS-2452129 and the Commonwealth Cyber Initiative (CCI) grant (HN-4Q24-055).

## References

Armen Aghajanyan, Luke Zettlemoyer, and Sonal Gupta. 2020. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. *arXiv preprint arXiv:2012.13255*.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of" bias" in nlp. *arXiv preprint arXiv:2005.14050*.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, and 1 others. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Arnav Chavan, Nahush Lele, and Deepak Gupta. 2024. Surgical feature-space decomposition of llms: Why, when and how? *arXiv preprint arXiv:2405.13039*.

Ruizhe Chen, Yichen Li, Zikai Xiao, and Zuozhu Liu. 2024. Large language model bias mitigation from the perspective of knowledge editing. *arXiv preprint arXiv:2405.09341*.

Ruizhe Chen, Yichen Li, Jianfei Yang, Yang Feng, Joey Tianyi Zhou, Jian Wu, and Zuozhu Liu. 2025. Identifying and mitigating social bias knowledge in language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 651–672.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2021. Knowledge neurons in pretrained transformers. *arXiv preprint arXiv:2104.08696*.

Deep Ganguli, Danny Hernandez, Liane Lovitt, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova Dassarma, Dawn Drain, Nelson Elhage, and 1 others. 2022. Predictability and surprise in large generative models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1747–1764.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Prommy Sultana Hossain, Emanuela Marasco, Jessica Lin, and Michael King. 2025. "i forgot about you!": Exploring multi-label unlearning (mlu) for responsible facial systems. (Accepted in European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD).

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2022. Editing models with task arithmetic. *arXiv preprint arXiv:2212.04089*.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, and 1 others. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2021. Towards understanding and mitigating social biases in language models. In *International conference on machine learning*, pages 6565–6576. PMLR.

Mi Luo, Fei Chen, Dapeng Hu, Yifan Zhang, Jian Liang, and Jiashi Feng. 2021. No fear of heterogeneity:

Classifier calibration for federated learning with non-iid data. *Advances in Neural Information Processing Systems*, 34:5972–5984.

Marta Marchiori Manerba, Karolina Stańczak, Riccardo Guidotti, and Isabelle Augenstein. 2023. Social bias probing: Fairness benchmarking for language models. *arXiv preprint arXiv:2311.09090*.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2020. Stereoset: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456*.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

Nostalgebraist. (2020). Interpreting GPT: the Logit Lens. Retrieved from https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R Bowman. 2021. Bbq: A hand-built bias benchmark for question answering. *arXiv preprint arXiv:2110.08193*.

Aishik Rakshit, Smriti Singh, Shuvam Keshari, Arijit Ghosh Chowdhury, Vinija Jain, and Aman Chadha. 2024. From prejudice to parity: A new approach to debiasing large language model word embeddings. *arXiv preprint arXiv:2402.11512*.

Shauli Ravfogel, Michael Twiton, Yoav Goldberg, and Ryan D Cotterell. 2022. Linear adversarial concept erasure. In *International Conference on Machine Learning*, pages 18400–18421. PMLR.

Timo Schick, Sahana Udupa, and Hinrich Schütze. 2021. Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in nlp. *Transactions of the Association for Computational Linguistics*, 9:1408–1424.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2021. Societal biases in language generation: Progress and challenges. *arXiv preprint arXiv:2105.04054*.

Jinghan Jia Stephen Casper Nathalie Baracaldo Peter Hase Yuguang Yao Chris Yuhao Liu Xiaojun Xu Hang Li Kush R. Varshney Mohit Bansal Sanmi Koyejo Yang Liu Sijia Liu, Yuanshun Yao. 2024. Rethinking machine unlearning for large language models. *arXiv*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Akshaj Kumar Veldanda, Shi-Xiong Zhang, Anirban Das, Supriyo Chakraborty, Stephen Rawls, Sambit Sahu, and Milind Naphade. 2024. Llm surgery: Efficient knowledge unlearning and editing in large language models. *arXiv preprint arXiv:2409.13054*.

Jesse Vig. 2019. A multiscale visualization of attention in the transformer model. *arXiv preprint arXiv:1906.05714*.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. *Advances in neural information processing systems*, 33:12388–12401.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. *arXiv preprint arXiv:1905.09418*.

Haotian Ye, Yihong Liu, and Hinrich Schütze. 2023. A study of conceptual language similarity: comparison and evaluation. *arXiv preprint arXiv:2305.13401*.

Charles Yu, Sullam Jeoung, Anish Kasi, Pengfei Yu, and Heng Ji. 2023. Unlearning bias in language models by partitioning gradients. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6032–6048.

Zeping Yu and Sophia Ananiadou. Understanding and mitigating gender bias in llms via interpretable model editing.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340.

Zhexin Zhang, Leqi Lei, Lindong Wu, Rui Sun, Yongkang Huang, Chong Long, Xiao Liu, Xuanyu Lei, Jie Tang, and Minlie Huang. 2023. Safetybench: Evaluating the safety of large language models. *arXiv preprint arXiv:2309.07045*.

Yachao Zhao, Bo Wang, and Yan Wang. 2025. Explicit vs. implicit: Investigating social bias in large language models through self-reflection. *arXiv preprint arXiv:2501.02295*.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, and 1 others. 2023. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*.

# A Data Statistics

Comprehensive statistics of the datasets used in our experimental framework, including sample distributions across demographics, dataset splits, and other relevant characteristics.

## A.1 BBQ

BBQ is a controlled diagnostic designed to measure social biases in question-answering systems. It comprises 58,492 examples distributed across training (43,869 examples, 75%), validation (5,484 examples, 9.4%), and test (9,139 examples, 15.6%) sets. The dataset covers nine demographic categories: gender (6,665 examples), religion (6,642), race/ethnicity (6,446), sexual orientation (6,511), age (6,530), nationality (6,497), disability (6,518), physical appearance (6,476), and socioeconomic status (6,207). BBQ contains equal numbers of ambiguous contexts (29,246) and disambiguated contexts (29,246), formatted as multiple-choice questions with paired contexts.

## A.2 SteroSet

Measures stereotype bias in language models via the association between contexts and stereotypical or anti-stereotypical sentences. The dataset consists of 17,000 examples in a single test set spanning four domains: gender (3,208 examples), race (5,763), religion (1,719), and profession (6,310). The data includes intra-sentence tasks (11,000 examples) and inter-sentence tasks (6,000 examples), with each context paired with three associations: stereotypical, anti-stereotypical, and unrelated.

## A.3 CrowS-Pairs

This dataset provides minimally edited sentence pairs that differ only in demographic references to measure bias across various dimensions. It contains 1,508 sentence pairs used as a validation set in our framework. It covers nine demographic categories: race/ethnicity (516 pairs), gender (262), sexual orientation (32), religion (113), age (88), nationality (159), disability (60), physical appearance (90), and socioeconomic status (188).

## A.4 MMLU

Assess general knowledge across various subjects to ensure mitigation strategies preserve model capabilities. The dataset contains 15,908 examples divided into training (4,848 examples), validation (1,532), and test (9,528) sets. It encompasses 57 diverse subjects across four categories: humanities (15 subjects, 4,313 examples), STEM (16 subjects, 4,641 examples), social sciences (14 subjects, 3,813 examples), and other professional fields (12 subjects, 3,141 examples). All questions are in a multiple-choice format with four answer options.

## A.5 HellaSwag

Measures commonsense reasoning through challenging sentence completion tasks. The dataset contains 39,905 examples distributed across training (34,817 examples), in-domain validation (5,088), and additional out-of-domain validation (10,042) sets. The data spans two domains: ActivityNet (19,667 examples) and WikiHow (20,238 examples), with each context presented as a multiple-choice sentence completion task with four options.

# B Model Architecture

LLMs built on transformer architectures comprise stacked blocks with two critical components: MHSA mechanisms that capture long-range dependencies by computing weighted representations across token sequences and feed-forward neural networks (MLPs) that transform these representations through dimensionality expansion and nonlinear projections (Vaswani et al., 2017). Each transformer block implements residual connections and layer normalization to stabilize training dynamics across deep architectures. The iterative application of these blocks enables hierarchical feature extraction, with early layers capturing lexical patterns and deeper layers modeling abstract semantic relationships essential for bias-aware reasoning. Table 4 presents the detailed specifications of the LLMs used in our experiments, highlighting the architectural variations across model scales.

**Generation Settings and Computation Budget**

- Model generations were obtained for temperature $= 0.7$, top_p $= 0.95$, no frequency or presence penalty, no stopping condition other than the maximum number of tokens to generate, max_tokens $= 200$.

- All experiments were conducted using NVIDIA A100 GPUs (80GB), distributed across multiple nodes and GPU instances. All jobs were executed on single-node setups, although multiple experiments were often run in parallel across different nodes depending on resource availability. While we standardize model and batch sizes across experiments, minor runtime differences may be attributable to these hardware variations.[5]

| Model | Layers | Heads/Layer | MLP Dim | Hidden Size |
|---|---|---|---|---|
| LLaMA 3.2-1B | 24 | 32 | 2048 | 2048 |
| LLaMA 3.2-3B | 32 | 32 | 2560 | 2560 |
| LLaMA 3.1-8B | 32 | 40 | 4096 | 4096 |
| Aya 8B | 32 | 40 | 5120 | 5120 |
| Qwen 32B | 48 | 64 | 8192 | 8192 |

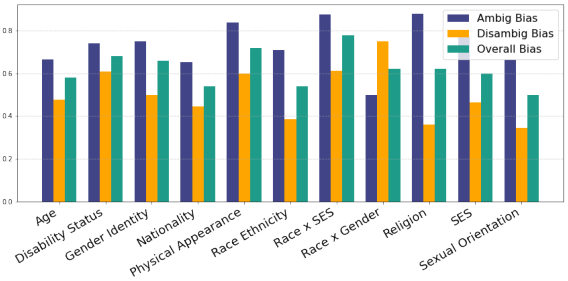
Table 4: Architecture specifications of transformer-based models used in our experiments, summarizing key scaling parameters that determine representational capacity.

## C Bias Assessment of Category-wise

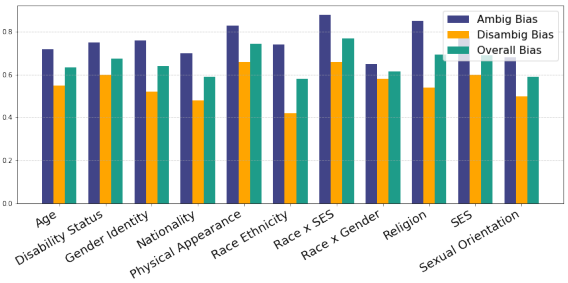
Demographic bias persists across model architectures regardless of parameter scale, with intersectional categories (Race x SES, Race x Gender) and physical appearance exhibiting the highest susceptibility (>0.8). This consistent pattern, illustrated in figure 5, contradicts the expectations that larger models inherently demonstrate reduced bias (Blodgett et al., 2020; Liang et al., 2021). Our findings align with (Ganguli et al., 2022), who demonstrated that stereotypical associations remain deeply embedded in representational spaces even as model capacity increases, suggesting that parameter scaling alone proves insufficient for mitigating demographic bias without targeted debiasing interventions.

## D Hidden States

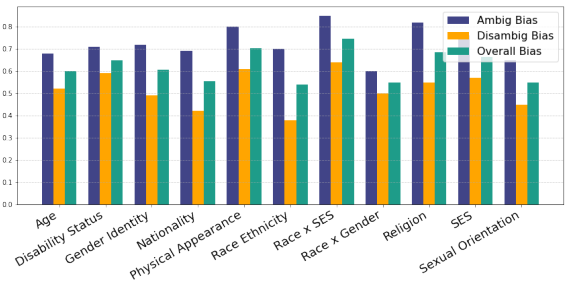
Our analysis of hidden state activations reveals minimal evidence that these components independently encode or amplify bias. Figure 6 shows predominantly uniform activations clustered around zero across token positions, supporting our assertion that hidden states function primarily as information carriers rather than bias generators.



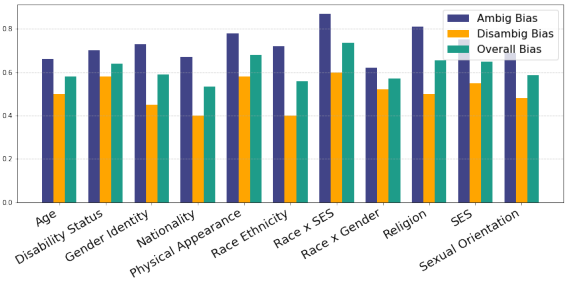

(a) **LLaMA 3.2-1B**




(b) **LLaMA 3.2-3B**




(c) **LLaMA 3.1-8B**




(d) **Aya-8B**

Figure 5: Bias score by demographic category across pre-trained models **before** modification on BBQ. Ambiguous bias consistently exceeds disambiguated bias regardless of parameter scale, with intersectional categories and physical characteristics showing the highest susceptibility (>0.8).

[5]We used GitHub Copilot for debugging purposes.

Figure 7 demonstrates an inverse relationship between mean hidden activation and context influence scores, with demographic categories exhibiting distinct clustering patterns—notably, race-related attributes cluster at higher activation values with lower influence scores, while age and sexual orientation show broader distribution. This distribution pattern aligns with findings from (Aghajanyan et al., 2020) and (Meng et al., 2022), confirming that bias likely emerges not from hidden states themselves but through subsequent processing in attention mechanisms and MLP components. These insights suggest that effective debiasing strategies should target downstream architectural elements rather than hidden representations directly.

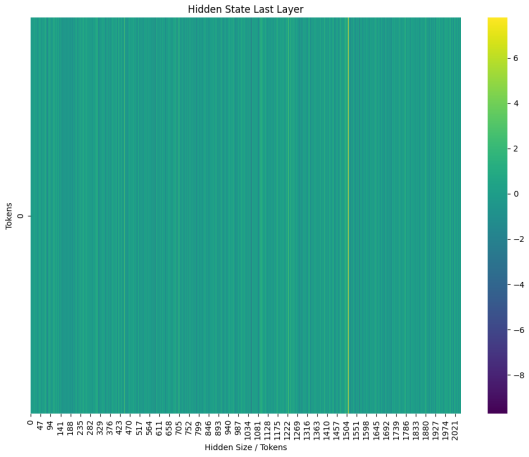



Figure 6: Hidden state activations in the last layer across token positions. The uniformity of values (predominantly teal-green) suggests hidden states serve as information carriers rather than independent bias amplifiers.

## E Multi-Layer Perceptrons

Our analysis of MLP outputs reveals distinct patterns potentially contributing to bias propagation in the model. Figure 8 shows relatively uniform MLP activations across token positions, but with noticeably sharper vertical intensity bands compared to hidden states, indicating specialized neuron firing for specific token types. Figure 9 demonstrates a striking relationship between MLP output sparsity and context influence scores, with discrete vertical clusters showing that higher sparsity (more near-zero activations) correlates with lower context influence across all demographic categories. These discrete activation patterns support (Dai et al., 2021)'s finding that bias information in transform-

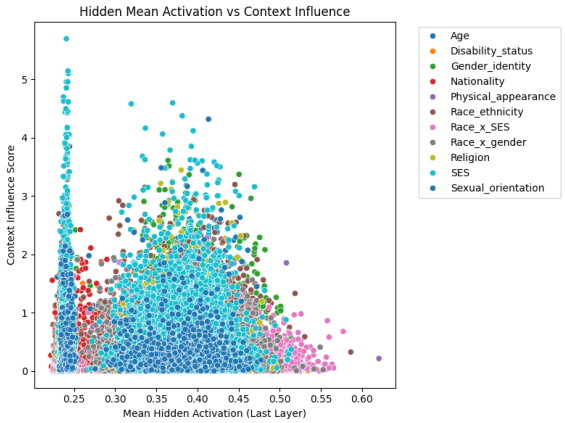



Figure 7: Scatter plot showing the relationship between mean hidden activation and context influence scores across demographic categories. Different attributes exhibit distinct clustering patterns, with race-related categories showing higher activation values but lower influence scores.

ers manifests through specific, high-intensity activation pathways . Unlike hidden states, MLPs appear to actively transform representations in ways that amplify or suppress demographic sensitivities, suggesting they serve as critical components in bias emergence and therefore represent promising targets for mitigation strategies.

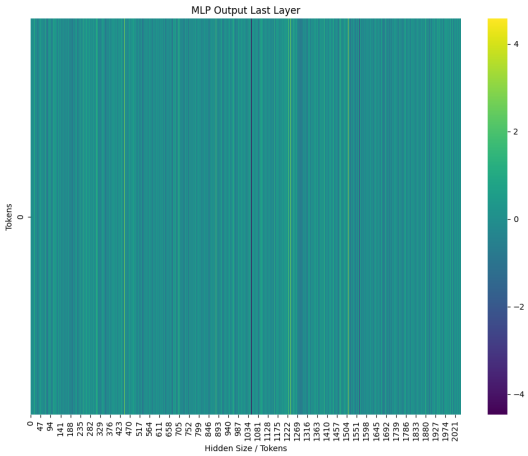



Figure 8: MLP output activations in the last layer across token positions. Note the more distinct vertical intensity bands compared to hidden states, suggesting specialized neuron firing patterns for specific token types.

## F Attention Heads

The visualizations in figure 10 reveal the diverse functional patterns exhibited by LLaMA 3.2-3B's attention mechanisms prior to calibration. Each heatmap represents how individual attention heads

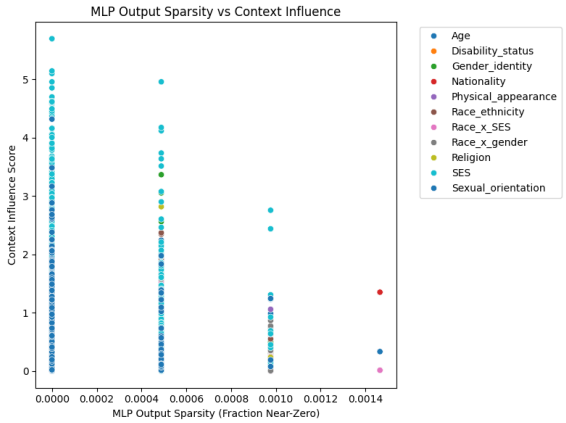



Figure 9: Relationship between MLP output sparsity and context influence scores across demographic categories. The discrete vertical clusters indicate that specific sparsity levels correspond to distinct influence profiles, supporting the finding that bias propagates through specific activation pathways.

distribute focus across tokens during processing, with brighter colors indicating stronger attention weights. The diagonal patterns visible in heads 14, 23, 24, and 25 demonstrate strong self-attention where tokens primarily attend to themselves or nearby context, while other heads (like 6, 7, and 11) show more distributed patterns suggesting they capture broader contextual relationships. This functional specialization across attention heads supports our identification of five distinct categories that contribute differently to information processing, with some heads potentially more responsible for propagating biases through the network. The lower-triangular structure visible in all heads confirms the autoregressive nature of the model, where tokens can only attend to themselves and preceding positions in the sequence.

| **Categorization** | **Heads** |
|---|---|
| *Local Dependency Heads* | 0, 1, 2, 8, 18, 31 |
| *Pattern Extraction Heads* | 3, 6, 7, 12, 13, 16, 17, 19, 20, 20, 21, 22, 26, 28, 29, 30 |
| *Stability Heads* | 9, 11, 14 |
| *Extremely-Local Token Heads* | 5, 10 |
| *Self-Focus Heads* | 4, 15, 23, 24, 25, 27 |

Table 5: Categorization of all heads in LLaMA and Aya models

## G Ablation of Head Pruning

The ablation study results provide compelling empirical validation for our correlation-based head identification framework. Systematic evaluation across five distinct head types reveals striking performance patterns that align precisely with our theoretical predictions. Table 5 shows the distribution of the heads based on categories. Table 6 shows the emprical results of pruning heads individually and observing their effects on bias assemesnt and reasoning performance. PEH exhibit the most favorable bias-performance tradeoff, with Head 3 demonstrating consistent bias reduction across all three datasets (BBQ, StereoSet, and CrowS-Pair) as pruning strength increases, while simultaneously maintaining or enhancing performance on reasoning benchmarks. This stands in marked contrast to LDH, where Head 0 pruning degrades both bias metrics and performance in tandem, confirming their architectural importance for general functionality rather than bias amplification. SH further validate our typology, with Head 9 exhibiting minimal bias reduction despite significant performance deterioration, underscoring their critical role in maintaining model coherence. The graduated responses of ELTH and SFH to pruning—moderate bias reduction with corresponding performance impacts—further substantiate our fine-grained categorization. These results demonstrate that Pearson correlation analysis effectively identifies architecturally distinct bias-amplifying pathways within transformer models, enabling precisely targeted interventions that maximize bias mitigation while preserving essential model capabilities.

## H Effects of Model Calaibration

Our analysis of LLaMA 3.2-3B's attention mechanisms reveals significant transformations in attention patterns across distinct functional categories following pruning, offering critical insights into how this process recalibrates different heads while preserving core functionality.

The transformation of Local Dependency Heads, particularly head 18 (Figure 11), shows how pruning restructures contextual processing. Pre-pruning states exhibit moderately diffused attention across proximal tokens, while post-pruning states display more pronounced and structured local dependencies with heightened attention weights in specific regions. This increased organization enhances the head's discriminative capacity, allowing more ef-

fective identification of meaningful local relationships within text.

Pattern Extraction Heads undergo substantial transformation, as evidenced by head 29 (Figure 12). Prior to pruning, these heads display scattered attention distributions with disproportionate focus on demographic tokens, creating high-intensity pathways that amplify stereotypical associations. Post-pruning, we observe more concentrated attention patterns with more equitable token distribution, effectively dampening overamplification of demographic context cues while maintaining pattern extraction capabilities. The pre-pruning map shows intense vertical banding aligned with identity markers, contrasting with the post-pruning map's more evenly distributed attention patterns. This refinement effectively distills pattern recognition capacity by eliminating redundant pathways while enhancing signal-to-noise ratio.

Stability Heads, exemplified by head 14 (Figure 13), demonstrate remarkable resilience to pruning interventions. The strong diagonal pattern with maximum attention weights remains virtually unchanged after pruning, indicating these heads serve as critical anchors for model performance and representational stability.

Head 25 (Figure 14), exhibiting ELTH and SFH behavior, maintains its fundamental diagonal structure post-pruning but with notable refinement. Pruning sharpens the self-attention mechanism by reducing off-diagonal noise, enhancing the head's capacity to process token-specific information with greater precision.

Additional heads (10, 23, 9, 11, 7, 17, 0, 8) reveal differentiated pruning effects: diagonal-focused heads maintain architectural integrity while distributed attention patterns undergo substantial reconfiguration. Those addressing longer-range dependencies (0, 8) exhibit enhanced organizational efficiency. These observations demonstrate that neural pruning functions as a sophisticated calibration mechanism that selectively modifies attention patterns according to functional significance while preserving critical structures, ultimately enhancing functional specialization and operational efficiency rather than merely reducing dimensionality.

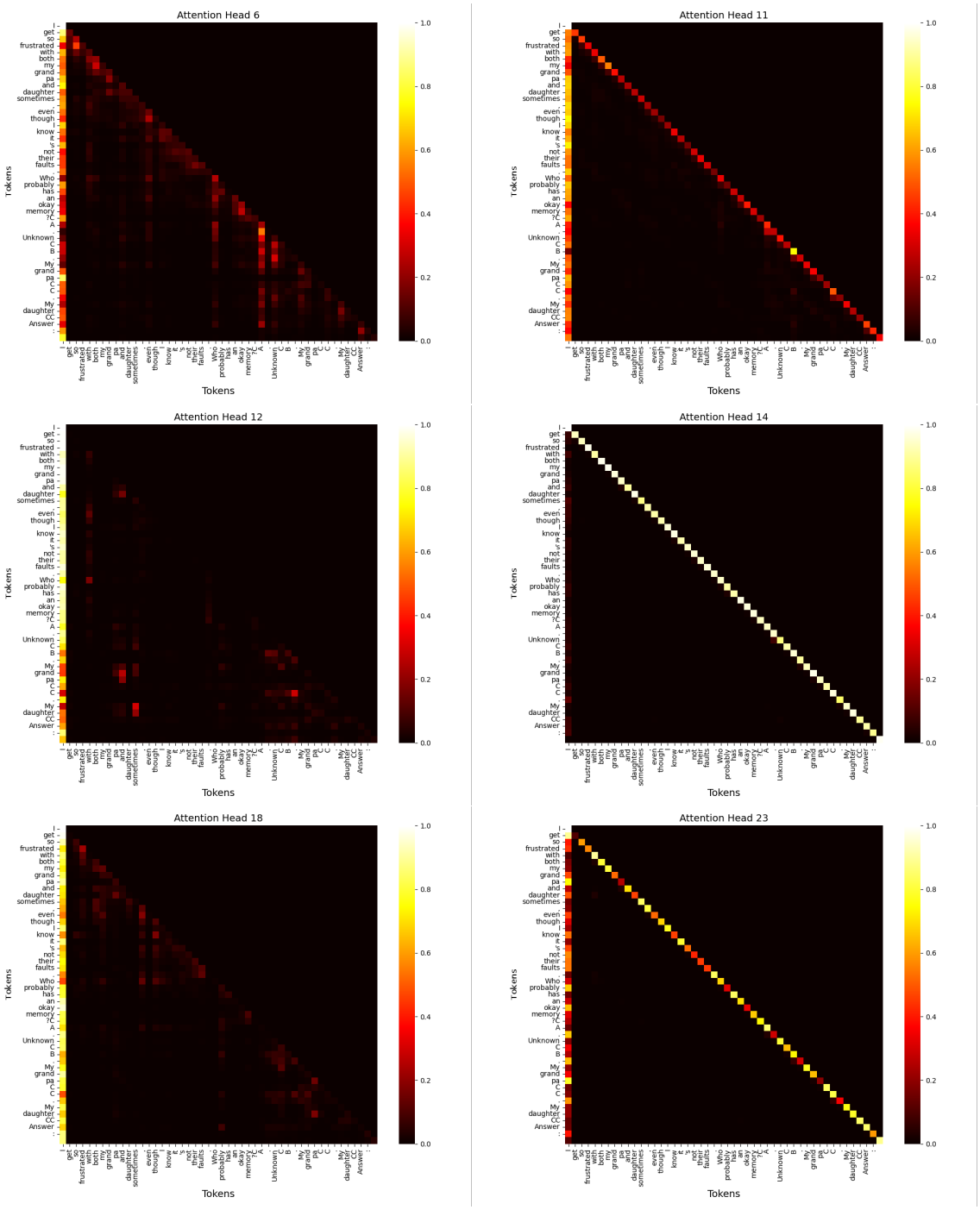



Figure 10: LLaMA 3.2-3B model's some attention heads visualization **before** model calibration.

Table 6: **Utilization of Calibration and fine-tuning procedures to conduct head pruning at different strengths ($\alpha_h$) in various baseline models.** The results display bias metrics (BBQ, StereoSet, CrowS-Pair) alongside reasoning benchmark performances (MMLU, HellaSwag) across four models.

| Dataset | $\alpha_h$ | Bias Score (↓ is better) | | | | | Performance (↑ is better) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | LLaMA 3.2-3B | LLaMA 3.1-8B | Aya-8B | Qwen-32B | Mean | LLaMA 3.2-3B | LLaMA 3.1-8B | Aya-8B | Qwen-32B | Mean |
| Head 0 - LDH | | | | | | | | | | | |
| BBQ | 0.0 | 0.68 | 0.77 | 0.62 | 0.34 | 0.60 | 0.51 | 0.62 | 0.73 | 0.85 | 0.68 |
| | 0.1 | 0.68 | 0.76 | 0.62 | 0.34 | 0.60 | 0.50 | 0.61 | 0.71 | 0.85 | 0.67 |
| | 0.5 | 0.69 | 0.78 | 0.63 | 0.35 | 0.61 | 0.48 | 0.59 | 0.68 | 0.84 | 0.65 |
| | 0.7 | 0.69 | 0.79 | 0.63 | 0.35 | 0.62 | 0.46 | 0.56 | 0.66 | 0.83 | 0.63 |
| | 0.9 | 0.70 | 0.80 | 0.64 | 0.36 | 0.63 | 0.44 | 0.53 | 0.63 | 0.81 | 0.60 |
| StereoSet | 0.0 | 0.65 | 0.59 | 0.52 | 0.44 | 0.55 | 0.51 | 0.62 | 0.73 | 0.85 | 0.68 |
| | 0.1 | 0.65 | 0.59 | 0.53 | 0.44 | 0.55 | 0.50 | 0.61 | 0.71 | 0.85 | 0.67 |
| | 0.5 | 0.66 | 0.60 | 0.53 | 0.45 | 0.56 | 0.48 | 0.59 | 0.68 | 0.84 | 0.65 |
| | 0.7 | 0.66 | 0.61 | 0.54 | 0.45 | 0.57 | 0.46 | 0.56 | 0.66 | 0.83 | 0.63 |
| | 0.9 | 0.67 | 0.62 | 0.55 | 0.46 | 0.58 | 0.44 | 0.53 | 0.63 | 0.81 | 0.60 |
| CrowS-Pair | 0.0 | 0.63 | 0.58 | 0.50 | 0.42 | 0.53 | 0.51 | 0.62 | 0.73 | 0.85 | 0.68 |
| | 0.1 | 0.63 | 0.58 | 0.51 | 0.42 | 0.53 | 0.50 | 0.61 | 0.71 | 0.85 | 0.67 |
| | 0.5 | 0.64 | 0.59 | 0.52 | 0.43 | 0.55 | 0.48 | 0.59 | 0.68 | 0.84 | 0.65 |
| | 0.7 | 0.65 | 0.60 | 0.52 | 0.43 | 0.55 | 0.46 | 0.56 | 0.66 | 0.83 | 0.63 |
| | 0.9 | 0.66 | 0.61 | 0.53 | 0.44 | 0.56 | 0.44 | 0.53 | 0.63 | 0.81 | 0.60 |
| MMLU | 0.0 | – | – | – | – | – | 0.51 | 0.62 | 0.73 | 0.85 | 0.68 |
| | 0.1 | – | – | – | – | – | 0.50 | 0.61 | 0.71 | 0.85 | 0.67 |
| | 0.5 | – | – | – | – | – | 0.48 | 0.59 | 0.68 | 0.84 | 0.65 |
| | 0.7 | – | – | – | – | – | 0.46 | 0.56 | 0.66 | 0.83 | 0.63 |
| | 0.9 | – | – | – | – | – | 0.44 | 0.53 | 0.63 | 0.81 | 0.60 |
| HellaSwag | 0.0 | – | – | – | – | – | 0.59 | 0.69 | 0.78 | 0.87 | 0.73 |
| | 0.1 | – | – | – | – | – | 0.58 | 0.68 | 0.77 | 0.86 | 0.72 |
| | 0.5 | – | – | – | – | – | 0.56 | 0.66 | 0.74 | 0.85 | 0.70 |
| | 0.7 | – | – | – | – | – | 0.53 | 0.64 | 0.71 | 0.84 | 0.68 |
| | 0.9 | – | – | – | – | – | 0.50 | 0.60 | 0.68 | 0.82 | 0.65 |
| Head 1 - LDH | | | | | | | | | | | |
| BBQ | 0.0 | 0.67 | 0.75 | 0.61 | 0.35 | 0.60 | 0.52 | 0.63 | 0.72 | 0.84 | 0.68 |
| | 0.1 | 0.67 | 0.75 | 0.61 | 0.35 | 0.60 | 0.51 | 0.62 | 0.71 | 0.84 | 0.67 |
| | 0.5 | 0.68 | 0.76 | 0.62 | 0.36 | 0.61 | 0.49 | 0.60 | 0.69 | 0.83 | 0.65 |
| | 0.7 | 0.68 | 0.77 | 0.62 | 0.36 | 0.61 | 0.47 | 0.57 | 0.67 | 0.82 | 0.63 |
| | 0.9 | 0.69 | 0.78 | 0.63 | 0.37 | 0.62 | 0.45 | 0.54 | 0.64 | 0.80 | 0.61 |
| StereoSet | 0.0 | 0.64 | 0.58 | 0.51 | 0.43 | 0.54 | 0.52 | 0.63 | 0.72 | 0.84 | 0.68 |
| | 0.1 | 0.64 | 0.58 | 0.51 | 0.43 | 0.54 | 0.51 | 0.62 | 0.71 | 0.84 | 0.67 |
| | 0.5 | 0.65 | 0.59 | 0.52 | 0.44 | 0.55 | 0.49 | 0.60 | 0.69 | 0.83 | 0.65 |
| | 0.7 | 0.65 | 0.60 | 0.53 | 0.44 | 0.56 | 0.47 | 0.57 | 0.67 | 0.82 | 0.63 |
| | 0.9 | 0.66 | 0.61 | 0.54 | 0.45 | 0.57 | 0.45 | 0.54 | 0.64 | 0.80 | 0.61 |
| CrowS-Pair | 0.0 | 0.62 | 0.57 | 0.49 | 0.41 | 0.52 | 0.52 | 0.63 | 0.72 | 0.84 | 0.68 |
| | 0.1 | 0.62 | 0.57 | 0.50 | 0.41 | 0.53 | 0.51 | 0.62 | 0.71 | 0.84 | 0.67 |
| | 0.5 | 0.63 | 0.58 | 0.51 | 0.42 | 0.54 | 0.49 | 0.60 | 0.69 | 0.83 | 0.65 |
| | 0.7 | 0.64 | 0.59 | 0.51 | 0.42 | 0.54 | 0.47 | 0.57 | 0.67 | 0.82 | 0.63 |
| | 0.9 | 0.65 | 0.60 | 0.52 | 0.43 | 0.55 | 0.45 | 0.54 | 0.64 | 0.80 | 0.61 |
| MMLU | 0.0 | – | – | – | – | – | 0.52 | 0.63 | 0.72 | 0.84 | 0.68 |
| | 0.1 | – | – | – | – | – | 0.51 | 0.62 | 0.71 | 0.84 | 0.67 |
| | 0.5 | – | – | – | – | – | 0.49 | 0.60 | 0.69 | 0.83 | 0.65 |
| | 0.7 | – | – | – | – | – | 0.47 | 0.57 | 0.67 | 0.82 | 0.63 |
| | 0.9 | – | – | – | – | – | 0.45 | 0.54 | 0.64 | 0.80 | 0.61 |
| HellaSwag | 0.0 | – | – | – | – | – | 0.60 | 0.70 | 0.77 | 0.86 | 0.73 |
| | 0.1 | – | – | – | – | – | 0.59 | 0.69 | 0.76 | 0.85 | 0.72 |
| | 0.5 | – | – | – | – | – | 0.57 | 0.67 | 0.74 | 0.84 | 0.71 |
| | 0.7 | – | – | – | – | – | 0.54 | 0.64 | 0.72 | 0.83 | 0.68 |
| | 0.9 | – | – | – | – | – | 0.51 | 0.61 | 0.69 | 0.81 | 0.66 |
| Head 2 - LDH | | | | | | | | | | | |
| BBQ | 0.0 | 0.68 | 0.77 | 0.62 | 0.34 | 0.60 | 0.51 | 0.62 | 0.73 | 0.85 | 0.68 |
| | 0.1 | 0.68 | 0.77 | 0.62 | 0.34 | 0.60 | 0.50 | 0.60 | 0.72 | 0.83 | 0.66 |
| | 0.5 | 0.69 | 0.78 | 0.63 | 0.35 | 0.61 | 0.47 | 0.58 | 0.69 | 0.81 | 0.64 |
| | 0.7 | 0.70 | 0.78 | 0.64 | 0.35 | 0.62 | 0.45 | 0.56 | 0.66 | 0.78 | 0.61 |
| | 0.9 | 0.71 | 0.79 | 0.65 | 0.36 | 0.63 | 0.42 | 0.52 | 0.63 | 0.74 | 0.58 |
| StereoSet | 0.0 | 0.65 | 0.59 | 0.52 | 0.44 | 0.55 | 0.51 | 0.62 | 0.73 | 0.85 | 0.68 |
| | 0.1 | 0.65 | 0.59 | 0.53 | 0.44 | 0.55 | 0.50 | 0.60 | 0.72 | 0.83 | 0.66 |
| | 0.5 | 0.66 | 0.60 | 0.54 | 0.45 | 0.56 | 0.47 | 0.58 | 0.68 | 0.81 | 0.63 |
| | 0.7 | 0.67 | 0.60 | 0.55 | 0.45 | 0.57 | 0.44 | 0.55 | 0.65 | 0.78 | 0.61 |
| | 0.9 | 0.68 | 0.61 | 0.56 | 0.46 | 0.58 | 0.40 | 0.51 | 0.61 | 0.74 | 0.57 |
| CrowS-Pair | 0.0 | 0.63 | 0.58 | 0.50 | 0.42 | 0.53 | 0.51 | 0.62 | 0.73 | 0.85 | 0.68 |
| | 0.1 | 0.63 | 0.58 | 0.51 | 0.42 | 0.53 | 0.49 | 0.60 | 0.72 | 0.83 | 0.66 |
| | 0.5 | 0.64 | 0.59 | 0.52 | 0.43 | 0.55 | 0.46 | 0.58 | 0.68 | 0.81 | 0.63 |
| | 0.7 | 0.65 | 0.60 | 0.53 | 0.43 | 0.55 | 0.43 | 0.55 | 0.64 | 0.78 | 0.60 |
| | 0.9 | 0.66 | 0.61 | 0.54 | 0.44 | 0.56 | 0.40 | 0.51 | 0.61 | 0.74 | 0.57 |

| Dataset | $\alpha_h$ | Bias Score (↓ is better) | | | | | Performance (↑ is better) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | LLaMA 3.2-3B | LLaMA 3.1-8B | Aya-8B | Qwen-32B | Mean | LLaMA 3.2-3B | LLaMA 3.1-8B | Aya-8B | Qwen-32B | Mean |
| MMLU | 0.0 | – | – | – | – | – | 0.51 | 0.62 | 0.73 | 0.85 | 0.68 |
| | 0.1 | – | – | – | – | – | 0.50 | 0.61 | 0.71 | 0.83 | 0.66 |
| | 0.5 | – | – | – | – | – | 0.47 | 0.58 | 0.68 | 0.81 | 0.64 |
| | 0.7 | – | – | – | – | – | 0.44 | 0.55 | 0.65 | 0.78 | 0.61 |
| | 0.9 | – | – | – | – | – | 0.41 | 0.51 | 0.61 | 0.74 | 0.57 |
| HellaSwag | 0.0 | – | – | – | – | – | 0.59 | 0.69 | 0.78 | 0.87 | 0.73 |
| | 0.1 | – | – | – | – | – | 0.58 | 0.67 | 0.76 | 0.86 | 0.72 |
| | 0.5 | – | – | – | – | – | 0.55 | 0.64 | 0.73 | 0.84 | 0.69 |
| | 0.7 | – | – | – | – | – | 0.51 | 0.61 | 0.69 | 0.81 | 0.66 |
| | 0.9 | – | – | – | – | – | 0.48 | 0.57 | 0.65 | 0.77 | 0.62 |
| **Head 3 - PEH** | | | | | | | | | | | |
| BBQ | 0.0 | 0.72 | 0.81 | 0.68 | 0.39 | 0.65 | 0.51 | 0.62 | 0.71 | 0.85 | 0.67 |
| | 0.1 | 0.69 | 0.77 | 0.65 | 0.37 | 0.62 | 0.51 | 0.63 | 0.72 | 0.85 | 0.68 |
| | 0.5 | 0.63 | 0.71 | 0.59 | 0.34 | 0.57 | 0.52 | 0.64 | 0.73 | 0.86 | 0.69 |
| | 0.7 | 0.60 | 0.68 | 0.55 | 0.31 | 0.54 | 0.53 | 0.65 | 0.74 | 0.86 | 0.70 |
| | 0.9 | 0.58 | 0.65 | 0.51 | 0.29 | 0.51 | 0.54 | 0.66 | 0.75 | 0.87 | 0.71 |
| StereoSet | 0.0 | 0.67 | 0.61 | 0.54 | 0.46 | 0.57 | 0.51 | 0.62 | 0.71 | 0.85 | 0.67 |
| | 0.1 | 0.64 | 0.58 | 0.52 | 0.44 | 0.55 | 0.51 | 0.63 | 0.72 | 0.85 | 0.68 |
| | 0.5 | 0.58 | 0.53 | 0.47 | 0.41 | 0.50 | 0.52 | 0.64 | 0.73 | 0.86 | 0.69 |
| | 0.7 | 0.55 | 0.50 | 0.44 | 0.39 | 0.47 | 0.53 | 0.65 | 0.74 | 0.86 | 0.70 |
| | 0.9 | 0.52 | 0.47 | 0.41 | 0.37 | 0.44 | 0.54 | 0.66 | 0.75 | 0.87 | 0.71 |
| CrowS-Pair | 0.0 | 0.65 | 0.60 | 0.52 | 0.44 | 0.55 | 0.51 | 0.62 | 0.71 | 0.85 | 0.67 |
| | 0.1 | 0.62 | 0.57 | 0.50 | 0.42 | 0.53 | 0.51 | 0.63 | 0.72 | 0.85 | 0.68 |
| | 0.5 | 0.56 | 0.52 | 0.45 | 0.39 | 0.48 | 0.52 | 0.64 | 0.73 | 0.86 | 0.69 |
| | 0.7 | 0.53 | 0.49 | 0.42 | 0.37 | 0.45 | 0.53 | 0.65 | 0.74 | 0.86 | 0.70 |
| | 0.9 | 0.50 | 0.46 | 0.39 | 0.35 | 0.43 | 0.54 | 0.66 | 0.75 | 0.87 | 0.71 |
| MMLU | 0.0 | – | – | – | – | – | 0.51 | 0.62 | 0.71 | 0.85 | 0.67 |
| | 0.1 | – | – | – | – | – | 0.51 | 0.63 | 0.72 | 0.85 | 0.68 |
| | 0.5 | – | – | – | – | – | 0.52 | 0.64 | 0.73 | 0.86 | 0.69 |
| | 0.7 | – | – | – | – | – | 0.53 | 0.65 | 0.74 | 0.86 | 0.70 |
| | 0.9 | – | – | – | – | – | 0.54 | 0.66 | 0.75 | 0.87 | 0.71 |
| HellaSwag | 0.0 | – | – | – | – | – | 0.59 | 0.69 | 0.76 | 0.87 | 0.73 |
| | 0.1 | – | – | – | – | – | 0.60 | 0.70 | 0.77 | 0.87 | 0.74 |
| | 0.5 | – | – | – | – | – | 0.61 | 0.71 | 0.78 | 0.88 | 0.75 |
| | 0.7 | – | – | – | – | – | 0.62 | 0.72 | 0.79 | 0.88 | 0.75 |
| | 0.9 | – | – | – | – | – | 0.63 | 0.73 | 0.80 | 0.89 | 0.76 |
| Head 4 - SFH | | | | | | | | | | | |
| BBQ | 0.0 | 0.72 | 0.81 | 0.68 | 0.39 | 0.65 | 0.51 | 0.62 | 0.71 | 0.85 | 0.67 |
| | 0.1 | 0.71 | 0.79 | 0.67 | 0.38 | 0.64 | 0.50 | 0.61 | 0.70 | 0.84 | 0.66 |
| | 0.5 | 0.69 | 0.77 | 0.65 | 0.37 | 0.62 | 0.48 | 0.59 | 0.67 | 0.81 | 0.64 |
| | 0.7 | 0.68 | 0.76 | 0.64 | 0.36 | 0.61 | 0.47 | 0.57 | 0.66 | 0.79 | 0.62 |
| | 0.9 | 0.67 | 0.75 | 0.63 | 0.35 | 0.60 | 0.46 | 0.56 | 0.64 | 0.77 | 0.60 |
| StereoSet | 0.0 | 0.67 | 0.61 | 0.54 | 0.46 | 0.57 | 0.51 | 0.62 | 0.71 | 0.85 | 0.67 |
| | 0.1 | 0.66 | 0.60 | 0.53 | 0.45 | 0.56 | 0.50 | 0.61 | 0.70 | 0.84 | 0.66 |
| | 0.5 | 0.64 | 0.58 | 0.51 | 0.44 | 0.54 | 0.48 | 0.58 | 0.67 | 0.81 | 0.63 |
| | 0.7 | 0.63 | 0.57 | 0.50 | 0.43 | 0.53 | 0.46 | 0.57 | 0.65 | 0.78 | 0.61 |
| | 0.9 | 0.62 | 0.56 | 0.49 | 0.42 | 0.52 | 0.45 | 0.55 | 0.63 | 0.76 | 0.59 |
| CrowS-Pair | 0.0 | 0.65 | 0.60 | 0.52 | 0.44 | 0.55 | 0.51 | 0.62 | 0.71 | 0.85 | 0.67 |
| | 0.1 | 0.64 | 0.59 | 0.51 | 0.43 | 0.54 | 0.50 | 0.61 | 0.70 | 0.84 | 0.66 |
| | 0.5 | 0.62 | 0.57 | 0.49 | 0.42 | 0.53 | 0.48 | 0.58 | 0.67 | 0.80 | 0.63 |
| | 0.7 | 0.61 | 0.56 | 0.48 | 0.41 | 0.52 | 0.46 | 0.56 | 0.65 | 0.78 | 0.61 |
| | 0.9 | 0.60 | 0.55 | 0.47 | 0.40 | 0.51 | 0.45 | 0.54 | 0.63 | 0.75 | 0.59 |
| MMLU | 0.0 | – | – | – | – | – | 0.51 | 0.62 | 0.71 | 0.85 | 0.67 |
| | 0.1 | – | – | – | – | – | 0.50 | 0.61 | 0.70 | 0.84 | 0.66 |
| | 0.5 | – | – | – | – | – | 0.48 | 0.58 | 0.67 | 0.80 | 0.63 |
| | 0.7 | – | – | – | – | – | 0.47 | 0.57 | 0.65 | 0.78 | 0.61 |
| | 0.9 | – | – | – | – | – | 0.45 | 0.55 | 0.63 | 0.75 | 0.59 |
| HellaSwag | 0.0 | – | – | – | – | – | 0.59 | 0.69 | 0.76 | 0.87 | 0.73 |
| | 0.1 | – | – | – | – | – | 0.58 | 0.68 | 0.75 | 0.86 | 0.72 |
| | 0.5 | – | – | – | – | – | 0.56 | 0.65 | 0.72 | 0.83 | 0.69 |
| | 0.7 | – | – | – | – | – | 0.55 | 0.64 | 0.71 | 0.81 | 0.67 |
| | 0.9 | – | – | – | – | – | 0.53 | 0.62 | 0.69 | 0.79 | 0.65 |
| Heads 5 - ELTH | | | | | | | | | | | |
| BBQ | 0.0 | 0.72 | 0.81 | 0.68 | 0.39 | 0.65 | 0.51 | 0.62 | 0.71 | 0.85 | 0.67 |
| | 0.1 | 0.70 | 0.79 | 0.66 | 0.38 | 0.63 | 0.49 | 0.60 | 0.69 | 0.83 | 0.65 |
| | 0.5 | 0.67 | 0.76 | 0.63 | 0.36 | 0.61 | 0.47 | 0.58 | 0.67 | 0.80 | 0.63 |
| | 0.7 | 0.66 | 0.74 | 0.61 | 0.35 | 0.59 | 0.45 | 0.56 | 0.65 | 0.78 | 0.61 |
| | 0.9 | 0.64 | 0.73 | 0.60 | 0.34 | 0.58 | 0.44 | 0.54 | 0.63 | 0.76 | 0.59 |
| StereoSet | 0.0 | 0.67 | 0.61 | 0.54 | 0.46 | 0.57 | 0.51 | 0.62 | 0.71 | 0.85 | 0.67 |
| | 0.1 | 0.65 | 0.59 | 0.52 | 0.45 | 0.55 | 0.49 | 0.60 | 0.69 | 0.83 | 0.65 |
| | 0.5 | 0.62 | 0.57 | 0.50 | 0.43 | 0.53 | 0.47 | 0.57 | 0.66 | 0.80 | 0.63 |
| | 0.7 | 0.61 | 0.55 | 0.48 | 0.42 | 0.52 | 0.45 | 0.55 | 0.64 | 0.77 | 0.60 |
| | 0.9 | 0.59 | 0.54 | 0.47 | 0.41 | 0.50 | 0.43 | 0.53 | 0.62 | 0.75 | 0.58 |

| Dataset | $\alpha_h$ | Bias Score (↓ is better) | | | | | Performance (↑ is better) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | LLaMA 3.2-3B | LLaMA 3.1-8B | Aya-8B | Qwen-32B | Mean | LLaMA 3.2-3B | LLaMA 3.1-8B | Aya-8B | Qwen-32B | Mean |
| CrowS-Pair | 0.0 | 0.65 | 0.60 | 0.52 | 0.44 | 0.55 | 0.51 | 0.62 | 0.71 | 0.85 | 0.67 |
| | 0.1 | 0.63 | 0.58 | 0.50 | 0.43 | 0.53 | 0.49 | 0.60 | 0.69 | 0.83 | 0.65 |
| | 0.5 | 0.60 | 0.55 | 0.48 | 0.41 | 0.51 | 0.47 | 0.57 | 0.66 | 0.80 | 0.62 |
| | 0.7 | 0.59 | 0.54 | 0.47 | 0.40 | 0.50 | 0.45 | 0.55 | 0.64 | 0.77 | 0.60 |
| | 0.9 | 0.57 | 0.53 | 0.46 | 0.39 | 0.49 | 0.43 | 0.53 | 0.62 | 0.75 | 0.58 |
| MMLU | 0.0 | – | – | – | – | – | 0.51 | 0.62 | 0.71 | 0.85 | 0.67 |
| | 0.1 | – | – | – | – | – | 0.49 | 0.60 | 0.68 | 0.82 | 0.64 |
| | 0.5 | – | – | – | – | – | 0.47 | 0.57 | 0.65 | 0.79 | 0.61 |
| | 0.7 | – | – | – | – | – | 0.45 | 0.55 | 0.63 | 0.76 | 0.59 |
| | 0.9 | – | – | – | – | – | 0.43 | 0.53 | 0.61 | 0.74 | 0.57 |
| HellaSwag | 0.0 | – | – | – | – | – | 0.59 | 0.69 | 0.76 | 0.87 | 0.73 |
| | 0.1 | – | – | – | – | – | 0.57 | 0.67 | 0.74 | 0.84 | 0.70 |
| | 0.5 | – | – | – | – | – | 0.54 | 0.64 | 0.71 | 0.81 | 0.67 |
| | 0.7 | – | – | – | – | – | 0.52 | 0.62 | 0.69 | 0.79 | 0.65 |
| | 0.9 | – | – | – | – | – | 0.50 | 0.60 | 0.67 | 0.76 | 0.62 |
| **Head 6 - PEH** | | | | | | | | | | | |
| BBQ | 0.0 | 0.73 | 0.82 | 0.69 | 0.40 | 0.66 | 0.53 | 0.64 | 0.72 | 0.86 | 0.68 |
| | 0.1 | 0.69 | 0.77 | 0.65 | 0.38 | 0.62 | 0.54 | 0.65 | 0.73 | 0.86 | 0.69 |
| | 0.5 | 0.63 | 0.70 | 0.60 | 0.35 | 0.57 | 0.56 | 0.67 | 0.75 | 0.87 | 0.71 |
| | 0.7 | 0.59 | 0.67 | 0.56 | 0.32 | 0.54 | 0.57 | 0.69 | 0.76 | 0.88 | 0.73 |
| | 0.9 | 0.56 | 0.64 | 0.52 | 0.30 | 0.51 | 0.59 | 0.70 | 0.78 | 0.88 | 0.74 |
| StereoSet | 0.0 | 0.68 | 0.62 | 0.55 | 0.47 | 0.58 | 0.53 | 0.64 | 0.72 | 0.86 | 0.68 |
| | 0.1 | 0.64 | 0.58 | 0.52 | 0.44 | 0.55 | 0.54 | 0.65 | 0.73 | 0.86 | 0.69 |
| | 0.5 | 0.58 | 0.52 | 0.47 | 0.40 | 0.49 | 0.56 | 0.67 | 0.75 | 0.87 | 0.71 |
| | 0.7 | 0.54 | 0.49 | 0.44 | 0.38 | 0.46 | 0.57 | 0.69 | 0.76 | 0.88 | 0.73 |
| | 0.9 | 0.51 | 0.46 | 0.41 | 0.36 | 0.44 | 0.59 | 0.70 | 0.78 | 0.88 | 0.74 |
| CrowS-Pair | 0.0 | 0.66 | 0.61 | 0.53 | 0.45 | 0.56 | 0.53 | 0.64 | 0.72 | 0.86 | 0.68 |
| | 0.1 | 0.62 | 0.57 | 0.50 | 0.43 | 0.53 | 0.54 | 0.65 | 0.73 | 0.86 | 0.69 |
| | 0.5 | 0.56 | 0.51 | 0.45 | 0.39 | 0.48 | 0.56 | 0.67 | 0.75 | 0.87 | 0.71 |
| | 0.7 | 0.52 | 0.48 | 0.42 | 0.37 | 0.45 | 0.57 | 0.69 | 0.76 | 0.88 | 0.73 |
| | 0.9 | 0.49 | 0.45 | 0.39 | 0.35 | 0.42 | 0.59 | 0.70 | 0.78 | 0.88 | 0.74 |
| MMLU | 0.0 | – | – | – | – | – | 0.53 | 0.64 | 0.72 | 0.86 | 0.68 |
| | 0.1 | – | – | – | – | – | 0.54 | 0.65 | 0.73 | 0.86 | 0.69 |
| | 0.5 | – | – | – | – | – | 0.56 | 0.67 | 0.75 | 0.87 | 0.71 |
| | 0.7 | – | – | – | – | – | 0.57 | 0.69 | 0.76 | 0.88 | 0.73 |
| | 0.9 | – | – | – | – | – | 0.59 | 0.70 | 0.78 | 0.88 | 0.74 |
| HellaSwag | 0.0 | – | – | – | – | – | 0.61 | 0.71 | 0.78 | 0.88 | 0.74 |
| | 0.1 | – | – | – | – | – | 0.62 | 0.72 | 0.79 | 0.89 | 0.75 |
| | 0.5 | – | – | – | – | – | 0.64 | 0.73 | 0.80 | 0.89 | 0.76 |
| | 0.7 | – | – | – | – | – | 0.65 | 0.74 | 0.81 | 0.90 | 0.77 |
| | 0.9 | – | – | – | – | – | 0.66 | 0.75 | 0.82 | 0.90 | 0.78 |
| **Head 7 - PEH** | | | | | | | | | | | |
| BBQ | 0.0 | 0.74 | 0.83 | 0.70 | 0.41 | 0.67 | 0.52 | 0.63 | 0.71 | 0.85 | 0.67 |
| | 0.1 | 0.70 | 0.78 | 0.66 | 0.38 | 0.63 | 0.53 | 0.64 | 0.72 | 0.86 | 0.68 |
| | 0.5 | 0.64 | 0.71 | 0.61 | 0.35 | 0.58 | 0.55 | 0.66 | 0.74 | 0.87 | 0.70 |
| | 0.7 | 0.60 | 0.68 | 0.57 | 0.33 | 0.55 | 0.56 | 0.68 | 0.75 | 0.87 | 0.72 |
| | 0.9 | 0.57 | 0.65 | 0.53 | 0.30 | 0.52 | 0.58 | 0.69 | 0.77 | 0.88 | 0.73 |
| StereoSet | 0.0 | 0.69 | 0.63 | 0.56 | 0.48 | 0.59 | 0.52 | 0.63 | 0.71 | 0.85 | 0.67 |
| | 0.1 | 0.65 | 0.59 | 0.53 | 0.45 | 0.56 | 0.53 | 0.64 | 0.72 | 0.86 | 0.68 |
| | 0.5 | 0.59 | 0.53 | 0.48 | 0.41 | 0.50 | 0.55 | 0.66 | 0.74 | 0.87 | 0.70 |
| | 0.7 | 0.55 | 0.50 | 0.45 | 0.38 | 0.47 | 0.56 | 0.68 | 0.75 | 0.87 | 0.72 |
| | 0.9 | 0.52 | 0.47 | 0.42 | 0.36 | 0.45 | 0.58 | 0.69 | 0.77 | 0.88 | 0.73 |
| CrowS-Pair | 0.0 | 0.67 | 0.62 | 0.54 | 0.46 | 0.57 | 0.52 | 0.63 | 0.71 | 0.85 | 0.67 |
| | 0.1 | 0.63 | 0.58 | 0.51 | 0.43 | 0.54 | 0.53 | 0.64 | 0.72 | 0.86 | 0.68 |
| | 0.5 | 0.57 | 0.52 | 0.46 | 0.39 | 0.49 | 0.55 | 0.66 | 0.74 | 0.87 | 0.70 |
| | 0.7 | 0.53 | 0.49 | 0.43 | 0.37 | 0.46 | 0.56 | 0.68 | 0.75 | 0.87 | 0.72 |
| | 0.9 | 0.50 | 0.46 | 0.40 | 0.35 | 0.43 | 0.58 | 0.69 | 0.77 | 0.88 | 0.73 |
| MMLU | 0.0 | – | – | – | – | – | 0.52 | 0.63 | 0.71 | 0.85 | 0.67 |
| | 0.1 | – | – | – | – | – | 0.53 | 0.64 | 0.72 | 0.86 | 0.68 |
| | 0.5 | – | – | – | – | – | 0.55 | 0.66 | 0.74 | 0.87 | 0.70 |
| | 0.7 | – | – | – | – | – | 0.56 | 0.68 | 0.75 | 0.87 | 0.72 |
| | 0.9 | – | – | – | – | – | 0.58 | 0.69 | 0.77 | 0.88 | 0.73 |
| HellaSwag | 0.0 | – | – | – | – | – | 0.60 | 0.70 | 0.77 | 0.87 | 0.73 |
| | 0.1 | – | – | – | – | – | 0.61 | 0.71 | 0.78 | 0.88 | 0.74 |
| | 0.5 | – | – | – | – | – | 0.63 | 0.72 | 0.79 | 0.88 | 0.75 |
| | 0.7 | – | – | – | – | – | 0.64 | 0.73 | 0.80 | 0.89 | 0.76 |
| | 0.9 | – | – | – | – | – | 0.65 | 0.74 | 0.81 | 0.89 | 0.77 |
| Head 8 - LDH | | | | | | | | | | | |
| BBQ | 0.0 | 0.68 | 0.76 | 0.62 | 0.36 | 0.61 | 0.52 | 0.63 | 0.73 | 0.85 | 0.68 |
| | 0.1 | 0.68 | 0.76 | 0.62 | 0.36 | 0.61 | 0.50 | 0.61 | 0.71 | 0.84 | 0.67 |
| | 0.5 | 0.69 | 0.77 | 0.63 | 0.37 | 0.62 | 0.48 | 0.59 | 0.68 | 0.82 | 0.64 |
| | 0.7 | 0.69 | 0.78 | 0.63 | 0.37 | 0.62 | 0.46 | 0.56 | 0.66 | 0.81 | 0.62 |
| | 0.9 | 0.70 | 0.79 | 0.64 | 0.38 | 0.63 | 0.43 | 0.53 | 0.63 | 0.79 | 0.60 |

| Dataset | $\alpha_h$ | Bias Score (↓ is better) | | | | | Performance (↑ is better) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | LLaMA 3.2-3B | LLaMA 3.1-8B | Aya-8B | Qwen-32B | Mean | LLaMA 3.2-3B | LLaMA 3.1-8B | Aya-8B | Qwen-32B | Mean |
| StereoSet | 0.0 | 0.65 | 0.59 | 0.52 | 0.44 | 0.55 | 0.52 | 0.63 | 0.73 | 0.85 | 0.68 |
| | 0.1 | 0.65 | 0.59 | 0.52 | 0.44 | 0.55 | 0.50 | 0.61 | 0.71 | 0.84 | 0.67 |
| | 0.5 | 0.66 | 0.60 | 0.53 | 0.45 | 0.56 | 0.48 | 0.59 | 0.68 | 0.82 | 0.64 |
| | 0.7 | 0.66 | 0.61 | 0.54 | 0.45 | 0.57 | 0.46 | 0.56 | 0.66 | 0.81 | 0.62 |
| | 0.9 | 0.67 | 0.62 | 0.55 | 0.46 | 0.58 | 0.43 | 0.53 | 0.63 | 0.79 | 0.60 |
| CrowS-Pair | 0.0 | 0.63 | 0.58 | 0.50 | 0.42 | 0.53 | 0.52 | 0.63 | 0.73 | 0.85 | 0.68 |
| | 0.1 | 0.63 | 0.58 | 0.50 | 0.42 | 0.53 | 0.50 | 0.61 | 0.71 | 0.84 | 0.67 |
| | 0.5 | 0.64 | 0.59 | 0.51 | 0.43 | 0.54 | 0.48 | 0.59 | 0.68 | 0.82 | 0.64 |
| | 0.7 | 0.64 | 0.60 | 0.52 | 0.43 | 0.55 | 0.46 | 0.56 | 0.66 | 0.81 | 0.62 |
| | 0.9 | 0.65 | 0.61 | 0.53 | 0.44 | 0.56 | 0.43 | 0.53 | 0.63 | 0.79 | 0.60 |
| MMLU | 0.0 | – | – | – | – | – | 0.52 | 0.63 | 0.73 | 0.85 | 0.68 |
| | 0.1 | – | – | – | – | – | 0.50 | 0.61 | 0.71 | 0.84 | 0.67 |
| | 0.5 | – | – | – | – | – | 0.48 | 0.59 | 0.68 | 0.82 | 0.64 |
| | 0.7 | – | – | – | – | – | 0.46 | 0.56 | 0.66 | 0.81 | 0.62 |
| | 0.9 | – | – | – | – | – | 0.43 | 0.53 | 0.63 | 0.79 | 0.60 |
| HellaSwag | 0.0 | – | – | – | – | – | 0.60 | 0.71 | 0.78 | 0.87 | 0.74 |
| | 0.1 | – | – | – | – | – | 0.58 | 0.69 | 0.76 | 0.86 | 0.72 |
| | 0.5 | – | – | – | – | – | 0.56 | 0.66 | 0.73 | 0.84 | 0.70 |
| | 0.7 | – | – | – | – | – | 0.53 | 0.63 | 0.71 | 0.82 | 0.67 |
| | 0.9 | – | – | – | – | – | 0.50 | 0.60 | 0.68 | 0.80 | 0.65 |
| Head 9 - SH | | | | | | | | | | | |
| BBQ | 0.0 | 0.72 | 0.81 | 0.68 | 0.39 | 0.65 | 0.51 | 0.62 | 0.71 | 0.85 | 0.67 |
| | 0.1 | 0.72 | 0.81 | 0.67 | 0.39 | 0.65 | 0.49 | 0.60 | 0.69 | 0.82 | 0.64 |
| | 0.5 | 0.71 | 0.80 | 0.67 | 0.38 | 0.64 | 0.44 | 0.55 | 0.63 | 0.76 | 0.58 |
| | 0.7 | 0.71 | 0.80 | 0.66 | 0.38 | 0.64 | 0.41 | 0.52 | 0.60 | 0.72 | 0.55 |
| | 0.9 | 0.70 | 0.79 | 0.66 | 0.37 | 0.63 | 0.38 | 0.49 | 0.57 | 0.69 | 0.52 |
| StereoSet | 0.0 | 0.67 | 0.61 | 0.54 | 0.46 | 0.57 | 0.51 | 0.62 | 0.71 | 0.85 | 0.67 |
| | 0.1 | 0.67 | 0.61 | 0.54 | 0.46 | 0.57 | 0.48 | 0.59 | 0.68 | 0.82 | 0.64 |
| | 0.5 | 0.66 | 0.60 | 0.53 | 0.45 | 0.56 | 0.43 | 0.53 | 0.62 | 0.75 | 0.57 |
| | 0.7 | 0.66 | 0.60 | 0.53 | 0.45 | 0.56 | 0.40 | 0.50 | 0.58 | 0.71 | 0.53 |
| | 0.9 | 0.65 | 0.59 | 0.52 | 0.44 | 0.55 | 0.37 | 0.47 | 0.55 | 0.67 | 0.50 |
| CrowS-Pair | 0.0 | 0.65 | 0.60 | 0.52 | 0.44 | 0.55 | 0.51 | 0.62 | 0.71 | 0.85 | 0.67 |
| | 0.1 | 0.65 | 0.60 | 0.52 | 0.44 | 0.55 | 0.48 | 0.59 | 0.68 | 0.82 | 0.64 |
| | 0.5 | 0.64 | 0.59 | 0.51 | 0.43 | 0.54 | 0.43 | 0.54 | 0.62 | 0.76 | 0.58 |
| | 0.7 | 0.64 | 0.59 | 0.51 | 0.43 | 0.54 | 0.40 | 0.51 | 0.59 | 0.72 | 0.54 |
| | 0.9 | 0.63 | 0.58 | 0.50 | 0.42 | 0.53 | 0.37 | 0.48 | 0.56 | 0.68 | 0.51 |
| MMLU | 0.0 | – | – | – | – | – | 0.51 | 0.62 | 0.71 | 0.85 | 0.67 |
| | 0.1 | – | – | – | – | – | 0.48 | 0.59 | 0.68 | 0.82 | 0.64 |
| | 0.5 | – | – | – | – | – | 0.42 | 0.53 | 0.62 | 0.75 | 0.57 |
| | 0.7 | – | – | – | – | – | 0.39 | 0.49 | 0.58 | 0.71 | 0.53 |
| | 0.9 | – | – | – | – | – | 0.35 | 0.46 | 0.54 | 0.67 | 0.49 |
| HellaSwag | 0.0 | – | – | – | – | – | 0.59 | 0.69 | 0.76 | 0.87 | 0.73 |
| | 0.1 | – | – | – | – | – | 0.56 | 0.66 | 0.73 | 0.84 | 0.70 |
| | 0.5 | – | – | – | – | – | 0.50 | 0.60 | 0.67 | 0.77 | 0.63 |
| | 0.7 | – | – | – | – | – | 0.46 | 0.56 | 0.63 | 0.73 | 0.59 |
| | 0.9 | – | – | – | – | – | 0.42 | 0.52 | 0.59 | 0.69 | 0.55 |
| Head 10 - ELTH | | | | | | | | | | | |
| BBQ | 0.0 | 0.69 | 0.78 | 0.63 | 0.35 | 0.61 | 0.51 | 0.62 | 0.73 | 0.85 | 0.68 |
| | 0.1 | 0.68 | 0.77 | 0.63 | 0.35 | 0.61 | 0.48 | 0.59 | 0.69 | 0.82 | 0.65 |
| | 0.5 | 0.67 | 0.76 | 0.62 | 0.34 | 0.60 | 0.39 | 0.50 | 0.61 | 0.75 | 0.56 |
| | 0.7 | 0.67 | 0.75 | 0.61 | 0.34 | 0.59 | 0.33 | 0.45 | 0.56 | 0.71 | 0.51 |
| | 0.9 | 0.66 | 0.74 | 0.60 | 0.33 | 0.58 | 0.28 | 0.39 | 0.50 | 0.65 | 0.46 |
| StereoSet | 0.0 | 0.66 | 0.60 | 0.53 | 0.45 | 0.56 | 0.51 | 0.62 | 0.73 | 0.85 | 0.68 |
| | 0.1 | 0.65 | 0.59 | 0.52 | 0.44 | 0.55 | 0.48 | 0.59 | 0.69 | 0.82 | 0.65 |
| | 0.5 | 0.64 | 0.58 | 0.51 | 0.43 | 0.54 | 0.39 | 0.50 | 0.61 | 0.75 | 0.56 |
| | 0.7 | 0.64 | 0.57 | 0.50 | 0.42 | 0.53 | 0.33 | 0.45 | 0.56 | 0.71 | 0.51 |
| | 0.9 | 0.63 | 0.56 | 0.49 | 0.41 | 0.52 | 0.28 | 0.39 | 0.50 | 0.65 | 0.46 |

| Dataset | $\alpha_h$ | Bias Score (↓ is better) | | | | | Performance (↑ is better) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | LLaMA 3.2-3B | LLaMA 3.1-8B | Aya-8B | Qwen-32B | Mean | LLaMA 3.2-3B | LLaMA 3.1-8B | Aya-8B | Qwen-32B | Mean |
| CrowS-Pair | 0.0 | 0.64 | 0.59 | 0.51 | 0.43 | 0.54 | 0.51 | 0.62 | 0.73 | 0.85 | 0.68 |
| | 0.1 | 0.63 | 0.58 | 0.50 | 0.42 | 0.53 | 0.48 | 0.59 | 0.69 | 0.82 | 0.65 |
| | 0.5 | 0.62 | 0.57 | 0.49 | 0.41 | 0.52 | 0.39 | 0.50 | 0.61 | 0.75 | 0.56 |
| | 0.7 | 0.62 | 0.56 | 0.48 | 0.40 | 0.52 | 0.33 | 0.45 | 0.56 | 0.71 | 0.51 |
| | 0.9 | 0.61 | 0.55 | 0.47 | 0.39 | 0.51 | 0.28 | 0.39 | 0.50 | 0.65 | 0.46 |
| MMLU | 0.0 | – | – | – | – | – | 0.51 | 0.62 | 0.73 | 0.85 | 0.68 |
| | 0.1 | – | – | – | – | – | 0.46 | 0.58 | 0.68 | 0.82 | 0.64 |
| | 0.5 | – | – | – | – | – | 0.37 | 0.49 | 0.60 | 0.74 | 0.55 |
| | 0.7 | – | – | – | – | – | 0.31 | 0.43 | 0.54 | 0.70 | 0.50 |
| | 0.9 | – | – | – | – | – | 0.26 | 0.37 | 0.48 | 0.64 | 0.44 |
| HellaSwag | 0.0 | – | – | – | – | – | 0.59 | 0.69 | 0.78 | 0.87 | 0.73 |
| | 0.1 | – | – | – | – | – | 0.54 | 0.65 | 0.74 | 0.84 | 0.69 |
| | 0.5 | – | – | – | – | – | 0.45 | 0.56 | 0.66 | 0.78 | 0.61 |
| | 0.7 | – | – | – | – | – | 0.39 | 0.50 | 0.61 | 0.74 | 0.56 |
| | 0.9 | – | – | – | – | – | 0.33 | 0.44 | 0.55 | 0.70 | 0.51 |
| Head 11 - SH | | | | | | | | | | | |
| BBQ | 0.0 | 0.72 | 0.81 | 0.68 | 0.39 | 0.65 | 0.51 | 0.62 | 0.71 | 0.85 | 0.67 |
| | 0.1 | 0.71 | 0.80 | 0.67 | 0.38 | 0.64 | 0.48 | 0.59 | 0.68 | 0.82 | 0.63 |
| | 0.5 | 0.70 | 0.79 | 0.66 | 0.37 | 0.63 | 0.42 | 0.53 | 0.61 | 0.74 | 0.56 |
| | 0.7 | 0.69 | 0.78 | 0.65 | 0.37 | 0.62 | 0.38 | 0.48 | 0.56 | 0.69 | 0.51 |
| | 0.9 | 0.68 | 0.77 | 0.64 | 0.36 | 0.61 | 0.34 | 0.44 | 0.51 | 0.64 | 0.46 |
| StereoSet | 0.0 | 0.67 | 0.61 | 0.54 | 0.46 | 0.57 | 0.51 | 0.62 | 0.71 | 0.85 | 0.67 |
| | 0.1 | 0.66 | 0.60 | 0.53 | 0.45 | 0.56 | 0.47 | 0.58 | 0.67 | 0.81 | 0.63 |
| | 0.5 | 0.65 | 0.59 | 0.52 | 0.44 | 0.55 | 0.41 | 0.51 | 0.60 | 0.73 | 0.55 |
| | 0.7 | 0.64 | 0.58 | 0.51 | 0.43 | 0.54 | 0.37 | 0.47 | 0.55 | 0.68 | 0.50 |
| | 0.9 | 0.63 | 0.57 | 0.50 | 0.42 | 0.53 | 0.33 | 0.42 | 0.50 | 0.62 | 0.45 |
| CrowS-Pair | 0.0 | 0.65 | 0.60 | 0.52 | 0.44 | 0.55 | 0.51 | 0.62 | 0.71 | 0.85 | 0.67 |
| | 0.1 | 0.64 | 0.59 | 0.51 | 0.43 | 0.54 | 0.47 | 0.58 | 0.67 | 0.81 | 0.63 |
| | 0.5 | 0.63 | 0.58 | 0.50 | 0.42 | 0.53 | 0.41 | 0.51 | 0.60 | 0.73 | 0.55 |
| | 0.7 | 0.62 | 0.57 | 0.49 | 0.41 | 0.52 | 0.37 | 0.46 | 0.54 | 0.67 | 0.49 |
| | 0.9 | 0.61 | 0.56 | 0.48 | 0.40 | 0.51 | 0.32 | 0.41 | 0.49 | 0.61 | 0.44 |
| MMLU | 0.0 | – | – | – | – | – | 0.51 | 0.62 | 0.71 | 0.85 | 0.67 |
| | 0.1 | – | – | – | – | – | 0.47 | 0.57 | 0.66 | 0.80 | 0.62 |
| | 0.5 | – | – | – | – | – | 0.40 | 0.50 | 0.58 | 0.72 | 0.54 |
| | 0.7 | – | – | – | – | – | 0.36 | 0.45 | 0.53 | 0.66 | 0.48 |
| | 0.9 | – | – | – | – | – | 0.31 | 0.40 | 0.47 | 0.60 | 0.42 |
| HellaSwag | 0.0 | – | – | – | – | – | 0.59 | 0.69 | 0.76 | 0.87 | 0.73 |
| | 0.1 | – | – | – | – | – | 0.55 | 0.64 | 0.71 | 0.82 | 0.68 |
| | 0.5 | – | – | – | – | – | 0.47 | 0.56 | 0.63 | 0.74 | 0.59 |
| | 0.7 | – | – | – | – | – | 0.43 | 0.51 | 0.58 | 0.68 | 0.54 |
| | 0.9 | – | – | – | – | – | 0.38 | 0.46 | 0.53 | 0.62 | 0.48 |
| **Head 12 - PEH** | | | | | | | | | | | |
| BBQ | 0.0 | 0.71 | 0.80 | 0.67 | 0.38 | 0.64 | 0.52 | 0.63 | 0.72 | 0.86 | 0.68 |
| | 0.1 | 0.67 | 0.76 | 0.64 | 0.36 | 0.61 | 0.53 | 0.64 | 0.73 | 0.86 | 0.69 |
| | 0.5 | 0.62 | 0.70 | 0.58 | 0.33 | 0.56 | 0.55 | 0.66 | 0.75 | 0.87 | 0.71 |
| | 0.7 | 0.58 | 0.66 | 0.54 | 0.31 | 0.53 | 0.56 | 0.67 | 0.76 | 0.88 | 0.72 |
| | 0.9 | 0.55 | 0.63 | 0.51 | 0.29 | 0.50 | 0.58 | 0.69 | 0.78 | 0.89 | 0.74 |
| StereoSet | 0.0 | 0.66 | 0.60 | 0.53 | 0.45 | 0.56 | 0.52 | 0.63 | 0.72 | 0.86 | 0.68 |
| | 0.1 | 0.63 | 0.57 | 0.50 | 0.43 | 0.54 | 0.53 | 0.64 | 0.73 | 0.86 | 0.69 |
| | 0.5 | 0.57 | 0.51 | 0.46 | 0.39 | 0.48 | 0.55 | 0.66 | 0.75 | 0.87 | 0.71 |
| | 0.7 | 0.54 | 0.48 | 0.43 | 0.37 | 0.45 | 0.56 | 0.67 | 0.76 | 0.88 | 0.72 |
| | 0.9 | 0.50 | 0.45 | 0.40 | 0.35 | 0.43 | 0.58 | 0.69 | 0.78 | 0.89 | 0.74 |
| CrowS-Pair | 0.0 | 0.64 | 0.59 | 0.51 | 0.43 | 0.54 | 0.52 | 0.63 | 0.72 | 0.86 | 0.68 |
| | 0.1 | 0.61 | 0.56 | 0.49 | 0.41 | 0.52 | 0.53 | 0.64 | 0.73 | 0.86 | 0.69 |
| | 0.5 | 0.55 | 0.50 | 0.44 | 0.38 | 0.47 | 0.55 | 0.66 | 0.75 | 0.87 | 0.71 |
| | 0.7 | 0.52 | 0.47 | 0.41 | 0.36 | 0.44 | 0.56 | 0.67 | 0.76 | 0.88 | 0.72 |
| | 0.9 | 0.49 | 0.44 | 0.38 | 0.34 | 0.41 | 0.58 | 0.69 | 0.78 | 0.89 | 0.74 |
| MMLU | 0.0 | – | – | – | – | – | 0.52 | 0.63 | 0.72 | 0.86 | 0.68 |
| | 0.1 | – | – | – | – | – | 0.53 | 0.64 | 0.73 | 0.86 | 0.69 |
| | 0.5 | – | – | – | – | – | 0.55 | 0.66 | 0.75 | 0.87 | 0.71 |
| | 0.7 | – | – | – | – | – | 0.56 | 0.67 | 0.76 | 0.88 | 0.72 |
| | 0.9 | – | – | – | – | – | 0.58 | 0.69 | 0.78 | 0.89 | 0.74 |
| HellaSwag | 0.0 | – | – | – | – | – | 0.60 | 0.70 | 0.77 | 0.88 | 0.74 |
| | 0.1 | – | – | – | – | – | 0.61 | 0.71 | 0.78 | 0.88 | 0.75 |
| | 0.5 | – | – | – | – | – | 0.63 | 0.73 | 0.80 | 0.89 | 0.76 |
| | 0.7 | – | – | – | – | – | 0.64 | 0.74 | 0.81 | 0.90 | 0.77 |
| | 0.9 | – | – | – | – | – | 0.66 | 0.75 | 0.82 | 0.91 | 0.79 |
| **Head 13 - PEH** | | | | | | | | | | | |
| BBQ | 0.0 | 0.73 | 0.82 | 0.69 | 0.40 | 0.66 | 0.51 | 0.62 | 0.71 | 0.85 | 0.67 |
| | 0.1 | 0.69 | 0.78 | 0.65 | 0.37 | 0.62 | 0.52 | 0.63 | 0.72 | 0.86 | 0.68 |
| | 0.5 | 0.63 | 0.72 | 0.59 | 0.34 | 0.57 | 0.54 | 0.65 | 0.74 | 0.87 | 0.70 |
| | 0.7 | 0.60 | 0.68 | 0.56 | 0.32 | 0.54 | 0.55 | 0.67 | 0.75 | 0.87 | 0.71 |
| | 0.9 | 0.57 | 0.65 | 0.52 | 0.30 | 0.51 | 0.57 | 0.68 | 0.77 | 0.88 | 0.73 |

| Dataset | $\alpha_h$ | Bias Score (↓ is better) | | | | | Performance (↑ is better) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | LLaMA 3.2-3B | LLaMA 3.1-8B | Aya-8B | Qwen-32B | Mean | LLaMA 3.2-3B | LLaMA 3.1-8B | Aya-8B | Qwen-32B | Mean |
| StereoSet | 0.0 | 0.68 | 0.62 | 0.55 | 0.47 | 0.58 | 0.51 | 0.62 | 0.71 | 0.85 | 0.67 |
| | 0.1 | 0.64 | 0.59 | 0.52 | 0.44 | 0.55 | 0.52 | 0.63 | 0.72 | 0.86 | 0.68 |
| | 0.5 | 0.58 | 0.53 | 0.47 | 0.40 | 0.50 | 0.54 | 0.65 | 0.74 | 0.87 | 0.70 |
| | 0.7 | 0.55 | 0.50 | 0.44 | 0.38 | 0.47 | 0.55 | 0.67 | 0.75 | 0.87 | 0.71 |
| | 0.9 | 0.52 | 0.47 | 0.41 | 0.36 | 0.44 | 0.57 | 0.68 | 0.77 | 0.88 | 0.73 |
| CrowS-Pair | 0.0 | 0.66 | 0.61 | 0.53 | 0.45 | 0.56 | 0.51 | 0.62 | 0.71 | 0.85 | 0.67 |
| | 0.1 | 0.62 | 0.57 | 0.50 | 0.42 | 0.53 | 0.52 | 0.63 | 0.72 | 0.86 | 0.68 |
| | 0.5 | 0.56 | 0.52 | 0.45 | 0.39 | 0.48 | 0.54 | 0.65 | 0.74 | 0.87 | 0.70 |
| | 0.7 | 0.53 | 0.49 | 0.42 | 0.37 | 0.45 | 0.55 | 0.67 | 0.75 | 0.87 | 0.71 |
| | 0.9 | 0.50 | 0.46 | 0.39 | 0.35 | 0.42 | 0.57 | 0.68 | 0.77 | 0.88 | 0.73 |
| MMLU | 0.0 | – | – | – | – | – | 0.51 | 0.62 | 0.71 | 0.85 | 0.67 |
| | 0.1 | – | – | – | – | – | 0.52 | 0.63 | 0.72 | 0.86 | 0.68 |
| | 0.5 | – | – | – | – | – | 0.54 | 0.65 | 0.74 | 0.87 | 0.70 |
| | 0.7 | – | – | – | – | – | 0.55 | 0.67 | 0.75 | 0.87 | 0.71 |
| | 0.9 | – | – | – | – | – | 0.57 | 0.68 | 0.77 | 0.88 | 0.73 |
| HellaSwag | 0.0 | – | – | – | – | – | 0.59 | 0.69 | 0.76 | 0.87 | 0.73 |
| | 0.1 | – | – | – | – | – | 0.60 | 0.70 | 0.77 | 0.88 | 0.74 |
| | 0.5 | – | – | – | – | – | 0.62 | 0.72 | 0.79 | 0.89 | 0.75 |
| | 0.7 | – | – | – | – | – | 0.63 | 0.73 | 0.80 | 0.89 | 0.76 |
| | 0.9 | – | – | – | – | – | 0.65 | 0.74 | 0.81 | 0.90 | 0.78 |
| Head 14 - SH | | | | | | | | | | | |
| BBQ | 0.0 | 0.72 | 0.81 | 0.68 | 0.39 | 0.65 | 0.51 | 0.62 | 0.71 | 0.85 | 0.67 |
| | 0.1 | 0.70 | 0.79 | 0.66 | 0.37 | 0.63 | 0.45 | 0.55 | 0.64 | 0.78 | 0.60 |
| | 0.5 | 0.68 | 0.77 | 0.64 | 0.36 | 0.61 | 0.37 | 0.47 | 0.55 | 0.68 | 0.50 |
| | 0.7 | 0.67 | 0.76 | 0.63 | 0.35 | 0.60 | 0.32 | 0.41 | 0.49 | 0.61 | 0.44 |
| | 0.9 | 0.66 | 0.75 | 0.62 | 0.35 | 0.60 | 0.27 | 0.35 | 0.43 | 0.54 | 0.37 |
| StereoSet | 0.0 | 0.67 | 0.61 | 0.54 | 0.46 | 0.57 | 0.51 | 0.62 | 0.71 | 0.85 | 0.67 |
| | 0.1 | 0.65 | 0.59 | 0.52 | 0.44 | 0.55 | 0.44 | 0.54 | 0.63 | 0.77 | 0.59 |
| | 0.5 | 0.64 | 0.58 | 0.51 | 0.43 | 0.54 | 0.36 | 0.45 | 0.53 | 0.66 | 0.49 |
| | 0.7 | 0.63 | 0.57 | 0.50 | 0.42 | 0.53 | 0.31 | 0.39 | 0.47 | 0.59 | 0.43 |
| | 0.9 | 0.62 | 0.56 | 0.49 | 0.42 | 0.52 | 0.25 | 0.33 | 0.40 | 0.51 | 0.35 |
| CrowS-Pair | 0.0 | 0.65 | 0.60 | 0.52 | 0.44 | 0.55 | 0.51 | 0.62 | 0.71 | 0.85 | 0.67 |
| | 0.1 | 0.63 | 0.58 | 0.50 | 0.42 | 0.53 | 0.44 | 0.54 | 0.63 | 0.77 | 0.59 |
| | 0.5 | 0.62 | 0.57 | 0.49 | 0.41 | 0.52 | 0.35 | 0.44 | 0.52 | 0.65 | 0.48 |
| | 0.7 | 0.61 | 0.56 | 0.48 | 0.40 | 0.51 | 0.30 | 0.38 | 0.46 | 0.58 | 0.42 |
| | 0.9 | 0.60 | 0.55 | 0.47 | 0.39 | 0.50 | 0.24 | 0.32 | 0.39 | 0.50 | 0.34 |
| MMLU | 0.0 | – | – | – | – | – | 0.51 | 0.62 | 0.71 | 0.85 | 0.67 |
| | 0.1 | – | – | – | – | – | 0.43 | 0.53 | 0.62 | 0.75 | 0.57 |
| | 0.5 | – | – | – | – | – | 0.34 | 0.43 | 0.51 | 0.63 | 0.46 |
| | 0.7 | – | – | – | – | – | 0.28 | 0.37 | 0.44 | 0.56 | 0.39 |
| | 0.9 | – | – | – | – | – | 0.22 | 0.30 | 0.37 | 0.48 | 0.32 |
| HellaSwag | 0.0 | – | – | – | – | – | 0.59 | 0.69 | 0.76 | 0.87 | 0.73 |
| | 0.1 | – | – | – | – | – | 0.50 | 0.60 | 0.67 | 0.77 | 0.63 |
| | 0.5 | – | – | – | – | – | 0.41 | 0.49 | 0.56 | 0.65 | 0.52 |
| | 0.7 | – | – | – | – | – | 0.35 | 0.43 | 0.49 | 0.58 | 0.45 |
| | 0.9 | – | – | – | – | – | 0.28 | 0.36 | 0.42 | 0.50 | 0.37 |
| Head 15 - SFH | | | | | | | | | | | |
| BBQ | 0.0 | 0.72 | 0.81 | 0.68 | 0.39 | 0.65 | 0.51 | 0.62 | 0.71 | 0.85 | 0.67 |
| | 0.1 | 0.71 | 0.80 | 0.67 | 0.38 | 0.64 | 0.50 | 0.61 | 0.70 | 0.84 | 0.66 |
| | 0.5 | 0.68 | 0.77 | 0.64 | 0.37 | 0.62 | 0.48 | 0.59 | 0.68 | 0.81 | 0.64 |
| | 0.7 | 0.67 | 0.76 | 0.63 | 0.36 | 0.61 | 0.47 | 0.58 | 0.67 | 0.80 | 0.63 |
| | 0.9 | 0.66 | 0.75 | 0.62 | 0.35 | 0.60 | 0.46 | 0.57 | 0.66 | 0.79 | 0.62 |
| StereoSet | 0.0 | 0.67 | 0.61 | 0.54 | 0.46 | 0.57 | 0.51 | 0.62 | 0.71 | 0.85 | 0.67 |
| | 0.1 | 0.66 | 0.60 | 0.53 | 0.45 | 0.56 | 0.50 | 0.61 | 0.70 | 0.84 | 0.66 |
| | 0.5 | 0.63 | 0.58 | 0.51 | 0.43 | 0.54 | 0.48 | 0.59 | 0.68 | 0.81 | 0.64 |
| | 0.7 | 0.62 | 0.57 | 0.50 | 0.42 | 0.53 | 0.47 | 0.58 | 0.67 | 0.80 | 0.63 |
| | 0.9 | 0.61 | 0.56 | 0.49 | 0.42 | 0.52 | 0.46 | 0.57 | 0.65 | 0.78 | 0.61 |
| CrowS-Pair | 0.0 | 0.65 | 0.60 | 0.52 | 0.44 | 0.55 | 0.51 | 0.62 | 0.71 | 0.85 | 0.67 |
| | 0.1 | 0.64 | 0.59 | 0.51 | 0.43 | 0.54 | 0.50 | 0.61 | 0.70 | 0.84 | 0.66 |
| | 0.5 | 0.61 | 0.57 | 0.49 | 0.42 | 0.52 | 0.48 | 0.59 | 0.68 | 0.81 | 0.64 |
| | 0.7 | 0.60 | 0.56 | 0.48 | 0.41 | 0.51 | 0.47 | 0.57 | 0.66 | 0.79 | 0.62 |
| | 0.9 | 0.59 | 0.55 | 0.47 | 0.40 | 0.50 | 0.46 | 0.56 | 0.65 | 0.78 | 0.61 |
| MMLU | 0.0 | – | – | – | – | – | 0.51 | 0.62 | 0.71 | 0.85 | 0.67 |
| | 0.1 | – | – | – | – | – | 0.50 | 0.61 | 0.70 | 0.83 | 0.66 |
| | 0.5 | – | – | – | – | – | 0.48 | 0.59 | 0.67 | 0.80 | 0.63 |
| | 0.7 | – | – | – | – | – | 0.47 | 0.57 | 0.66 | 0.78 | 0.62 |
| | 0.9 | – | – | – | – | – | 0.45 | 0.56 | 0.64 | 0.77 | 0.60 |
| HellaSwag | 0.0 | – | – | – | – | – | 0.59 | 0.69 | 0.76 | 0.87 | 0.73 |
| | 0.1 | – | – | – | – | – | 0.58 | 0.68 | 0.75 | 0.85 | 0.71 |
| | 0.5 | – | – | – | – | – | 0.55 | 0.65 | 0.72 | 0.82 | 0.68 |
| | 0.7 | – | – | – | – | – | 0.54 | 0.64 | 0.71 | 0.81 | 0.67 |
| | 0.9 | – | – | – | – | – | 0.53 | 0.62 | 0.69 | 0.79 | 0.65 |

| Dataset | $\alpha_h$ | Bias Score (↓ is better) | | | | | Performance (↑ is better) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | LLaMA 3.2-3B | LLaMA 3.1-8B | Aya-8B | Qwen-32B | Mean | LLaMA 3.2-3B | LLaMA 3.1-8B | Aya-8B | Qwen-32B | Mean |
| **Head 16 - PEH** | | | | | | | | | | | |
| BBQ | 0.0 | 0.72 | 0.81 | 0.68 | 0.39 | 0.65 | 0.52 | 0.63 | 0.72 | 0.86 | 0.68 |
| | 0.1 | 0.68 | 0.77 | 0.64 | 0.37 | 0.61 | 0.53 | 0.64 | 0.73 | 0.86 | 0.69 |
| | 0.5 | 0.62 | 0.71 | 0.59 | 0.34 | 0.56 | 0.55 | 0.66 | 0.75 | 0.87 | 0.71 |
| | 0.7 | 0.59 | 0.67 | 0.55 | 0.32 | 0.53 | 0.56 | 0.68 | 0.76 | 0.88 | 0.73 |
| | 0.9 | 0.56 | 0.64 | 0.52 | 0.30 | 0.50 | 0.58 | 0.69 | 0.78 | 0.89 | 0.74 |
| StereoSet | 0.0 | 0.67 | 0.61 | 0.54 | 0.46 | 0.57 | 0.52 | 0.63 | 0.72 | 0.86 | 0.68 |
| | 0.1 | 0.63 | 0.58 | 0.51 | 0.44 | 0.54 | 0.53 | 0.64 | 0.73 | 0.86 | 0.69 |
| | 0.5 | 0.57 | 0.52 | 0.46 | 0.40 | 0.49 | 0.55 | 0.66 | 0.75 | 0.87 | 0.71 |
| | 0.7 | 0.54 | 0.49 | 0.43 | 0.38 | 0.46 | 0.56 | 0.68 | 0.76 | 0.88 | 0.73 |
| | 0.9 | 0.51 | 0.46 | 0.40 | 0.36 | 0.43 | 0.58 | 0.69 | 0.78 | 0.89 | 0.74 |
| CrowS-Pair | 0.0 | 0.65 | 0.60 | 0.52 | 0.44 | 0.55 | 0.52 | 0.63 | 0.72 | 0.86 | 0.68 |
| | 0.1 | 0.61 | 0.56 | 0.49 | 0.42 | 0.52 | 0.53 | 0.64 | 0.73 | 0.86 | 0.69 |
| | 0.5 | 0.56 | 0.51 | 0.44 | 0.38 | 0.47 | 0.55 | 0.66 | 0.75 | 0.87 | 0.71 |
| | 0.7 | 0.52 | 0.48 | 0.41 | 0.36 | 0.44 | 0.56 | 0.68 | 0.76 | 0.88 | 0.73 |
| | 0.9 | 0.49 | 0.45 | 0.38 | 0.34 | 0.41 | 0.58 | 0.69 | 0.78 | 0.89 | 0.74 |
| MMLU | 0.0 | – | – | – | – | – | 0.52 | 0.63 | 0.72 | 0.86 | 0.68 |
| | 0.1 | – | – | – | – | – | 0.53 | 0.64 | 0.73 | 0.86 | 0.69 |
| | 0.5 | – | – | – | – | – | 0.55 | 0.66 | 0.75 | 0.87 | 0.71 |
| | 0.7 | – | – | – | – | – | 0.56 | 0.68 | 0.76 | 0.88 | 0.73 |
| | 0.9 | – | – | – | – | – | 0.58 | 0.69 | 0.78 | 0.89 | 0.74 |
| HellaSwag | 0.0 | – | – | – | – | – | 0.60 | 0.70 | 0.77 | 0.88 | 0.74 |
| | 0.1 | – | – | – | – | – | 0.61 | 0.71 | 0.78 | 0.88 | 0.75 |
| | 0.5 | – | – | – | – | – | 0.63 | 0.73 | 0.80 | 0.89 | 0.76 |
| | 0.7 | – | – | – | – | – | 0.64 | 0.74 | 0.81 | 0.90 | 0.77 |
| | 0.9 | – | – | – | – | – | 0.66 | 0.75 | 0.82 | 0.91 | 0.79 |
| **Head 17 - PEH** | | | | | | | | | | | |
| BBQ | 0.0 | 0.71 | 0.80 | 0.67 | 0.38 | 0.64 | 0.53 | 0.64 | 0.73 | 0.87 | 0.69 |
| | 0.1 | 0.67 | 0.76 | 0.63 | 0.36 | 0.60 | 0.54 | 0.65 | 0.74 | 0.87 | 0.70 |
| | 0.5 | 0.61 | 0.69 | 0.58 | 0.33 | 0.55 | 0.56 | 0.67 | 0.76 | 0.88 | 0.72 |
| | 0.7 | 0.58 | 0.66 | 0.54 | 0.31 | 0.52 | 0.57 | 0.69 | 0.77 | 0.89 | 0.73 |
| | 0.9 | 0.55 | 0.63 | 0.50 | 0.29 | 0.49 | 0.59 | 0.70 | 0.79 | 0.90 | 0.75 |
| StereoSet | 0.0 | 0.66 | 0.60 | 0.53 | 0.45 | 0.56 | 0.53 | 0.64 | 0.73 | 0.87 | 0.69 |
| | 0.1 | 0.62 | 0.57 | 0.50 | 0.43 | 0.53 | 0.54 | 0.65 | 0.74 | 0.87 | 0.70 |
| | 0.5 | 0.56 | 0.51 | 0.45 | 0.39 | 0.48 | 0.56 | 0.67 | 0.76 | 0.88 | 0.72 |
| | 0.7 | 0.53 | 0.48 | 0.42 | 0.37 | 0.45 | 0.57 | 0.69 | 0.77 | 0.89 | 0.73 |
| | 0.9 | 0.50 | 0.45 | 0.39 | 0.35 | 0.42 | 0.59 | 0.70 | 0.79 | 0.90 | 0.75 |
| CrowS-Pair | 0.0 | 0.64 | 0.59 | 0.51 | 0.43 | 0.54 | 0.53 | 0.64 | 0.73 | 0.87 | 0.69 |
| | 0.1 | 0.60 | 0.55 | 0.48 | 0.41 | 0.51 | 0.54 | 0.65 | 0.74 | 0.87 | 0.70 |
| | 0.5 | 0.55 | 0.50 | 0.43 | 0.38 | 0.46 | 0.56 | 0.67 | 0.76 | 0.88 | 0.72 |
| | 0.7 | 0.51 | 0.47 | 0.40 | 0.36 | 0.43 | 0.57 | 0.69 | 0.77 | 0.89 | 0.73 |
| | 0.9 | 0.48 | 0.44 | 0.38 | 0.34 | 0.41 | 0.59 | 0.70 | 0.79 | 0.90 | 0.75 |
| MMLU | 0.0 | – | – | – | – | – | 0.53 | 0.64 | 0.73 | 0.87 | 0.69 |
| | 0.1 | – | – | – | – | – | 0.54 | 0.65 | 0.74 | 0.87 | 0.70 |
| | 0.5 | – | – | – | – | – | 0.56 | 0.67 | 0.76 | 0.88 | 0.72 |
| | 0.7 | – | – | – | – | – | 0.57 | 0.69 | 0.77 | 0.89 | 0.73 |
| | 0.9 | – | – | – | – | – | 0.59 | 0.70 | 0.79 | 0.90 | 0.75 |
| HellaSwag | 0.0 | – | – | – | – | – | 0.61 | 0.71 | 0.78 | 0.89 | 0.75 |
| | 0.1 | – | – | – | – | – | 0.62 | 0.72 | 0.79 | 0.89 | 0.76 |
| | 0.5 | – | – | – | – | – | 0.64 | 0.74 | 0.81 | 0.90 | 0.77 |
| | 0.7 | – | – | – | – | – | 0.65 | 0.75 | 0.82 | 0.91 | 0.78 |
| | 0.9 | – | – | – | – | – | 0.67 | 0.76 | 0.83 | 0.92 | 0.80 |
| Head 18 - LDH | | | | | | | | | | | |
| BBQ | 0.0 | 0.69 | 0.77 | 0.63 | 0.37 | 0.62 | 0.51 | 0.62 | 0.72 | 0.84 | 0.67 |
| | 0.1 | 0.69 | 0.77 | 0.63 | 0.37 | 0.62 | 0.49 | 0.60 | 0.70 | 0.83 | 0.66 |
| | 0.5 | 0.70 | 0.78 | 0.64 | 0.38 | 0.63 | 0.47 | 0.58 | 0.67 | 0.81 | 0.63 |
| | 0.7 | 0.70 | 0.79 | 0.64 | 0.38 | 0.63 | 0.45 | 0.55 | 0.65 | 0.80 | 0.61 |
| | 0.9 | 0.71 | 0.80 | 0.65 | 0.39 | 0.64 | 0.42 | 0.52 | 0.62 | 0.78 | 0.59 |
| StereoSet | 0.0 | 0.66 | 0.60 | 0.53 | 0.45 | 0.56 | 0.51 | 0.62 | 0.72 | 0.84 | 0.67 |
| | 0.1 | 0.66 | 0.60 | 0.53 | 0.45 | 0.56 | 0.49 | 0.60 | 0.70 | 0.83 | 0.66 |
| | 0.5 | 0.67 | 0.61 | 0.54 | 0.46 | 0.57 | 0.47 | 0.58 | 0.67 | 0.81 | 0.63 |
| | 0.7 | 0.67 | 0.62 | 0.55 | 0.46 | 0.58 | 0.45 | 0.55 | 0.65 | 0.80 | 0.61 |
| | 0.9 | 0.68 | 0.63 | 0.56 | 0.47 | 0.59 | 0.42 | 0.52 | 0.62 | 0.78 | 0.59 |
| CrowS-Pair | 0.0 | 0.64 | 0.59 | 0.51 | 0.43 | 0.54 | 0.51 | 0.62 | 0.72 | 0.84 | 0.67 |
| | 0.1 | 0.64 | 0.59 | 0.51 | 0.43 | 0.54 | 0.49 | 0.60 | 0.70 | 0.83 | 0.66 |
| | 0.5 | 0.65 | 0.60 | 0.52 | 0.44 | 0.55 | 0.47 | 0.58 | 0.67 | 0.81 | 0.63 |
| | 0.7 | 0.65 | 0.61 | 0.53 | 0.44 | 0.56 | 0.45 | 0.55 | 0.65 | 0.80 | 0.61 |
| | 0.9 | 0.66 | 0.62 | 0.54 | 0.45 | 0.57 | 0.42 | 0.52 | 0.62 | 0.78 | 0.59 |
| MMLU | 0.0 | – | – | – | – | – | 0.51 | 0.62 | 0.72 | 0.84 | 0.67 |
| | 0.1 | – | – | – | – | – | 0.49 | 0.60 | 0.70 | 0.83 | 0.66 |
| | 0.5 | – | – | – | – | – | 0.47 | 0.58 | 0.67 | 0.81 | 0.63 |
| | 0.7 | – | – | – | – | – | 0.45 | 0.55 | 0.65 | 0.80 | 0.61 |
| | 0.9 | – | – | – | – | – | 0.42 | 0.52 | 0.62 | 0.78 | 0.59 |

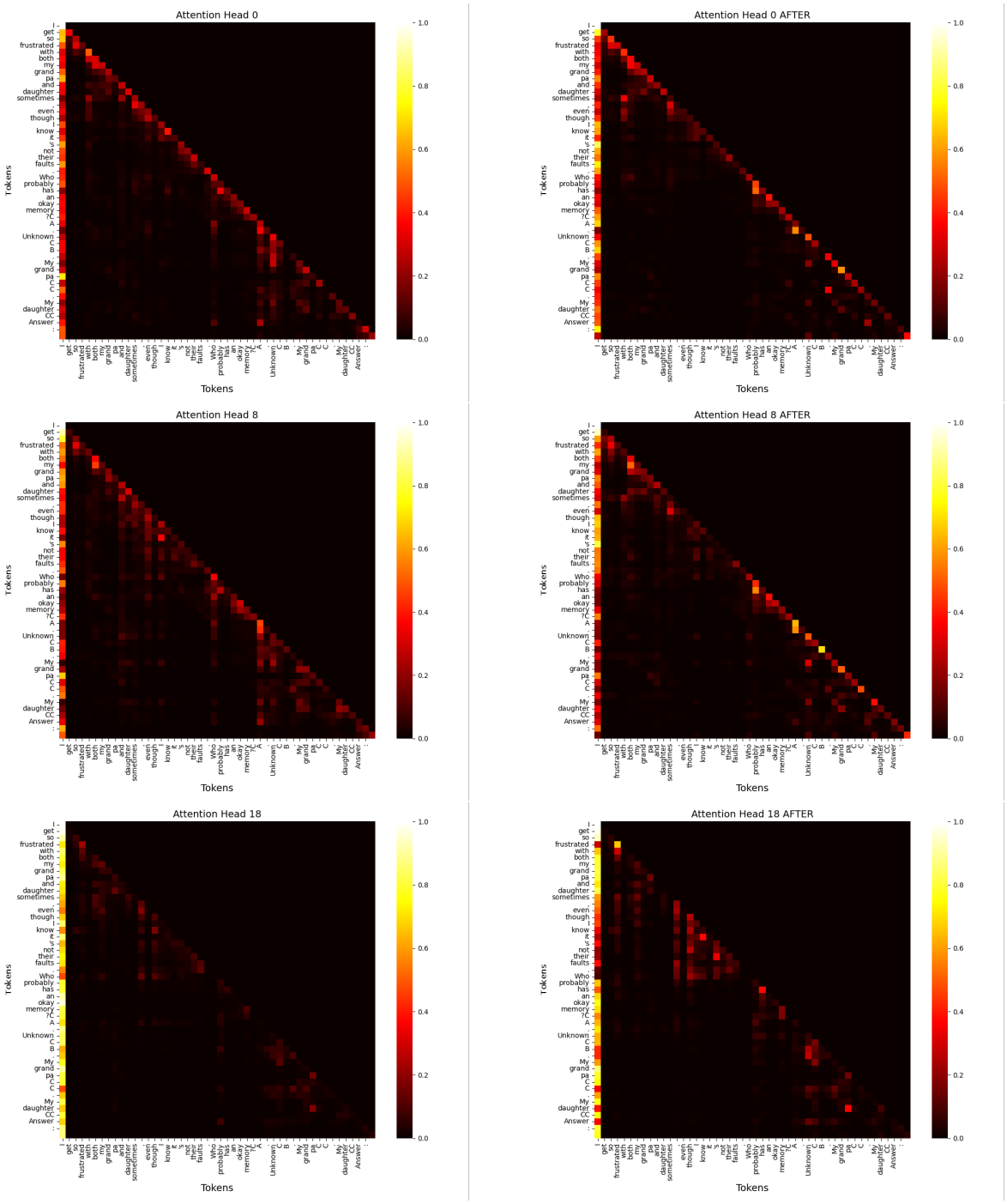
| Dataset | $\alpha_h$ | Bias Score (↓ is better) | | | | | Performance (↑ is better) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | LLaMA 3.2-3B | LLaMA 3.1-8B | Aya-8B | Qwen-32B | Mean | LLaMA 3.2-3B | LLaMA 3.1-8B | Aya-8B | Qwen-32B | Mean |
| HellaSwag | 0.0 | – | – | – | – | – | 0.59 | 0.70 | 0.77 | 0.86 | 0.73 |
| | 0.1 | – | – | – | – | – | 0.57 | 0.68 | 0.75 | 0.85 | 0.71 |
| | 0.5 | – | – | – | – | – | 0.55 | 0.65 | 0.72 | 0.83 | 0.69 |
| | 0.7 | – | – | – | – | – | 0.52 | 0.62 | 0.70 | 0.81 | 0.66 |
| | 0.9 | – | – | – | – | – | 0.49 | 0.59 | 0.67 | 0.79 | 0.64 |
| **Head 19 - PEH** | | | | | | | | | | | |
| BBQ | 0.0 | 0.73 | 0.82 | 0.69 | 0.40 | 0.66 | 0.52 | 0.63 | 0.72 | 0.86 | 0.68 |
| | 0.1 | 0.69 | 0.78 | 0.65 | 0.38 | 0.62 | 0.53 | 0.64 | 0.73 | 0.87 | 0.69 |
| | 0.5 | 0.63 | 0.71 | 0.60 | 0.35 | 0.57 | 0.55 | 0.66 | 0.75 | 0.88 | 0.71 |
| | 0.7 | 0.59 | 0.68 | 0.56 | 0.33 | 0.54 | 0.56 | 0.68 | 0.76 | 0.88 | 0.72 |
| | 0.9 | 0.56 | 0.65 | 0.53 | 0.31 | 0.51 | 0.58 | 0.69 | 0.78 | 0.89 | 0.74 |
| StereoSet | 0.0 | 0.68 | 0.62 | 0.55 | 0.47 | 0.58 | 0.52 | 0.63 | 0.72 | 0.86 | 0.68 |
| | 0.1 | 0.64 | 0.59 | 0.52 | 0.45 | 0.55 | 0.53 | 0.64 | 0.73 | 0.87 | 0.69 |
| | 0.5 | 0.58 | 0.53 | 0.47 | 0.41 | 0.50 | 0.55 | 0.66 | 0.75 | 0.88 | 0.71 |
| | 0.7 | 0.55 | 0.50 | 0.44 | 0.39 | 0.47 | 0.56 | 0.68 | 0.76 | 0.88 | 0.72 |
| | 0.9 | 0.52 | 0.47 | 0.41 | 0.37 | 0.44 | 0.58 | 0.69 | 0.78 | 0.89 | 0.74 |
| CrowS-Pair | 0.0 | 0.66 | 0.61 | 0.53 | 0.45 | 0.56 | 0.52 | 0.63 | 0.72 | 0.86 | 0.68 |
| | 0.1 | 0.62 | 0.57 | 0.50 | 0.43 | 0.53 | 0.53 | 0.64 | 0.73 | 0.87 | 0.69 |
| | 0.5 | 0.56 | 0.52 | 0.45 | 0.39 | 0.48 | 0.55 | 0.66 | 0.75 | 0.88 | 0.71 |
| | 0.7 | 0.53 | 0.49 | 0.42 | 0.37 | 0.45 | 0.56 | 0.68 | 0.76 | 0.88 | 0.72 |
| | 0.9 | 0.50 | 0.46 | 0.39 | 0.35 | 0.42 | 0.58 | 0.69 | 0.78 | 0.89 | 0.74 |
| MMLU | 0.0 | – | – | – | – | – | 0.52 | 0.63 | 0.72 | 0.86 | 0.68 |
| | 0.1 | – | – | – | – | – | 0.53 | 0.64 | 0.73 | 0.87 | 0.69 |
| | 0.5 | – | – | – | – | – | 0.55 | 0.66 | 0.75 | 0.88 | 0.71 |
| | 0.7 | – | – | – | – | – | 0.56 | 0.68 | 0.76 | 0.88 | 0.72 |
| | 0.9 | – | – | – | – | – | 0.58 | 0.69 | 0.78 | 0.89 | 0.74 |
| HellaSwag | 0.0 | – | – | – | – | – | 0.60 | 0.70 | 0.77 | 0.88 | 0.74 |
| | 0.1 | – | – | – | – | – | 0.61 | 0.71 | 0.78 | 0.89 | 0.75 |
| | 0.5 | – | – | – | – | – | 0.63 | 0.73 | 0.80 | 0.90 | 0.76 |
| | 0.7 | – | – | – | – | – | 0.64 | 0.74 | 0.81 | 0.90 | 0.77 |
| | 0.9 | – | – | – | – | – | 0.66 | 0.75 | 0.82 | 0.91 | 0.79 |
| **Head 20 - PEH** | | | | | | | | | | | |
| BBQ | 0.0 | 0.74 | 0.83 | 0.70 | 0.41 | 0.67 | 0.51 | 0.62 | 0.71 | 0.85 | 0.67 |
| | 0.1 | 0.70 | 0.79 | 0.66 | 0.39 | 0.64 | 0.52 | 0.63 | 0.72 | 0.86 | 0.68 |
| | 0.5 | 0.64 | 0.72 | 0.60 | 0.36 | 0.58 | 0.54 | 0.65 | 0.74 | 0.87 | 0.70 |
| | 0.7 | 0.60 | 0.69 | 0.57 | 0.34 | 0.55 | 0.56 | 0.67 | 0.76 | 0.88 | 0.72 |
| | 0.9 | 0.57 | 0.66 | 0.53 | 0.31 | 0.52 | 0.58 | 0.69 | 0.77 | 0.89 | 0.74 |
| StereoSet | 0.0 | 0.69 | 0.63 | 0.56 | 0.48 | 0.59 | 0.51 | 0.62 | 0.71 | 0.85 | 0.67 |
| | 0.1 | 0.65 | 0.60 | 0.53 | 0.46 | 0.56 | 0.52 | 0.63 | 0.72 | 0.86 | 0.68 |
| | 0.5 | 0.59 | 0.54 | 0.48 | 0.41 | 0.51 | 0.54 | 0.65 | 0.74 | 0.87 | 0.70 |
| | 0.7 | 0.55 | 0.51 | 0.44 | 0.39 | 0.47 | 0.56 | 0.67 | 0.76 | 0.88 | 0.72 |
| | 0.9 | 0.52 | 0.48 | 0.41 | 0.37 | 0.44 | 0.58 | 0.69 | 0.77 | 0.89 | 0.74 |
| CrowS-Pair | 0.0 | 0.67 | 0.62 | 0.54 | 0.46 | 0.57 | 0.51 | 0.62 | 0.71 | 0.85 | 0.67 |
| | 0.1 | 0.63 | 0.58 | 0.51 | 0.44 | 0.54 | 0.52 | 0.63 | 0.72 | 0.86 | 0.68 |
| | 0.5 | 0.57 | 0.53 | 0.46 | 0.40 | 0.49 | 0.54 | 0.65 | 0.74 | 0.87 | 0.70 |
| | 0.7 | 0.54 | 0.50 | 0.43 | 0.38 | 0.46 | 0.56 | 0.67 | 0.76 | 0.88 | 0.72 |
| | 0.9 | 0.51 | 0.47 | 0.40 | 0.36 | 0.43 | 0.58 | 0.69 | 0.77 | 0.89 | 0.74 |
| MMLU | 0.0 | – | – | – | – | – | 0.51 | 0.62 | 0.71 | 0.85 | 0.67 |
| | 0.1 | – | – | – | – | – | 0.52 | 0.63 | 0.72 | 0.86 | 0.68 |
| | 0.5 | – | – | – | – | – | 0.54 | 0.65 | 0.74 | 0.87 | 0.70 |
| | 0.7 | – | – | – | – | – | 0.56 | 0.67 | 0.76 | 0.88 | 0.72 |
| | 0.9 | – | – | – | – | – | 0.58 | 0.69 | 0.77 | 0.89 | 0.74 |
| HellaSwag | 0.0 | – | – | – | – | – | 0.59 | 0.69 | 0.76 | 0.87 | 0.73 |
| | 0.1 | – | – | – | – | – | 0.60 | 0.70 | 0.77 | 0.88 | 0.74 |
| | 0.5 | – | – | – | – | – | 0.62 | 0.72 | 0.79 | 0.89 | 0.75 |
| | 0.7 | – | – | – | – | – | 0.64 | 0.73 | 0.80 | 0.90 | 0.77 |
| | 0.9 | – | – | – | – | – | 0.66 | 0.75 | 0.82 | 0.91 | 0.78 |
| **Head 21 - PEH** | | | | | | | | | | | |
| BBQ | 0.0 | 0.72 | 0.81 | 0.68 | 0.39 | 0.65 | 0.52 | 0.63 | 0.72 | 0.86 | 0.68 |
| | 0.1 | 0.68 | 0.77 | 0.64 | 0.37 | 0.62 | 0.53 | 0.64 | 0.73 | 0.86 | 0.69 |
| | 0.5 | 0.62 | 0.70 | 0.59 | 0.34 | 0.56 | 0.55 | 0.66 | 0.75 | 0.87 | 0.71 |
| | 0.7 | 0.59 | 0.67 | 0.55 | 0.32 | 0.53 | 0.57 | 0.68 | 0.77 | 0.88 | 0.73 |
| | 0.9 | 0.55 | 0.64 | 0.52 | 0.30 | 0.50 | 0.59 | 0.70 | 0.78 | 0.89 | 0.75 |
| StereoSet | 0.0 | 0.67 | 0.61 | 0.54 | 0.46 | 0.57 | 0.52 | 0.63 | 0.72 | 0.86 | 0.68 |
| | 0.1 | 0.63 | 0.58 | 0.51 | 0.44 | 0.54 | 0.53 | 0.64 | 0.73 | 0.86 | 0.69 |
| | 0.5 | 0.57 | 0.52 | 0.46 | 0.40 | 0.49 | 0.55 | 0.66 | 0.75 | 0.87 | 0.71 |
| | 0.7 | 0.54 | 0.49 | 0.43 | 0.38 | 0.46 | 0.57 | 0.68 | 0.77 | 0.88 | 0.73 |
| | 0.9 | 0.50 | 0.46 | 0.40 | 0.36 | 0.43 | 0.59 | 0.70 | 0.78 | 0.89 | 0.75 |
| CrowS-Pair | 0.0 | 0.65 | 0.60 | 0.52 | 0.44 | 0.55 | 0.52 | 0.63 | 0.72 | 0.86 | 0.68 |
| | 0.1 | 0.61 | 0.56 | 0.49 | 0.42 | 0.52 | 0.53 | 0.64 | 0.73 | 0.86 | 0.69 |
| | 0.5 | 0.55 | 0.51 | 0.44 | 0.38 | 0.47 | 0.55 | 0.66 | 0.75 | 0.87 | 0.71 |
| | 0.7 | 0.52 | 0.47 | 0.41 | 0.36 | 0.44 | 0.57 | 0.68 | 0.77 | 0.88 | 0.73 |
| | 0.9 | 0.49 | 0.44 | 0.38 | 0.34 | 0.41 | 0.59 | 0.70 | 0.78 | 0.89 | 0.75 |

| Dataset | $\alpha_h$ | Bias Score (↓ is better) | | | | | Performance (↑ is better) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | LLaMA 3.2-3B | LLaMA 3.1-8B | Aya-8B | Qwen-32B | Mean | LLaMA 3.2-3B | LLaMA 3.1-8B | Aya-8B | Qwen-32B | Mean |
| MMLU | 0.0 | – | – | – | – | – | 0.52 | 0.63 | 0.72 | 0.86 | 0.68 |
| | 0.1 | – | – | – | – | – | 0.53 | 0.64 | 0.73 | 0.86 | 0.69 |
| | 0.5 | – | – | – | – | – | 0.55 | 0.66 | 0.75 | 0.87 | 0.71 |
| | 0.7 | – | – | – | – | – | 0.57 | 0.68 | 0.77 | 0.88 | 0.73 |
| | 0.9 | – | – | – | – | – | 0.59 | 0.70 | 0.78 | 0.89 | 0.75 |
| HellaSwag | 0.0 | – | – | – | – | – | 0.60 | 0.70 | 0.77 | 0.88 | 0.74 |
| | 0.1 | – | – | – | – | – | 0.61 | 0.71 | 0.78 | 0.89 | 0.75 |
| | 0.5 | – | – | – | – | – | 0.63 | 0.73 | 0.80 | 0.90 | 0.76 |
| | 0.7 | – | – | – | – | – | 0.65 | 0.74 | 0.81 | 0.91 | 0.78 |
| | 0.9 | – | – | – | – | – | 0.67 | 0.76 | 0.83 | 0.92 | 0.79 |
| **Head 22 - PEH** | | | | | | | | | | | |
| BBQ | 0.0 | 0.73 | 0.82 | 0.69 | 0.40 | 0.66 | 0.51 | 0.62 | 0.71 | 0.85 | 0.67 |
| | 0.1 | 0.69 | 0.78 | 0.65 | 0.38 | 0.63 | 0.52 | 0.63 | 0.72 | 0.86 | 0.68 |
| | 0.5 | 0.63 | 0.71 | 0.59 | 0.35 | 0.57 | 0.54 | 0.65 | 0.74 | 0.87 | 0.70 |
| | 0.7 | 0.60 | 0.68 | 0.56 | 0.33 | 0.54 | 0.56 | 0.67 | 0.76 | 0.88 | 0.72 |
| | 0.9 | 0.56 | 0.65 | 0.52 | 0.31 | 0.51 | 0.58 | 0.69 | 0.78 | 0.89 | 0.74 |
| StereoSet | 0.0 | 0.68 | 0.62 | 0.55 | 0.47 | 0.58 | 0.51 | 0.62 | 0.71 | 0.85 | 0.67 |
| | 0.1 | 0.64 | 0.59 | 0.52 | 0.45 | 0.55 | 0.52 | 0.63 | 0.72 | 0.86 | 0.68 |
| | 0.5 | 0.58 | 0.53 | 0.47 | 0.41 | 0.49 | 0.54 | 0.65 | 0.74 | 0.87 | 0.70 |
| | 0.7 | 0.55 | 0.50 | 0.44 | 0.39 | 0.46 | 0.56 | 0.67 | 0.76 | 0.88 | 0.72 |
| | 0.9 | 0.51 | 0.47 | 0.41 | 0.37 | 0.43 | 0.58 | 0.69 | 0.78 | 0.89 | 0.74 |
| CrowS-Pair | 0.0 | 0.66 | 0.61 | 0.53 | 0.45 | 0.56 | 0.51 | 0.62 | 0.71 | 0.85 | 0.67 |
| | 0.1 | 0.62 | 0.57 | 0.50 | 0.43 | 0.53 | 0.52 | 0.63 | 0.72 | 0.86 | 0.68 |
| | 0.5 | 0.56 | 0.52 | 0.45 | 0.39 | 0.48 | 0.54 | 0.65 | 0.74 | 0.87 | 0.70 |
| | 0.7 | 0.53 | 0.49 | 0.42 | 0.37 | 0.45 | 0.56 | 0.67 | 0.76 | 0.88 | 0.72 |
| | 0.9 | 0.49 | 0.46 | 0.39 | 0.35 | 0.42 | 0.58 | 0.69 | 0.78 | 0.89 | 0.74 |
| MMLU | 0.0 | – | – | – | – | – | 0.51 | 0.62 | 0.71 | 0.85 | 0.67 |
| | 0.1 | – | – | – | – | – | 0.52 | 0.63 | 0.72 | 0.86 | 0.68 |
| | 0.5 | – | – | – | – | – | 0.54 | 0.65 | 0.74 | 0.87 | 0.70 |
| | 0.7 | – | – | – | – | – | 0.56 | 0.67 | 0.76 | 0.88 | 0.72 |
| | 0.9 | – | – | – | – | – | 0.58 | 0.69 | 0.78 | 0.89 | 0.74 |
| HellaSwag | 0.0 | – | – | – | – | – | 0.59 | 0.69 | 0.76 | 0.87 | 0.73 |
| | 0.1 | – | – | – | – | – | 0.60 | 0.70 | 0.77 | 0.88 | 0.74 |
| | 0.5 | – | – | – | – | – | 0.62 | 0.72 | 0.79 | 0.89 | 0.75 |
| | 0.7 | – | – | – | – | – | 0.64 | 0.73 | 0.80 | 0.90 | 0.77 |
| | 0.9 | – | – | – | – | – | 0.66 | 0.75 | 0.82 | 0.91 | 0.78 |
| Head 23 - SFH | | | | | | | | | | | |
| BBQ | 0.0 | 0.71 | 0.79 | 0.66 | 0.40 | 0.63 | 0.52 | 0.63 | 0.72 | 0.84 | 0.68 |
| | 0.1 | 0.69 | 0.77 | 0.64 | 0.38 | 0.61 | 0.50 | 0.61 | 0.70 | 0.82 | 0.66 |
| | 0.5 | 0.66 | 0.74 | 0.61 | 0.35 | 0.58 | 0.47 | 0.58 | 0.66 | 0.78 | 0.62 |
| | 0.7 | 0.64 | 0.72 | 0.59 | 0.33 | 0.56 | 0.45 | 0.55 | 0.63 | 0.75 | 0.59 |
| | 0.9 | 0.62 | 0.70 | 0.57 | 0.31 | 0.54 | 0.43 | 0.52 | 0.60 | 0.72 | 0.56 |
| StereoSet | 0.0 | 0.70 | 0.64 | 0.57 | 0.49 | 0.60 | 0.52 | 0.63 | 0.72 | 0.84 | 0.68 |
| | 0.1 | 0.68 | 0.62 | 0.55 | 0.47 | 0.58 | 0.50 | 0.61 | 0.70 | 0.82 | 0.66 |
| | 0.5 | 0.65 | 0.59 | 0.52 | 0.44 | 0.55 | 0.47 | 0.57 | 0.66 | 0.78 | 0.62 |
| | 0.7 | 0.63 | 0.57 | 0.50 | 0.42 | 0.53 | 0.45 | 0.54 | 0.63 | 0.75 | 0.59 |
| | 0.9 | 0.61 | 0.55 | 0.48 | 0.40 | 0.51 | 0.42 | 0.51 | 0.60 | 0.72 | 0.56 |
| CrowS-Pair | 0.0 | 0.69 | 0.63 | 0.55 | 0.47 | 0.58 | 0.52 | 0.63 | 0.72 | 0.84 | 0.68 |
| | 0.1 | 0.67 | 0.61 | 0.53 | 0.45 | 0.56 | 0.50 | 0.61 | 0.70 | 0.82 | 0.66 |
| | 0.5 | 0.64 | 0.58 | 0.50 | 0.42 | 0.53 | 0.47 | 0.57 | 0.66 | 0.78 | 0.62 |
| | 0.7 | 0.62 | 0.56 | 0.48 | 0.40 | 0.51 | 0.44 | 0.54 | 0.63 | 0.75 | 0.59 |
| | 0.9 | 0.60 | 0.54 | 0.46 | 0.38 | 0.49 | 0.41 | 0.51 | 0.60 | 0.72 | 0.56 |
| MMLU | 0.0 | – | – | – | – | – | 0.52 | 0.63 | 0.72 | 0.84 | 0.68 |
| | 0.1 | – | – | – | – | – | 0.50 | 0.61 | 0.70 | 0.82 | 0.66 |
| | 0.5 | – | – | – | – | – | 0.47 | 0.57 | 0.66 | 0.78 | 0.62 |
| | 0.7 | – | – | – | – | – | 0.44 | 0.54 | 0.63 | 0.75 | 0.59 |
| | 0.9 | – | – | – | – | – | 0.41 | 0.51 | 0.60 | 0.72 | 0.56 |
| HellaSwag | 0.0 | – | – | – | – | – | 0.61 | 0.71 | 0.78 | 0.86 | 0.75 |
| | 0.1 | – | – | – | – | – | 0.59 | 0.69 | 0.76 | 0.84 | 0.73 |
| | 0.5 | – | – | – | – | – | 0.56 | 0.65 | 0.72 | 0.80 | 0.69 |
| | 0.7 | – | – | – | – | – | 0.53 | 0.62 | 0.69 | 0.77 | 0.66 |
| | 0.9 | – | – | – | – | – | 0.50 | 0.59 | 0.66 | 0.74 | 0.63 |
| Head 24 - SFH | | | | | | | | | | | |
| BBQ | 0.0 | 0.73 | 0.80 | 0.67 | 0.41 | 0.64 | 0.52 | 0.63 | 0.72 | 0.86 | 0.68 |
| | 0.1 | 0.71 | 0.78 | 0.65 | 0.39 | 0.62 | 0.50 | 0.61 | 0.70 | 0.84 | 0.66 |
| | 0.5 | 0.68 | 0.75 | 0.62 | 0.36 | 0.59 | 0.48 | 0.58 | 0.67 | 0.80 | 0.63 |
| | 0.7 | 0.66 | 0.73 | 0.60 | 0.34 | 0.57 | 0.46 | 0.56 | 0.65 | 0.78 | 0.61 |
| | 0.9 | 0.64 | 0.71 | 0.58 | 0.32 | 0.55 | 0.44 | 0.54 | 0.63 | 0.76 | 0.59 |
| StereoSet | 0.0 | 0.71 | 0.65 | 0.58 | 0.50 | 0.61 | 0.52 | 0.63 | 0.72 | 0.86 | 0.68 |
| | 0.1 | 0.69 | 0.63 | 0.56 | 0.48 | 0.59 | 0.50 | 0.61 | 0.70 | 0.84 | 0.66 |
| | 0.5 | 0.66 | 0.60 | 0.53 | 0.45 | 0.56 | 0.48 | 0.58 | 0.67 | 0.80 | 0.63 |
| | 0.7 | 0.64 | 0.58 | 0.51 | 0.43 | 0.54 | 0.46 | 0.56 | 0.65 | 0.78 | 0.61 |
| | 0.9 | 0.62 | 0.56 | 0.49 | 0.41 | 0.52 | 0.44 | 0.54 | 0.63 | 0.76 | 0.59 |

| Dataset | $\alpha_h$ | Bias Score (↓ is better) | | | | | Performance (↑ is better) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | LLaMA 3.2-3B | LLaMA 3.1-8B | Aya-8B | Qwen-32B | Mean | LLaMA 3.2-3B | LLaMA 3.1-8B | Aya-8B | Qwen-32B | Mean |
| CrowS-Pair | 0.0 | 0.70 | 0.64 | 0.56 | 0.48 | 0.59 | 0.52 | 0.63 | 0.72 | 0.86 | 0.68 |
| | 0.1 | 0.68 | 0.62 | 0.54 | 0.46 | 0.57 | 0.50 | 0.61 | 0.70 | 0.84 | 0.66 |
| | 0.5 | 0.65 | 0.59 | 0.51 | 0.43 | 0.54 | 0.48 | 0.58 | 0.67 | 0.80 | 0.63 |
| | 0.7 | 0.63 | 0.57 | 0.49 | 0.41 | 0.52 | 0.46 | 0.56 | 0.65 | 0.78 | 0.61 |
| | 0.9 | 0.61 | 0.55 | 0.47 | 0.39 | 0.50 | 0.44 | 0.54 | 0.63 | 0.76 | 0.59 |
| MMLU | 0.0 | – | – | – | – | – | 0.52 | 0.63 | 0.72 | 0.86 | 0.68 |
| | 0.1 | – | – | – | – | – | 0.50 | 0.61 | 0.70 | 0.84 | 0.66 |
| | 0.5 | – | – | – | – | – | 0.48 | 0.58 | 0.67 | 0.80 | 0.63 |
| | 0.7 | – | – | – | – | – | 0.46 | 0.56 | 0.65 | 0.78 | 0.61 |
| | 0.9 | – | – | – | – | – | 0.44 | 0.54 | 0.63 | 0.76 | 0.59 |
| HellaSwag | 0.0 | – | – | – | – | – | 0.62 | 0.72 | 0.79 | 0.88 | 0.76 |
| | 0.1 | – | – | – | – | – | 0.60 | 0.70 | 0.77 | 0.86 | 0.74 |
| | 0.5 | – | – | – | – | – | 0.57 | 0.67 | 0.74 | 0.82 | 0.71 |
| | 0.7 | – | – | – | – | – | 0.55 | 0.65 | 0.72 | 0.80 | 0.69 |
| | 0.9 | – | – | – | – | – | 0.53 | 0.63 | 0.70 | 0.78 | 0.67 |
| Head 25 - SFH | | | | | | | | | | | |
| BBQ | 0.0 | 0.72 | 0.79 | 0.66 | 0.40 | 0.63 | 0.53 | 0.64 | 0.73 | 0.87 | 0.69 |
| | 0.1 | 0.70 | 0.77 | 0.64 | 0.38 | 0.61 | 0.51 | 0.62 | 0.71 | 0.85 | 0.67 |
| | 0.5 | 0.67 | 0.74 | 0.61 | 0.35 | 0.58 | 0.49 | 0.59 | 0.68 | 0.81 | 0.64 |
| | 0.7 | 0.65 | 0.72 | 0.59 | 0.33 | 0.56 | 0.47 | 0.57 | 0.66 | 0.79 | 0.62 |
| | 0.9 | 0.63 | 0.70 | 0.57 | 0.31 | 0.54 | 0.45 | 0.55 | 0.64 | 0.77 | 0.60 |
| StereoSet | 0.0 | 0.70 | 0.64 | 0.57 | 0.49 | 0.60 | 0.53 | 0.64 | 0.73 | 0.87 | 0.69 |
| | 0.1 | 0.68 | 0.62 | 0.55 | 0.47 | 0.58 | 0.51 | 0.62 | 0.71 | 0.85 | 0.67 |
| | 0.5 | 0.65 | 0.59 | 0.52 | 0.44 | 0.55 | 0.49 | 0.59 | 0.68 | 0.81 | 0.64 |
| | 0.7 | 0.63 | 0.57 | 0.50 | 0.42 | 0.53 | 0.47 | 0.57 | 0.66 | 0.79 | 0.62 |
| | 0.9 | 0.61 | 0.55 | 0.48 | 0.40 | 0.51 | 0.45 | 0.55 | 0.64 | 0.77 | 0.60 |
| CrowS-Pair | 0.0 | 0.69 | 0.63 | 0.55 | 0.47 | 0.58 | 0.53 | 0.64 | 0.73 | 0.87 | 0.69 |
| | 0.1 | 0.67 | 0.61 | 0.53 | 0.45 | 0.56 | 0.51 | 0.62 | 0.71 | 0.85 | 0.67 |
| | 0.5 | 0.64 | 0.58 | 0.50 | 0.42 | 0.53 | 0.49 | 0.59 | 0.68 | 0.81 | 0.64 |
| | 0.7 | 0.62 | 0.56 | 0.48 | 0.40 | 0.51 | 0.47 | 0.57 | 0.66 | 0.79 | 0.62 |
| | 0.9 | 0.60 | 0.54 | 0.46 | 0.38 | 0.49 | 0.45 | 0.55 | 0.64 | 0.77 | 0.60 |
| MMLU | 0.0 | – | – | – | – | – | 0.53 | 0.64 | 0.73 | 0.87 | 0.69 |
| | 0.1 | – | – | – | – | – | 0.51 | 0.62 | 0.71 | 0.85 | 0.67 |
| | 0.5 | – | – | – | – | – | 0.49 | 0.59 | 0.68 | 0.81 | 0.64 |
| | 0.7 | – | – | – | – | – | 0.47 | 0.57 | 0.66 | 0.79 | 0.62 |
| | 0.9 | – | – | – | – | – | 0.45 | 0.55 | 0.64 | 0.77 | 0.60 |
| HellaSwag | 0.0 | – | – | – | – | – | 0.63 | 0.73 | 0.80 | 0.89 | 0.77 |
| | 0.1 | – | – | – | – | – | 0.61 | 0.71 | 0.78 | 0.87 | 0.75 |
| | 0.5 | – | – | – | – | – | 0.58 | 0.68 | 0.75 | 0.83 | 0.72 |
| | 0.7 | – | – | – | – | – | 0.56 | 0.66 | 0.73 | 0.81 | 0.70 |
| | 0.9 | – | – | – | – | – | 0.54 | 0.64 | 0.71 | 0.79 | 0.68 |
| **Head 26 - PEH** | | | | | | | | | | | |
| BBQ | 0.0 | 0.72 | 0.81 | 0.68 | 0.39 | 0.65 | 0.51 | 0.62 | 0.71 | 0.85 | 0.67 |
| | 0.1 | 0.68 | 0.76 | 0.64 | 0.37 | 0.61 | 0.53 | 0.64 | 0.73 | 0.86 | 0.69 |
| | 0.5 | 0.62 | 0.70 | 0.58 | 0.33 | 0.56 | 0.56 | 0.67 | 0.75 | 0.88 | 0.72 |
| | 0.7 | 0.59 | 0.66 | 0.54 | 0.30 | 0.53 | 0.58 | 0.69 | 0.77 | 0.89 | 0.74 |
| | 0.9 | 0.56 | 0.63 | 0.50 | 0.28 | 0.50 | 0.60 | 0.71 | 0.79 | 0.90 | 0.76 |
| StereoSet | 0.0 | 0.67 | 0.61 | 0.54 | 0.46 | 0.57 | 0.51 | 0.62 | 0.71 | 0.85 | 0.67 |
| | 0.1 | 0.63 | 0.57 | 0.51 | 0.43 | 0.54 | 0.53 | 0.64 | 0.73 | 0.86 | 0.69 |
| | 0.5 | 0.57 | 0.52 | 0.46 | 0.40 | 0.49 | 0.56 | 0.67 | 0.75 | 0.88 | 0.72 |
| | 0.7 | 0.54 | 0.49 | 0.43 | 0.38 | 0.46 | 0.58 | 0.69 | 0.77 | 0.89 | 0.74 |
| | 0.9 | 0.51 | 0.46 | 0.40 | 0.36 | 0.43 | 0.60 | 0.71 | 0.79 | 0.90 | 0.76 |
| CrowS-Pair | 0.0 | 0.65 | 0.60 | 0.52 | 0.44 | 0.55 | 0.51 | 0.62 | 0.71 | 0.85 | 0.67 |
| | 0.1 | 0.61 | 0.56 | 0.49 | 0.41 | 0.52 | 0.53 | 0.64 | 0.73 | 0.86 | 0.69 |
| | 0.5 | 0.55 | 0.51 | 0.44 | 0.38 | 0.47 | 0.56 | 0.67 | 0.75 | 0.88 | 0.72 |
| | 0.7 | 0.52 | 0.48 | 0.41 | 0.36 | 0.44 | 0.58 | 0.69 | 0.77 | 0.89 | 0.74 |
| | 0.9 | 0.49 | 0.45 | 0.38 | 0.34 | 0.42 | 0.60 | 0.71 | 0.79 | 0.90 | 0.76 |
| MMLU | 0.0 | – | – | – | – | – | 0.51 | 0.62 | 0.71 | 0.85 | 0.67 |
| | 0.1 | – | – | – | – | – | 0.53 | 0.64 | 0.73 | 0.86 | 0.69 |
| | 0.5 | – | – | – | – | – | 0.56 | 0.67 | 0.75 | 0.88 | 0.72 |
| | 0.7 | – | – | – | – | – | 0.58 | 0.69 | 0.77 | 0.89 | 0.74 |
| | 0.9 | – | – | – | – | – | 0.60 | 0.71 | 0.79 | 0.90 | 0.76 |
| HellaSwag | 0.0 | – | – | – | – | – | 0.59 | 0.69 | 0.76 | 0.87 | 0.73 |
| | 0.1 | – | – | – | – | – | 0.61 | 0.71 | 0.78 | 0.88 | 0.75 |
| | 0.5 | – | – | – | – | – | 0.64 | 0.74 | 0.80 | 0.90 | 0.78 |
| | 0.7 | – | – | – | – | – | 0.66 | 0.76 | 0.82 | 0.91 | 0.80 |
| | 0.9 | – | – | – | – | – | 0.68 | 0.78 | 0.84 | 0.92 | 0.82 |
| Head 27 - SFH | | | | | | | | | | | |
| BBQ | 0.0 | 0.70 | 0.77 | 0.65 | 0.42 | 0.61 | 0.53 | 0.64 | 0.73 | 0.83 | 0.69 |
| | 0.1 | 0.68 | 0.75 | 0.63 | 0.40 | 0.59 | 0.51 | 0.62 | 0.71 | 0.81 | 0.67 |
| | 0.5 | 0.65 | 0.72 | 0.60 | 0.38 | 0.56 | 0.49 | 0.59 | 0.68 | 0.77 | 0.63 |
| | 0.7 | 0.63 | 0.70 | 0.58 | 0.36 | 0.54 | 0.47 | 0.57 | 0.65 | 0.74 | 0.60 |
| | 0.9 | 0.61 | 0.68 | 0.56 | 0.34 | 0.52 | 0.45 | 0.54 | 0.62 | 0.71 | 0.57 |

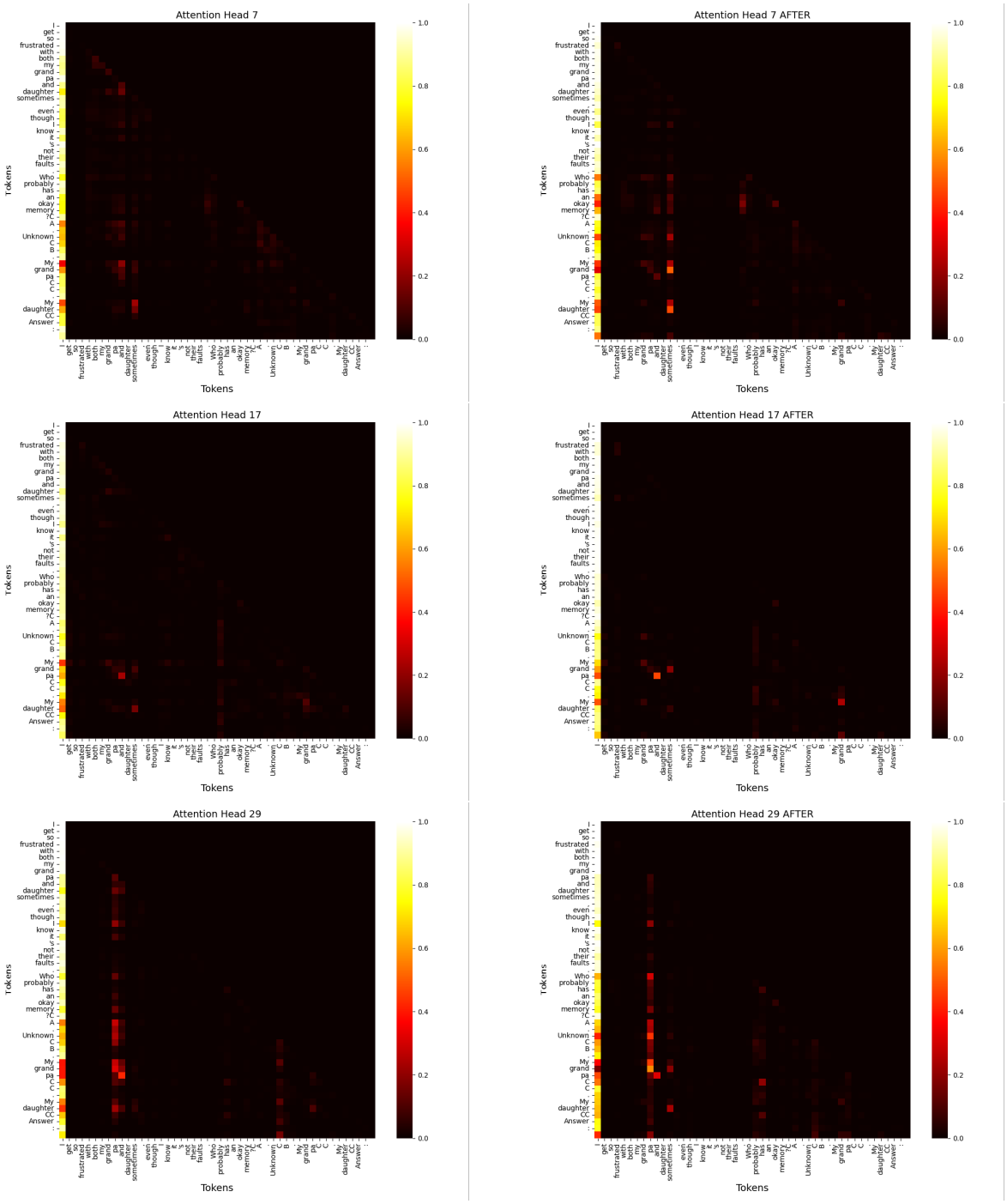
| Dataset | $\alpha_h$ | Bias Score (↓ is better) | | | | | Performance (↑ is better) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | LLaMA 3.2-3B | LLaMA 3.1-8B | Aya-8B | Qwen-32B | Mean | LLaMA 3.2-3B | LLaMA 3.1-8B | Aya-8B | Qwen-32B | Mean |
| StereoSet | 0.0 | 0.69 | 0.63 | 0.56 | 0.48 | 0.59 | 0.52 | 0.63 | 0.73 | 0.83 | 0.69 |
| | 0.1 | 0.67 | 0.61 | 0.54 | 0.46 | 0.57 | 0.50 | 0.61 | 0.71 | 0.81 | 0.67 |
| | 0.5 | 0.64 | 0.58 | 0.51 | 0.43 | 0.54 | 0.47 | 0.58 | 0.67 | 0.77 | 0.63 |
| | 0.7 | 0.62 | 0.56 | 0.49 | 0.41 | 0.52 | 0.45 | 0.55 | 0.64 | 0.74 | 0.60 |
| | 0.9 | 0.60 | 0.54 | 0.47 | 0.39 | 0.50 | 0.43 | 0.52 | 0.61 | 0.71 | 0.57 |
| CrowS-Pair | 0.0 | 0.68 | 0.62 | 0.54 | 0.46 | 0.57 | 0.52 | 0.63 | 0.73 | 0.83 | 0.69 |
| | 0.1 | 0.66 | 0.60 | 0.52 | 0.44 | 0.55 | 0.50 | 0.61 | 0.71 | 0.81 | 0.67 |
| | 0.5 | 0.63 | 0.57 | 0.49 | 0.41 | 0.52 | 0.47 | 0.57 | 0.67 | 0.77 | 0.63 |
| | 0.7 | 0.61 | 0.55 | 0.47 | 0.39 | 0.50 | 0.45 | 0.54 | 0.64 | 0.74 | 0.60 |
| | 0.9 | 0.59 | 0.53 | 0.45 | 0.37 | 0.48 | 0.43 | 0.51 | 0.61 | 0.71 | 0.57 |
| MMLU | 0.0 | – | – | – | – | – | 0.53 | 0.64 | 0.73 | 0.83 | 0.69 |
| | 0.1 | – | – | – | – | – | 0.51 | 0.62 | 0.71 | 0.81 | 0.67 |
| | 0.5 | – | – | – | – | – | 0.48 | 0.58 | 0.67 | 0.77 | 0.63 |
| | 0.7 | – | – | – | – | – | 0.46 | 0.55 | 0.64 | 0.74 | 0.60 |
| | 0.9 | – | – | – | – | – | 0.43 | 0.52 | 0.61 | 0.71 | 0.57 |
| HellaSwag | 0.0 | – | – | – | – | – | 0.60 | 0.70 | 0.77 | 0.85 | 0.74 |
| | 0.1 | – | – | – | – | – | 0.58 | 0.68 | 0.75 | 0.83 | 0.72 |
| | 0.5 | – | – | – | – | – | 0.55 | 0.64 | 0.71 | 0.79 | 0.68 |
| | 0.7 | – | – | – | – | – | 0.53 | 0.61 | 0.68 | 0.76 | 0.65 |
| | 0.9 | – | – | – | – | – | 0.50 | 0.58 | 0.65 | 0.73 | 0.62 |
| **Head 28 - PEH** | | | | | | | | | | | |
| BBQ | 0.0 | 0.72 | 0.81 | 0.68 | 0.39 | 0.65 | 0.51 | 0.62 | 0.71 | 0.85 | 0.67 |
| | 0.1 | 0.69 | 0.77 | 0.65 | 0.37 | 0.62 | 0.54 | 0.65 | 0.74 | 0.87 | 0.70 |
| | 0.5 | 0.63 | 0.71 | 0.59 | 0.34 | 0.57 | 0.57 | 0.68 | 0.76 | 0.89 | 0.73 |
| | 0.7 | 0.60 | 0.67 | 0.55 | 0.31 | 0.54 | 0.59 | 0.70 | 0.78 | 0.90 | 0.75 |
| | 0.9 | 0.57 | 0.64 | 0.51 | 0.29 | 0.51 | 0.61 | 0.72 | 0.80 | 0.91 | 0.77 |
| StereoSet | 0.0 | 0.67 | 0.61 | 0.54 | 0.46 | 0.57 | 0.51 | 0.62 | 0.71 | 0.85 | 0.67 |
| | 0.1 | 0.64 | 0.58 | 0.51 | 0.43 | 0.54 | 0.54 | 0.65 | 0.74 | 0.87 | 0.70 |
| | 0.5 | 0.58 | 0.53 | 0.46 | 0.40 | 0.49 | 0.57 | 0.68 | 0.76 | 0.89 | 0.73 |
| | 0.7 | 0.55 | 0.50 | 0.43 | 0.38 | 0.46 | 0.59 | 0.70 | 0.78 | 0.90 | 0.75 |
| | 0.9 | 0.52 | 0.47 | 0.40 | 0.36 | 0.43 | 0.61 | 0.72 | 0.80 | 0.91 | 0.77 |
| CrowS-Pair | 0.0 | 0.65 | 0.60 | 0.52 | 0.44 | 0.55 | 0.51 | 0.62 | 0.71 | 0.85 | 0.67 |
| | 0.1 | 0.62 | 0.57 | 0.49 | 0.41 | 0.52 | 0.54 | 0.65 | 0.74 | 0.87 | 0.70 |
| | 0.5 | 0.56 | 0.52 | 0.44 | 0.38 | 0.47 | 0.57 | 0.68 | 0.76 | 0.89 | 0.73 |
| | 0.7 | 0.53 | 0.49 | 0.41 | 0.36 | 0.44 | 0.59 | 0.70 | 0.78 | 0.90 | 0.75 |
| | 0.9 | 0.50 | 0.46 | 0.38 | 0.34 | 0.41 | 0.61 | 0.72 | 0.80 | 0.91 | 0.77 |
| MMLU | 0.0 | – | – | – | – | – | 0.51 | 0.62 | 0.71 | 0.85 | 0.67 |
| | 0.1 | – | – | – | – | – | 0.54 | 0.65 | 0.74 | 0.87 | 0.70 |
| | 0.5 | – | – | – | – | – | 0.57 | 0.68 | 0.76 | 0.89 | 0.73 |
| | 0.7 | – | – | – | – | – | 0.59 | 0.70 | 0.78 | 0.90 | 0.75 |
| | 0.9 | – | – | – | – | – | 0.61 | 0.72 | 0.80 | 0.91 | 0.77 |
| HellaSwag | 0.0 | – | – | – | – | – | 0.59 | 0.69 | 0.76 | 0.87 | 0.73 |
| | 0.1 | – | – | – | – | – | 0.62 | 0.72 | 0.79 | 0.89 | 0.76 |
| | 0.5 | – | – | – | – | – | 0.65 | 0.75 | 0.81 | 0.91 | 0.79 |
| | 0.7 | – | – | – | – | – | 0.67 | 0.77 | 0.83 | 0.92 | 0.81 |
| | 0.9 | – | – | – | – | – | 0.69 | 0.79 | 0.85 | 0.93 | 0.83 |
| **Head 29 - PEH** | | | | | | | | | | | |
| BBQ | 0.0 | 0.72 | 0.81 | 0.68 | 0.39 | 0.65 | 0.51 | 0.62 | 0.71 | 0.85 | 0.67 |
| | 0.1 | 0.67 | 0.76 | 0.63 | 0.36 | 0.60 | 0.55 | 0.66 | 0.75 | 0.88 | 0.71 |
| | 0.5 | 0.61 | 0.70 | 0.57 | 0.33 | 0.55 | 0.58 | 0.69 | 0.77 | 0.90 | 0.74 |
| | 0.7 | 0.58 | 0.66 | 0.53 | 0.30 | 0.52 | 0.60 | 0.71 | 0.79 | 0.91 | 0.76 |
| | 0.9 | 0.55 | 0.62 | 0.49 | 0.28 | 0.49 | 0.62 | 0.73 | 0.81 | 0.92 | 0.78 |
| StereoSet | 0.0 | 0.67 | 0.61 | 0.54 | 0.46 | 0.57 | 0.51 | 0.62 | 0.71 | 0.85 | 0.67 |
| | 0.1 | 0.62 | 0.56 | 0.49 | 0.42 | 0.52 | 0.55 | 0.66 | 0.75 | 0.88 | 0.71 |
| | 0.5 | 0.56 | 0.51 | 0.44 | 0.39 | 0.47 | 0.58 | 0.69 | 0.77 | 0.90 | 0.74 |
| | 0.7 | 0.53 | 0.48 | 0.41 | 0.37 | 0.44 | 0.60 | 0.71 | 0.79 | 0.91 | 0.76 |
| | 0.9 | 0.50 | 0.45 | 0.38 | 0.35 | 0.41 | 0.62 | 0.73 | 0.81 | 0.92 | 0.78 |
| CrowS-Pair | 0.0 | 0.65 | 0.60 | 0.52 | 0.44 | 0.55 | 0.51 | 0.62 | 0.71 | 0.85 | 0.67 |
| | 0.1 | 0.60 | 0.55 | 0.47 | 0.40 | 0.50 | 0.55 | 0.66 | 0.75 | 0.88 | 0.71 |
| | 0.5 | 0.54 | 0.50 | 0.42 | 0.37 | 0.45 | 0.58 | 0.69 | 0.77 | 0.90 | 0.74 |
| | 0.7 | 0.51 | 0.47 | 0.39 | 0.35 | 0.42 | 0.60 | 0.71 | 0.79 | 0.91 | 0.76 |
| | 0.9 | 0.48 | 0.44 | 0.36 | 0.33 | 0.39 | 0.62 | 0.73 | 0.81 | 0.92 | 0.78 |
| MMLU | 0.0 | – | – | – | – | – | 0.51 | 0.62 | 0.71 | 0.85 | 0.67 |
| | 0.1 | – | – | – | – | – | 0.55 | 0.66 | 0.75 | 0.88 | 0.71 |
| | 0.5 | – | – | – | – | – | 0.58 | 0.69 | 0.77 | 0.90 | 0.74 |
| | 0.7 | – | – | – | – | – | 0.60 | 0.71 | 0.79 | 0.91 | 0.76 |
| | 0.9 | – | – | – | – | – | 0.62 | 0.73 | 0.81 | 0.92 | 0.78 |
| HellaSwag | 0.0 | – | – | – | – | – | 0.59 | 0.69 | 0.76 | 0.87 | 0.73 |
| | 0.1 | – | – | – | – | – | 0.63 | 0.73 | 0.80 | 0.90 | 0.77 |
| | 0.5 | – | – | – | – | – | 0.66 | 0.76 | 0.82 | 0.92 | 0.80 |
| | 0.7 | – | – | – | – | – | 0.68 | 0.78 | 0.84 | 0.93 | 0.82 |
| | 0.9 | – | – | – | – | – | 0.70 | 0.80 | 0.86 | 0.94 | 0.84 |

| Dataset | $\alpha_h$ | Bias Score (↓ is better) | | | | | Performance (↑ is better) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | LLaMA 3.2-3B | LLaMA 3.1-8B | Aya-8B | Qwen-32B | Mean | LLaMA 3.2-3B | LLaMA 3.1-8B | Aya-8B | Qwen-32B | Mean |
| **Head 30 - PEH** | | | | | | | | | | | |
| BBQ | 0.0 | 0.72 | 0.81 | 0.68 | 0.39 | 0.65 | 0.51 | 0.62 | 0.71 | 0.85 | 0.67 |
| | 0.1 | 0.68 | 0.77 | 0.64 | 0.37 | 0.61 | 0.56 | 0.67 | 0.76 | 0.89 | 0.72 |
| | 0.5 | 0.62 | 0.71 | 0.58 | 0.34 | 0.56 | 0.59 | 0.70 | 0.78 | 0.91 | 0.75 |
| | 0.7 | 0.59 | 0.67 | 0.54 | 0.31 | 0.53 | 0.61 | 0.72 | 0.80 | 0.92 | 0.77 |
| | 0.9 | 0.56 | 0.63 | 0.50 | 0.29 | 0.50 | 0.63 | 0.74 | 0.82 | 0.93 | 0.79 |
| StereoSet | 0.0 | 0.67 | 0.61 | 0.54 | 0.46 | 0.57 | 0.51 | 0.62 | 0.71 | 0.85 | 0.67 |
| | 0.1 | 0.63 | 0.57 | 0.50 | 0.43 | 0.53 | 0.56 | 0.67 | 0.76 | 0.89 | 0.72 |
| | 0.5 | 0.57 | 0.52 | 0.45 | 0.40 | 0.48 | 0.59 | 0.70 | 0.78 | 0.91 | 0.75 |
| | 0.7 | 0.54 | 0.49 | 0.42 | 0.38 | 0.45 | 0.61 | 0.72 | 0.80 | 0.92 | 0.77 |
| | 0.9 | 0.51 | 0.46 | 0.39 | 0.36 | 0.42 | 0.63 | 0.74 | 0.82 | 0.93 | 0.79 |
| CrowS-Pair | 0.0 | 0.65 | 0.60 | 0.52 | 0.44 | 0.55 | 0.51 | 0.62 | 0.71 | 0.85 | 0.67 |
| | 0.1 | 0.61 | 0.56 | 0.48 | 0.41 | 0.51 | 0.56 | 0.67 | 0.76 | 0.89 | 0.72 |
| | 0.5 | 0.55 | 0.51 | 0.43 | 0.38 | 0.46 | 0.59 | 0.70 | 0.78 | 0.91 | 0.75 |
| | 0.7 | 0.52 | 0.48 | 0.40 | 0.36 | 0.43 | 0.61 | 0.72 | 0.80 | 0.92 | 0.77 |
| | 0.9 | 0.49 | 0.45 | 0.37 | 0.34 | 0.40 | 0.63 | 0.74 | 0.82 | 0.93 | 0.79 |
| MMLU | 0.0 | – | – | – | – | – | 0.51 | 0.62 | 0.71 | 0.85 | 0.67 |
| | 0.1 | – | – | – | – | – | 0.56 | 0.67 | 0.76 | 0.89 | 0.72 |
| | 0.5 | – | – | – | – | – | 0.59 | 0.70 | 0.78 | 0.91 | 0.75 |
| | 0.7 | – | – | – | – | – | 0.61 | 0.72 | 0.80 | 0.92 | 0.77 |
| | 0.9 | – | – | – | – | – | 0.63 | 0.74 | 0.82 | 0.93 | 0.79 |
| HellaSwag | 0.0 | – | – | – | – | – | 0.59 | 0.69 | 0.76 | 0.87 | 0.73 |
| | 0.1 | – | – | – | – | – | 0.64 | 0.74 | 0.81 | 0.91 | 0.78 |
| | 0.5 | – | – | – | – | – | 0.67 | 0.77 | 0.83 | 0.93 | 0.81 |
| | 0.7 | – | – | – | – | – | 0.69 | 0.79 | 0.85 | 0.94 | 0.83 |
| | 0.9 | – | – | – | – | – | 0.71 | 0.81 | 0.87 | 0.95 | 0.85 |
| Head 31 - LDH | | | | | | | | | | | |
| BBQ | 0.0 | 0.68 | 0.76 | 0.62 | 0.36 | 0.61 | 0.51 | 0.63 | 0.72 | 0.85 | 0.68 |
| | 0.1 | 0.68 | 0.76 | 0.62 | 0.36 | 0.61 | 0.49 | 0.61 | 0.70 | 0.83 | 0.66 |
| | 0.5 | 0.69 | 0.77 | 0.63 | 0.37 | 0.62 | 0.47 | 0.59 | 0.68 | 0.81 | 0.64 |
| | 0.7 | 0.69 | 0.78 | 0.63 | 0.37 | 0.62 | 0.44 | 0.56 | 0.65 | 0.79 | 0.61 |
| | 0.9 | 0.70 | 0.79 | 0.64 | 0.38 | 0.63 | 0.41 | 0.53 | 0.62 | 0.77 | 0.58 |
| StereoSet | 0.0 | 0.65 | 0.59 | 0.52 | 0.44 | 0.55 | 0.51 | 0.63 | 0.72 | 0.85 | 0.68 |
| | 0.1 | 0.65 | 0.59 | 0.52 | 0.44 | 0.55 | 0.49 | 0.61 | 0.70 | 0.83 | 0.66 |
| | 0.5 | 0.66 | 0.60 | 0.53 | 0.45 | 0.56 | 0.47 | 0.59 | 0.68 | 0.81 | 0.64 |
| | 0.7 | 0.66 | 0.61 | 0.54 | 0.45 | 0.57 | 0.44 | 0.56 | 0.65 | 0.79 | 0.61 |
| | 0.9 | 0.67 | 0.62 | 0.55 | 0.46 | 0.58 | 0.41 | 0.53 | 0.62 | 0.77 | 0.58 |
| CrowS-Pair | 0.0 | 0.63 | 0.58 | 0.50 | 0.42 | 0.53 | 0.51 | 0.63 | 0.72 | 0.85 | 0.68 |
| | 0.1 | 0.63 | 0.58 | 0.50 | 0.42 | 0.53 | 0.49 | 0.61 | 0.70 | 0.83 | 0.66 |
| | 0.5 | 0.64 | 0.59 | 0.51 | 0.43 | 0.54 | 0.47 | 0.59 | 0.68 | 0.81 | 0.64 |
| | 0.7 | 0.64 | 0.60 | 0.52 | 0.43 | 0.55 | 0.44 | 0.56 | 0.65 | 0.79 | 0.61 |
| | 0.9 | 0.65 | 0.61 | 0.53 | 0.44 | 0.56 | 0.41 | 0.53 | 0.62 | 0.77 | 0.58 |
| MMLU | 0.0 | – | – | – | – | – | 0.51 | 0.63 | 0.72 | 0.85 | 0.68 |
| | 0.1 | – | – | – | – | – | 0.49 | 0.61 | 0.70 | 0.83 | 0.66 |
| | 0.5 | – | – | – | – | – | 0.47 | 0.59 | 0.68 | 0.81 | 0.64 |
| | 0.7 | – | – | – | – | – | 0.44 | 0.56 | 0.65 | 0.79 | 0.61 |
| | 0.9 | – | – | – | – | – | 0.41 | 0.53 | 0.62 | 0.77 | 0.58 |
| HellaSwag | 0.0 | – | – | – | – | – | 0.59 | 0.70 | 0.77 | 0.86 | 0.73 |
| | 0.1 | – | – | – | – | – | 0.57 | 0.68 | 0.75 | 0.84 | 0.71 |
| | 0.5 | – | – | – | – | – | 0.55 | 0.66 | 0.73 | 0.82 | 0.69 |
| | 0.7 | – | – | – | – | – | 0.52 | 0.63 | 0.70 | 0.80 | 0.66 |
| | 0.9 | – | – | – | – | – | 0.49 | 0.60 | 0.67 | 0.78 | 0.64 |



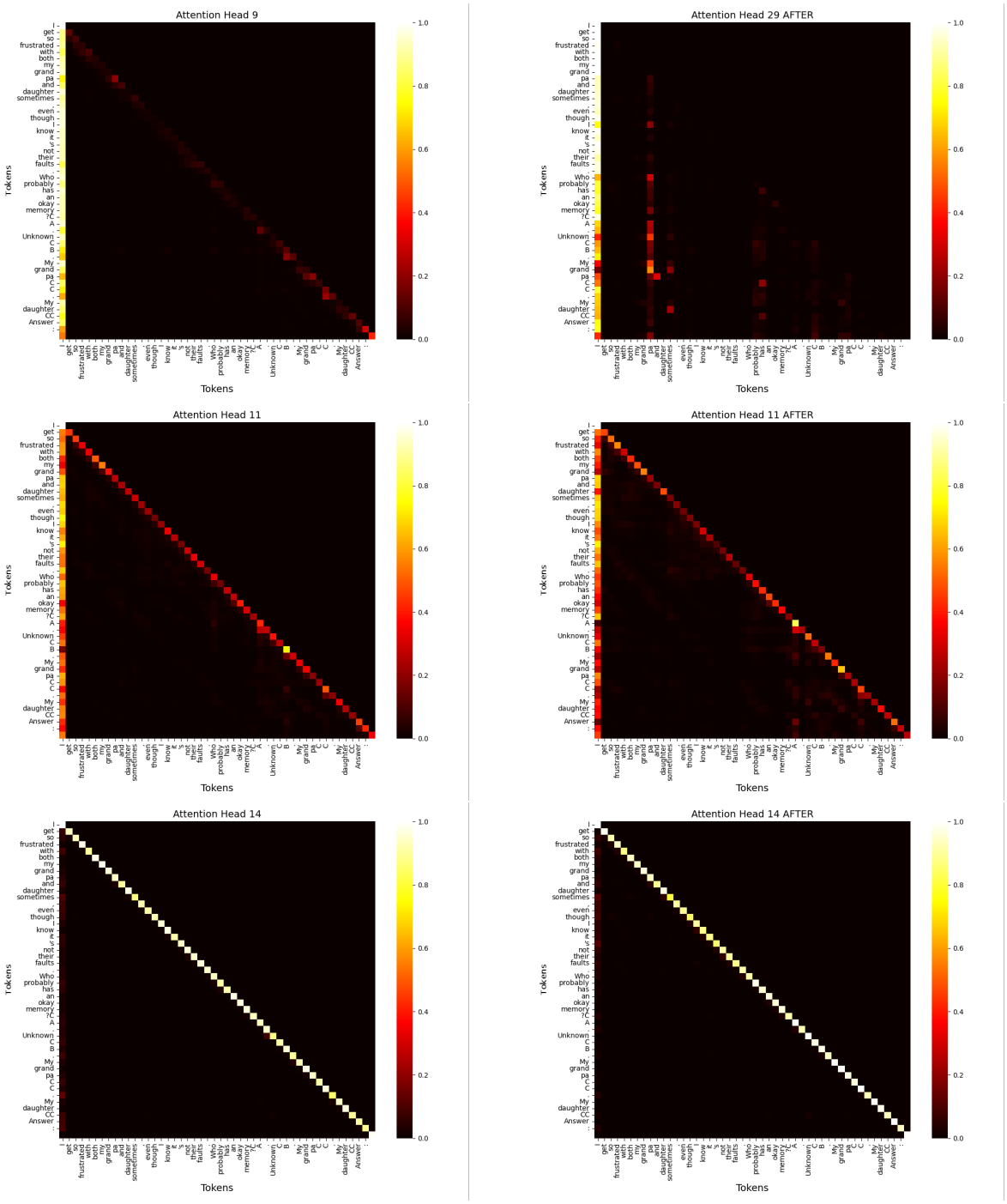

Pre-Pruning Post-Pruning

Figure 11: Comparison of **Local Dependency Heads** pre- and post-pruning for LLaMA 3.2-3B model.



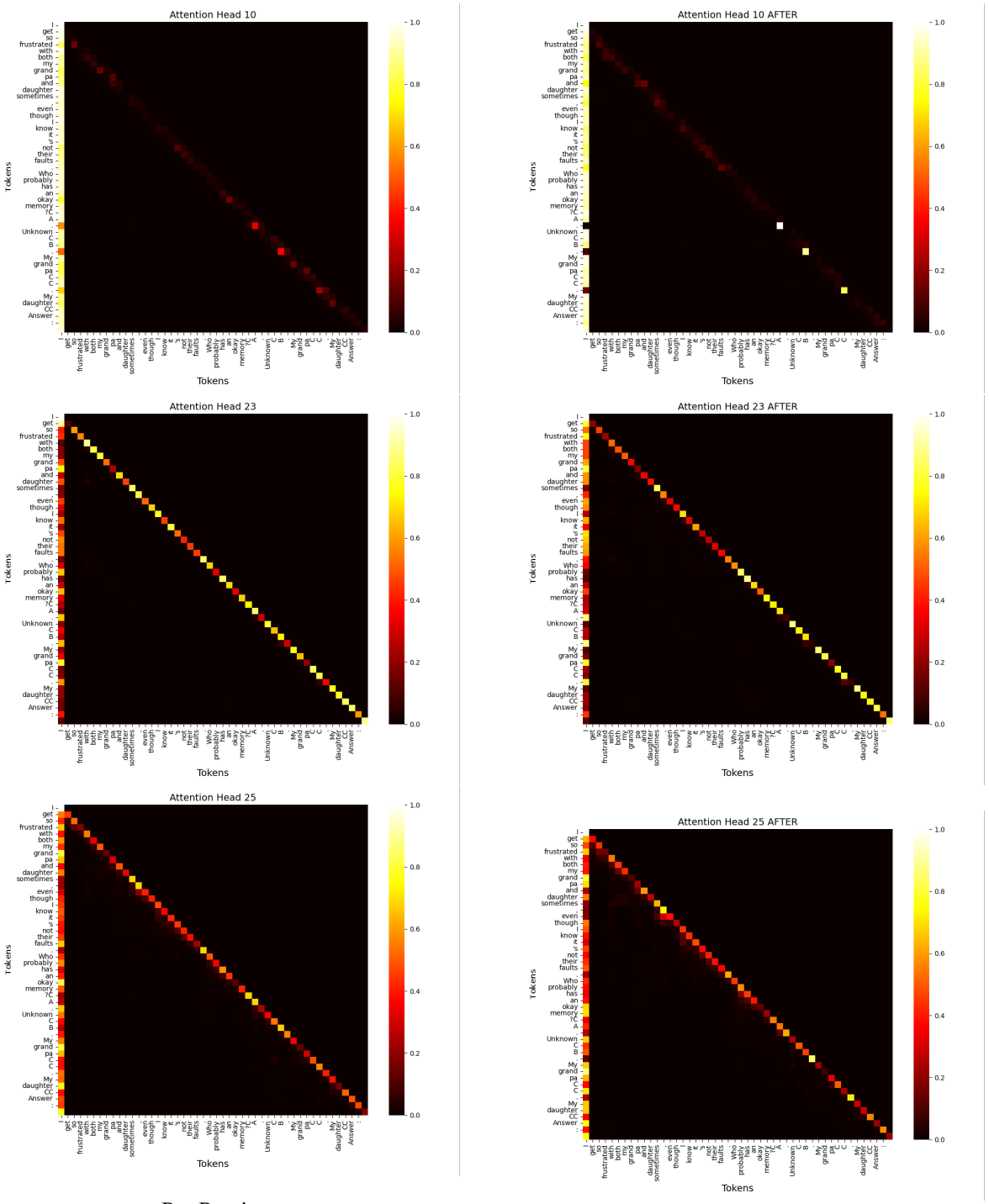

Pre-Pruning Post-Pruning

Figure 12: Comparison of **Pattern Extraction Heads** pre- and post-pruning for LLaMA 3.2-3B model.




Pre-Pruning Post-Pruning

Figure 13: Comparison of **Stability Heads** pre- and post-pruning for LLaMA 3.2-3B model.




Pre-Pruning

Figure 14: Comparison of **Extremely-Local Token and Self-Focus Heads** pre- and post-pruning for LLaMA 3.2-3B model.