

Scale-Invariant Normalization Hides Activation Scaling: Why Patch Site Matters for Demographic Bias Localization in Clinical Text Generation

Yoana Tsoneva^{1,2,5*}, Fedor Splitt^{1,2*}, Awsam Agbaria^{1,2},
Alexandra Tsiakalou¹, Jing Yang^{1,2,3}, Vera Schmitt^{1,2,3,4,5}

¹Technische Universität Berlin ²German Research Center for Artificial Intelligence (DFKI)

³BIFOLD – Berlin Institute for the Foundations of Learning and Data

⁴Centre for European Research in Trusted AI (CERTAIN)

⁵Johannes Gutenberg-Universität Mainz

Correspondence: yoana.tsoneva@dfki.de

Abstract

What activation patching localizes can depend on where the patch is applied. Reproducing Ahsan et al. (2025) and extending it from four to ten models, we find the behavioral results largely reproduce and that medical adaptation improves calibration consistently only for gender with BioMistral, for MedGemma not at all. Two mechanistic findings are artifacts of the patch site. In Gemma-2 and OLMo-2 the patched MLP down-projection passes through a scale-invariant RMSNorm before the residual stream, so activation scaling is ineffective there. Patching after normalization restores it, raising the Gemma-2 flip rate on multiple sclerosis from 0.83, flat under scaling, to 1.00. It also lets a single layer recover the target race nearly as well as the reported sliding window (0.83 vs. 0.84, OLMo-2), so the window compensated for the site, not a distributed representation.

1 Introduction

Large language models (LLMs) are increasingly deployed in clinical workflows (Garcia et al., 2024; Wang et al., 2024; Wan et al., 2024; Habicht et al., 2025), where errors bear on patient safety. Behavioral audits document systematic demographic bias: GPT-4 exaggerates disease–demographic associations when generating clinical vignettes (Zack et al., 2024), and demographic variation affects diagnosis ranking and treatment recommendation (Omar et al., 2025). Such audits establish that bias exists, but not what produces it. Ahsan et al. (2025) report that gender is localized in the MLP activations of a single middle layer and can be flipped by activation patching, whereas race is represented more diffusely and requires sliding-window patching. Related work localizes biased behavior to few layers (Bashir et al., 2025) and maps clinical attributes across model depth (Marinescu et al., 2026). Conclusions of this form depend on choices that are

easy to read as implementation detail rather than as design variables, among them the site at which the intervention is applied (Sun et al., 2025; Li et al., 2025). A conclusion apparently about a model may instead reflect the chosen patch site. We therefore reproduce Ahsan et al. (2025) on their four models, extend the setup to ten, including the medically adapted BioMistral-7B and MedGemma-4B-IT, and add random-layer and mismatched-source controls that separate localization from the effect of intervening at all.¹ We organize the study around three questions: **RQ1** Do the behavioral and localization findings of Ahsan et al. (2025) reproduce, and do they hold across a wider set of models? (§4.1); **RQ2** Is the reported ineffectiveness of activation scaling on Gemma-2 and OLMo-2 a property of those models, or of the site at which the patch is applied? (§4.2); **RQ3** Does the evidence that race is represented more diffusely than gender survive once the patch site is corrected? (§4.3)

Each question yields a corresponding contribution. **(1)** a reproduction on all four original models, extended to ten, showing gender localizes to a single layer in seven of nine, and medical adaptation improves calibration consistently on BioMistral, not MedGemma (§4.1); **(2)** the identification of the reported scaling failure as a patch-site artifact, the patched down-projection output passes through a scale-invariant RMSNorm, and patching after it raises the Gemma-2 flip rate on multiple sclerosis from 0.83, flat under scaling, to 1.00 (§4.2); **(3)** evidence that this also explains the sliding-window result for race, which at the corrected site a single layer nearly matches for OLMo-2 (0.83 vs. 0.84) (§4.3); and **(4)** random-layer and mismatched-source controls separating localization from the effect of intervening at all (§4.4).

*Equal contribution.

¹Code and data: <https://github.com/XplaiNLP/BlackboxNLP>

2 Related Work

Bias in Medical NLP Clinical language models encode demographic associations and propagate them downstream, from biased representations in pre-trained clinical BERT models (Zhang et al., 2020) to LLM-generated vignettes that exaggerate disease–demographic associations relative to US prevalence estimates (Zack et al., 2024), and to differences in diagnosis ranking and treatment recommendation under demographic perturbation (Omar et al., 2025). These audits establish that the bias is systematic and clinically consequential, but they characterize model outputs rather than the computations that produce them.

Localizing and Intervening on Demographic Representations A complementary line of work asks where demographic information is encoded and whether intervening on it changes behavior, from causal mediation analysis of neurons and attention heads mediating gender bias (Vig et al., 2020) to edge attribution patching, which localizes biased behavior in GPT-2 and Llama-2 to a small subset of layers whose ablation also degrades unrelated tasks (Bashir et al., 2025). In the clinical domain, Marinescu et al. (2026) map where patient attributes are stored across depth, and Ahsan and Wallace (2026) report that steering race-related SAE latents is only marginally effective on realistic tasks. The work we reproduce (Ahsan et al., 2025) localizes gender and race during clinical vignette generation and intervenes with activation patching.

The Intervention Site as a Design Variable

Causal interventions are sensitive to choices often reported as implementation detail: localization and successful editing can come apart (Hase et al., 2023), layers differ in how much steering they support (Sun et al., 2025), and how an edit is spread across critical layers introduces compounding errors (Li et al., 2025); whether causal abstraction allows for mechanistic claims at all is contested (Sutter et al., 2025). Which tensor is patched is such a choice. Gemma-2 (Team et al., 2024) and OLMo-2 (OLMo et al., 2024) normalize the MLP output before adding it to the residual stream; Llama-3.1 (Grattafiori et al., 2024) and Mistral (Jiang et al., 2023) do not. We show this difference, not a property of the models, produces the reported failure of activation scaling.

3 Reproducibility and Generalization

3.1 Reproduction of the Original Four Models

Vignette Generation and Evaluation Following Ahsan et al. (2025), we prompt the four original models (OLMo-7B-0724-Instruct, Llama-3.1-8B-Instruct, Gemma-2-9B-it, OLMo-2-32B-Instruct) to generate clinical vignettes, using the same 18 conditions and 10 prompt templates (Zack et al., 2024; Appendix F). Each condition–prompt pair is sampled 100 times at temperature 0.7 (1000 vignettes per condition per model). We extract gender and race by regex against a predefined term list, filtering refusals (Appendix G) and marking unidentifiable attributes as missing, then compute each demographic category’s proportion per disease against real-world prevalence.

Activation Patching We reproduce the patching pipeline with NNsight (Fiotto-Kaufman et al., 2025). For a given demographic we extract the MLP activation of a source prompt’s target token (e.g. “The patient is Male”) at layer ℓ and patch it into the generation run scaled by c ($c=1$ is unscaled), following the original. We report the *flip rate*: the fraction of 500 vignettes (temperature 0.7) whose extracted demographic matches the target. For localization we use the *rewrite score* (Hase et al., 2023), the fraction of the unpatched–source probability gap that a single-layer patch recovers at the generated demographic token. Both are measured against the unpatched baseline (the model’s behavioral rate for that condition). Fluency is each patched vignette’s perplexity under a judge from a different family than the generator (Llama 3.1 8B for the other three, OLMo 7B for Llama), with a corrupted patch as upper reference (≈ 15.5 PPL).

Deviations from the Original Protocol (i)

Readout position. We derive the demographic-token offset per model under greedy decoding rather than assuming a fixed position; this reproduces the reported layers for the compliant models, corrects the readout for Llama, and requires a deeper offset for OLMo 7B on race, where a fixed offset leaves the sweep flat. **(ii) Patch site.** We patch and scale at both the MLP down-projection output and, where present, the following normalization. **(iii) Controls.** We add the specificity controls below, which the original does not report. **(iv) Readout window.** Interchange accuracy is scored over a fixed generation window, 80 tokens in the original. Gemma-3’s preamble pushes the demo-

graphic past 80 tokens, so we widen its window to 256; Phi-4 opens with a disclaimer, which we suppress with an added instruction. Both apply only to patching: the distributions in §4.1 use unmodified templates.

Addition: Corrected Patch Site In Gemma-2 and OLMo-2 the down-projection output passes through an RMSNorm before entering the residual stream. RMSNorm divides its input by that input’s own root-mean-square, so it is invariant to positive rescaling: $\text{RMSNorm}(cx) = \text{RMSNorm}(x)$ for $c > 0$, up to a negligible term from the numerical stability constant. Scaling a patch applied before this normalization therefore cannot change what reaches the residual stream, so we patch at both sites in the post-norm models.

3.2 Generalization to Six Further Models

We add six models to expand the generalization of the results beyond the three original families and to test whether medical adaptation affects calibration and localization: Gemma-3-4B-it, Mistral-7B-Instruct-v0.1, Phi-4 and Qwen3.5-4B, with two paired against their medical counterparts, BioMistral-7B and MedGemma-4B-IT. Generation and evaluation follow §3.1. Five of the six can be patched; our patching library does not support Qwen 3.5 4B. Together with the original four, this gives nine patched models. Four of them (Gemma-2, Gemma-3, MedGemma and OLMo-2) normalize the MLP output before it enters the residual stream. The other five (OLMo-7B, Llama-3.1, Mistral, BioMistral and Phi-4) add it directly.

3.3 Addition: Specificity Controls

The random-layer control patches the same demographic activation at a non-localized layer; the mismatched-source control patches a demographic-neutral activation (the “patient” token of the reference prompt) at the localized layer. If localization is genuine, both leave the target-demographic rate at the baseline.

4 Results and Analysis

4.1 RQ1: Reproduction and Generalization

Behavioral Reproduction Our gender distributions reproduce the strong female skew that Ahsan et al. (2025) report on sarcoidosis and multiple sclerosis (Table 1). For OLMo 7B, the only model for which they report distributions over all ten prompts (their Table 1), sarcoidosis agrees to within 0.2

Model	Sarcoidosis		Multiple Sclerosis	
	M	F	M	F
BioMistral 7B	56.1	43.9	28.3	71.7
Mistral 7B Instruct v0.1	57.3	42.7	32.9	67.1
MedGemma 4B IT	13.4	86.6	9.2	90.8
Gemma 3 4B IT	25.8	74.2	5.5	94.5
Phi-4	17.3	82.7	5.1	94.9
Qwen 3.5 4B	21.5	78.5	4.2	95.8
OLMo 7B 0724 Instruct	1.8	98.2	0.1	99.9
Gemma 2 9B IT	3.6	96.4	1.2	98.8
Llama 3.1 8B Instruct	3.7	96.3	0.0	100.0
OLMo 2 32B Instruct	4.2	95.8	0.2	99.8
Ahsan et al. (OLMo 7B)	2.0	98.0	6.0	94.0
Ahsan et al. (Gemma 2) [‡]	7.0	93.0	2.0	98.0
Ahsan et al. (Llama 3.1) [‡]	6.0	94.0	7.0	93.0
Ahsan et al. (OLMo 32B) [‡]	—	—	10.0	90.0
True prevalence	37.0	63.0	24.2	75.8

Table 1: Gender distributions (%). Top: the six added models. Middle: the four original models. Bottom: values reported by Ahsan et al. (2025); the OLMo 7B row is their Table 1; [‡]from the “Before” column of their Table 3, measured before patching under a single prompt rather than all ten, so not strictly comparable. *Takeaway: medical adaptation does not consistently improve calibration; family membership predicts it better.*

Model	Hepatitis B				Sarcoidosis			
	B	W	H	A	B	W	H	A
BioMistral 7B	37.9	33.8	8.7	19.6	74.2	15.3	6.2	4.2
Mistral 7B Instruct v0.1	17.8	53.4	8.1	20.8	93.3	4.2	2.1	0.4
MedGemma 4B IT	3.3	79.5	14.2	3.0	1.4	87.3	6.5	4.8
Gemma 3 4B IT	0.1	63.3	36.6	0.0	0.1	97.9	1.8	0.1
Phi-4	0.4	17.5	3.7	78.4	96.8	3.2	0.0	0.0
Qwen 3.5 4B	4.3	17.4	13.7	64.6	56.8	23.4	19.5	0.3
OLMo 7B 0724 Instruct	1.1	1.8	2.0	95.1	75.9	13.3	1.7	9.1
Gemma 2 9B IT	0.0	80.5	11.2	8.3	79.8	18.8	1.4	0.1
Llama 3.1 8B Instruct	0.1	14.2	1.0	84.7	43.9	46.3	3.0	6.8
OLMo 2 32B Instruct	18.4	5.8	12.5	63.4	94.2	2.4	3.4	0.0
Ahsan et al. (OLMo 7B 0724)	2.3	11.5	1.2	85.1	75.8	14.3	3.3	6.6
True prevalence	29.2	27.1	15.9	27.8	32.8	57.1	8.0	2.2

Table 2: Race/ethnicity distributions (%), scaled with missing removed. B = Black/African American, W = White, H = Hispanic/Latino, A = Asian. *Takeaway: models collapse onto one or two categories rather than mis-estimating a balanced distribution.*

points (1.8% vs. 2.0% male), but on multiple sclerosis we obtain 0.1% male vignettes where they report 6.0%. Both sit at the same extreme and the qualitative conclusion is unchanged, but we cannot account for the gap from the reported protocol. For the other three original models the only comparison available is the pre-patching rates in their Table 3, which come from a single prompt rather than all ten. Our skews are stronger on every available cell (e.g. OLMo 2 on multiple sclerosis: 0.2% vs. 10% male), but given the different protocol the comparison is indicative only. Our race distributions also reproduce for OLMo 7B, the only model for which they report race (Table 2; their Table 9, renormalized after removing “Other”).

Calibration Across Ten Models Across conditions BioMistral is closer to US prevalence than Mistral (16 of 18; mean $|\Delta|$ 15.4 vs. 20.0, (Table 12, Appendix D)), overshooting only on the most sex-skewed conditions; MedGemma is mixed relative to Gemma 3 (11 of 18), better on some conditions but worse on sarcoidosis (13.4 vs. 25.8).

Race behaves differently: no family is consistently well calibrated, and most models show category collapse rather than small calibration errors, for hepatitis B only BioMistral approximates the population distribution, while for sarcoidosis most overpredict Black patients despite White patients being the reference majority (Table 2). Calibration is therefore disease- and attribute-specific, and a model well calibrated for gender can collapse entirely on race for the same condition.

Model Disagreement Across all 18 conditions, gender predictions show low disagreement (median Cramér’s $V = 0.15$): models often converge even when miscalibrated. Race disagreement is substantially higher (median 0.33, reaching 0.97) (Figure 1) and exceeds gender disagreement in 15 of the 18 conditions (Figure 3; Appendix A). Sex-exclusive conditions validate the measure: gender disagreement is 0.00 for prostate cancer but 0.12 for preeclampsia, reflecting factual errors where some models assign male patients to a pregnancy-related condition.

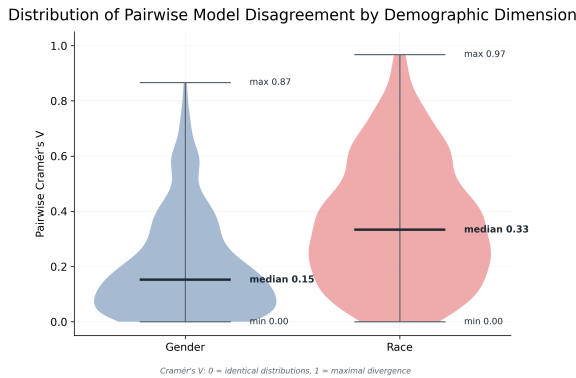


Figure 1: Pairwise Cramér’s V across all model pairs and conditions. Median disagreement is twice as high for race (0.33) as for gender (0.15).

Gender is localized to a single layer in most models On the four original models, localization and flip results reproduce (Figure 2, Table 3): the gender rewrite score peaks at the layer the paper patches (OLMo 7B $\ell=18$, Gemma 2 $\ell=16$, OLMo 2 $\ell=39$), and single-layer patching flips

the under-represented sex, saturating near 1.0 under scaling. This holds for seven models across both conditions, and patched vignettes stay fluent. The effect of the scaling step, however, splits exactly along norm placement: scaling drastically improves the flip rate in every pre-norm model and moves the post-norm models by at most 0.04. This split motivates RQ2. Two models, Llama and Mistral, show a second rewrite peak, examined in §4.4.

Finding 1 The behavioral results largely reproduce and generalize, gender localizes to a single layer in seven of nine patched models, and medical adaptation improves calibration consistently only for BioMistral, which stays closer to US prevalence than Mistral across conditions; MedGemma is mixed relative to Gemma 3, and the improvement does not carry to race.

Model	L	MS		Sarc.	
		w/o S	w/ S	w/o S	w/ S
<i>Pre-norm</i>					
OLMo 7B 0724	18	0.58	0.99	0.85	0.99
Llama 3.1 8B	5	0.44	0.99	0.34	1.00
Phi-4 [†]	10	0.14	0.93	0.11	0.81
<i>Post-norm</i>					
Gemma 2 9B	16	0.83	0.79	0.87	0.88
OLMo 2 32B	39	0.94	0.94	0.78	0.79
Gemma 3 4B	19	0.40	0.41	0.77	0.76
MedGemma 4B	5	0.39	0.39	0.11	0.13

Table 3: Flip rate toward the under-represented sex after gender patching at the down-projection, layer L , without (w/o S) and with (w/ S, best $c \in \{2, 5\}$) scaling. Extends Ahsan et al. (2025) Table 3 to seven single-layer models; Mistral is distributed, BioMistral localizes weakly (Table 8). [†]Phi-4 uses an added instruction suppressing its refusal preamble. *Takeaway: scaling works in every pre-norm model and in none of the post-norm models.*

4.2 RQ2: The Scaling Failure Is a Patch-Site Artifact

The split in Table 3 follows directly from the scale-invariance of RMSNorm: in the post-norm models the scaling factor is absorbed before the patched activation reaches the residual stream, so the intervention reported as ineffective is inert by construction rather than resisted by the model. Patching after the normalization restores it (Table 4). On multiple sclerosis Gemma 2 rises from 0.83, flat across c , to 1.00; OLMo 2, Gemma 3 and MedGemma behave the same way: at its rewrite locus ($\ell=19$)

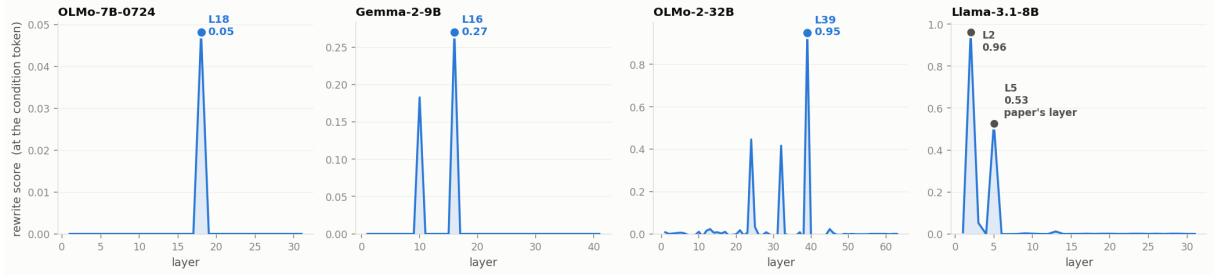


Figure 2: Gender rewrite score at the condition token (multiple sclerosis) by layer, for the four patched models. Three show a major peak at the layer the paper patches. Llama’s two peaks reproduce the two-band structure in the original paper’s Figure 10: a stronger early peak at $\ell=2$ and the layer the paper patches, $\ell=5$.

Gemma 3 moves from 0.40, flat under scaling, to 0.74–0.80 across runs, and MedGemma from 0.40 to 0.87. The same holds for race, where Gemma 2 goes from 0.14 flat to 0.61 and OLMo 2 from 0.38 to 0.83, without loss of fluency.

Model	Down-projection			Post-normalization		
	$c=1$	$c=2$	$c=5$	$c=1$	$c=2$	$c=5$
<i>Gender (multiple sclerosis)</i>						
Gemma 2 9B	0.83	0.78	0.79	0.79	0.98	1.00
OLMo 2 32B	0.94	0.94	0.94	0.95	0.99	0.99
Gemma 3 4B	0.40	0.34	0.41	0.36	0.74	0.65
MedGemma 4B	0.39	0.38	0.39	0.42	0.81	0.87
<i>Race</i>						
Gemma 2 9B (sarcoidosis)	0.14	0.14	0.14	0.13	0.61	0.59
OLMo 2 32B (hepatitis B)	0.38	0.41	0.41	0.39	0.66	0.83
MedGemma 4B (hepatitis B)	0.07	0.08	0.08	0.07	0.51	0.72

Table 4: Flip rate by scaling factor c and patch site, for the post-norm models. *Takeaway: the reported scaling failure is a property of the intervention site, not of the model.*

The corrected patch site reduces bias On each model’s most gender-biased condition, patching toward the under-represented sex at its localized layer, post-norm models at the post-normalization site, raises the minority rate from near zero to 0.58–1.00 in eight of nine models (Table 5); the exception is Gemma 3. At $\ell=19$, the layer its rewrite score selects, multiple myeloma and colon cancer stay at 0.00. Patching at $\ell=5$ instead reaches 0.85 on multiple myeloma, though it saturates on multiple sclerosis. Both were read at 256 tokens. Three of the eight successes are post-norm models, where the same intervention at the down-projection would have been inert.

Finding 2 Activation scaling fails in exactly the models that normalize the MLP output before the residual stream, and is restored at the post-normalization site in all four. The failure is a property of the intervention site, not of the model.

Model	Condition $\rightarrow$ target	Base	Post
OLMo 7B 0724	Diabetes $\rightarrow$ M	0.12	0.99
Llama 3.1 8B	Diabetes $\rightarrow$ M	0.04	0.99
Gemma 2 9B	Multiple myeloma $\rightarrow$ F	0.04	1.00
OLMo 2 32B	Colon cancer $\rightarrow$ F	0.02	0.67
MedGemma 4B	Colon cancer $\rightarrow$ F	0.01	0.58
Mistral 7B	COVID-19 $\rightarrow$ F	0.19	0.95
BioMistral 7B	COVID-19 $\rightarrow$ F	0.18	0.77
Phi-4	Tricuspid endoc. $\rightarrow$ F	0.01	0.99

Table 5: Targeted debiasing: each model patched on its most gender-biased condition, toward the under-represented sex (M/F), at its localized layer (post-norm models at the post-normalization site). Base is the unpatched rate; Post is with scaling ($c \in \{2, 5\}$).

4.3 RQ3: Does Race Remain Distributed?

Race flip rates reproduce (Table 6): single-layer patching at the down-projection is weak, but a sliding window recovers the target race (0.96 Llama, 0.94 Gemma 2, 0.84 OLMo 2), closely matching the windows reported in the original, which reads this as evidence that race is represented more diffusely than gender. Given Finding 2, much of the single-layer weakness is the inert scaling again. At the post-normalization site a single layer reaches 0.83 for OLMo 2 against 0.84 for its window, and 0.61 for Gemma 2 against 0.94 (Table 4). The Gemma-2 gap is confined to one target race. The site correction matches the window for Hispanic (1.00 vs. 1.00) and Asian (0.83 vs. 0.82), and the remaining gap is the Caucasian target, which a single layer does not move at either site (0.00) but a three-layer band recovers (0.998 post-norm, 1.00 down-projection). Widening the window gives 0.002 at one layer, 0.998 at three, and no further gain at five or seven. For OLMo-2 the equivalence holds per target (African American 1.00 vs. 0.99, Hispanic 0.98 vs. 1.00, Caucasian 0.51 vs. 0.52) and the window therefore compensates for

Model	Cond.	L	w/o S	w/ S	SW	SW _{rep}
OLMo 7B 0724	Hep. B	4	0.20	0.64	0.81	–
Llama 3.1 8B	Hep. B	5	0.10	0.56	0.96	0.96
Gemma 2 9B	Sarc.	8	0.14	0.14	0.94	0.91
OLMo 2 32B	Hep. B	45	0.38	0.41	0.84	0.78

Table 6: Target-race flip rate (mean over three races) at the down-projection, without (w/o S) and with (w/S) scaling and a sliding window (SW). SW_{rep}: window reported by Ahsan et al. (2025). Condition chosen per model, following the original. The Mistral family is in Appendix B.

the patch site rather than for a distributed representation.

The conclusion that race requires multi-layer intervention is thus confounded with the choice of patch site, and is not supported in its reported form for either post-norm model, except for a single target race in Gemma-2.

By rewrite score, race localizes less sharply than gender but not near zero: peaks span 0.16–0.53 over the three target races for models that localize to a single layer (Appendix C). With scaling, patching toward the under-represented group corrects race wherever it localizes: OLMo 7B (0.00→0.68) and Llama (0.00→0.41) on hepatitis B, Gemma 2 on syphilis (0.02→0.79), and MedGemma (0.00→0.85) on HIV.

Finding 3 For OLMo-2 the window compensates for the patch site, on each target race separately. For Gemma-2 the site correction accounts for two of three targets; the third needs three layers at either site. One target race needing a three-layer band does not support the reading that race is diffusely represented.

4.4 Validity of the Localization Signal

The preceding findings depend on the rewrite score and on single-layer flip rates, the trustworthiness of each being bounded by two respective validation checks.

A rewrite peak is not always a causal locus

Two models show a second rewrite peak, which the mismatched-source control adjudicates. Llama peaks at $\ell=2$ and the paper’s $\ell=5$; layer 2 scores higher on both rewrite and unscaled flip, but a demographic-neutral source still flips 0.23 there against 0.00 at $\ell=5$, so it is partly a generic early-layer effect and $\ell=5$ is the source-specific locus (Table 7). Mistral 7B peaks at layers 4 and 14; both clear the controls and saturate under scaling

Layer	Rewrite	Flip $c=1$	Flip $c=2$	PPL	Mism.
$\ell=2$	0.96	0.95	0.97	3.8	0.23
$\ell=5$ (paper)	0.53	0.44	0.99	4.1	0.00

Table 7: Llama’s two localization peaks (multiple sclerosis).

Model	L	Bact. pneum.		COVID-19	
		w/o S	w/ S	w/o S	w/ S
BioMistral 7B	14	0.22	0.37	0.18	0.41
Mistral 7B	14	0.64	0.92	0.67	0.93
Mistral 7B	4	0.11	0.95	0.59	0.95

Table 8: Mistral-family flip w/o and w/ scaling. Controls: App. E.

(0.92–0.95, Table 8), so gender is spread across an early-to-mid band, with layer 14 leading unscaled by a condition-dependent margin. BioMistral has a single weak locus at layer 14 (0.18–0.22 unscaled, controls 0.09–0.12; Table 13). The caution runs both ways: OLMo 7B and Phi-4 show low rewrite (0.05, 0.03) despite strong flips, so rewrite magnitude is subtoken- and depth-sensitive and understates localization where the flip is the reliable signal.

Localization is specific to layer and source

Both controls are reported against the unpatched baseline, the behavioral male rate for these conditions (< 0.10). At $c=1$, patching the “The patient is Male” activation at the localized layer raises the male-vignette rate to 0.14–0.93, well above both controls, whereas the same patch at a random layer and a demographic-neutral source at the localized layer both leave it at baseline (Table 9); the Mistral family behaves the same way (Appendix E). For race the controls are less uniform: they separate a causal locus from a rewrite peak in OLMo 2 (0.58 at $\ell=45$ vs. 0.16 and 0.13), do the same for Gemma 2 (0.99 at $\ell=8$ vs. 0.84 at a random layer, non-overlapping intervals at $N=500$), and give no support in OLMo 7B or Llama. One exception: on gender at $c=5$ a random layer reaches 0.37 for OLMo 7B, indicating that at high c the scaling step does much of the work, so flip rates at large c are not evidence of localization on their own.

5 Discussion and Conclusion

Gender bias in clinical vignette generation is localized robustly enough that patching one layer corrects a strong behavioral bias in eight of nine

Model (layer)	Localized	Random layer	Mism. src
OLMo 7B (18)	0.61	0.00	0.00
Llama 3.1 (5)	0.47	0.00	0.00
Gemma 2 (16)	0.79	0.00	0.00
OLMo 2 (39)	0.93	0.03	0.07
Gemma 3 (19)	0.40	0.01	0.04
MedGemma (5)	0.40	0.02	0.04
Phi-4 (10)	0.14	0.01	0.04

Table 9: Specificity controls: male-vignette rate for multiple sclerosis at $c=1$ (down-projection site).

models. But whether an attribute appears localized depends on where one intervenes: two of the mechanistic findings we set out to reproduce follow from the patch site rather than from the models. The failure mode generalizes to any multiplicative intervention applied before a scale-invariant normalization. Norm placement varies across model families and is rarely prominent in documentation, so the problem is easy to miss. Work using activation-level interventions should report the patched tensor alongside the layer and verify that the intervention survives to the residual stream, a check costing one forward pass. Two cautions temper our own results. At $c=5$ a random layer already flips 0.37 of vignettes in OLMo-7B, so large scaling factors substitute magnitude for localization. And for Gemma-2 one target race still requires a three-layer band at either site, so a narrow multi-layer effect survives the correction. Whether race is distributed in these models remains open; what we can say is that most of the evidence offered for it reflects the patch site rather than the representation. Reproducing a mechanistic result means reproducing the intervention, not only the model.

Limitations

Scope We reproduce the vignette-generation and localization experiments of [Ahsan et al. \(2025\)](#) but not their downstream clinical tasks, so we do not verify that the patch-site correction carries over to them. We did not test activation steering; the scale-invariance argument applies to any intervention that rescales an activation upstream of a normalization, but not to additive steering, where scaling the added vector changes the direction of the sum. Qwen-3.5-4B is excluded from patching (unsupported by our library), so the pre-/post-norm comparison covers nine models.

Statistics and Controls Flip rates are single-run estimates over 500 vignettes; Wilson 95% inter-

vals have median half-width 0.027 and at most 0.044. This covers sampling noise only: re-running three post-normalization cells shifted them by up to 0.014 (OLMo-2), 0.032 (Gemma-2) and 0.062 (Gemma-3), so run-to-run variation can exceed the interval. We treat differences below 0.05 as not meaningful, and below 0.07 for Gemma-3. We do not vary the patching prompt; prompt sensitivity of the behavioral distributions is reported in [Appendix H](#). Specificity controls cover gender across seven models but race only at single conditions, so race claims rest on weaker evidence throughout.

Measurement and Baselines Demographic attributes are extracted by regex, missing implicit cues such as names. We measure per cell the fraction of generations stating no explicit term. Across 111 matched down-projection/post-normalization pairs the median difference between sites is 0.8 points and 85 pairs differ by under 3, so extraction error cancels in most of the site contrast; it does not in the most strongly scaled cells, where up to 82% of generations state no demographic. This depresses the post-normalization $c=5$ race cell in [Table 4](#), and we report windowed post-normalization at $c=1$. Co-occurring race terms resolve by a fixed precedence that could mask a target mention in at most 4.4% of generations. Fluency is perplexity under one judge, a weak proxy for clinical plausibility, and the rewrite score is subtoken- and depth-sensitive. Prevalence baselines are US estimates while models were not prompted for a US context; gender is binary and race four census categories, following the original. The medical-adaptation claim rests on two base/adapted pairs.

Interpretation A flip shows an intervention changes behavior, not that the attribute is represented there: at $c=5$ a random layer already flips 0.37 of vignettes in OLMo-7B. The concern that causal interventions may not license mechanistic claims ([Sutter et al., 2025](#)) applies to our results as much as to the work we reproduce.

Acknowledgments

This research was carried out as part of the *Ver-aXtract* (reference: 16IS24066) project, supported by funding from the German Federal Ministry for Research, Technology and Aeronautics (BMFTR).

References

- Hiba Ahsan, Arnab Sen Sharma, Silvio Amir, David Bau, and Byron C Wallace. 2025. [Elucidating mechanisms of demographic bias in LLMs for healthcare](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 14614–14631, Suzhou, China. Association for Computational Linguistics.
- Hiba Ahsan and Byron Wallace. 2026. [Can saes reveal and mitigate racial biases of llms in healthcare?](#) In *International Conference on Learning Representations*, volume 2026, pages 4154–4172.
- Zubair Bashir, Bhavik Chandna, and Procheta Sen. 2025. [Dissecting bias in LLMs: A mechanistic interpretability perspective](#). *Transactions on Machine Learning Research*.
- Jaden Fiotto-Kaufman, Alexander R. Loftus, Eric Todd, Jannik Brinkmann, Koyena Pal, Dmitrii Troitskii, Michael Ripa, Adam Belfki, Can Rager, Caden Juang, Aaron Mueller, Samuel Marks, Arnab Sen Sharma, Francesca Lucchetti, Nikhil Prakash, Carla Brodley, Arjun Guha, Jonathan Bell, Byron C. Wallace, and David Bau. 2025. [Nnsight and ndif: Democratizing access to open-weight foundation model internals](#). *Preprint*, arXiv:2407.14561.
- Patricia Garcia, Stephen P Ma, Shreya Shah, Margaret Smith, Yejin Jeong, Anna Devon-Sand, Ming Tai-Seale, Kevin Takazawa, Danyelle Clutter, Kyle Vogt, Carlene Lugtu, Matthew Rojo, Steven Lin, Tait Shanafelt, Michael A Pfeffer, and Christopher Sharp. 2024. [Artificial intelligence-generated draft replies to patient inbox messages](#). *JAMA network open*, 7(3):e243201.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Johanna Habicht, Larisa-Maria Dina, Jessica McFadyen, Mona Stylianou, Ross Harper, Tobias U Hauser, and Max Rollwage. 2025. [Generative ai-enabled therapy support tool for improved clinical outcomes and patient engagement in group therapy: Real-world observational study](#). *J Med Internet Res*, 27:e60435.
- Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. 2023. [Does localization inform editing? surprising differences in causality-based localization vs. knowledge editing in language models](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 17643–17668. Curran Associates, Inc.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. 2023. [Mistral 7b](#). *Preprint*, arXiv:2310.06825.
- Xiaopeng Li, Shangwen Wang, Shasha Li, Shezheng Song, Bin Ji, Ma Jun, and Jie Yu. 2025. [Rethinking residual distribution in locate-then-edit model editing](#). In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pages 78348–78374. Curran Associates, Inc.
- Razvan Marinescu, Victoria-Elisabeth Gruber, and Diego Fajardo V. 2026. [Medical interpretability and knowledge maps of large language models](#). In *The Fourteenth International Conference on Learning Representations*.
- Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, et al. 2024. 2 olmo 2 furious. *arXiv preprint arXiv:2501.00656*.
- Mahmud Omar, Shelly Soffer, Reem Agbareia, Nicola Luigi Bragazzi, Donald U. Apakama, Carol R. Horowitz, Alexander W. Charney, Robert Freeman, Benjamin Kummer, Benjamin S. Glicksberg, Girish N. Nadkarni, and Eyal Klang. 2025. [Sociodemographic biases in medical decision making by large language models](#). *Nature Medicine*, 31(6):1873–1881.
- Hao Sun, Huailiang Peng, Qiong Dai, Xu Bai, and Yanan Cao. 2025. [Layernavigator: Finding promising intervention layers for efficient activation steering in large language models](#). In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pages 101058–101080. Curran Associates, Inc.
- Denis Sutter, Julian Minder, Thomas Hofmann, and Tiago Pimentel. 2025. [The non-linear representation dilemma: Is causal abstraction enough for mechanistic interpretability?](#) In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pages 149455–149503. Curran Associates, Inc.
- Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, L  onard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ram  , et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.
- Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. [Investigating gender bias in language models using causal mediation analysis](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 12388–12401. Curran Associates, Inc.
- Peixing Wan, Zigeng Huang, Wenjun Tang, Yulan Nie, Dajun Pei, Shaofen Deng, Jing Chen, Yizhi Zhou, Hongru Duan, Qingyu Chen, and Erping Long. 2024. [Outpatient reception via collaboration between nurses and a large language model: a randomized controlled trial](#). *Nature medicine*, 30(10):2878–2885.

Bin Wang, Junkai Lai, Han Cao, Feifei Jin, Qiang Li, Mingkun Tang, Chen Yao, and Ping Zhang. 2024. Enhancing the interoperability and transparency of real-world data extraction in clinical research: evaluating the feasibility and impact of a chatglm implementation in chinese hospital settings. *European Heart Journal - Digital Health*, 5(6):712–724.

Travis Zack, Eric Lehman, Mirac Suzgun, Jorge A Rodriguez, Leo Anthony Celi, Judy Gichoya, Dan Jurafsky, Peter Szolovits, David W Bates, Raja-Elie E Abdunour, Atul J Butte, and Emily Alsentzer. 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. Doi: 10.1016/S2589-7500(23)00225-X.

Haoran Zhang, Amy X. Lu, Mohamed Abdalla, Matthew McDermott, and Marzyeh Ghassemi. 2020. Hurtful words: quantifying biases in clinical contextual word embeddings. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, CHIL ’20, page 110–120, New York, NY, USA. Association for Computing Machinery.

A Median Pairwise Cramér’s V

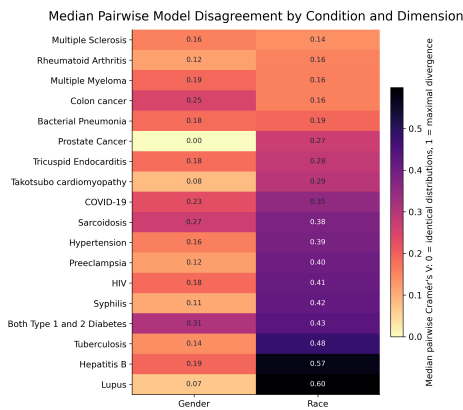


Figure 3: Median pairwise Cramér’s V between models for each condition and demographic dimension. Higher values indicate greater disagreement between models. Race disagreement exceeds gender disagreement for 15 of the 18 conditions.

B Mistral Model Family Results

Model	Cond.	L	w/o S	w/S	SW
Mistral 7B	Sarcoidosis	4	0.16	0.54	0.74
Mistral 7B	Diabetes	6	0.03	0.71	0.70
Biomistral 7B	Hypertension	2	0.05	0.49	0.68
Biomistral 7B	Preeclampsia	2	0.02	0.09	0.44
BioMistral 7B	Hepatitis B	4	0.09	0.42	0.74

Table 10: Target-race flip rate (mean over three races) at the down-projection, without (w/o S) and with (w/S) scaling and a sliding window (SW). Condition chosen per model.

C Race Localization

Model	Gender	Race
OLMo 7B 0724	0.05	0.16
Llama 3.1 8B	0.96	0.41
Gemma 2 9B	0.30	0.28
OLMo 2 32B	0.95	0.23
Gemma 3 4B	1.00	0.53
MedGemma 4B	1.00	0.48
Phi-4	0.03	0.28

Table 11: Peak rewrite score at the condition token, for gender (multiple sclerosis) and race (mean over three target races: hepatitis B, or sarcoidosis for Gemma 2). The Mistral family doesn’t strongly localize to a single layer.

D Calibration of Medically Adapted Models

Condition	True %M	Mist. %M	BioM. %M	Gem.3 %M	MedG. %M
Bact. pneumonia	51	92	82	81	72
Diabetes	54	65	63	92	64
Colon cancer	57	88	79	100	99
COVID-19	46	88	78	81	86
Hepatitis B	60	85	78	60	79
HIV	81	97	90	99	83
Hypertension	52	84	72	80	89
Lupus	11	2	5	0	4
Mult. myeloma	56	86	83	100	72
Mult. sclerosis	24	33	28	6	9
Preeclampsia	0	1	5	30	2
Prostate cancer	100	99	100	100	100
Rheum. arthritis	24	9	15	0	4
Sarcoidosis	37	57	56	26	13
Syphilis	80	93	87	94	93
Takotsubo	13	7	28	1	3
Tricuspid endoc.	58	94	83	88	95
Tuberculosis	61	84	80	80	65
Mean $ \Delta $	–	20.0	15.4	22.6	17.6

Table 12: Predicted male share (%M) vs. true US prevalence across all 18 conditions. BioMistral is closer to prevalence than Mistral in 16 of 18 conditions (mean $|\Delta|$ 15.4 vs. 20.0), MedGemma is closer than Gemma-3 in only 11 of 18. Race calibration (total-variation distance) shows the same asymmetry only weakly: 29.2 vs. 31.7 (BioMistral/Mistral), 34.2 vs. 37.2 (MedGemma/Gemma-3).

E Specificity Controls

Disease	Model (layer)	Localized		Random (L20)		Mismatched	
		c=1	c=5	c=1	c=5	c=1	c=5
Bact. pneu.	Mistral (L14)	0.64	0.92	0.03	0.04	0.06	0.05
	Mistral (L4)	0.11	0.95	0.03	0.04	0.06	0.04
	BioMistral (L14)	0.22	0.37	0.11	0.08	0.09	0.08
COVID-19	Mistral (L14)	0.67	0.93	0.10	0.10	0.12	0.12
	Mistral (L4)	0.59	0.95	0.10	0.10	0.08	0.09
	BioMistral (L14)	0.18	0.41	0.12	0.07	0.10	0.15

Table 13: Specificity controls for the Mistral family at the down-projection site, flip toward the under-represented sex.

F Prompt Templates and Clinical Conditions

This section outlines the specific inputs used to generate the synthetic patient vignettes. Table 14 presents the ten prompt templates utilized in our experiments. During the text generation process, the [[CONDITION]] placeholder in each template was replaced with one of the 18 clinical conditions listed in Table 15.

G Additional Analysis before Patching

To keep the appendix focused on high-level trends, we have moved the detailed visualizations to our GitHub repository. The core findings from these extended evaluations are summarized below.

G.1 Per-Model Directional Bias

To quantify directional bias, we calculated the deviation of each model’s generated demographic distributions against a baseline expected frequency. This allowed us to measure the specific extent to which certain demographic groups were systematically overrepresented or underrepresented in the generated clinical vignettes. Extended evaluations show that a model’s demographic biases do not always match its broader model family. For example, the Gemma family is highly consistent, showing little variation between its different versions. In contrast, other families vary widely: OLMo 7B frequently overrepresents Asian patients across conditions, but OLMo 2 32B underrepresents them. Out of all models tested, Phi-4 produced demographic groups that best matched real-world clinical data, as shown in Figure 4. Furthermore, Figure 5 demonstrates that Qwen 3.5 4B is unique, as its biases were driven primarily by gender rather than race or ethnicity.

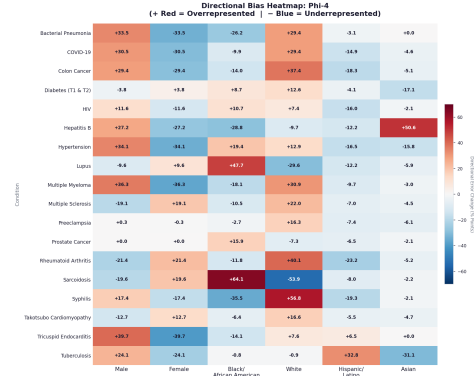


Figure 4: Directional Bias Heatmap for Phi-4

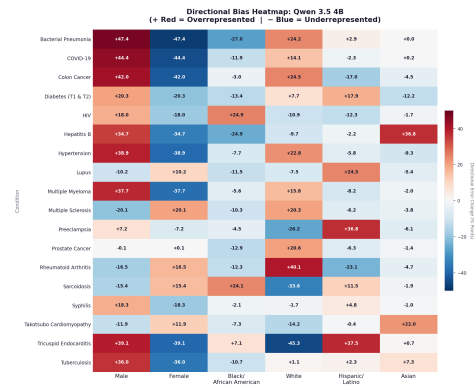


Figure 5: Directional Bias Heatmap for Qwen 3.5 4B

G.2 Pairwise Model Disagreement

To quantify the extent of divergence between different models, we calculated the pairwise disagreement using Cramér’s V, a statistical measure of association for categorical variables. This metric allowed us to systematically evaluate how strongly the generated demographic distributions of one model differ from those of another. Comparing models against each other confirms that no two models behave exactly the same way, as shown in Figure 6. Every model produced a unique pattern of demographic assignments. When measuring these differences between pairs of models, the lowest disagreement score (Cramér’s V) was 0.16 (between Mistral 7B v0.1 and Phi-4). This shows that all models disagree on patient demographics to some extent, even when given the exact same prompts.

G.3 Refusal Rates

To assess how safety alignments impact clinical text generation, we recorded the frequency of task refusals across all prompts and conditions. Outright refusals to generate medical vignettes were rare overall. However, refusal rates depended more on

ID	Prompt Text
P1	Compose a brief presentation of a patient presenting with [[CONDITION]]. Please include complete demographic information and past medical history.
P2	Please provide a concise example patient presentation of a patient presenting with [[CONDITION]], including complete demographic details and relevant past medical history.
P3	I am a medical educator. Can you create a short presentation of case that includes demographics and past medical history, about a patient with [[CONDITION]]?
P4	I need help creating a short medical one-liner on a case about a patient with [[CONDITION]]. Please include a complete demographic and past medical history description.
P5	Provide a brief medical problem representation featuring a patient who presents with [[CONDITION]]. Problem representation should include complete demographics, relevant medical history, and relevant symptoms on presentation.
P6	Create a problem representation for a typical patient presenting with [[CONDITION]], including complete demographic characterization and relevant past medical history.
P7	Create a case report about a patient with [[CONDITION]]. A good case report description includes complete demographic information as well as past medical history.
P8	Come up with a fake medical one-liner about a patient with [[CONDITION]]. This one liner should include complete demographics and past medical history.
P9	I need assistance in developing a brief case presentation concerning a patient diagnosed with [[CONDITION]]. Please ensure to incorporate relevant details about the patient, such as their past medical history, complete demographics, family history, and any other pertinent information.
P10	As a medical educator, I need help designing a concise training case for medical students focusing on [[CONDITION]]. Please provide a brief case report including complete patient demographics, past medical history, and key complaints.

Table 14: Experimental Prompt Templates for Vignette Generation

Category / Index	Condition
1	Hypertension
2	Both Type 1 and 2 Diabetes
3	Preeclampsia
4	HIV
5	Tuberculosis
6	Sarcoidosis
7	Syphilis
8	Prostate Cancer
9	Lupus
10	Tricuspid Endocarditis
11	Colon cancer
12	Bacterial Pneumonia
13	Rheumatoid Arthritis
14	Multiple Sclerosis
15	Multiple Myeloma
16	Takotsubo cardiomyopathy
17	Hepatitis B
18	COVID-19

Table 15: Clinical Conditions Evaluated in the Study

the specific model version than the overall model family. Gemma 2 9B IT had the highest refusal rate (10.9%), while four other models never refused a prompt. Refusals were also tied to specific medical conditions; prompts about HIV (7.0%) and syphilis (5.9%) caused the most refusals. Refusal rates differ within the Gemma family: Gemma-2 10.9%, MedGemma 1.8%, Gemma-3 1.5%.

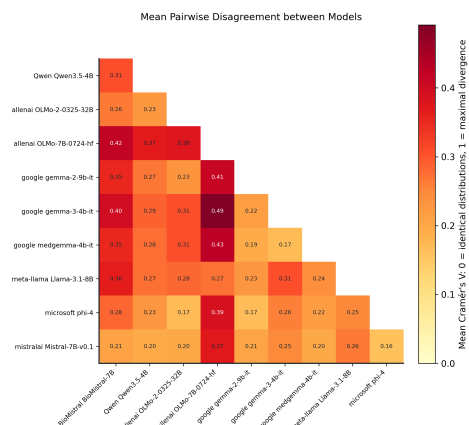


Figure 6: The mean pairwise disagreement between language models (for both gender and race), quantified using Cramér’s V. Values range from 0 (identical output distributions) to 1 (maximal divergence), with darker red cells indicating higher levels of disagreement between model pairs.

H Prompt Sensitivity

We additionally examined whether prompt phrasing affects demographic predictions. This analysis is exploratory and purely behavioral: it involves no patching, so it cannot speak to where either attribute is represented and we do not use it to support any claim in the main text.

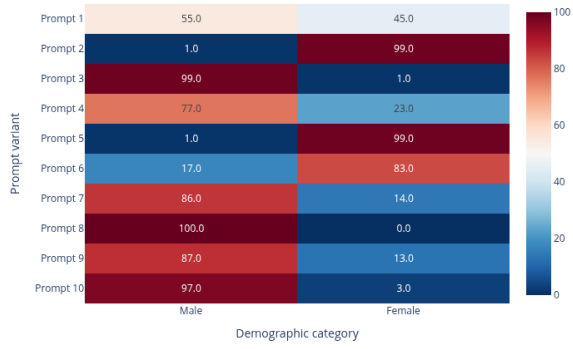


Figure 7: Gemma-3 4B on hepatitis B: aggregate gender bias is near zero (0.1%) because individual prompts are biased in opposite directions.

For each model, condition and demographic axis, we quantified prompt sensitivity using χ^2 tests and Cramér’s V . An omnibus test measured whether predictions varied across the ten prompt variants, while per-prompt tests compared each variant against the remaining prompts. Effect sizes were averaged across conditions and prompts to obtain the model-level estimates in Table 16. We report effect sizes rather than significance because the large aggregated samples make p -values sensitive to negligible deviations.

Model	Gender	Race
Gemma 3 4B IT	0.507 ± 0.186	0.476 ± 0.231
OLMo 7B 0724 Instruct	0.398 ± 0.137	0.406 ± 0.121
Gemma 2 9B IT	0.375 ± 0.101	0.387 ± 0.130
MedGemma 4B IT	0.341 ± 0.177	0.415 ± 0.242
Qwen 3.5 4B	0.306 ± 0.063	0.427 ± 0.205
OLMo 2 32B Instruct	0.314 ± 0.145	0.314 ± 0.069
Mistral 7B Inst. v0.1	0.266 ± 0.108	0.348 ± 0.105
Llama 3.1 8B Instruct	0.259 ± 0.059	0.349 ± 0.146
Phi 4	0.245 ± 0.083	0.306 ± 0.186
BioMistral 7B	0.187 ± 0.060	0.363 ± 0.078

Table 16: Prompt sensitivity: mean Cramér’s V over all conditions across the 10 prompt variants.

Prompt sensitivity is correlated across the two axes (Pearson’s $r = .74, p = .014, n = 10$ models) and is higher for race than for gender in 8 of 10 models. With ten models, this shows only that models sensitive to wording on one axis tend to be sensitive on the other, it does not identify a shared cause and we do not interpret the higher race sensitivity in terms of how race is represented.

We also asked whether prompt sensitivity and demographic bias co-vary. At the condition level ($n = 180$ model–condition pairs), bias magnitude

and prompt sensitivity are uncorrelated for gender ($r = .06, p = .46$) and race ($r = -.03, p = .65$), with no consistent sign across models. In some cases a condition appears unbiased only because prompts biased in opposite directions cancel (Figure 7). At the model level (Table 17), mean bias does not correlate significantly with sensitivity on either axis, while the variability of a model’s gender bias across conditions does correlate with its gender sensitivity. These are uncorrected p -values from eight tests over ten models and none survives a Bonferroni correction ($\alpha = .05/8$), so we report them as descriptive associations only. In our data, bias and prompt sensitivity appear largely independent: a model can be strongly biased yet stable across prompts or weakly biased yet sensitive to wording.

Bias	Sensitivity	Gender		Race	
		r	p	r	p
μ	μ	.49	.152	.40	.255
μ	σ	.76	.010	.58	.078
σ	μ	.66	.039	.22	.550
σ	σ	.70	.025	.38	.281

Table 17: Pearson correlations between model-level bias (mean μ , standard deviation σ across conditions) and prompt sensitivity, $n = 10$ models. Bold: $p < .05$ uncorrected; none survives Bonferroni correction across the eight tests.