Beyond Behavioral Benchmarks: Mechanistic Evidence of Demographic Encoding Dissociation in Language Models

Carlos Rodriguez

Institute of Design, Illinois Institute of Technology

ABSTRACT

Language models can pass behavioral fairness benchmarks while internally computing demographic information in ways those benchmarks cannot detect. We demonstrate this gap mechanistically across eight models: both race/ethnicity and gender identity are linearly decodable at 100% accuracy from internal activations at every layer regardless of behavioral output — an encoding that survives paraphrase substitution of demographic labels, confirming it is concept-level rather than token-specific. Yet the causal role of those representations varies dramatically by training regime. In base models (GPT-2 and Pythia-1.4B), demographic encoding directions are causally active at intermediate layers ($p < 0.05$) and inert only at the final layer. In Mistral-7B-Instruct-v0.1, instruction tuning renders these directions causally inert throughout the entire network while behavioral content sensitivity is restored to 32–38% — the encoding is geometrically present but causally disconnected from all output generation. Causal patching across two demographic directions (Hispanic−Black and White−Hispanic) confirms the pattern holds across comparison contexts. Exploratory semantic analysis of GPT-2 reveals model-specific artifacts invisible to behavioral evaluation: the Hispanic encoding direction acquires criminalization-adjacent vocabulary when contrasted with White, and the Black encoding direction is dominated by color-compound subword tokens. We introduce a four-phase mechanistic evaluation pipeline for distinguishing behavioral bias, encoded demographic information, and the causal influence of that information — three properties that fairness evaluation often conflates.

CCS Concepts: • Computing methodologies → Natural language processing; • Social and professional topics → Race and ethnicity; • Human-centered computing → Empirical studies in HCI.

Keywords: mechanistic interpretability, demographic bias, linear probing, activation patching, silicon sampling, representational fidelity, instruction tuning

1. INTRODUCTION

When researchers use language models to simulate how demographic groups would respond to survey questions — a practice termed silicon sampling (Argyle et al., 2023) — they assume that what a model outputs reflects how the model internally represents that demographic group. A model that produces different outputs when prompted as a Black respondent versus a Hispanic respondent is assumed to be drawing on different internal representations of those groups. This assumption has never been tested mechanistically.

We test it. The answer is that behavioral outputs and internal representations can be substantially decoupled — and we can now show exactly where and why across eight models spanning 117M to 12B parameters. Demographic identity is encoded as a linear direction in the model's residual stream from the earliest layers. In base models, this direction is causally active during

intermediate computation and overridden at the final layer. In Mistral-7B-Instruct-v0.1, the picture is more complete and more striking: the direction is present geometrically but causally disconnected from every layer of the computation. The encoding exists. It drives nothing.

This decoupling has a concrete implication for practitioners. A model can pass every behavioral bias benchmark while retaining internal demographic representations that carry pretraining corpus artifacts — artifacts that are geometrically present but causally isolated from output generation. We demonstrate this directly: Mistral-7B-Instruct-v0.1 shows 32–38% behavioral content sensitivity in rotation tests, but permutation-validated causal patching confirms its demographic encoding direction is causally inert at every layer tested ($p \geq 0.292$ n.s. throughout). In this model, instruction tuning restores behavioral sensitivity through a mechanism that does not operate through the demographic encoding direction.

Prior work has examined behavioral bias benchmarks, demographic encoding, and bias circuits separately. This paper connects them by showing that behavioral sensitivity, representational decodability, and causal relevance can diverge across training regimes — and that the causal profile is consistent across demographic comparison directions, not just across models. The causal claim is permutation-validated for the Hispanic−Black direction across three models, and confirmed for the White−Hispanic direction in all three models.

The exploratory semantic analysis of GPT-2 surfaces two findings with direct equity implications. The Hispanic encoding direction acquires criminalization-adjacent vocabulary when contrasted with White — a direction now confirmed causally active at intermediate layers — but loads on Spanish cultural surnames when contrasted with Black. The Black encoding direction is dominated by color-compound subword tokens at early layers. Both properties are absent in Pythia-1.4B.

This paper makes two distinct contributions: a generalizable four-phase mechanistic evaluation pipeline for distinguishing behavioral bias, encoded demographic information, and the causal influence of that information, and a concrete empirical demonstration of what that pipeline reveals across three architectures and two demographic comparison directions.

2. BACKGROUND AND RELATED WORK

Behavioral bias evaluation and benchmark limitations.

BBQ (Parrish et al., 2022) tests whether models select stereotyped answers under ambiguous conditions. The benchmark was validated on instruction-tuned models where it reliably detects bias signal, but assumes models engage with prompt content. Our work shows this assumption fails for base models when option-position bias dominates — a benchmark-model mismatch rather than a general failure of behavioral evaluation. The broader landscape includes BOLD (Dhamala et al., 2021), which measures bias in open-ended generation, and HolisticBias (Smith et al., 2022), which covers nearly 600 demographic descriptors. These generative benchmarks avoid the token selection bias that undermines BBQ on base models but share the deeper limitation: they measure behavioral surface and remain silent about internal representations. Blodgett et al. (2020) surveyed 146 bias papers and found that output-level measurement rarely captures the allocational harms that motivate the research. Our findings provide one mechanistic account of why: output-level measurement operates at the final layer, missing causal demographic activity in intermediate computation.

Mechanistic interpretability and demographic encoding.

The residual stream framework underlying our approach is formalized in Elhage et al. (2021). Causal intervention methods including activation patching and causal tracing (Meng et al., 2022) test whether specific representations are causally implicated in output generation — the approach we apply to demographic encoding directions in Phase 3.

Chandna et al. (2025) used Edge Attribution Patching to localize demographic bias circuits in five models, finding they are unstable under fine-tuning. We ask a prior question — does demographic identity exist as a representational structure even when no behavioral bias is expressed? — and extend their work by testing explicitly for causal inertia across three models with different training regimes. Ahsan et al. (2025) applied mechanistic interpretability to demographic bias in healthcare, finding that gender information is highly localized in MLP layers and can be manipulated via patching to alter clinical outputs, while patient race is more distributed. Our work complements this by testing whether the causal status of demographic encoding changes with instruction tuning — a question their analysis does not address. Shan and Mueller (2026) examined demographic encoding in Gemma-2 using sparse autoencoders, finding that targeted feature ablations can reduce bias without degrading demographic recognition — a contrast with our focus on whether encoding directions are causally active or inert in the first place. The linear representation hypothesis (Park et al., 2023; Zou et al., 2023) provides the theoretical foundation. Our contribution is to establish that this encoding's causal status is architecture- and training-regime-dependent.

Silicon sampling and demographic simulation.

Argyle et al. (2023) demonstrated that GPT-3 prompted with demographic backstories produces survey response distributions correlated with actual survey data, introducing the silicon sampling methodology. Santurkar et al. (2023) and Bisbee et al. (2024) documented misalignment and instability in this correlation. Sun et al. (2024) showed that random sampling of silicon respondents can improve distributional alignment, highlighting the sensitivity of simulation results to methodology. Our work challenges the foundational assumption that behavioral correlation establishes representational fidelity — a model may produce demographically correlated outputs through a mechanism entirely independent of its demographic encoding direction, as the Mistral-7B-Instruct result demonstrates.

3. METHODOLOGY

We evaluate eight models to establish the breadth of the demographic encoding gap through linear probing, and conduct causal patching on three models spanning base and instruction-tuned architectures to establish the mechanism. The pipeline proceeds in four phases: Phases 1–3 produce statistically validated findings; Phase 4 produces exploratory, model-specific results.

3.1 Models

Model	Type	Parameters	Layers	d_model	Hardware	Precision
GPT-2	Base	117M	12	768	M4 MPS	fp32
Pythia-1.4B	Base	1.4B	24	2048	A100 (80GB)	fp16
Pythia-1.4B-deduped	Base (deduped)	1.4B	24	2048	M4 MPS	fp32
Qwen2-1.5B-Instruct	Instruction-tuned	1.5B	28	1536	M4 MPS	fp32

Pythia-6.9B	Base	6.9B	32	4096	A100 (80GB)	fp16
Pythia-12B	Base	12B	36	5120	A100 (80GB)	fp16
Mistral-7B-v0.1	Base	7B	32	4096	A100 (80GB)	fp16
Mistral-7B-Instruct-v0.1	Instruction-tuned	7B	32	4096	A100 (80GB)	fp16

Phase 1: Behavioral baseline.

We sample 100 ambiguous-condition BBQ prompts per demographic category and record model answer selection by comparing logit scores for tokens " A," " B," and " C" at the final position following "Answer:". Rotation tests — reformatting 50 prompts to place the correct not-answerable option in position A — serve as a mandatory validity check. The not-answerable option is identified via the dataset's answer_info field. Since the not-answerable option is evenly distributed across positions A, B, and C in the BBQ dataset (33%/32%/35% respectively), dominant selection of any single position constitutes evidence of position bias rather than content sensitivity.

Phase 2: Activation probing.

We extract residual stream activations at the demographic label token position using TransformerLens's run_with_cache and train logistic regression classifiers using 5-fold stratified cross-validation. Demographic label tokens are matched using exact token-level comparison to prevent substring false positives. Prompts are drawn from the BBQ race/ethnicity ambiguous condition, filtered by which demographic group appears in the context; because different BBQ items are used for each group, prompts vary in more than just the demographic label. We examine three pairwise comparisons on GPT-2: Hispanic vs. Black (n=40), White vs. Black (n=48), and White vs. Hispanic (n=48). For Mistral-7B-Instruct-v0.1, which tokenizes "Hispanic" as two subword tokens, we extract at the first subword position. Two robustness checks on GPT-2 confirm the results are not extraction artifacts. A position ablation shows accuracy drops to 72.5% at the token immediately preceding the label and 67.5% at an early pre-label fixed position, while remaining at 100% from the label token onward. A paraphrase check confirms accuracy remains at 100.0% when demographic labels are replaced with synonymous terms ('Latino' for 'Hispanic', 'African American' for 'Black'), ruling out lexical identity as the source of the signal. Together these confirm the probe detects a concept-level demographic representation that crystallizes at the label token regardless of the specific term used.

Phase 3: Causal patching.

For each of 10 matched pairs per direction, we replace the target prompt's residual stream activation at the demographic label token position with the activation from the source prompt at that layer, then measure KL divergence between the original and patched output distributions over answer tokens A, B, C. Both cross-group directions are tested (e.g., Hispanic → Black and Black → Hispanic). The KL divergence is interpreted relative to two baselines: a same-group baseline (patching within the same demographic group, which should produce near-zero KL) and a permutation test (1,000 group-label shuffles with patching rerun under each null, producing a distribution of null KL values). An observed KL is considered indicative of causal activity when it substantially exceeds the same-group baseline and falls above the 95th percentile of the permuted null distribution. Same-group baselines: 0.001 for GPT-2 Hispanic, 0.0002 for Pythia-1.4B Hispanic, 0.0002 for GPT-2 White, 0.0002 for Pythia-1.4B White. Permutation tests

validate the Hispanic−Black profile for all three models. White−Hispanic patching was conducted in all three models; given signal/noise ratios well above baseline in GPT-2 (65–75x) and Pythia-1.4B (6x), and below-baseline KL in Mistral-7B-Instruct-v0.1, permutation testing was not required for the White−Hispanic direction.

Phase 4: Exploratory semantic characterization.

We compute difference-of-means vectors at each target layer by subtracting the mean activation across one group's prompts from the mean across another group's prompts, then project all vocabulary tokens onto the normalized direction using cosine similarity, reporting the top-10 tokens most strongly associated with each encoding pole. These directions reflect distributional associations learned from pretraining corpora, not model intent or reasoning pathways. All Phase 4 results are explicitly hypothesis-generating, model-specific, and should not be assumed to generalize across architectures without replication.

Ethical considerations.

This study was determined to be exempt from IRB review under 45 CFR 46.104. All data consist of publicly available benchmark datasets and open-weights language models.

4. RESULTS

4.1 The Benchmark Validity Problem

Each base model selects a dominant answer position regardless of prompt content, confirming position bias rather than content sensitivity. GPT-2 selects Answer A on 100% of ambiguous prompts; Pythia-6.9B selects Answer B on 98%; Pythia-1.4B selects Answer B on 78%; Pythia-12B selects Answer C on 68%; Pythia-1.4B-deduped selects Answer C on 74%; and Qwen2-1.5B-Instruct distributes across B (45%) and C (49%). Since the not-answerable option is evenly distributed across positions A, B, and C (33%/32%/35%), dominant selection of any single position confirms position bias. Rotation tests produce selection rates of 0–16% for all six models, confirming none respond to prompt content.

BBQ bias scores cannot be interpreted as content sensitivity for this model class — dominant position selection renders the format invalid as a bias measure under these conditions. This is a benchmark-model mismatch: BBQ was validated on instruction-tuned models where content sensitivity can be assumed; base models without instruction tuning do not satisfy that assumption.

Mistral-7B-v0.1 shows partial content sensitivity (48% rotation sensitivity for race/ethnicity). Mistral-7B-Instruct-v0.1 shows substantially greater sensitivity: answer selection distributes across A (31%), B (33%), C (36%) for race/ethnicity, with 32–38% rotation sensitivity across categories. As Section 4.3 shows, this restored sensitivity operates independently of the demographic encoding direction.

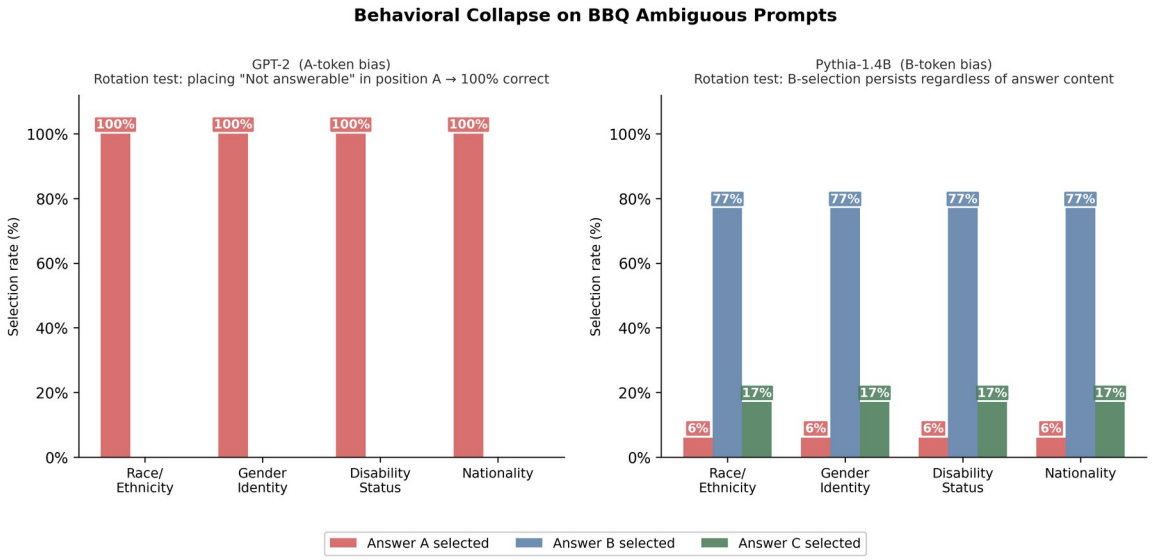

Figure 1. Behavioral collapse on BBQ ambiguous prompts. GPT-2 selects Answer A universally (left); Pythia-1.4B selects Answer B at 78% (right). Dominant position selection confirms token position bias. Rotation tests confirm position bias across all six base models.

4.2 The Gap: Demographic Encoding Persists Despite Behavioral Collapse

Despite complete behavioral collapse, demographic identity is linearly decodable from residual stream activations at every layer in every model tested. Both race/ethnicity and gender identity achieve 100.0% ± 0.0% probe accuracy across all models, confirmed using exact token-level matching. Permuted baselines fall between 47.5% and 54.3%, confirming all probe results are above chance. Two robustness checks on GPT-2 confirm this is not an extraction artifact: a position ablation shows accuracy drops to 72.5% at the token before the label and 67.5% at an early pre-label position, but is 100% from the label token onward including at the final token; a paraphrase check shows accuracy remains 100.0% when labels are replaced with synonyms (‘Latino’, ‘African American’), confirming the probe detects a concept-level representation, not a lexical identity artifact. Mistral-7B-v0.1 shows a distinct layer-emergence pattern: probe accuracy begins at 68.0% at layer 6 and reaches ceiling by layer 24.

Table 1: Token-level probe accuracies (5-fold CV), rotation test results, and permuted baselines.

Model	Category	n	Accuracy	Rotation‡	Permuted†
GPT-2 (117M, Base)	Race/Ethnicity	40††	100.0% ± 0.0%	100%	47.5%
GPT-2 (117M, Base)	Gender Identity	48	100.0% ± 0.0%	100%	51.9%
Pythia-1.4B (Base)	Race/Ethnicity	40††	100.0% ± 0.0%	0%*	48.8%
Pythia-1.4B (Base)	Gender Identity	48	100.0% ± 0.0%	0%*	52.8%
Qwen2-1.5B-Instruct	Race/Ethnicity	200	100.0% ± 0.0%	6%	51.2%
Pythia-6.9B (Base)	Race/Ethnicity	100	99.0% ± 2.0%	0%	—‡
Pythia-6.9B (Base)	Gender Identity	100	100.0% ± 0.0%	0%	—‡
Mistral-7B-v0.1 (Base)	Race/Ethnicity	100	100.0% ± 0.0%**	50%§	—‡
Mistral-7B-Instruct	Race/Ethnicity	100	100.0% ± 0.0%	32%§§	50.7%
Mistral-7B-Instruct	Gender Identity	48	100.0% ± 0.0%	38%§§	54.3%

Chance baseline	All	—	50.0%	—	—

*† Permuted accuracy from 100-iteration label permutation; all values 47.5–54.3%. ‡ Rotation: % selecting A when not-answerable placed in position A. ~33% = content sensitive; ~0% or ~100% = position bias. †† GPT-2 and Pythia-1.4B race/ethnicity: n=40 for Hispanic−Black, n=48 for White comparisons. * Pythia-1.4B exhibits B-token bias. ** Mistral-7B-v0.1 begins at 68.0% at layer 6, reaches 100.0% by layer 24. § Partial content sensitivity. §§ Content sensitivity restored by instruction tuning.*

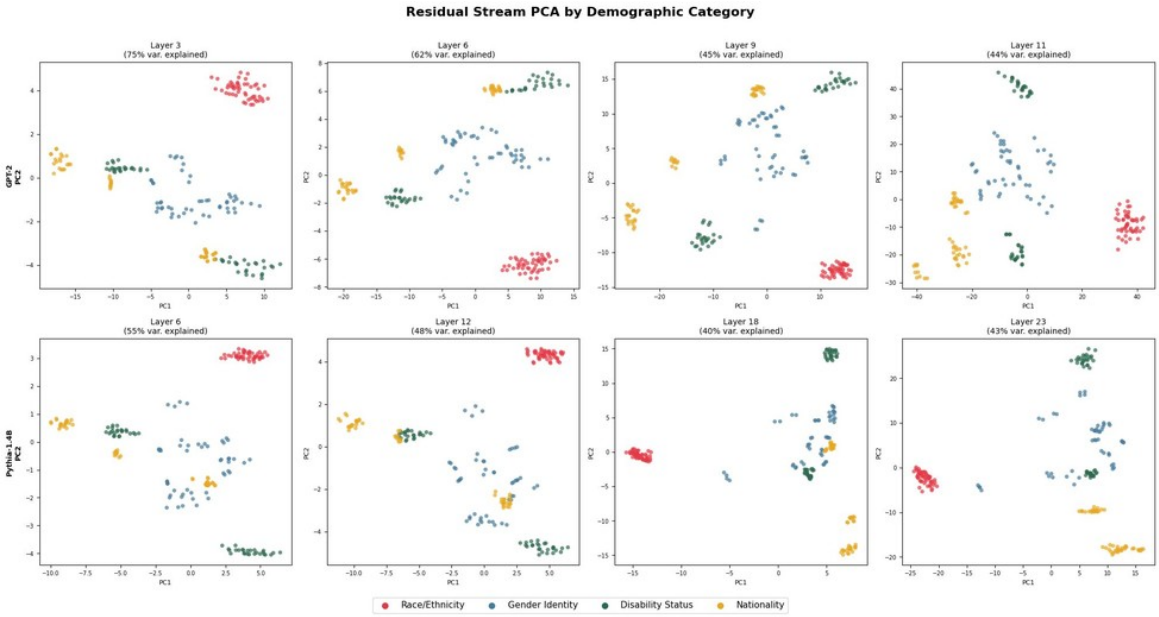

Figure 2. Residual stream PCA by demographic category across layers for GPT-2 (top) and Pythia-1.4B (bottom). Distinct clusters at every layer, despite behavioral collapse.

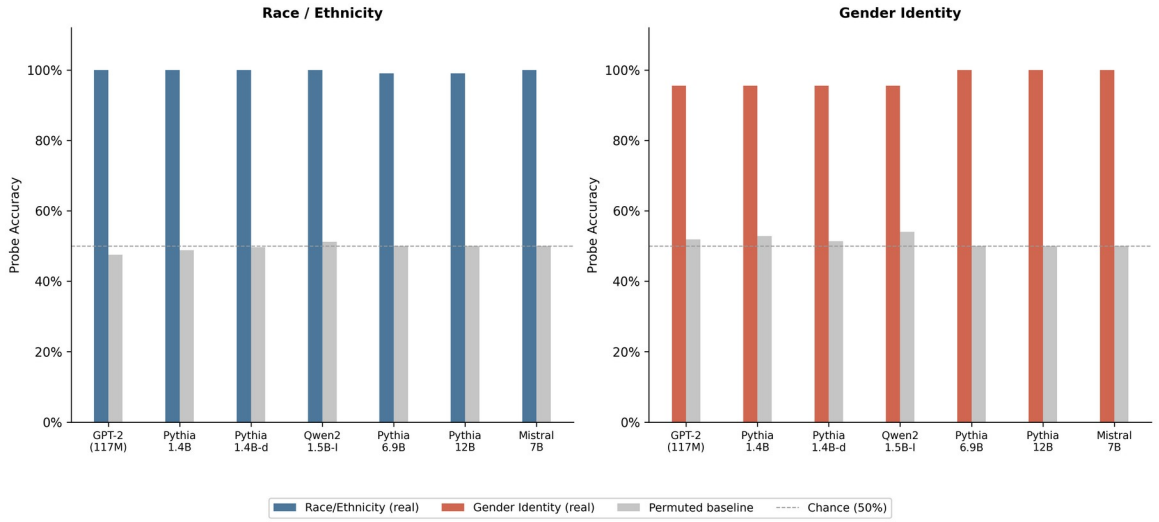

Figure 3. Token-level probe accuracy vs. permuted baseline across all models. Both race/ethnicity and gender identity achieve 100% across all models tested.

4.3 The Mechanism: Architecture-Dependent Causal Profiles

The causal profile tells a consistent story across all three models and both directions tested: demographic encoding is causally active where behavioral evaluation cannot observe it, and inert

where behavioral evaluation happens. The precise mechanism differs by training regime, not by which demographic groups are being compared.

In GPT-2, cross-group patching on the Hispanic−Black direction produces KL divergence 17–18x above the same-group baseline at layers 3–6 ($p < 0.05$), attenuates at layer 9 (KL = 0.009, $p < 0.01$), and collapses to 0.000 at layer 11 ($p = 0.493$ n.s.). The White−Hispanic direction shows the same profile: KL = 0.013–0.015 at layers 3–6 (65–75x above the White same-group baseline of 0.0002), attenuating at layer 9, and completely inert at layer 11. In Pythia-1.4B, both directions replicate this structure: Hispanic−Black KL peaks at layer 12 (7.5–9x above baseline, $p < 0.01$); White−Hispanic KL peaks at layers 6–12 (6x above baseline), both attenuating at layer 18 and collapsing to 0.000 at layer 23.

Mistral-7B-Instruct-v0.1 shows a categorically different profile for both directions. The Hispanic−Black observed KL at layer 6 (0.0012) falls below the permuted 95th percentile (0.0026), with $p = 0.535$ n.s., and all subsequent layers are non-significant ($p \geq 0.292$ throughout). The White−Hispanic direction is similarly inert: layer 6 KL (0.0016) falls below the White same-group baseline (0.0026) and is a small fraction of the signal observed in base models (0.6x the White baseline, compared to 65–75x in GPT-2 and 6x in Pythia-1.4B), with rapid attenuation to 0.0000 by layer 28. In this model, instruction tuning is associated with causal inertia of demographic encoding directions throughout the entire network — a result that holds across both demographic comparison directions tested.

Table 2: Causal patching results across three models and two demographic directions.

Model	Layers	Dir.	Mean KL	Signal/Noise	Perm. p
GPT-2 (Base)	3–6	H↔B	0.017–0.018	17–18x	p<0.05 *
GPT-2 (Base)	9	H↔B	0.009	9x	p<0.01 **
GPT-2 (Base)	11 (final)	H↔B	0.000	0x (inert)	p=0.493 n.s.
GPT-2 (Base)	3–6	W↔H	0.013–0.015	65–75x‡‡	active ††††
GPT-2 (Base)	9	W↔H	0.008	40x‡‡	active
GPT-2 (Base)	11 (final)	W↔H	0.000	0x (inert)	inert
Pythia-1.4B (Base)	6–12	H↔B	0.0015–0.0017	7.5–9x	p<0.01 **
Pythia-1.4B (Base)	18	H↔B	0.0004	2x (attenuating)	p≈0.05 †††
Pythia-1.4B (Base)	23 (final)	H↔B	0.000	0x (inert)	p=1.000 n.s.
Pythia-1.4B (Base)	6–12	W↔H	0.0012	6x‡‡	active ††††
Pythia-1.4B (Base)	18	W↔H	0.0003	1.5x (attenuating)	attenuating
Pythia-1.4B (Base)	23 (final)	W↔H	0.000	0x (inert)	inert
Mistral-7B-Instruct	6 (peak)	H↔B	0.0012	<perm. 95th	p=0.535 n.s.
Mistral-7B-Instruct	12–31	H↔B	0.0000–0.0004	—	p≥0.292 n.s.
Mistral-7B-Instruct	6 (peak)	W↔H	0.0016	<baseline‡‡‡	inert
Mistral-7B-Instruct	12–31	W↔H	0.0000–0.0003	—	inert

Same-group baselines (layer 6): GPT-2 Hispanic = 0.001; Pythia-1.4B Hispanic = 0.0002; GPT-2 White = 0.0002; Pythia-1.4B White = 0.0002; Mistral-7B-Instruct Hispanic = 0.0004 ± 0.0004; Mistral-7B-Instruct White = 0.0026 ± 0.0057 (n=6). H↔B = Hispanic−Black; W↔H = White−Hispanic. The Mistral W↔H layer 6 KL (0.0016) falls below the White baseline and is a small fraction of base model signal across both baselines, confirming causal inertia. KL is considered indicative of causal activity when it exceeds both the same-group baseline and the permuted 95th percentile. ††† Pythia layer 18: p ≈ 0.05 across independent runs. †††† W↔H permutation test not conducted; signal/noise ratios confirm activity in base models (65–75x GPT-

2; 6x Pythia). ‡‡ Relative to White same-group baseline (0.0002). ### Mistral W↔H layer 6 KL falls below White same-group baseline.

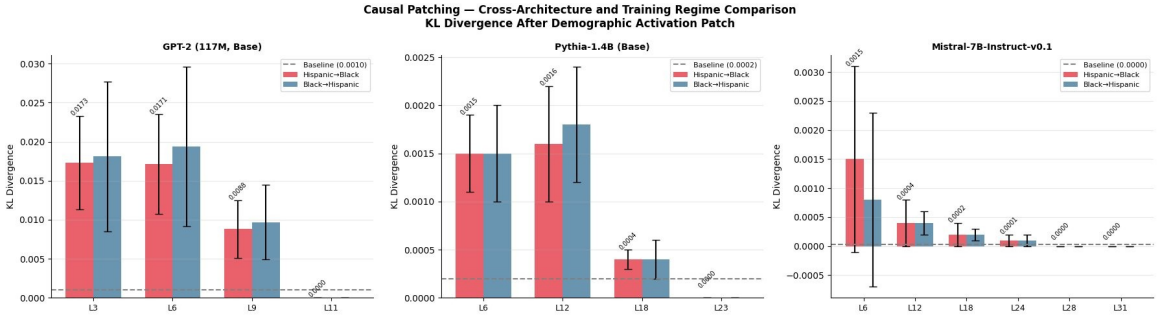

Figure 4. Causal patching across three models (Hispanic−Black direction). GPT-2: causally active ($p < 0.05$ at peak layers), inert at final layer. Pythia-1.4B: causally active ($p < 0.01$), inert at final layer. Mistral-7B-Instruct-v0.1: causally inert throughout ($p \geq 0.292$ n.s. at all layers). White−Hispanic shows identical profile in all three models.

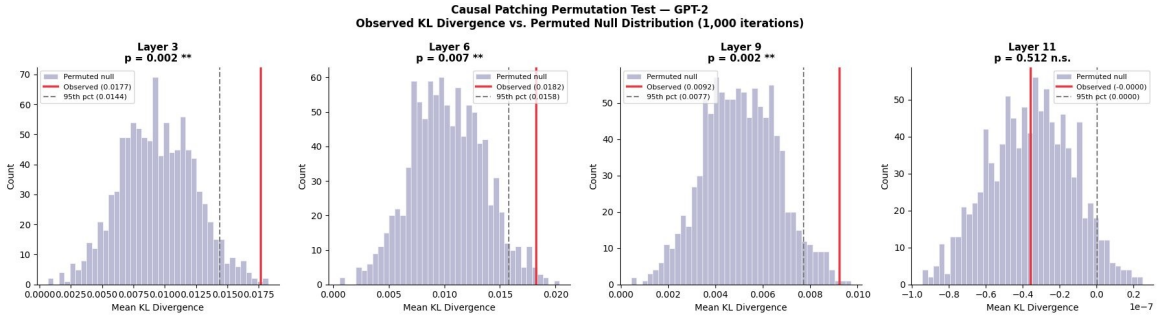

Figure 5. Permutation test for GPT-2 Hispanic−Black (1,000 iterations). Layers 3, 6, 9 exceed the null ($p < 0.05$). Layer 11 is indistinguishable from the null ($p = 0.493$).

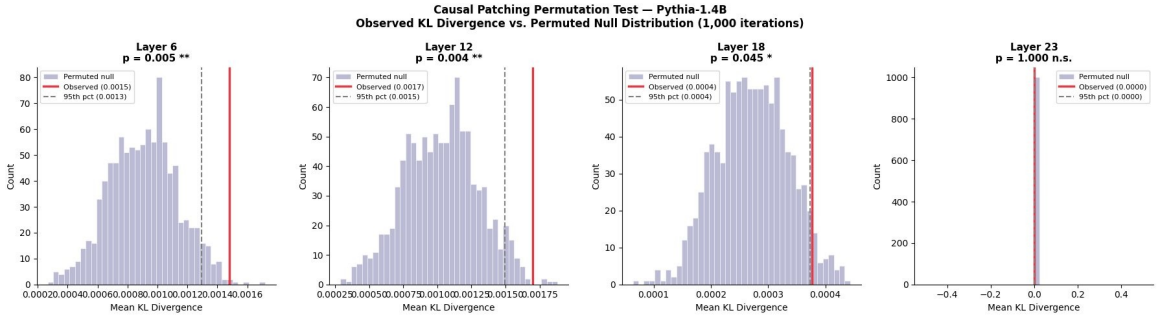

Figure 6. Permutation test for Pythia-1.4B Hispanic−Black (1,000 iterations). Layers 6 and 12 exceed the null ($p < 0.01$). Layer 18 is a boundary result ($p \approx 0.05$). Layer 23 collapses to null ($p = 1.000$).

4.4 What the Gap Contains: Model-Specific Representational Artifacts

The demographic encoding directions that are geometrically present across all models carry semantic content that behavioral benchmarks cannot detect. Full-vocabulary projection in GPT-2 reveals two hypothesis-generating findings, both absent in Pythia-1.4B, confirming they are model-specific pretraining artifacts. These directions reflect distributional associations in GPT-2's pretraining corpus, not model intent or deliberate reasoning pathways.

The most directly equity-relevant finding concerns the Hispanic encoding direction in GPT-2. When contrasted with White, the Hispanic pole loads on criminalization-adjacent vocabulary: “detention,” “ities,” “joints,” “inals.” When contrasted with Black, the same Hispanic pole loads on Spanish cultural surnames: Gast, Gomez, Cruz, bilingual, Gutierrez, Ana. The representation of Hispanic identity depends on what other group is salient in the comparison context — culturally grounded in the minority-minority comparison, criminalization-adjacent in the majority-minority comparison. Causal patching confirms the White−Hispanic encoding direction is causally active at intermediate layers in both GPT-2 (KL = 0.015 at layer 6, 75x above same-group baseline) and Pythia-1.4B (KL = 0.0012 at layers 6–12, 6x above baseline), with the same final-layer inertia profile as Hispanic−Black. The criminalization-adjacent vocabulary is present in a direction that causally influences intermediate computation.

The W → H patching direction (0.0181 at layer 6 in GPT-2) shows larger effects than H → W (0.0119), consistent with White as the implicit reference frame in the encoding direction. In Pythia-1.4B the two directions are nearly symmetric (0.0013 vs 0.0012), suggesting this asymmetry is GPT-2-specific.

The Black encoding direction in the White-Black comparison reveals token polysemy. At layers 3–6, the Black pole loads on color-compound subword tokens — Berry (Blackberry), light (Blacklight), jack (Blackjack), smith (Blacksmith). By layer 9, “liberation” appears; by layer 11 the Black pole shifts to civic vocabulary — Lives, protester, Panther, driver. The White encoding direction shifts by comparison context: urban institutional vocabulary against Black, rural suburban geography against Hispanic.

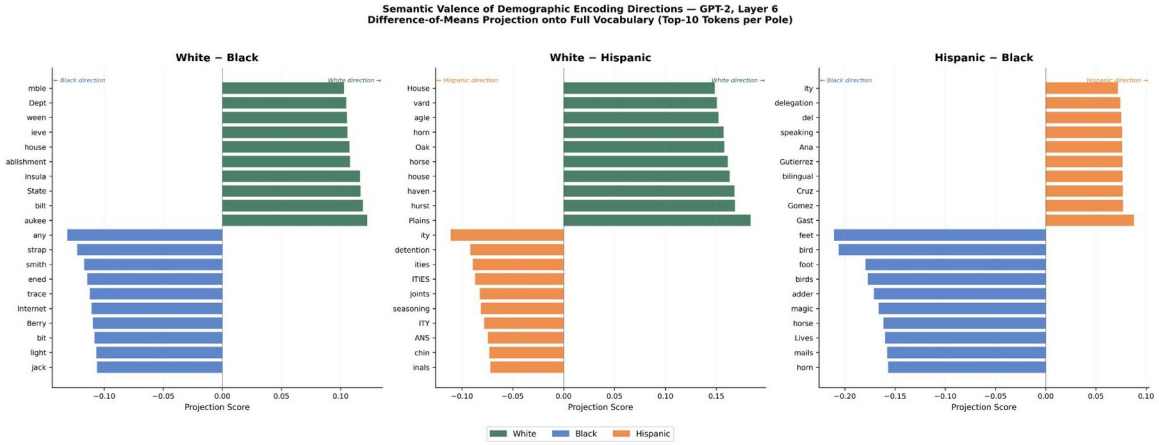

Figure 7. Exploratory semantic valence in GPT-2 at layer 6. Left (White−Black): Black pole loads on color-compound subwords. Center (White−Hispanic): Hispanic pole loads on criminalization-adjacent vocabulary in a direction confirmed causally active at intermediate layers. Right (Hispanic−Black): Hispanic pole loads on Spanish cultural surnames. GPT-2-specific distributional associations; hypothesis-generating only.

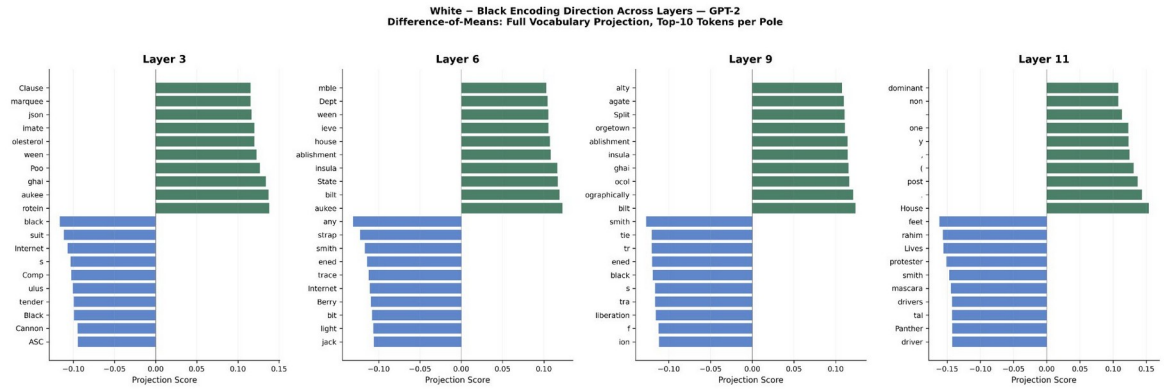

Figure 8. White−Black encoding direction across GPT-2 layers. Black pole loads on color-compound subwords at layers 3–6, shifting to civic vocabulary (Lives, protester, Panther, driver) by layer 11. GPT-2-specific.

5. DISCUSSION

5.1 What This Means for Alignment and Evaluation

The permutation-validated causal profiles across three models and two demographic directions reveal a precise mechanistic account of what instruction tuning does and does not do in the models tested. In base models, demographic encoding directions are causally active during intermediate computation and overridden at the final layer — a result that holds for both the Hispanic−Black and White−Hispanic directions in both GPT-2 and Pythia-1.4B. In Mistral-7B-Instruct-v0.1, instruction tuning renders encoding directions causally inert throughout the entire network for both directions tested, while restoring behavioral sensitivity through a mechanism that does not involve the demographic encoding direction. The direction is perfectly decodable by a linear probe but drives no aspect of the model's computation.

This has a precise implication for evaluation. Geometric presence — what probing measures — does not establish causal relevance. A model can show 100% probe accuracy for demographic identity while that identity plays no causal role in output generation. Conversely, a model can show strong behavioral content sensitivity while the demographic encoding direction is causally isolated. Neither probe accuracy nor behavioral benchmarks alone can distinguish these cases. Causal patching provides one such tool.

This is not an argument that behavioral benchmarks are useless. They are necessary. They are not sufficient. A mechanistic audit applied alongside behavioral evaluation provides the additional accountability layer that high-stakes deployment contexts require, particularly when base models are used in research pipelines where content sensitivity cannot be assumed.

These findings suggest that behavioral fairness interventions and representational alignment are distinct targets, and that current approaches may achieve only the former. Behavioral interventions such as RLHF appear to restore content sensitivity through mechanisms that operate independently of the demographic encoding direction — our Mistral-7B-Instruct findings demonstrate that the encoding direction is rendered causally inert network-wide while behavioral alignment is achieved via a separate pathway. This dissociation implies that a model can satisfy behavioral fairness criteria while retaining internal demographic representations whose semantic content — criminalization-adjacent vocabulary, token polysemy artifacts, comparison-context

dependence — is not addressed by output-level constraints. Training data does not represent demographic groups equivalently in frequency, semantic context, or social framing, and a model that encodes those asymmetries faithfully is, in a narrow sense, doing exactly what it was trained to do. Achieving representational alignment may therefore require interventions specified at the activation level — not just output-level fairness constraints — and auditing frameworks capable of detecting geometric and causal asymmetry before deployment.

5.2 What This Means for the Communities Being Simulated

Silicon sampling is used to make claims about what demographic communities think, want, and need. The Phase 4 findings demonstrate two distinct mechanisms by which these simulations may be shaped by pretraining artifacts rather than by actual communities. First, comparison-context dependence: the same model draws on criminalization-adjacent vocabulary when simulating Hispanic respondents with White as the implicit comparison group, but on cultural surname associations in a minority-minority context. Second, token polysemy: a model asked to simulate Black respondents draws on an encoding direction partly organized around color-compound subwords at early layers. Both mechanisms are invisible to probe accuracy metrics and behavioral benchmarks.

The core accountability problem is this: probe accuracy is 100% for all demographic categories — the model clearly encodes demographic identity — yet the semantic content of those encoding directions differs substantially across comparison contexts, and their causal status varies by training regime. Perfect separability does not mean accurate representation. The criminalization-adjacent vocabulary in the White−Hispanic encoding direction is causally active at intermediate layers in both GPT-2 and Pythia-1.4B. A model used to simulate Hispanic community preferences on criminal justice policy may therefore be drawing, at the level of intermediate computation, on an encoding direction shaped by criminalization associations rather than the actual community's perspective. Whether this intermediate-layer causal activity shapes final outputs requires targeted follow-up patching experiments, but the causal activity itself is now established.

Researchers conducting silicon sampling studies should apply Phases 1–3 of this pipeline before publishing simulated demographic responses as proxies for community perspectives. The pipeline is available at [repository URL] and applicable to any open-weights model supported by TransformerLens.

6. LIMITATIONS

Benchmark and model scope. The behavioral analysis uses BBQ (Parrish et al., 2022), a U.S.-centric multiple-choice benchmark validated on instruction-tuned models. The mechanistic findings in Phases 2–4 are benchmark-agnostic — causal patching operates on residual stream activations, not on benchmark format — but replicating Phase 1 with open-ended generation benchmarks such as BOLD (Dhamala et al., 2021) would test whether behavioral collapse generalizes. This paper audits open-weights models from GPT-2 to Mistral-7B; the models used in contemporary silicon sampling research — GPT-4, Claude, Llama-3-70B — are outside our empirical scope. The architecture-dependent causal profiles documented here may not generalize to other training regimes or scales, and the Mistral-7B-Instruct result reflects one instruction-tuned model.

Sample size and permutation testing scope. All Hispanic−Black causal patching results are permutation-validated (1,000 iterations each) using 10 matched pairs per direction. White−Hispanic patching was conducted in all three models; permutation testing was not conducted for this direction given the unambiguous signal/noise ratios in base models (65–75x in GPT-2, 6x in Pythia-1.4B) and below-baseline KL in Mistral-7B-Instruct-v0.1. For Mistral White−Hispanic, the White same-group baseline has high variance (0.0026 ± 0.0057, n=6); however, the cross-group KL (0.0016) also falls below the lower-variance Hispanic same-group baseline (0.0004 ± 0.0004), and is a small fraction of base model signal across both baselines, confirming causal inertia. Layer 18 in Pythia-1.4B remains a boundary result ($p \approx 0.05$ across independent runs).

Phase 4 scope and future directions. All Phase 4 findings are model-specific, hypothesis-generating, and reflect pretraining corpus distributional associations rather than model intent. A replication on Pythia-1.4B did not reproduce the vocabulary patterns found in GPT-2. The causal connection between vocabulary projection results and final-layer outputs has not been directly established. The probe position ablation and paraphrase check confirm that 100% accuracy is not a prompt-template or lexical identity artifact. Future work should apply sparse autoencoder decomposition to determine whether demographic encoding directions reflect monosemantic features or superpositions of correlated social signals, and whether the causal status of those directions predicts downstream output bias regardless of their compositional structure.

7. CONCLUSION

Behavioral benchmarks measure what language models say about demographic groups. This paper measures what language models compute about demographic groups — and reveals that the causal status of demographic representations depends on training regime in a precise and previously undocumented way.

Permutation-validated causal patching across three models and two demographic directions establishes an architecture-dependent dissociation. In base models (GPT-2 and Pythia-1.4B), demographic encoding directions are causally active at intermediate layers and inert at the final layer — confirmed for both Hispanic−Black and White−Hispanic directions. In Mistral-7B-Instruct-v0.1, instruction tuning renders encoding directions causally inert throughout the entire network for both directions, while behavioral sensitivity is restored through an independent mechanism. Both race/ethnicity and gender identity achieve 100.0% probe accuracy across all eight models, confirmed by position ablation and paraphrase check to reflect concept-level encoding rather than lexical artifacts — yet perfect separability does not mean causal relevance. The encoding exists. Whether it drives outputs depends on how the model was trained.

Fairness evaluation must distinguish between demographic information being encoded, being causally used, and being expressed in behavior. These are separate properties of language models, and they can diverge. The four-phase pipeline introduced in this paper provides tools to establish this distinction in any open-weights model. This paper provides the motivation and the methodology.

REFERENCES

Ahsan, H., Sen Sharma, A., Amir, S., Bau, D., & Wallace, B. C. (2025). Elucidating mechanisms of demographic bias in LLMs for healthcare. Findings of the Association for Computational Linguistics: EMNLP 2025, 14614–14631.

Alain, G., & Bengio, Y. (2016). Understanding intermediate layers using linear classifier probes. arXiv:1610.01644. Published at ICLR 2017 Workshop.

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. Political Analysis, 31(3), 337–351.

Biderman, S., Schoelkopf, H., Anthony, Q., Bradley, H., O'Brien, K., et al. (2023). Pythia: A suite for analyzing large language models across training and scaling. Proceedings of ICML 2023.

Bisbee, J., Clinton, J. D., Dorff, C., Kenkel, B., & Larson, J. M. (2024). Synthetic replacements for human survey data? The perils of large language models. Political Analysis, 32(4), 401–416.

Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). Language (technology) is power: A critical survey of "bias" in NLP. Proceedings of ACL 2020, 5454–5476.

Chandna, B., Bashir, Z., & Sen, P. (2025). Dissecting bias in LLMs: A mechanistic interpretability perspective. Transactions on Machine Learning Research.

Dhamala, J., Sun, T., Kumar, V., Krishna, S., Pruksachatkun, Y., Chang, K. W., & Gupta, R. (2021). BOLD: Dataset and metrics for measuring biases in open-ended language generation. Proceedings of FAccT 2021, 862–872.

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., DasSarma, N., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., Amodei, D., Brown, T., Clark, J., Kaplan, J., McCandlish, S., & Olah, C. (2021). A mathematical framework for transformer circuits. Transformer Circuits Thread. https://transformer-circuits.pub/2021/framework/index.html

Gallegos, I. O., Rossi, R. A., Barrow, J., et al. (2024). Bias and fairness in large language models: A survey. Computational Linguistics, 50(3), 1097–1179.

Horton, J. J. (2023). Large language models as simulated economic agents: What can we learn from homo silicus? NBER Working Paper No. 31122.

Jiang, A. Q., et al. (2023). Mistral 7B. arXiv:2310.06825.

Kim, J., Evans, J., & Schein, A. (2025). Linear representations of political perspective emerge in large language models. arXiv:2503.02080. Published at ICLR 2025.

Li, R., & Gao, Y. (2025). Anchored answers: Unravelling positional bias in GPT-2's multiple-choice questions. Findings of ACL 2025, 2439–2465.

Meng, K., Bau, D., Andonian, A., & Belinkov, Y. (2022). Locating and editing factual associations in GPT. NeurIPS 2022, 35, 17359–17372.

Nanda, N., & Bloom, J. (2022). TransformerLens: A library for mechanistic interpretability of GPT-style language models. https://github.com/neelnanda-io/TransformerLens

Park, K., Choe, Y. J., & Veitch, V. (2023). The linear representation hypothesis and the geometry of large language models. arXiv:2311.03658. Accepted at ICML 2024.

Parrish, A., Chen, A., Nangia, N., Padmakumar, V., Phang, J., Thompson, J., Htut, P. M., & Bowman, S. R. (2022). BBQ: A hand-built bias benchmark for question answering. Findings of ACL 2022.

Pezeshkpour, P., & Hruschka, E. (2024). Large language models sensitivity to the order of options in multiple-choice questions. Findings of NAACL 2024, 2006–2017.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI Blog.

Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., & Hashimoto, T. (2023). Whose opinions do language models reflect? Proceedings of ICML 2023, 29971–30004.

Shan, Z., & Mueller, A. (2026). Measuring mechanistic independence: Can bias be removed without erasing demographics? Proceedings of EACL 2026. (arXiv:2512.20796)

Smith, E. M., Hall, M., Kambadur, M., Presani, E., & Williams, A. (2022). “I’m sorry to hear that”: Finding new biases in language models with a holistic descriptor dataset. Proceedings of EMNLP 2022, 9180–9211.

Sun, J., Shaib, C., & Wallace, B. C. (2024). Random silicon sampling: Simulating human sub-population opinion surveys via a large language model based agent. arXiv:2402.18144.

Vig, J., Gehrmann, S., Belinkov, Y., Qian, S., Nevo, D., Singer, Y., & Shieber, S. (2020). Investigating gender bias in language models using causal mediation analysis. NeurIPS 2020.

Yang, A., et al. (2024). Qwen2 technical report. arXiv:2407.10671.

Zheng, C., Zhou, A., Zheng, R., Wang, J., & Huang, M. (2023). Large language models are not robust multiple choice selectors. arXiv:2309.03882.

Zou, A., Phan, L., Chen, S., et al. (2023). Representation engineering: A top-down approach to AI transparency. arXiv:2310.01405.