Race, Ethnicity and Their Implication on Bias in Large Language Models

Shiyue Hu[1,2] **Ruizhe Li**[3,*] **Yanjun Gao**[1,*]
[1]University of Colorado Anschutz
[2]University of Colorado Boulder
[3]University of Aberdeen
shiyue.hu@colorado.edu, ruizhe.li@abdn.ac.uk, yanjun.gao@cuanschutz.edu
[*]Co-senior authors

Abstract

Large language models (LLMs) increasingly operate in high-stakes settings including healthcare and medicine, where demographic attributes such as race and ethnicity may be explicitly stated or implicitly inferred from text. However, existing studies primarily document outcome-level disparities, offering limited insight into internal mechanisms underlying these effects. We present a mechanistic study of how race and ethnicity are represented and operationalized within LLMs. Using two publicly available datasets spanning toxicity-related generation and clinical narrative understanding tasks, we analyze three open-source models with a reproducible interpretability pipeline combining probing, neuron-level attribution, and targeted intervention. We find that demographic information is distributed across internal units with substantial cross-model variation. Although some units encode sensitive or stereotype-related associations from pretraining, identical demographic cues can induce qualitatively different behaviors. Interventions suppressing such neurons reduce bias but leave substantial residual effects, suggesting behavioral rather than representational change and motivating more systematic mitigation.

1 Introduction

Large language models (LLMs) are increasingly used in high-stakes domains such as healthcare, where demographic attributes (e.g., race, ethnicity, gender) may be explicitly stated or implicitly inferred from text. Prior work shows that LLMs can condition their outputs on demographic information even when it is not task-relevant (Zack et al., 2024; Kim et al., 2023; Fraser and Kiritchenko, 2024; Zhao et al., 2025), therefore can induce misattribution on model output with undesirable or biased behavior (Demchak et al., 2024; Levartovsky et al., 2025; Zack et al., 2024).

Most prior studies on demographic bias focus on outcome-level effects, evaluating disparities in generated responses, accuracy, calibration, or toxicity scores across demographic groups (Tan and Lee, 2025; Hartvigsen et al., 2022; Guan et al., 2025; Wang et al., 2025). While these analyses are essential for documenting harm, they treat LLMs as black boxes, offering limited insight into whether demographic attributes are encoded as high-level semantic features, task-relevant representations, or spurious shortcuts during prediction. In parallel, recent works in mechanistic interpretability demonstrated how LLMs encode demographic information and manipulated internal LLMs' states to ensure fairness (Yu and Ananiadou, 2025; Ahsan and Wallace, 2025; Karvonen and Marks, 2025), yet these tools have rarely been applied to demographic bias in a systematic and task-diverse manner.

A central challenge is that demographic attributes interact with language in complex ways. In many real-world settings, demographic information may be explicitly stated (e.g., "a Black patient," "a Hispanic speaker") or implicitly conveyed through linguistic, cultural, or geographical cues, i.e. the "proxy" cues. Moreover, the same demographic signal can have qualitatively different effects depending on the task: it may alter predicted medical risk in a clinical scenario, while simultaneously modulating perceived toxicity, credibility, or intent in open-ended generation. Existing evaluation typically isolate a single task or domain (Hartvigsen et al., 2022; Zack et al., 2024; Levartovsky et al., 2025), making it difficult to assess whether demographic sensitivity reflects general representational mechanisms or task-specific heuristics.

In this work, we investigate how demographic attributes influence LLM behavior, with a focus on mechanistic explanations rather than surface-level disparities. We examine *race* and *ethnicity* as commonly occurring coarse-grained categories (e.g. White, Black, Asian, Hispanic and Latino) as they appear in the studied datasets, rather than attempting to model the full sociological complexity of these constructs. Using two publicly available datasets, we study: 1) toxicity-related generation tasks (Hartvigsen et al., 2022), where the same attributes may alter the likelihood, tone, or framing of model outputs, and 2) clinical narrative tasks (Bear Don't Walk IV et al., 2024), where the same attributes appear through explicit or indirect cues in medical text and modulate model behavior despite identical clinical evidence.

We adopt a mechanistic interpretability (MI) framework to study how lexical cues of race and ethnicity are encoded and propagated within three open-source LLMs that are widely used: Qwen2.5-7B (Team, 2024), Mistral-7B (Jiang et al., 2023), and Llama-3.1-8B (Grattafiori et al., 2024). Our contributions are threefold:

- a reproducible MI **pipeline** that combines multi-class probing, neuron-level attribution, and targeted intervention to identify internal units associated with demographic attributes and to examine their functional relevance across tasks. The proposed framework is applicable to other social variables beyond race and ethnicity.
- a fine-grained **characterization** of race and ethnicity representations across LLMs, revealing the distributed nature of demographic information and model-specific emphasis on semantic facets such as geography, language, culture, or historical context.
- a mechanistic **analysis** of how demographic representation influences model behaviors. Although internal features encode sensitive or harmful stereotype-related concepts present in pretraining data, these representations are unevenly activated by direct and indirect demographic cues.

Our findings show that while race- and ethnicity-associated representations can be identified and analyzed at the neuron level, their associations with biased model behavior persist even when highly active neurons are surpressed. This indicates that biased behavior in LLMs cannot be fully explained or controlled by manipulating a small set of identifiable neurons alone.

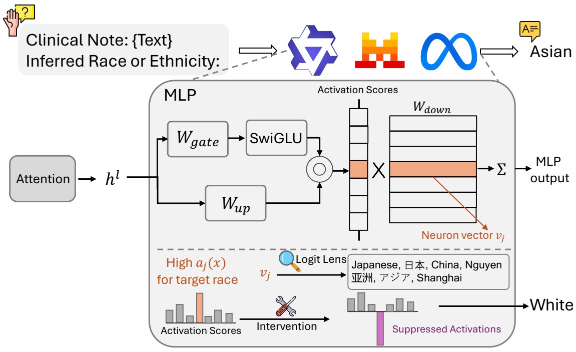

Figure 1: With MLP, we locate neurons relevant to race information and inspect them via Logit Lens. For the higher activation score for target race, we adjust its value to steer model's behavior.

2 Related Work

Mechanistic Interpretability of Bias in LLMs. Recent works in mechanistic interpretability have begun to locate where demographic information is encoded. Our approach of using probe-based neuron localization aligns with emerging research in this space. Yu and Ananiadou (2025) utilized neuron editing to understand and mitigate gender bias, identifying specific "gender neurons" within the MLP layers. In the medical domain, Ahsan et al. (2025) investigated the mechanisms of demographic bias specifically for healthcare tasks, suggesting that certain internal representations are disproportionately sensitive to racial identifiers. More recently, Ahsan and Wallace (2025) explored the use of Sparse Autoencoders (SAEs) to reveal clinical racial biases. Our work extends these findings by demonstrating that race-specific neurons are not only present in general datasets but are consistently activated and influential on domain-specific text.

Internal Bias Mitigation and Steering within LLMs. Beyond identification, recent work focuses on manipulating internal model states to ensure fairness. Zhou et al. (2024) proposed the UniBias framework, which mitigates bias by manipulating attention heads and MLP components. For real-time applications, Li et al. (2025) introduced *FairSteer*, a dynamic activation steering method that adjusts model behavior during inference. Karvonen and Marks (2025) further demonstrated that interpretability-based interventions can improve fairness more robustly than traditional fine-tuning in realistic settings. Our methodology contributes to this line of work by providing a targeted intervention strategy, specifically sign-flipping and amplification, to suppress biased pathways. This builds upon the "context-aware" fairness frameworks suggested by Nadeem et al. (2025), ensuring that mitigation is grounded in the semantic understanding of the racial directions we extract.

Racial Bias in Clinical LLMs. Extensive research has documented that LLMs inherit and propagate racial biases when applied to clinical decision support. Zhang et al. (2023) demonstrated that ChatGPT exhibits disparate treatment recommendations for acute coronary syndrome based on racial and gender cues. Similarly, Zack et al. (2024) evaluated GPT-4, finding that model frequently perpetuates harmful stereotypes that could lead to inequitable health outcomes. Poulain et al. (2024) further expanded this analysis across various clinical decision-support tasks, highlighting that bias patterns are not idiosyncratic but systematic across model families. While these studies establish the existence of bias, they largely treat model as a black box. Our work seeks to uncover the internal mechanisms driving these disparate outputs.

3 Background

MLP Layers and Neuron Activation. Modern Transformer-based LLMs process information through a *residual stream*. In this framework, the residual stream acts as a communication channel, while MLP layers function as key-value memories that store and inject factual associations into the stream (Geva et al., 2021a). Contemporary models like Llama 3.1, Mistral, and Qwen 2.5 utilize the *SwiGLU* gated architecture (Shazeer, 2020). The output of an MLP block with input x is defined as:

$$\text{MLP}(x) = \left(\text{SwiGLU}(xW_{\text{gate}}) \odot (xW_{\text{up}})\right) W_{\text{down}} \quad (1)$$

where $\odot$ is the element-wise product. We define an individual *neuron* j as the j-th element of the intermediate gated state. The total MLP output is the sum of these neurons' contributions:

$$\text{MLP}(x) = \sum_{j=1}^{d_{mlp}} a_j(x) \cdot v_j \quad (2)$$

where $a_j(x)$ is the activation score (the product of the gate and up-projections) and v_j is the j-th row of W_{down}. Our method specifically probes these *output vectors* v_j to locate racial information.

Logit Lens. To interpret high-dimensional vectors in residual stream or neuron output vectors v_j, we use *Logit Lens* (nostalgebraist, 2020). This technique projects a vector h directly into vocabulary space using model's unembedding matrix W_U: logits $= hW_U$. By inspecting top-ranked tokens in the resulting distribution, we can decode the semantic concepts encoded within specific neurons.

4 Methodology

We propose a mechanistic interpretability framework to determine *where* and *how* race information is encoded within LLMs. Our approach progresses from identifying global race directions via multi-class probing to identifying the specific neurons responsible for these encodings.

4.1 Locating Race Directions via Multi-Class Probing

To extract race representations, we train linear probes W_{Race} to classify racial group membership for each model. The probe is trained on the final-layer residual stream $\bar{h}^{L-1}$, averaged across all token positions:

$$P(\text{race} = c \mid \bar{h}^{L-1}) = \text{softmax}(W_{\text{Race}}^\top \bar{h}^{L-1} + b)_c \quad (3)$$

where $W_{\text{Race}} \in \mathbb{R}^{d \times |\mathcal{C}|}$ is the learned probe matrix, b is bias vector, and $\mathcal{C}$ denotes the set of racial super-categories (e.g., Asian, Black, Middle Eastern). Each column w_c of W_{Race} represents *race direction* for group c in model's representation space.

4.2 From Race Directions to Neurons

Having identified the global race directions w_c, we locate the MLP neurons that write to these directions, motivated by prior work showing MLPs act as key−value memories (Geva et al., 2021b).

Interpreting the Probe Direction. We first verify that our learned directions w_c capture meaningful racial semantics. Using Logit Lens, we project each direction into vocabulary space via the model's unembedding matrix W_U:

$$z_{\text{probe}} = W_U w_c \quad (4)$$

The top-k tokens (we use $k = 20$) with the highest values in z_{probe} serve as a semantic fingerprint for each racial group.

Identifying Candidate Neurons. To locate neurons that write to race direction, we compute the cosine similarity between each MLP neuron's output vector v_j^l at layer l and the probe direction w_c:

$$\text{Score}(l, j) = \frac{w_c \cdot v_j^l}{\|w_c\| \|v_j^l\|} \quad (5)$$

All neurons in final four MLP layers are ranked by this score, and the top 20 candidates are selected. Each candidate is selected by projecting its output vector into vocabulary space and inspecting top-20 tokens. Neurons are retained only if their tokens show clear alignment with the target racial group.

4.3 Validating Neurons via Activation Analysis and Intervention

To confirm that identified neurons encode meaningful racial information and causally influence model behavior, we design a two-stage validation pipeline.

Activation Analysis. We measure how strongly each neuron group activates across different inputs. For a given input text, we extract the activation score of each candidate neuron during the forward pass. For ToxiGen, we average activations across all token positions; for C-REACT, where the model must produce a classification, we extract activations at the final token position where the prediction is made. We then compare activation patterns across racial groups: if neurons identified for a particular race show higher activation scores when processing text about that group compared to other groups, this provides evidence that these neurons selectively respond to race-related content.

Targeted Intervention. To test whether identified neurons causally influence model outputs, we manipulate their activations during inference. We implement this using PyTorch forward hooks on the MLP W_{down}. During the forward pass, we intercept the activation score of each targeted neuron and force it into a strongly negative range. Specifically, we multiply positive activations by a negative factor (e.g., -5) and amplify negative activations by a positive factor (e.g., 5). This ensures that the neuron's output remains negative, effectively suppressing the encoded feature. This modified signal propagates through the remaining layers, allowing us to measure whether suppressing specific neurons changes model behavior and final outputs.

5 Data and Models

Datasets. We evaluate our framework on two domains to test the generalizability of race encoding mechanisms across different contexts.

ToxiGen (Hartvigsen et al., 2022) is a large-scale machine-generated dataset of toxic statements about minority groups. We use the annotated subset containing 9.9k samples across 13 target groups. Each sample consists of a statement targeting a specific group (e.g., "asians are trying to take over the world with their karate skills" is targeting Asian).

From the 13 target groups, we consolidate related categories into five racial categories: Asian (asian, chinese), Black (black), Latino (latino, mexican), Native American (native_american), and Middle Eastern (middle_east, jewish, muslim). We exclude non-racial categories (women, lgbtq, mental_dis, physical_dis).

Model	Group	Top tokens projected by probe
Qwen2.5-7B	Asian	Asian, 亚洲, Chinese, CJK, 东亚
	Latino	Mex, Mexico
	Native American	natives, native, Native, indigenous
Mistral-7B	Asian	Chinese, Asian, China, Korean, Taiwan
	Black	black, African, Black
	Latino	Mexico, Salvador, Colombia, Chile, Mexican
	Native American	Indians, trib, Native, tribes, Indian
	Middle Eastern	Islamic, Palestinian, Muhammad, Muslim, Israel
Llama-3.1-8B	Asian	Asian, Mandarin, CJK, asian, china
	Black	Black, _black, -black, .black, 黑
	Latino	Mundo, _BORDER
	Native American	Native, natives, Indians, indigenous, tribes
	Middle Eastern	Islamic, Middle, ISIL, Christian

Table 1: Top tokens by race group across models (ToxiGen). Full results in Appendix A.2. Translations: 亚洲 (Asia), 东亚 (East Asia), 黑 (Black).

C-REACT (Contextualized Race and Ethnicity Annotations for Clinical Text (Bear Don't Walk IV et al., 2024)) provides race and ethnicity annotations for 17,281 sentences drawn from clinical notes in the MIMIC-III database. C-REACT contains real clinical text where race information appears in two forms: direct mentions that explicitly state race (e.g., *"Pt is 42 yo AA female"*) and indirect mentions that imply race through associated attributes such as spoken language or country of origin (e.g., *"Pt required a Spanish interpreter"*, *"Pt is recently immigrated from France"*). C-REACT provides five racial categories: White, Black/African American (Black/AA), Asian, Native American or Alaska Native, and Native Hawaiian or Other Pacific Islander. However, the dataset is highly imbalanced: zero patients labeled as Native Hawaiian or Other Pacific Islander were found, and only three patients labeled as Native American or Alaska Native. We therefore use three racial categories with sufficient representation: White, Black/AA, and Asian.

Models. We study three instruction-tuned LLMs of comparable scale from different geographic and cultural training contexts: Llama-3.1-8B-IT (Grattafiori et al., 2024) (US), Mistral-7B-IT-v0.3 (Jiang et al., 2023) (France), and Qwen2.5-7B-IT (Team, 2024) (China). This selection allows us to investigate whether models trained on data from different linguistic and cultural contexts encode racial information differently, given that conceptions of race and ethnicity vary across societies.

Model	Group	Neuron	Top tokens
Qwen2.5-7B	Asian	$\text{MLP.v}^{28}_{13406}$	Japanese, 日本, Japan, Tokyo
		$\text{MLP.v}^{25}_{15029}$	Chinese, China, 中国, Asian
	Black	MLP.v^{27}_{2240}	black, 黑, Black, 黑色
	Latino	MLP.v^{28}_{4781}	Latin, 拉丁, latino, Latina
		$\text{MLP.v}^{27}_{18125}$	Spanish, Hispanic, Chile, Mexican
	Native Am.	MLP.v^{25}_{3458}	native, Native, indigenous
		$\text{MLP.v}^{25}_{11197}$	colonial, colon, colony, imperial
	Middle Eastern	MLP.v^{28}_{9988}	Israel, Jerusalem, Hebrew, Zion
		MLP.v^{26}_{3012}	Jew, Jewish, Judaism, Rabbi
Mistral-7B	Asian	MLP.v^{32}_{4453}	Japanese, Korean, Taiwan, Asian
	Black	MLP.v^{32}_{5923}	Black, black, Negro, African
		$\text{MLP.v}^{30}_{12572}$	Black, blacks, 黑, dark
	Native Am.	MLP.v^{31}_{3440}	colonial, colon
		$\text{MLP.v}^{29}_{12205}$	native, ind, igenous
	Middle Eastern	MLP.v^{32}_{5573}	Jewish, Jews, Jerusalem, Israel
		MLP.v^{28}_{147}	Mediterranean, Turkish, Egyptian, Turkey
Llama-3.1-8B	Asian	MLP.v^{32}_{5691}	Chinese, China, Beijing, 中国
		$\text{MLP.v}^{31}_{14299}$	Li, yuan, Dong, Huang, Wang
	Black	MLP.v^{30}_{7195}	Jamaica, Caribbean, Trinidad, Jazz
		$\text{MLP.v}^{29}_{13826}$	African, Afro, negro, blacks
	Latino	MLP.v^{32}_{9242}	Spanish, Hispanic, Mexican, Argentine
	Native Am.	MLP.v^{32}_{6893}	colon, colonial, colonization, colonies
		MLP.v^{31}_{1186}	native, Native, -native, natives
	Middle Eastern	$\text{MLP.v}^{30}_{11051}$	Arab, Arabic, Saudi, Muslim
		MLP.v^{29}_{2750}	Islamic, Islam, mosques, Muhammad

Table 2: Top race-encoding neurons identified via cosine similarity with probe directions. **WARNING: Some tokens reflect harmful stereotypes.** Full results in Appendix A.3. Translations: 日本 (Japan), 中国 (China), 黑 (Black), 黑色 (Black color), 拉丁 (Latin).

6 Experiments and Results

6.1 Toxigen

Table 1 lists top tokens projected by each race direction. Across models, probes reach similar performance on ToxiGen (around 75% accuracy/macro-F1; Appendix A.1). These tokens capture various facets of racial encoding, including geography, religion, demographic labels, and cultural terms. Across all three models, the learned directions identify tokens that align closely with the target racial groups. This confirms that LLMs store clear racial representations within their residual streams.

Table 2 presents race encoding neurons identified within the final four MLP layers. These neurons reveal that LLMs decompose racial concepts into distinct semantic dimensions. Some neurons encode broad demographic terminology that directly names groups, such as Mistral-7B's MLP.v^{32}_{5923} (*Black, black, African*) and Asian neurons across all models (*Asian, Chinese, Japanese*), functioning as explicit demographic classifiers. Others encode race through associated attributes: Llama-3.1-8B's MLP.v^{32}_{5691} links Asian identity to geographic terms (*Chinese, China, Beijing*), while $\text{MLP.v}^{31}_{14299}$ captures Chinese last names (*Li, yuan, Dong, Huang, Wang*); Middle Eastern neurons project to religious and regional identifiers (*Jewish, Judaism, Islam, Jerusalem, Saudi*). We also observe neurons that encode historically harmful associations. Native American neurons across all three models project to colonial terminology (*colonial, colony, colonization*), and neurons for Black identity recover offensive racial terms that persist across models despite different training corpora.

Neuron Activation Analysis To verify that identified neurons selectively respond to their target racial groups, we measure mean activation values when processing test samples from each group. Figure 2 displays these activation patterns as heatmaps, where diagonal cells represent neurons processing their target group. The results confirm that most race encoding neurons activate more strongly for their target group than for others. This is most pronounced for Latino and Middle Eastern neurons: Llama-3.1-8B achieves activation values of 0.83 and 0.90 respectively, while Qwen2.5-7B reaches 0.71 and 1.23. Black neurons also demonstrate consistent selectivity across all three models, with positive diagonal values compared to near zero or negative off diagonal values. Asian and Native American neurons exhibit weaker selectivity, likely reflecting sparser representation in training data. Nevertheless, the overall diagonal pattern validates our neuron identification method: neurons selected via probe direction alignment do preferentially activate for their target groups, validating their role in demographic encoding.

6.2 C-REACT

We train separate probes on direct and indirect mentions to evaluate whether each type captures distinct representations. Direct-mention probes achieve higher accuracy/F1 than indirect probes (Appendix A.1). This likely reflects that direct cues are explicit and indirect data are sparser. Table 3 compares tokens projected by each probe type. Direct and indirect probes capture semantically distinct representations. Direct probes recover general racial and ethnic terminology (*Asian, African, black,* 白), while indirect probes recover specific countries and regions associated with each group. For instance, White indirect probe projects strongly to Russia and Eastern European terms across all models, reflecting dataset composition where Russian is the most frequent language among White patients. Similarly, Black or African American indirect probes recover Caribbean and African nations (*Haiti, Caribbean, Nigeria*). This divergence confirms that LLMs en-

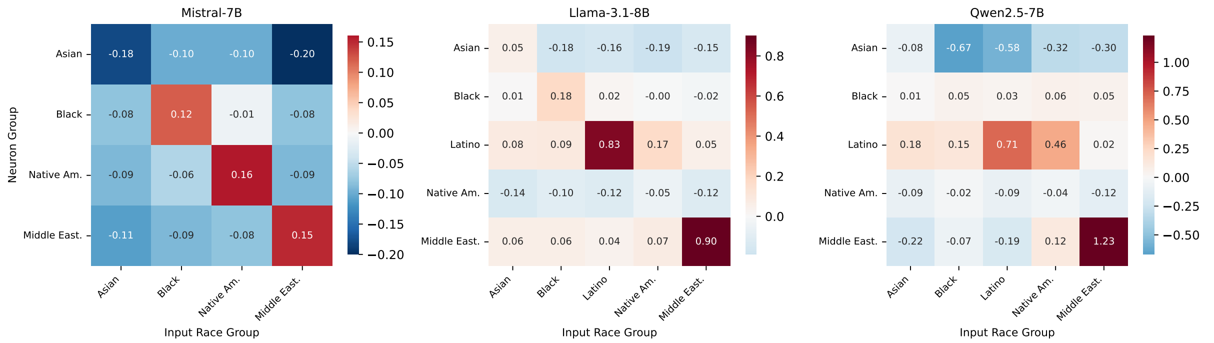

Figure 2: Mean activation values of race encoding neurons when processing text from each racial group (ToxiGen). Diagonal cells represent neurons processing their target group. Higher values (red) indicate stronger activation; lower values (blue) indicate suppression.

code race through multiple pathways: explicit demographic labels and associated geographic or linguistic attributes.

Tables 4 and 5 present race encoding neurons identified from direct and indirect probes respectively. The neuron projections mirror the probe token patterns: direct mention neurons recover explicit demographic terms (*Asian, Black, African,* 白 *(White)*), while indirect mention neurons recover geographic and cultural associations (*Russia, Moscow, Vietnam, Caribbean*). As in ToxiGen, we observe neurons encoding harmful associations. Qwen2.5-7B's MLP.v^{28}_{11088} projects to *racist, racism, Harlem, segregation*, and several Black/AA neurons encode terms related to slavery. The persistence of such encodings across both general and clinical domains confirms that these representations are inherited from pretraining rather than induced by domain specific data.

Model	Group	Direct	Indirect
Qwen2.5-7B	White	白	Russian, 俄罗斯, Russia
	Asian	Asian, Asia, Asians, 亚洲	Chinese, Xia, Tibetan, China
	Black/AA	African, 非洲, Africa, black	Hait, Haiti, Tropical, 热带
Mistral-7B	White		Moscow, Russian, Ukrain, Polish
	Asian	Asian, Taiwan, Japanese, Malays	Korea, Korean, Asian, Vietnam
	Black/AA	African, blacks, Negro, slavery	Caribbean, Cuba, Nigeria, Brazil
Llama-3.1-8B	White		Russia, Kremlin, Putin, Moscow
	Asian	Asian, Indonesian, Asia, Taiwanese	Cambodia, Chinese, wang, Buddhism
	Black/AA	black, African, Afro, negro	Haiti, Caribbean, Dominican, Bahamas

Table 3: Comparison of top tokens from direct (race/ethnicity) vs. indirect (language/country) mentions in C-REACT. **WARNING: Some tokens reflect harmful stereotypes.** Full results in Appendix A.4. Translations: 白 (white), 俄罗斯 (Russia), 非洲 (Africa), 热带 (tropical).

6.3 Neuron Intervention

To test whether the race encoding neurons we identified actually influence model behavior, we design an intervention experiment using C-REACT indirect mentions. Using a template prompt shown in Figure 3, we prompt each model to predict patient race based on clinical text containing only indirect cues such as language or country information, then

Model	Group	Neuron	Top tokens
Qwen2.5-7B	White	MLP.v^{28}_{16880} MLP.v^{27}_{17660}	英国, Dutch, French, Italian German, Germany, EU, euro
	Asian	MLP.v^{27}_{6943} MLP.v^{27}_{217}	Asian, Asia, Chinese, Indian Asian, 亚洲, Asia, アジア
	Black/AA	MLP.v^{28}_{11088} MLP.v^{27}_{2240}	racist, racism, Harlem, segregation black, 黑, Black, 黑色
Mistral-7B	White	MLP.v^{32}_{1606} MLP.v^{32}_{9831}	England, France, Europe, Switzerland European, Europe, EU, Euro
	Asian	MLP.v^{32}_{4453} MLP.v^{31}_{2346}	Japanese, Korean, Japan, Taiwan Japanese, anime, Japan, Tokyo
	Black/AA	MLP.v^{32}_{5923} MLP.v^{31}_{8715}	Black, 黑, Negro, African African, Africa, Kenya, Nigeria
Llama-3.1-8B	White	MLP.v^{31}_{9094}	White, white, WHITE, 白
	Asian	MLP.v^{32}_{5691} MLP.v^{29}_{5272}	Chinese, China, Beijing, Shanghai Asia, Asia, continent, 亚洲
	Black/AA	MLP.v^{30}_{7195} MLP.v^{29}_{13826}	Mississippi, Jamaica, Caribbean, Louisiana African, african, Afro, negro

Table 4: Top race-encoding neurons from C-REACT **direct** mentions (explicit race/ethnicity). **WARNING: Some tokens reflect harmful stereotypes.** Full results in Appendix A.5. Translations: 英国 (England), 亚洲 (Asia), アジア (Asia), 黑 (Black), 黑色 (Black color), 白 (White).

```
Please read the following sentence from a clinical
note. Based on the information contained within
the sentence, determine which of the following races
or ethnicities the patient belongs to. Please respond
with only one option.

Options: [White, Black or African American, Asian]

Clinical Note: "{text}"

Inferred Race or Ethnicity:
```

Figure 3: Prompt template for race prediction on C-REACT indirect mentions.

manipulate race encoding neurons to see if we can correct observed biases.

Baseline Classification. Table 6 shows misclassification patterns across the three models. The dominant error type varies by model: for Qwen2.5-7B and Llama-3.1-8B, White→Asian misclassification is the primary error, accounting for 75.0% and 95.6% of errors respectively. Llama's bias is the most pronounced, with 395 of 537 White patients incorrectly classified as Asian. In contrast, Mistral-7B shows a different pattern: its dominant error is White→Black/AA (76.2% of errors). This divergence suggests that models encode and apply

Model	Group	Neuron	Top tokens
Qwen2.5-7B	White	$\text{MLP.v}^{27}_{17660}$	German, Germany, 荷兰, euro
		MLP.v^{26}_{3382}	Russians, Russia, 俄罗斯, Moscow
	Asian	$\text{MLP.v}^{28}_{13406}$	Japanese, 日本, Tokyo, Osaka
		MLP.v^{25}_{2001}	Vietnam, Viet, Nguyen, Vietnamese
	Black/AA	$\text{MLP.v}^{25}_{10230}$	非洲, African, slave, slavery
		$\text{MLP.v}^{25}_{10739}$	African, Africa, Ghana, Nigerian
Mistral-7B	White	MLP.v^{32}_{2399}	Russian, Vlad, Moscow, Soviet
		MLP.v^{29}_{260}	Italian, Italy, Giovanni, Francesco
	Black/AA	MLP.v^{31}_{8715}	African, Africa, Kenya, Nigeria
Llama-3.1-8B	White	$\text{MLP.v}^{32}_{10606}$	Russian, Moscow, Soviet, Putin
		MLP.v^{29}_{4193}	Czech, Hungarian, Slovak, Budapest
	Asian	MLP.v^{32}_{5691}	Chinese, China, Beijing, Shanghai
		$\text{MLP.v}^{29}_{10616}$	Indian, India, Bollywood, Mumbai
	Black/AA	MLP.v^{29}_{6824}	tropical, jungle, Congo, Caribbean
		$\text{MLP.v}^{29}_{13826}$	African, african, Afro, negro

Table 5: Top race-encoding neurons from C-REACT **indirect** mentions (language/country). **WARNING: Some tokens reflect harmful stereotypes.** Full results in Appendix A.6. Translations: 荷兰 (Netherlands), 俄罗斯 (Russia), 日本 (Japan), 非洲 (Africa).

racial information differently during inference.

Error Type	Qwen2.5-7B	Mistral-7B	Llama-3.1-8B
White→Asian	**27**	4	**395**
White→Black/AA	4	**16**	5
Black/AA→White	4	0	0
Black/AA→Asian	0	0	10
Asian→White	1	1	1
Asian→Black/AA	0	0	2
Total Errors	36	21	413
Dominant Error %	**75.0%**	**76.2%**	**95.6%**

Table 6: Misclassification patterns on C-REACT indirect mentions. The dominant error type (bold) varies across models: White→Asian for Qwen and Llama, White→Black/AA for Mistral.

Activation Patterns. To investigate what drives these biases, we measure activation levels for all neuron groups across all classification outcomes (Table 7). We observe a strong correspondence between neuron groups exhibiting consistently high activation and the dominant error directions identified in Table 6.

For Qwen2.5-7B, which primarily misclassifies White patients as Asian, the Asian Direct neurons show consistently high positive activation regardless of ground truth or prediction. Similarly, Mistral-7B's tendency toward White → Black/AA errors aligns with elevated activity in Black/AA Direct neurons across most scenarios. Llama-3.1-8B presents a different pattern: while its dominant error is also White → Asian, Asian Indirect neurons show consistently high activation across scenarios.

These patterns reveal that neuron groups exhibiting high activation regardless of input correspond to the dominant misclassification directions, suggesting they may act as bias drivers.

Intervention Results. Having identified candidate bias drivers, we test whether suppressing these neurons can correct misclassification. Specifically, we evaluate the intervention using three amplification factors (5, 10, 20) to assess if these adjustments alter the model's predictions. Figure 4 compares correct prediction rates between Direct and Indirect neuron intervention, while Figure 5 shows the full prediction distribution across all conditions.

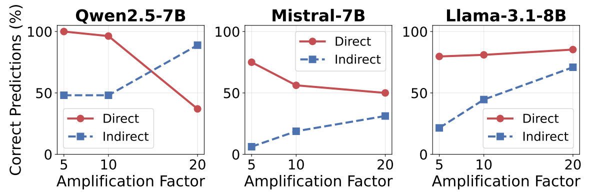

Figure 4: Correct prediction rates after neuron intervention across amplification factors. Direct neuron intervention (solid) generally outperforms Indirect intervention (dashed), demonstrating that neurons encoding explicit racial terminology have stronger causal influence on predictions.

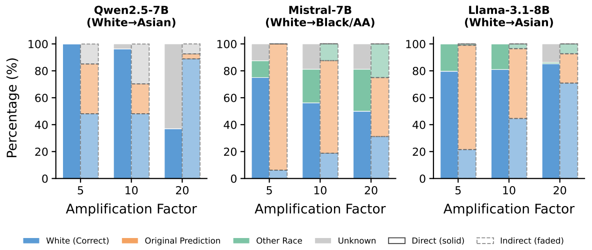

Figure 5: Prediction distribution after neuron intervention on misclassified samples. Direct intervention (solid bars) eliminates the original bias entirely (orange = 0%) across all models and factors, while Indirect intervention (faded bars) leaves residual bias. Higher amplification factors increase Unknown responses (gray)

Direct vs. Indirect Neurons. Across all three models, Direct neuron intervention demonstrates stronger causal efficacy than Indirect intervention (Figure 4). At factor 5, Direct intervention achieves substantially higher correct prediction rates across all models compared to Indirect intervention. More importantly, Direct intervention completely eliminates the original bias across all models and amplification factors (Figure 5), while Indirect intervention leaves residual bias. This gap is also consistent with the Llama-3.1-8B pattern: although the Asian Indirect group has higher mean activation in, but higher activation does not necessarily mean stronger causal influence on the final prediction. Indirect cues tend to reflect broad, proxy signals that can be supported by multiple parts of the network, so suppressing one indirect group may be partly compensated elsewhere and leads to a smaller behavioral change. By contrast, Direct neurons are more directly tied to producing the explicit race label, which makes intervening on them more effective.

Amplification Factor Selection. The choice of

Model	Neuron Group	Classification Outcome (Actual → Predicted)								
		W→W	W→B	W→A	B→W	B→B	B→A	A→W	A→B	A→A
Qwen2.5-7B	Asian Direct	**+6.04**	**+6.25**	**+5.85**	**+3.50**	**+5.71**	—	**+4.66**	—	**+3.86**
	Asian Indirect	−1.03	−1.02	−0.96	+0.31	−0.06	—	−1.55	—	−0.33
	Black/AA Direct	−0.74	−0.04	−0.40	+0.28	+0.83	—	+0.27	—	+0.27
	Black/AA Indirect	−0.57	+0.40	−0.63	+5.80	+3.30	—	−0.39	—	−0.78
	White Direct	−3.25	−3.49	−3.80	−2.06	−3.05	—	−4.87	—	−4.01
	White Indirect	+0.36	+0.31	−0.20	−0.90	−1.24	—	−0.33	—	−1.31
Mistral-7B	Asian Direct	+0.22	+0.24	+0.14	—	−0.15	—	+0.36	—	+0.19
	Black/AA Direct	**+0.82**	**+0.89**	**+0.64**	—	−0.07	—	**+1.00**	—	**+0.41**
	Black/AA Indirect	−0.56	−0.42	−0.55	—	+0.19	—	−0.52	—	−0.44
	White Direct	+0.10	−0.06	−0.02	—	−0.47	—	+0.33	—	+0.37
	White Indirect	−0.01	+0.24	−0.17	—	−0.27	—	−0.44	—	−0.32
Llama-3.1-8B	Asian Direct	−0.53	−0.59	−0.59	—	−0.57	−0.54	−0.41	−0.51	−0.55
	Asian Indirect	**+0.27**	**+0.28**	**+0.33**	—	**+0.36**	**+0.43**	**+0.29**	**+0.34**	**+0.47**
	Black/AA Direct	−1.07	−1.43	−1.20	—	−1.97	−1.66	−0.54	−1.60	−0.96
	Black/AA Indirect	−0.01	+0.01	−0.00	—	+0.04	+0.07	+0.03	+0.03	+0.03
	White Direct	+0.11	+0.37	+0.24	—	+0.51	+0.49	+0.62	+0.34	+0.70
	White Indirect	+0.22	+0.28	+0.32	—	+0.01	+0.04	−0.04	+0.01	−0.01

Table 7: Mean neuron activation scores on C-REACT indirect mention prompts, grouped by classification outcome (*actual → predicted*). **W/B/A** denote **White / Black/AA / Asian** (e.g., **W→B**: actual White, predicted Black/AA; **W→W**: correct). Positive values indicate the neuron group writes in the direction of its output vectors (strengthening the associated signal during generation), while negative values indicate writing in the opposite direction (weakening it). "—" indicates no samples for that outcome.

amplification factor involves a tradeoff between bias correction and model stability. We tested factors of 5, 10, and 20, representing increasingly aggressive intervention.

Across all three models, factor 5 yields the best balance. Qwen2.5-7B achieves 100% correct predictions with no Unknown outputs at factor 5, but destabilizes at factor 20 (63% Unknown responses). Mistral-7B reaches 75% correct predictions at factor 5, with higher factors increasingly shifting predictions toward Asian rather than the correct White label. Llama-3.1-8B performs similarly at factors 5 and 10 (approximately 80% correct), with factor 20 introducing Unknown responses. These results suggest that race encoding neurons exert strong influence on predictions, requiring only moderate suppression to alter model behavior. We adopt factor 5 as the default for subsequent analyses.

To illustrate, we provide example model outputs after intervention. Unknown outputs are often malformed, such as "[Yellow] [X] [Black or African" or "[Yellow] The provided options do not include". In contrast, successful corrections produce valid predictions such as "[White]".

7 Discussion

Our results indicate that the way race and ethnicity are internally represented in LLMs is central to understanding how demographic bias emerges across tasks. The first important finding is that **racial and ethnic concepts are distributed across many internal units rather than localized to a small set of neurons**. Importantly, this distribution is not arbitrary: models decompose race and ethnicity into multiple, interpretable semantic facets, such as explicit group labels and associated geographic or linguistic attributes. Across both ToxiGen and C-REACT, these facets appear as distinct internal representations rather than single abstract concepts (Tables 1, 2, 3). Notably, stereotype-related and historically harmful associations are present across models despite differences in training data and geographic origin, suggesting that bias mitigation cannot rely on a universal map of demographic features but requires model-specific localization.

Secondly, due to this representational structure, the same internal components can be reused across different task contexts, sometimes in ways that lead to biased behavior. **Neurons encoding racial concepts are present in all three models, yet their influence on predictions varies substantially depending on whether the input associates strongly with the proxy cues.** The same representations that benignly encode demographic information can drive biased predictions when activated in contexts where race is irrelevant.

Crucially, the presence of such representations is not inherently problematic. Rather, bias arises from how these representations are operationalized during inference. Our intervention *did not erase* racial knowledge from the models; instead, it modulated how this knowledge was reused in task-specific settings. This distinction is critical: pretrained representations reflect what models learn about the world, whereas task-dependent bias reflects when and how those representations are inappropriately applied.

8 Conclusion

We provide a mechanistic analysis of how race and ethnicity are represented and operationalized within LLMs. We show that demographic concepts are encoded as distributed, multi-faceted internal representations that can be selectively reused across tasks. These findings suggest that mitigating demographic bias in LLMs requires not only outcome-level interventions, but also a deeper examination of representational structure and task-dependent reuse.

Ethical Statement

This work examine how race and ethnicity are encoded with large language models, which necessarily involves sensitive content including stereotypes and historically offensive terminology. We present these findings to expose potential bias, not to amplify them. We acknowledge that racial categories are socially constructed and vary across cultures; our use of categories reflects the structure of the datasets rather than an endorsement of these taxonomies.

Acknowledgment

This work was supported by the U.S. National Library of Medicine (NLM), National Institutes of Health, under award number R00LM014308.

References

Hiba Ahsan, Arnab Sen Sharma, Silvio Amir, David Bau, and Byron C Wallace. 2025. Elucidating mechanisms of demographic bias in llms for healthcare. *arXiv preprint arXiv:2502.13319*.

Hiba Ahsan and Byron C Wallace. 2025. Can saes reveal and mitigate racial biases of llms in healthcare? *arXiv preprint arXiv:2511.00177*.

Oliver Bear Don't Walk IV, Adrienne Pichon, Harry Reyes Nieva, Tony Sun, Jaan Li, Joshua Winston Joseph, Sivan Kinberg, Lauren R. Richter, Salvatore Crusco, Kyle Kulas, Shaan Ahmed, Daniel Snyder, Ashkon Rahbari, Benjamin Ranard, Pallavi Juneja, Dina Demner-Fushman, and Noemie Elhadad. 2024. C-REACT: Contextualized race and ethnicity annotations for clinical text. PhysioNet. Version 1.0.0.

Nathaniel Demchak, Xin Guan, Zekun Wu, Ziyi Xu, Adriano Koshiyama, and Emre Kazim. 2024. Assessing bias in metric models for llm open-ended generation bias benchmarks. *arXiv preprint arXiv:2410.11059*.

Kathleen Fraser and Svetlana Kiritchenko. 2024. Examining gender and racial bias in large vision–language models using a novel dataset of parallel images. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 690–713, St. Julian's, Malta. Association for Computational Linguistics.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021a. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021b. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, and et al. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Xin Guan, Nate Demchak, Saloni Gupta, Ze Wang, Ediz Ertekin Jr., Adriano Koshiyama, Emre Kazim, and Zekun Wu. 2025. SAGED: A holistic bias-benchmarking pipeline for language models with customisable fairness calibration. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3002–3026, Abu Dhabi, UAE. Association for Computational Linguistics.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3326, Dublin, Ireland. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Adam Karvonen and Samuel Marks. 2025. Robustly improving llm fairness in realistic settings via interpretability. *arXiv preprint arXiv:2506.10922*.

Michelle Kim, Junghwan Kim, and Kristen Johnson. 2023. Race, gender, and age biases in biomedical masked language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11806–11815, Toronto, Canada. Association for Computational Linguistics.

Asaf Levartovsky, Mahmud Omar, Girish N Nadkarni, Uri Kopylov, and Eyal Klang. 2025. Sociodemographic bias in large language model–assisted gastroenterology. *JAMA Network Open*, 8(9):e2532692–e2532692.

Yichen Li, Zhiting Fan, Ruizhe Chen, Xiaotang Gai, Luqi Gong, Yan Zhang, and Zuozhu Liu. 2025. Fairsteer: Inference time debiasing for llms with dynamic activation steering. *arXiv preprint arXiv:2504.14492*.

Afrozah Nadeem, Mark Dras, and Usman Naseem. 2025. Context-aware fairness evaluation and mitigation in llms. *arXiv preprint arXiv:2510.18914*.

nostalgebraist. 2020. interpreting GPT: the logit lens. https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens. Accessed: 2025-12-23.

Raphael Poulain, Hamed Fayyaz, and Rahmatollah Beheshti. 2024. Bias patterns in the application of llms for clinical decision support: A comprehensive study. *arXiv preprint arXiv:2404.15149*.

Noam Shazeer. 2020. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*.

Bryan Chen Zhengyu Tan and Roy Ka-Wei Lee. 2025. Unmasking implicit bias: Evaluating persona-prompted LLM responses in power-disparate social scenarios. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1075–1108, Albuquerque, New Mexico. Association for Computational Linguistics.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Angelina Wang, Michelle Phan, Daniel E. Ho, and Sanmi Koyejo. 2025. Fairness through difference awareness: Measuring *Desired* group discrimination in LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6867–6893, Vienna, Austria. Association for Computational Linguistics.

Zeping Yu and Sophia Ananiadou. 2025. Understanding and mitigating gender bias in llms via interpretable neuron editing. *arXiv preprint arXiv:2501.14457*.

Travis Zack, Eric Lehman, Mirac Suzgun, Jorge A Rodriguez, Leo Anthony Celi, Judy Gichoya, Dan Jurafsky, Peter Szolovits, David W Bates, Raja-Elie E Abdulnour, and 1 others. 2024. Assessing the potential of gpt-4 to perpetuate racial and gender biases in health care: a model evaluation study. *The Lancet Digital Health*, 6(1):e12–e22.

Angela Zhang, Mert Yuksekgonul, Joshua Guild, James Zou, and Joseph C Wu. 2023. Chatgpt exhibits gender and racial biases in acute coronary syndrome management. *arXiv preprint arXiv:2311.14703*.

Yachao Zhao, Bo Wang, Yan Wang, Dongming Zhao, Ruifang He, and Yuexian Hou. 2025. Explicit vs. implicit: Investigating social bias in large language models through self-reflection. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 1–12, Vienna, Austria. Association for Computational Linguistics.

Hanzhang Zhou, Zijian Feng, Zixiao Zhu, Junlang Qian, and Kezhi Mao. 2024. Unibias: Unveiling and mitigating llm bias through internal attention and ffn manipulation. *Advances in Neural Information Processing Systems*, 37:102173–102196.

A Appendix

A.1 Probe Performance

Model	Acc. (%)	Macro-F1
Qwen2.5-7B-IT	74.88	0.74
Llama-3.1-8B-IT	77.98	0.77
Mistral-7B-IT	75.20	0.74

Table 8: Test set performance of multi-class linear probes trained on **ToxiGen** (5-way: asian/black/latino/native_american/middle_eastern). We report Accuracy and Macro-F1.

	Direct		Indirect	
Model	Acc. (%)	Macro-F1	Acc. (%)	Macro-F1
Qwen2.5-7B-IT	90.36	–	81.61	0.79
Llama-3.1-8B-IT	93.98	0.85	80.46	0.80
Mistral-7B-IT	90.66	0.70	75.86	0.74

Table 9: Test set performance of linear probes trained on **C-REACT** (3-way: White/Black/AA/Asian), using Direct vs. Indirect prompt variants.

A.2 Probe Token Projections (ToxiGen)

Table 10 presents the complete top-20 tokens projected by each race direction probe for all three models.

A.3 Race-Encoding Neurons (ToxiGen)

From Table 11 to Table 13 presents the complete list of race-encoding neurons identified from ToxiGen for the three models.

A.4 Probe Token Projections (C-REACT)

Table 14 and Table 15 present the complete top-20 tokens projected by each race direction probe for direct and indirect mentions respectively.

A.5 Race-Encoding Neurons (C-REACT Direct)

Table 16 presents the complete list of race-encoding neurons identified from C-REACT direct mentions (explicit race/ethnicity).

A.6 Race-Encoding Neurons (C-REACT Indirect)

Table 17 presents the complete list of race-encoding neurons identified from C-REACT indirect mentions (language/country).

Model	Group	Top 20 Tokens
Qwen2.5-7B	Asian	Asian, 亚洲, Asian, Chun, Chinese, CJK, Yuan, 东亚,ˈɔ, chinese, 霹, 恶, Chinese, ActivityCreated, chner, Chu, ObjectContext, lobals, Hong, chin
	Black	=”{}/>, 相应, orsk, 台阶, 截, либо, 没有想到, Rudd, 钮, .BLL, .GetDirectoryName,]interface, setError, 鹕, Beaut, 夙, :async, 翾, 绶, SingleOrDefault
	Latino	nesty, Mex, .Span, .Shared, 垭, .”/, /items, SetLastError, ucher, getTime, .onclick, enticate, ERVE, mex, .Stretch, mexico, tü, evenodd, asher, 年之久
	Native Am.	natives, native, .Native, native, Native, Native, _native, indigenous, coma, .native, -native, ””}, 達, ings, NAV, ITERAL, reservation, INGS, 킴, 葆
	Middle Eastern	avigate, wargs, 箴, bsolute, arma, /xhtml, %”><, elic, .clearRect, арамет, ouv, 役, 铈, λ, _Tick, achts, wares, hill, kap, lambda
Mistral-7B	Asian	Chinese, Asian, China, Korean, Taiwan, Hong, inese, Japanese, ’, Shanghai, Beijing, Asia, omi,ˈ, ß, bt, Roose, MMMM, omy, agh
	Black	/******/, corpor, Coff, Email, black, uh, African, publicly, spell, ta, white, email, black, Black, Palm, Parl, riel, Black, mask, rele
	Latino	Mexico, Salvador, jd, Colombia, Chile, ¡, Colomb, ass, Mexican, Skip, ranch, Mex, Santiago, ully, Argent, jal, Partido, sketch, skip, aca
	Native Am.	Indians, trib, Native, tribes, Indian, medicine, Native, ye, Medicine, AD, ingle, minister, ilo, Inga, tribe, iga, unda, Trib, uti, erset
	Middle Eastern	Islamic, Palestinian, Muhammad, Muslim, Hass, Israel, esa, Naj, Turkish, Jewish, Bere, Turk, Muslims, jew, Arab, hash, Arabia, Jews, triple, Middle
Llama-3.1-8B	Asian	Asian, Asian, Mandarin, CJK, erse, ležit, ibold, Ding, inese, asian, rysler, china, ern, Chinese, Asia, ihan, Asians, ertainment, ź, Euras
	Black	CELER, Black, isay, Black, _black, urgeon, -black, adeon, arp, cie, anja, .black, 黑, ورا, -hit, /black, جل, ouse, بلند, وتر
	Latino	ucwords, 妙, udos, Buchanan, .si, ”, oppos, aside, i, City, rip, Mundo, odí, _BORDER, mant, cloak, _gid, .ul, iyon, ivery
	Native Am.	Native, Native, native, natives, native, -native, Indians, _Native, _native, indigenous, .native, /native, Indigenous, tribes, RIPT, Indian, reservation, .Native, ative, Reservation
	Middle Eastern	Mev, LLL, šem, Islamic, Middle, ohon, WW, acket, Aber, esso, profits, сов, owy, uzz, removed, ’ ’, станов, ISIL, Christian, etus

Table 10: Full top-20 probe token projections for all models (ToxiGen). Tokens are listed in descending order of projection score.

Model	Group	Neuron	Top Tokens
Qwen2.5-7B	Asian	MLP.v^{28}_{13406}	在日本, Japanese, 日本, Japanese, Japan, Japan, 日本人, Tokyo, japan, 东京, japan, 일본, 日军, japanese, japon, Nhật, 日本の, Tok, Hiro, jap, Osaka
		MLP.v^{28}_{5983}	Chi, chi, Chin, Chi, chin, chi, _chi, Xi, Hu, xi, Hu, xi, 人民币, Xi, chin, Huawei, 华为, 华夏, (xi, hu
		MLP.v^{27}_{8641}	ch, 是中国, Ch, 为中国, Chinese, 成为中国, 由中国, 与中国, 对中国, China, china, _ch, 在中国, China, Ch, china, Chinese, -Ch, CH, chin
		MLP.v^{27}_{217}	Asian, 亚洲, Asia, Asian, Asians, Asia, asian, asia, アジア, 亚, 亞, Asi, asia, asiat, asi, 亚太, เอเช, _As, asi, AS
		MLP.v^{25}_{15029}	Chinese, China, China, Chinese, chinese, china, china, 中国, 中国的, -China, 중국, الصين, 中國, Asian, 在中国, 是中国, 亚洲, 由中国, Asian, Asia
	Black	MLP.v^{27}_{2240}	black, 黑, black, Black, Black, 黑色, -black, BLACK, BLACK, _black, blacks, .black, /black, , ブラック, .Black, 黑, 黑白, _BLACK, blacklist
	Latino	MLP.v^{28}_{4781}	Latin, Latin, 拉丁, latin, latin, LATIN, latino, 巴西, Latino, LAT, Latina, latina, Lat, Brazil, _LAT, 阿根廷, Lat, Brazil, lat, LAT
		MLP.v^{28}_{9876}	Spanish, 西班牙, 湘, Portuguese, Hispanic, Spanish, 贵州, Brazilian, ¡, ¡, 贵州省, 黔, 葡萄牙, Juan, Chile, spanish, 贵阳, Brazil, Juan, Spain
		MLP.v^{27}_{18125}	Spanish, 西班牙, Hispanic, Spanish, Chile, Mexican, 贵州, 墨西哥, Spain, Brazilian, Juan, Mex, Juan, 巴西, 阿根廷, 贵州省, Mexico, 湘, spanish, Mexico
	Middle East	MLP.v^{28}_{9988}	以色列, Israel, Jerusalem, Israel, Israeli, Noah, Hebrew, Moses, Zion, -Israel, Palestine, Biblical, biblical, Israeli, Israelis, Luke, Zionist, Palestinian, Jer, Nathan
		MLP.v^{27}_{9840}	伊斯兰, 穆, Mu, Ali, Ali, Must, mu, 阿拉伯, Mu, Ah, MU, Muhammad, Moh, Fat, MU, Must, .mu, _mu, Ab, Muslims
		MLP.v^{27}_{8005}	宗教, religious, religious, relig, religion, Religious, Religion, religions, 基督教, 佛教, theological, spiritual, 信仰, Christian, prayer, secular, spirituality, 耶稣, 虔, 圣经
		MLP.v^{26}_{3012}	犹, Jew, Jewish, 猶, Jews, Judaism, ев, jewish, Juda, Hebrew, евр, 以色列, Israel, kosher, synagogue, -J, ewish, jew, Rabbi, Israeli
	Native Am.	MLP.v^{25}_{3458}	native, native, 自然, natural, Native, Native, 本土, 天然, -native, indigenous, natural, natives, naturally, Natural, /native, nat, Natural, _native, .native, 自发
		MLP.v^{25}_{7087}	native, 故, native, Native, 故乡, 本土, 家乡, -native, Native, natives, hometown, .native, _native, home, homeland, , 祖国, /native, birth, quê
		MLP.v^{25}_{11197}	殖民, colonial, colon, colony, colon, colonies, Colonial, Colon, Colon, Colony, imperial, Imperial, olon, icolon, -col, 帝国, ocol, colonization, dec, _COL

Table 11: Full top-20 tokens for race-encoding neurons in Qwen2.5-7B (ToxiGen).

Model	Group	Neuron	Top Tokens
Mistral-7B	Asian	MLP.v^{32}_{4453}	Japanese, Korean, Kol, Japan, Kor, Kaz, Sak, Taiwan, Korea, Nak, Kur, Kom, Bangl, Pakistan, Kab, Kon, Pak, Tibet, Ku, Asian
	Black	MLP.v^{32}_{5923}	Black, black, Black, black, blacks, 黑, 黒, Negro, BL, blk, African, BL, чер, Afr, Dark, ч, dark, Чер, лек, 블
		MLP.v^{30}_{12572}	Black, black, Black, black, blacks, 黑, BL, 黒, BL, Negro, blk, чер, dark, African, Afr, 블, лек, dark, Чер, ч
		MLP.v^{29}_{13186}	black, black, Black, blacks, Black, white, 黑, white, whites, African, 黒, Negro, 白, Afr, White, White, BL, чер, Af, brown
	Middle East	MLP.v^{32}_{5573}	Jewish, Jews, Jerusalem, Israel, Israeli, JS, js, JavaScript, JS, Palest, JSON, JSON, js, json, Json, json, ajax, Palestinian, javascript, jQuery
		MLP.v^{29}_{3203}	Mediterranean, Turkish, Egyptian, Turkey, Israeli, Iran, Jordan, Israel, Arab, Egypt, Palestinian, Iraq, Tur, Leb, Greek, Jerusalem, Saudi, Gulf, Arabia, Palest
	Native Am.	MLP.v^{31}_{3440}	imperial, fasc, Imperial, colonial, militar, Kent, /******/, colon, popul, racist, gent, antal, neo, provinc, Emitter, colon, carriage, TES, omena, ounds
		MLP.v^{29}_{12205}	native, native, Native, Native, nat, ind, ab, igenous, Ma, nat, Ab, Ma, abor, Nat, primitive, Nav, born, Mas, nav, Ab

Table 12: Full top-20 tokens for race-encoding neurons in Mistral-7B (ToxiGen).

Model	Group	Neuron	Top Tokens
Llama-3.1-8B	Asian	MLP.v^{32}_{5691}	Chinese, China, Chinese, China, chinese, china, -China, Beijing, 中国, Çin, 중국, Shanghai, 中國, china, چین, 中国, Zhang, Jiang, Tencent, Guang
		MLP.v^{31}_{14299}	Li, yuan, Dong, Huang, Liu, Wang, Chen, Yang, Tian, Zhou, Ding, Wu, dong, dong, Feng, wang, Zhang, Qin, Jiang, Guang
		MLP.v^{30}_{12566}	East, East, EAST, Eastern, east, -East, eastern, 東, Eastern, east, 东, -east, 東, Doğu, 东, Đông, شرق, vých, вост, orient
	Latino	MLP.v^{32}_{9242}	Spanish, Hispanic, Spanish, spanish, Mexican, Argentine, Santiago, Mexico, Puerto, Madrid, pesos, Chavez, Ecuador, Spain, Juan, Hispan, Colombian, Hispanics, Carlos, Chile
	Native Am.	MLP.v^{32}_{6893}	colon, Colon, Colon, colon, colonial, Colonial, colonization, colonies, Colony, colony, Colonel, 殖, olon, icolon, Colin, OLON, kol, COL, -Col, colore
		MLP.v^{31}_{1186}	native, energy, Energy, native, energy, Native, Native, -native, Energy, natives, _native, .native, -energy, -energy, supply, 能源, culture, /native, nice, Higher
	Middle East	MLP.v^{30}_{11051}	Arab, Middle, arab, Arabic, Arabs, Middle, Arabian, Arabia, Saudi, Yemen, Egypt, Cairo, Palestinian, Bahrain, Kuwait, Egyptian, Saudi, Riyadh, عرب, Libya
		MLP.v^{29}_{2750}	Islamic, Islam, Islam, Arabic, Islamic, mosques, mosque, Muslim, Muhammad, Muslim, Muslims, Abdullah, Mosque, Quran, Islamist, Mohammad, Ramadan, isl, мус, muslim
	Black	MLP.v^{30}_{7195}	Mississippi, Jamaica, Jama, Caribbean, Louisiana, Trinidad, Haiti, LSU, Ghana, Hait, Baton, Harlem, Nigeria, negro, Negro, Jazz, Bahamas, Memphis, Nigerian, Zimbabwe
		MLP.v^{29}_{13826}	African, african, Africans, Africa, Africa, afr, Afro, Afr, frican, negro, Af, Afrika, Af, africa, Negro, blacks, black, af, frica, Blacks

Table 13: Full top-20 tokens for race-encoding neurons in Llama-3.1-8B (ToxiGen).

Model	Group	Top 20 Tokens
Qwen2.5-7B	White	白, generado, -Nazi, ksam, onn, pl, owl, ucas, *&, '#{, .toByteArray, _EXTENSIONS, avras, onFailure, 镪, ski, Nederland, 好运, Luft, Ski
	Asian	Asian, Asian, Asians, Asia, 亚洲, Asia, asia, Asi, アジア, asiat, asian, 舢, Taiwan, Singapore, 舶, Singapore, Tai, Canton, Taiwanese, ракти
	Black/AA	african, African, 非洲, -AA, Africans, Africa, ienda, Africa, السود, 疟, mongo, /black, aina, AA, .BL, .Black, .Mongo, Afro, Nigerian, Black
Mistral-7B	White	ogle, ASC, cip, Kurt, heid, unächst, kle, þ, criptor, ór, NOP, och, zym, hid, eu, cow, cí, zens, vas, awa
	Asian	Asian, Taiwan, Japanese, Hong, Malays, Japan, Singapore, Chinese, Malaysia, Pak, Indones, Korean, Tai, Asia, Philippines, Philipp, Tokyo, Tok, jap, Sri
	Black/AA	African, Afr, Africa, blacks, Niger, Jama, Negro, Nigeria, Black, Af, ament, slavery, sist, black, external, external, Af, AA, lando, slave
Llama-3.1-8B	White	ithe, lan, Fres, ycz, hs, Wake, Bread, bread, itra, bairro, Fans, 1, Josh, wie, Jackets, ジオ, Marina, γι, Josh, avan
	Asian	Asian, Asian, Asians, asian, Indonesian, Asia, Taiwanese, Asia, asiat, Vietnamese, Japanese, Korean, Oriental, Malaysian, Buddhist, 亚洲, orean, Filipino, Chinese, Indones
	Black/AA	black, African, Black, african, black, frican, Afro, /black, negro, Negro, 黑, blacks, BLACK, zwarte, 黑, Black, .Black, _black, Africa, Africans

Table 14: Full top-20 probe token projections for C-REACT direct mentions (explicit race/ethnicity).

Model	Group	Top 20 Tokens
Qwen2.5-7B	White	RaisePropertyChanged, Russian, 俄罗斯, 橼, 坧, ocaly, Russia, _backward, Kremlin, Russian, *)((, .defaultValue, RU, russian, egrator, ipsis, Rad, UCE, abis, ENCIL
	Asian	.Dao, 华人, Chinese, 嵴, .insertBefore, Xia, chinese, ettel, Chinese, Tibetan, China, Wong, /apache, 华南, dao, dx, 갔, Jun, utenant, stylesheet
	Black/AA	Hait, Haiti, Tropical,)__, eneg, %;", ($., 埴, %;">, 彗, tropical, 垓, ⌈, .iterator, 热带, %;">,];//, loo, estring, *****
Mistral-7B	White	Moscow, Russian, Ukrain, Ukraine, Russians, Polish, Ukr, Russia, vod, Soviet, Vlad, russ, Kaz, Bulgar, Lieutenant, Mik, icz, Roma, dou, Serge
	Asian	Korea, Korean, Asian, trag, apis, Shan, Vietnam, gram, sg, apore, ga, Aires, Assembly, WD, lag, Nor, Viet, Schw, gan, cent
	Black/AA	Caribbean, Jama, Braz, Af, Cuba, Niger, Nigeria, Cub, Bah, São, Core, Currency, Bras, island, Hy, hur, Curt, Af, Brazil, mont
Llama-3.1-8B	White	Russia, Kremlin, Russian, Russians, Putin, Moscow, Ukraine, Putin, Ukrainian, Russian, Russia, Ukrain, russian, Rus, Belarus, Rus, russe, russ, Kiev, Rusya
	Asian	Cambodia, Asian, Chung, Cheng, Chinese, Camb, Kang, wang, Chinese, chinese, Buddhism, asian, Hong, Bangalore, Buddhist, Bang, Asian, Malaysia, Korean, epr
	Black/AA	Hait, Haiti, Caribbean, Maurit, Dominican, Bahamas, hait, Cre, Cre, Jama, ibbean, Trinidad, ingt, Jean, Cameroon, François, Maurice, Jean, анка, Santo

Table 15: Full top-20 probe token projections for C-REACT indirect mentions (language/country).

Model	Group	Neuron	Top 20 Tokens
Qwen2.5-7B	White	MLP.v^{28}_{16880}	英国, Dutch, French, 法国, 意大利, Italian, German, 荷兰, France, British, Belgian, European, Britain, Germany, 德国, Spanish, Italy, Spain, 欧洲, French
		MLP.v^{27}_{17660}	German, 万欧元, ドイツ, 德国, German, Germany, Germans, EU, 荷兰, euro, €, 欧盟, Germany, Swiss, 欧元, Euro, Dutch, 瑞士, 欧, Switzerland
		MLP.v^{25}_{4157}	UEFA, 地中海, Mediterranean, Cyprus, Pane, Roman, 欧盟, Euro, Roman, chalk, 希腊, Europe, UES, Rom, 罗马, EU, Kn, Europe, 橡, 騰
		MLP.v^{25}_{8669}	fod,].', []>, sterdam, Ừ, Spain, Gaines, Spain, 比利时, , WRITE, 英国, europe, <!–[, ======, %@", 示, 西班牙, _IMPORTED, euros
	Asian	MLP.v^{28}_{13406}	在日本, Japanese, 日本, Japanese, Japan, Japan, 日本人, Tokyo, japan, 东京, 일본, 日军, japanese, japon, Nhật, Tok, Hiro, 日本の, Osaka, jap
		MLP.v^{28}_{16570}	习, 習, 惯, habit, 习惯, 的习惯, habitual, Habit, 習慣, habit, habits, accustomed, 习近平, 习惯了, 慣, 习近, 惯例, 习俗, 习近平总, 总书记
		MLP.v^{27}_{6943}	Asian, Asian, 亚洲, Asia, Chinese, Asia, Chinese, China, 印度, Indian, India, India, Asians, China, Indian, chinese, 中國, Oriental, Japanese, asian
		MLP.v^{27}_{217}	Asian, 亚洲, Asia, Asian, Asians, Asia, asian, asia, アジア, 亚, 亞, Asi, asia, asiat, asi, 亚太, เอเช, _As, asi, AS
		MLP.v^{26}_{5187}	Chinese, Italian, Chinese, Irish, Italian, Asian, Japanese, chinese, Mexican, Vietnamese, Korean, German, Greek, Asian, Portuguese, Latin, Indian, African, German, italian
		MLP.v^{26}_{8828}	亚洲, Asian, Asia, Asian, Latin, 欧, Asia, 亚太, 欧洲, 非洲, 东亚, Southeast, Europe, 拉丁, European, 东南亚, 美洲, アジア, Euras, Asians
		MLP.v^{25}_{15029}	Chinese, China, China, Chinese, chinese, china, china, 中国, 中国的, -China, 중국, الصين, 中國, Asian, 在中国, 是中国, 亚洲, 由中国, Asian, Asia
	Black/AA	MLP.v^{28}_{11088}	racial, racially, racist, 种族, racism, racial, Harlem, segregated, segregation, Rac, interracial, Ethnic, Afro, hetto, ethnic, rac, Interracial
		MLP.v^{28}_{10048}	μ, ctx, ctx, μ, (ctx, Ken, Ken, Ctx, Cape, _ctx, .ctx, ctx, μ, _CTX, 南非, context, CTX, Kenya, μ, context
		MLP.v^{27}_{2240}	black, 黑, black, Black, Black, 黑色, -black, BLACK, BLACK, _black, blacks, .black, /black, ブラック, .Black, 黑, 黑白, _BLACK, blacklist
		MLP.v^{27}_{16596}	Ken, Ken, 肯, Nam, Liber, Bot, Tanz, Burk, Bur, Chad, ken, 赞, Nam, Kenya, Ug, Jordan, ken, Maur, Per, Jordan
		MLP.v^{26}_{18261}	黑, black, black, 黑色, Black, -black, Black, _black, .black, dark, 黑, BLACK, /black, 黑暗, blacks, .Black, BLACK, đen, darken
		MLP.v^{26}_{1091}	黑, black, 黑色, Black, Black, black, -black, BLACK, BLACK, /black, blacks, _black, dark, Blacks, Dark, blacklist, 黑暗, 黑, .Black
		MLP.v^{25}_{10230}	Af, af, 非洲, Af, African, slave, AF, Africans, AF, african, slavery, 奴隶, 奴, slaves, Slave, slave, Afro, Slave, Africa, afr
		MLP.v^{25}_{10739}	非洲, African, Africa, Africa, Africans, african, Ghana, Nigerian, Nigeria, afr, Kenya, Nairobi, Niger, -Saharan, Nd, Tanzania, Uganda, africa, Nz, Lagos
Mistral-7B	White	MLP.v^{32}_{1606}	England, France, Europe, Switzerland, Britain, Holland, Spain, Australia, España, Austria, Germany, Denmark, Francia, Canada, Europa, America, Wales, Deutschland, Belgium, Sweden
		MLP.v^{32}_{12760}	€, €, Finn, Mediterranean, Belgium, Italy, Finland, Belg, Luxem, Netherlands, Milan, Italian, Portugal, Denmark, Czech, Madrid, Norway, Sweden, Spain, Amsterdam
		MLP.v^{32}_{9831}	European, Eu, Europe, Europe, EU, eu, Europ, Euro, europe, eu, Europa, europ, Евро, urope, UEFA, €, urop, , €, вро
		MLP.v^{30}_{1521}	Pennsylvania, Michigan, Kansas, Philadelphia, Verm, ahl, Detroit, rus, ники, zel, Vic, Buc, ru, dispos, Hein, gotta, xa, una, anne, onic
	Asian	MLP.v^{32}_{4453}	Japanese, Korean, Kol, Japan, Kor, Kaz, Sak, Taiwan, Korea, Nak, Kur, Kom, Bangl, Pakistan, Kab, Kon, Pak, Tibet, Ku, Asian
		MLP.v^{31}_{2346}	Japanese, anime, Japan, Tokyo, jap, Jap, Korean, oji, apan, Tok, Asian, Taiwan, Israeli, Chinese, Korea, Indones, aku, Viet, Sak, oshi
		MLP.v^{30}_{2137}	Wil, mal, Mal, Ker, Mal, mal, Malays, Malaysia, ker, <\, $\, -\, $\|\, Carm, ($\, /\, mals, -\, +\, >\
		MLP.v^{30}_{8986}	asis, asi, Ash, lic, ox, Asia, yk, ke, ash, ym, as, ass, Async, asia, ais, eper, ob, ash, ek
	Black/AA	MLP.v^{32}_{5923}	Black, black, Black, black, blacks, 黑, 黒, Negro, BL, blk, African, BL, чер, Afr, Dark, ч, dark, Чер, лек, 블
		MLP.v^{31}_{8715}	African, Africa, Afr, Niger, Af, frica, Kenya, Af, Nigeria, fr, Negro, af, Afghan, blacks, FR, airo, Caribbean, Johannes, ₦, Jama
		MLP.v^{30}_{3398}	African, Africa, Afr, Af, Af, Asian, fr, Niger, Aaron, Afghan, Kenya, frica, af, AF, Mexican, AF, European, Nigeria, af, Egyptian
		MLP.v^{30}_{12572}	Black, black, Black, black, blacks, 黑, BL, 黒, BL, Negro, blk, чер, dark, African, Afr, 블, лек, dark, Чер, ч
		MLP.v^{29}_{13186}	black, black, Black, blacks, Black, white, 黑, white, whites, African, 黑, Negro, 白, Afr, White, White, BL, чер, Af, brown
Llama-3.1-8B	White	MLP.v^{31}_{9094}	White, white, White, white, WHITE, -white, 白, WHITE, _white, .White, whites, 白, .white, _WHITE, Whites, beyaz, :white, trắng, سفید, .WHITE
	Asian	MLP.v^{32}_{5691}	Chinese, China, Chinese, China, chinese, china, -China, Beijing, 中国, Çin, 중국, Shanghai, 中國, china, چین, 中国, Zhang, Jiang, Tencent, Guang
		MLP.v^{31}_{14299}	Li, yuan, Dong, Huang, Liu, Wang, Chen, Yang, Tian, Zhou, Ding, Wu, dong, dong, Feng, wang, Zhang, Qin, Jiang, Guang
		MLP.v^{29}_{5272}	Asia, Asia, Asian, Asian, asia, continent, Africa, continental, asian, Africa, asia, Asians, Continental, 亚洲, Asi, Europe, Continent, Latin, asi, Europe
	Black/AA	MLP.v^{30}_{7195}	Mississippi, Jamaica, Jama, Caribbean, Louisiana, Trinidad, Haiti, LSU, Ghana, Hait, Baton, Harlem, Nigeria, negro, Negro, Jazz, Bahamas, Memphis, Nigerian, Zimbabwe
		MLP.v^{30}_{3868}	black, Black, BLACK, Black, black, /black, BLACK, -black, 黑, blacks, _black, .Black, blacklist, 黒, Blacks, .black, 黑, Blackburn, _BLACK, blackout
		MLP.v^{29}_{13826}	African, african, Africans, Africa, Africa, afr, Afro, Afr, frican, negro, Af, Afrika, Af, africa, Negro, blacks, black, af, frica, Blacks
		MLP.v^{29}_{6824}	tropical, jungle, Tropical, Jungle, jung, ropical, trop, Fiji, Belize, mango, Congo, Jama, Caribbean, Hait, Honduras, Mango, BSON, Haiti, Safari, Jamaica

Table 16: Full list of race-encoding neurons identified from C-REACT direct mentions.

Model	Group	Neuron	Top 20 Tokens
Qwen2.5-7B	White	MLP.v^{28}_{8780}	Maryland, Pennsylvania, Baltimore, Ohio, Wisconsin, Maine, Minnesota, Illinois, Pittsburgh, Iowa, Ohio, Connecticut, Philadelphia, Chicago, Michigan, Milwaukee, Nebraska, Detroit, Seattle, 江苏省
		MLP.v^{28}_{9988}	以色列, Israel, Jerusalem, Israel, Israeli, Noah, Hebrew, Zion, Moses, -Israel, Palestine, Israeli, Israelis, Biblical, biblical, Zionist, Luke, Palestinian, Jer, Nathan
		MLP.v^{28}_{4318}	Ohio, Columbus, sz, Cleveland, Ohio, (sz, sz, 浙江, 杭州, Sz, Sz, 浙, 浙江省, 杭州市, SZ, Cincinnati, 杭, _sz, Akron, Croatian
		$\text{MLP.v}^{27}_{17660}$	German, 万欧元, ドイツ, 德国, German, Germany, Germans, EU, 荷兰, euro, €, 欧盟, Germany, Swiss, 欧元, Euro, Dutch, 瑞士, 欧, Switzerland
		MLP.v^{26}_{3012}	犹, Jew, Jewish, 猶, Jews, Judaism, ев, jewish, Juda, Hebrew, евр, 以色列, Israel, kosher, synagogue, -J, ewish, jew, Rabbi, Israeli
		MLP.v^{26}_{3382}	俄, Rus, Russ, RU, rus, Russians, Russ, Russia, Russian, russe, 俄罗斯, russ, Russo, 俄国, 莫斯科, Mos, Russia, Moscow, -Russian, RU
		MLP.v^{25}_{4157}	UEFA, 地中海, Mediterranean, Cyprus, Pane, Roman, 欧盟, Euro, Roman, chalk, 希腊, Europe, UES, Rom, 罗马, EU, Kn, Europe, 橡, 騰
		MLP.v^{25}_{5123}	Ukraine, 乌, 乌克兰, Russia, Ukrainian, Russian, 俄罗斯, Ukrain, 烏, 乌克, Russia, war, Russian, -Russian, 冲突, russian, Russo, Conflict, conflict, 地
	Asian	$\text{MLP.v}^{28}_{13406}$	在日本, Japanese, 日本, Japanese, Japan, Japan, 日本人, Tokyo, japan, 东京, 일본, 日军, japanese, japon, Nhật, Tok, Hiro, 日本の, Osaka, jap
		MLP.v^{27}_{6943}	Asian, Asian, 亚洲, Asia, Chinese, Asia, Chinese, China, 印度, Indian, India, India, Asians, China, Indian, chinese, 中國, Oriental, Japanese, asian
		MLP.v^{27}_{229}	Mã, Opcode, (equalTo, Singapore, Singapore, ').", wię, (defvar, ,tp, Ла, Ново, yne, (ls, Tah, 海南, établissement, airobi, Î, 槟
		MLP.v^{26}_{9908}	الثال, الإثن, Malays, Indones, protester, HttpMethod, Zambia, الجها, Rodrig, Taiwanese, 犛, Jama, 加工厂, الاست, Karnataka, UserDao, Rousse, Böl, african
		$\text{MLP.v}^{26}_{13889}$	chop, Conf, Mand, Canton, 功夫, mand, Dragon, Chop, Dragon, Conf, mand, dragon, Cant, dragon, Fu, Portsmouth, Bruce, Cath, 洗衣, Fuk
		MLP.v^{25}_{2001}	越, Vietnam, 越南, Viet, Nguyen, Viet, Vietnamese, .vn, Ho, Ph, Vu, Tran, viet, 就越, Ph, Dien, 越大, 越好, anh, gia
		$\text{MLP.v}^{25}_{15029}$	Chinese, China, China, Chinese, chinese, china, china, 中国, 中国的, -China, 중국, الصين, 中國, Asian, 在中国, 是中国, 亚洲, 由中国, Asian, Asia
	Black/AA	$\text{MLP.v}^{25}_{10230}$	Af, af, 非洲, Af, African, slave, AF, Africans, AF, african, slavery, 奴隶, 奴, slaves, Slave, slave, Afro, Slave, Africa, afr
		$\text{MLP.v}^{25}_{10739}$	非洲, African, Africa, Africa, Africans, african, Ghana, Nigerian, Nigeria, afr, Kenya, Nairobi, Niger, -Saharan, Nd, Tanzania, Uganda, africa, Nz, Lagos
Mistral-7B	White	MLP.v^{32}_{2399}	Russian, Vlad, Moscow, Soviet, Russia, Ukrain, Russians, Ukraine, Czech, Ukr, Bulgar, Slov, vod, Cz, Aleks, Serge, sov, russ, Polish, Stalin
		MLP.v^{31}_{7356}	Russell, Rus, Russ, Russian, rus, Russia, russ, Russians, Soviet, Moscow, рус, rust, rus, rust, py, sov, Rug, Rud, ruin, Ruth
		MLP.v^{30}_{4487}	Russell, Russ, russ, Russian, Rus, Russia, Russians, rus, rus, Moscow, рус, Росси, России, uss, Soviet, ussia, ussian, Vlad, USS, sov
		MLP.v^{29}_{260}	Italian, Ital, Italy, ital, italien, Italia, italiano, ital, Giovanni, Francesco, Carlo, Gian, Sic, Gi, IT, Milan, Serie, acci, Giov, IT
	Black/AA	MLP.v^{31}_{8715}	African, Africa, Afr, Niger, Af, frica, Kenya, Af, Nigeria, fr, Negro, af, Afghan, blacks, FR, airo, Caribbean, Johannes, ₦, Jama
Llama-3.1-8B	White	$\text{MLP.v}^{32}_{10606}$	Russian, Russians, Russia, Russian, Moscow, Soviet, Russia, -Russian, Putin, russ, russian, Kremlin, 俄, russe, Russ, USSR, Russell, .ru, Vladimir, Sergei
		$\text{MLP.v}^{32}_{10409}$	Baltic, Sloven, Celtic, Gron, Flem, Slovenia, Austria, Sch, Luxembourg, Monaco, Gros, Croatia, Naples, Blond, Croatian, Baron, Austrian, Mont, Trent, Malta
		$\text{MLP.v}^{31}_{12584}$	Russell, Russ, Russ, Rus, RUS, Rus, russ, rus, Russo, Russia, Russian, Rusya, 俄, Russians, Рус, -Russian, rus, Russia, russe, Russian
		MLP.v^{29}_{4193}	Czech, Hungarian, Boh, Slovak, Budapest, Hungary, Hung, Slovakia, Prague, Poland, polish, Polish, Sloven, Croatian, Krak, lovak, Brno, Slovenia, Romania, Croatia
	Asian	MLP.v^{32}_{5691}	Chinese, China, Chinese, China, chinese, china, -China, Beijing, 中国, Çin, 중국, Shanghai, 中國, china, چین, 中国, Zhang, Jiang, Tencent, Guang
		MLP.v^{32}_{6950}	Bhar, Bh, Sinh, Nagar, Shah, Muk, Hindi, Allah, Tamil, Gujar, Bollywood, Kh, Dh, Gujarat, Punjab, Kh, Mumbai, Uttar, Maharashtra, Jain
		$\text{MLP.v}^{29}_{10616}$	Indian, Indian, Indians, India, indian, India, india, Ấn, Bollywood, الهند, Indi, Indianapolis, Mumbai, , Hindu, Bombay, Hindi, Hindus, 印, Delhi
	Black/AA	MLP.v^{30}_{7195}	Mississippi, Jamaica, Jama, Caribbean, Louisiana, Trinidad, Haiti, LSU, Ghana, Hait, Baton, Harlem, Nigeria, negro, Negro, Jazz, Bahamas, Memphis, Nigerian, Zimbabwe
		MLP.v^{29}_{6824}	tropical, jungle, Tropical, Jungle, jung, ropical, trop, Fiji, Belize, mango, Congo, Jama, Caribbean, Hait, Honduras, Mango, BSON, Haiti, Safari, Jamaica
		$\text{MLP.v}^{29}_{13826}$	African, african, Africans, Africa, Africa, afr, Afro, Afr, frican, negro, Af, Afrika, Af, africa, Negro, blacks, black, af, frica, Blacks

Table 17: Full list of race-encoding neurons identified from C-REACT indirect mentions (language/country).