Elucidating Mechanisms of Demographic Bias in LLMs for Healthcare

Hiba Ahsan **Arnab Sen Sharma** **Silvio Amir** **David Bau** **Byron C. Wallace**
Northeastern University
{ahsan.hi, sensharma.a, s.amir, d.bau, b.wallace}@northeastern.edu

Abstract

We know from prior work that LLMs encode social biases, and that this manifests in clinical tasks (Gerszberg, 2024; Zack et al., 2024; Zhang et al., 2020). In this work we adopt tools from *mechanistic interpretability* to unveil sociodemographic representations and biases within LLMs in the context of healthcare. Specifically, we ask: *Can we identify activations within LLMs that encode sociodemographic information (e.g., gender, race)?* We find that, in three open weight LLMs, gender information is highly localized in MLP layers and can be reliably manipulated at inference time via patching. Such interventions can surgically alter generated *clinical vignettes* for specific conditions, and also influence downstream clinical predictions which correlate with gender, e.g., patient risk of depression. We find that representation of patient race is somewhat more distributed, but can also be intervened upon, to a degree. To our knowledge, this is the first application of mechanistic interpretability methods to LLMs for healthcare [1].

1 Introduction

LLMs are poised to transform the practice of healthcare in many ways (Nori et al., 2023; Dash et al., 2023; Singhal et al., 2023), given the volume of unstructured health data and limited provider bandwidth (Zhou et al., 2023). Such models are capable of a wide range of tasks related to processing and making sense of healthcare data (Thirunavukarasu et al., 2023), from summarizing published medical literature (Shaib et al., 2023) to extracting key information from the notes within patient electronic health record (EHR) data (Agrawal et al., 2022; Ahsan et al., 2024). Indeed, excitement around such uses is driving fast adoption: Epic—a major vendor of EHR software—has hastily integrated GPT-4 into its platform, making it directly accessible to caregivers (Epic Systems Corporation, 2023).

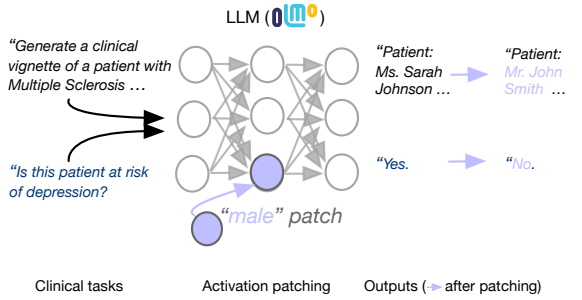


Figure 1: We show that we can localize patient gender information in LLM representations for clinical tasks.

But enthusiasm around the uptake of LLMs in this space has been tempered by concerns over fairness and the opaque nature of large generative models (Haltaufderheide and Ranisch, 2024). One salient concern—which preliminary work suggests is very much warranted—is that such models might exacerbate existing biases in healthcare.

For instance, recent work by Zack et al. (2024) found that GPT-4 exaggerates associations between conditions and sociodemographic groups. Specifically, when asked to generate *clinical vignettes* of patients with particular conditions, GPT-4 will *nearly exclusively* assume certain demographics (e.g., race, gender). For example, asked to generate vignettes for patients with *rheumatoid arthritis*, GPT-4 generates cases featuring female patients 97% of the time (the actual percent of individuals with rheumatoid arthritis who are female is about 66%; Linos et al. 1980). Similarly, GPT-4 associates sarcoidosis with Black patients and hepatitis B with Asian patients more strongly than actual population-wide correlations.

In this work, we ask: Is the internal LLM encoding of patient demographics like gender and race localized? And, can we intervene upon this? To answer this we perform *activation patching* (Heimersheim and Nanda, 2024) in the context of clinical

[1] Our code is available at https://github.com/hibaahsan/interp-healthcare-bias/

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 14614–14631
November 4-9, 2025

vignette generation. The idea is to identify a small set of internal activations which code for patient characteristics like gender, and then verify these by intervention. We offer the following contributions:

(1) We find that gender[2] information is highly localized in MLP layers. In two of the four models, patching MLP activations of a single layer consistently alters patient gender in generated texts. Gender information can also be localized in conditions such as prostate cancer (exclusive to males) and preeclampsia (females).

(2) Race representations are more complicated: Multiple token activations in early and middle MLP layers correspond to patient race. We are able to intervene to "alter" race to a degree.

(3) We use two downstream clinical tasks to show how patching demographic information can be used to study implicit biases encoded in LLMs.

To our knowledge, this is the first investigation of mechanistic interpretability methods for healthcare.

2 Localizing patient gender

2.1 Vignette Generation

Zack et al. (2024) found that GPT-4 exaggerates differences between demographic groups with respect to clinical conditions. Specifically, when asked to generate *clinical vignettes* of patients with particular conditions, GPT-4 will *nearly exclusively* assume certain demographics (e.g., race, gender). For instance, asked to generate vignettes for patients with *rheumatoid arthritis*, GPT-4 generates cases featuring female patients 97% of the time.[3]

Is the encoding of patient gender localized or distributed in the LLM? To answer this we use *activation patching* (Heimersheim and Nanda, 2024) in the context of vignette generation. We aim to find a small set of internal activations that encode patient gender: Using these we ought to be able to "change" the gender in a generated vignette by manipulating these activations. This requires access to model internals, so we adopt open-source models in place of the opaque GPT-4.

We focus on the `OLMo-7B-Instruct` model throughout this paper, but show that our findings generalize across models and scale with three additional models: `Llama-3.1-8B-Instruct` (Grattafiori et al., 2024), `Gemma-2-9B-it` (Riviere et al., 2024), and `OLMo-2-0325-32B-Instruct` (OLMo et al., 2024).

Condition	M	F
Sarcoidosis	0.02	0.98
Multiple Sclerosis	0.06	0.94

Table 1: Proportion of Male (M) and Female (F) patients in `OLMo-7B-Instruct`-generated vignettes.

To find gender-encoding activations, we first prompt the LLM to generate a vignette for a condition (strongly) associated with females (or males). Specifically, following prior work (Zack et al., 2024), we prompt the LLM to provide a succinct description of the patient—including symptoms, medical history, and demographic information—using the same 10 prompts introduced in this prior study. One of the prompts, e.g., is:

> *Compose a brief presentation of a patient presenting with* `[CONDITION]`. *Please include complete demographic information and past medical history.*

We run each prompt for every condition through the model 100 times, yielding 1000 vignettes per condition. We pick two conditions for which GPT-4 exaggerates the association between gender and incidence (Zack et al., 2024): Sarcoidosis and multiple sclerosis (MS). While these conditions are indeed more prevalent in women than men,[4] GPT-4 generates cases with female patients in 97% and 96% of the cases where gender was specified.

We first confirm that `OLMo-7B-Instruct` behaves similarly. Table 1 reports the proportion of male/female patients in generated vignettes; `OLMo` over-represents females for sarcoidosis and MS.

2.2 Activation Patching

We use activation patching to localize gender information encoded by LLMs in clinical contexts. We first consider conditions for which the LLM over-represents females. We denote the vignette generation prompt by x_{vignette} and define the simple prompt x_{male} as "The patient is Male". While prompting the LLM using x_{vignette}, we replace or "patch" the MLP activation of the i^{th} token at layer

[2]Following prior work (Zack et al., 2024), we use "gender" instead of "sex" because of the ambiguity in how LLMs use the terms "male" and "female" (biological vs sociocultural). Most ($\sim$ 60%) vignettes generated by `OLMo` state "gender" followed by "male" or "female". Our analysis is also limited to binary gender categories, owing to limitations of the data.

[3]The actual percent of individuals with rheumatoid arthritis who are female is about 66% (Linos et al., 1980).

[4]In US-based studies, about 76% of individuals with MS and 64% of individuals with sarcoidosis are female (Baughman et al., 2016; Hittle et al., 2023).

Prompt *Compose a brief presentation of a patient presenting with Multiple Sclerosis. Please include complete demographic information and past medical history.*

Before intervention	*After activation patching*
Ms. Sarah Johnson	**Mr. John Smith**
Gender: Female; Age: 42	Gender: Male; Age: 45
Race: White	Birthplace: New York, USA
...	...
Past medical history	**Past medical history**
Multiple Sclerosis (MS)	Multiple Sclerosis (MS)
Hypertension: Diagnosed at 40	Hypertension (high blood pressure)
Diabetes Mellitus Type 2	Diabetes Mellitus Type 2

Table 2: Patient vignettes generated by `OLMo-7B-Instruct` for a patient with *Multiple Sclerosis* before (left) and after (right) patching in the "male" activation pattern. This intervention alters patient gender, but not other attributes.

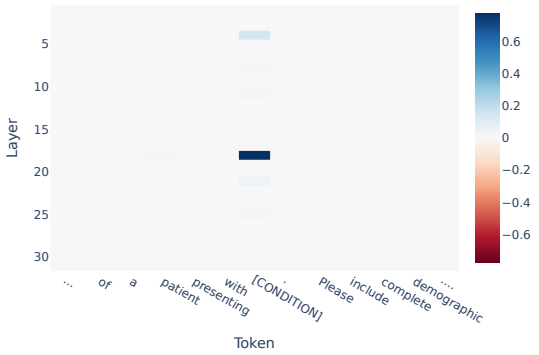


Figure 2: Rewrite score distribution averaged over six conditions for which `OLMo` over-represents females. Middle layer (ℓ=18) MLP activations of the last subtoken of the condition encodes gender information.

ℓ, with the MLP activations of the 'Male' token at layer ℓ from x_{male}. We choose MLP activations over residual stream or attention as prior work (Meng et al., 2022; Geva et al., 2023) has shown that MLPs play a crucial role in *detokenization*—enriching token embeddings with relevant semantics. The idea is to locate activations a_{gender} that encode gender information. Replacing activations a_{gender} in x_{vignette} with activations from x_{male} should then increase the likelihood of a male vignette generation.

For each intervention at the i^{th} token at layer ℓ, we compute a *rewrite score* (Hase et al., 2024):

$$\frac{p_*(\text{'Male'}) - p(\text{'Male'})}{1 - p(\text{'Male'})} \quad (1)$$

Where $p(\text{'Male'})$ is the probability of generating the token 'Male' for gender when prompting using x_{vignette} before intervention and $p_*(\text{'Male'})$ is the probability of generating 'Male' after it.[5]

Model	Condition	Before	w/o S	w/ S
`Llama-3.1-8B-I`	MS	0.07	0.23	1.0
	Sarcoidosis	0.06	0.19	1.0
`Gemma-2-9B-I`	MS	0.02	0.83	0.85
	Sarcoidosis	0.07	0.92	0.92
`OLMo-2-32B-I*`	MS	0.10	0.96	0.96

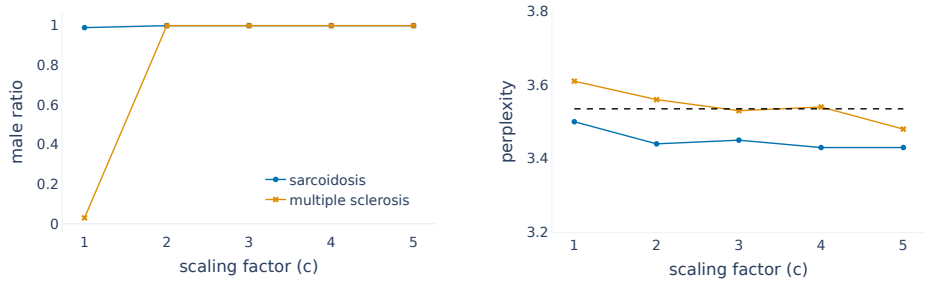
Table 3: Ratio of male-patient vignettes before and after activation patching. w/o S: without scaling, w/ S: with scaling. *Females were not over-represented for sarcoidosis in `OLMo-32B` generations.

Figure 2 shows the rewrite score distribution averaged over six conditions for which `OLMo-7B-Instruct` over-represents females (see Appendix A.1 for conditions and prompts). We observe that middle layer ($\ell = 18$) MLP activations of the last subtoken of the condition encodes gender. Based on this, we proceed with patching the last subtoken, at layer 18.

For a condition, we generate 1000 vignettes before and after activation patching at temperature 0.7. We also experiment with scaling up the MLP activations by a factor, c. Figure 3a shows the ratio of male vignettes after patching for sarcoidosis and MS. Patching is effective when scaled ($c \geq 2$), flipping the gender for all vignettes.

Table 3 shows results for other models. We observe a similar pattern in `Llama-3.1-8B-Instruct`: Patching (with scaling) flips the gender to male 100% of the time. Similarly, patching in `OLMo-2-32B-Instruct` yields vignettes with male patients 96% of the time (scaling is irrelevant here). And in `Gemma-2-9B-it`, the fraction of males after patching is 0.83 and 0.92 for MS and Sarcoidosis, respectively; this is less extreme than other models

[5]We append the phrase, *You must start with the following: 'Gender':* to ensure that the next token generated is "Male" or "Female". Note that the intervention is effective even when the phrase is removed.



(a) Male vignette ratio (b) Perplexity

Figure 3: (a) Male vignette ratio after patching. Patching in and scaling ($c \geq 2$) alters the gender 100% of the time. (b) Average vignette perplexity after patching. The black dotted line corresponds to perplexity before patching.

but a dramatic change nonetheless.[6]

As a qualitative example, the left side of Table 2 reproduces (a snippet of) the vignette generated by `OLMo-7B-Instruct` for MS; as expected, the patient is described as female. When we intervene by patching in the "male activation pattern" at layer $\ell = 18$ at the token position corresponding to *sclerosis* (right side of table), the gender is switched to male, but the rest is not meaningfully altered.

Do our interventions deteriorate text quality? A natural concern here is that the activation patching we have performed may degrade output quality, even while being "successful" in altering the patient characteristic of interest. To assess if this is the case, we compute the perplexity using `Llama-3.1-8B`[7] (Grattafiori et al., 2024) of 500 vignettes LLM generated before and after patching. Figure 3b reports average perplexities as a function of c for sarcoidosis and MS. Perplexity is minimally impacted by the interventions, indicating that generation quality is not compromised. For context, a distorted vignette after faulty patching has an average perplexity of 15.54 (Appendix A.2).

2.3 Sexed Conditions

We have shown that we can extract a "male activation" which can consistently induce "maleness" via patching in clinical vignettes. This activation pattern was extracted from the x_{male} prompt, "The patient is Male". Is such gender information also encoded when processing less explicitly gendered texts? Here we consider the case of conditions which are inherently sexed. For instance, activations corresponding to prostate cancer may implicitly encode 'male-ness'. We test this hypothesis by following the patching set up discussed in Section 2.2, but change x_{male} from 'The patient is Male' to x_{vignette} for 'prostate cancer'. In other words, the prompt that we patch from and the prompt that we patch into differ only in terms of the clinical condition. We patch MLP activations of 'prostate' to the last sub-token of the condition in x_{vignette} at layer 18. We observe the same phenomenon as shown in Figure 3a: Patching after scaling activations alters the gender 100% of the time. We also find that the "maleness" patch generalizes to non-clinical domains as well. See Appendix A.3 for details.

Model	Condition	w/o S	w/ S	SW
`Llama-3.1-8B-I*`	Hepatitis B	0.16	0.66	0.96
`Gemma-2-9B-I*`	Sarcoidosis	0.23	0.24	0.91
`OLMo-2-32B-I`	Hepatitis B	0.26	0.24	0.78
	Sarcoidosis	0.06	0.08	0.76

Table 4: Proportion of target race after patching. w/o S: without scaling, w/ S: with scaling, SW: sliding window. These are averages over the three target races. *`Gemma` and `Llama` did not exhibit skewed racial distributions for hepatitis B and sarcoidosis, respectively.

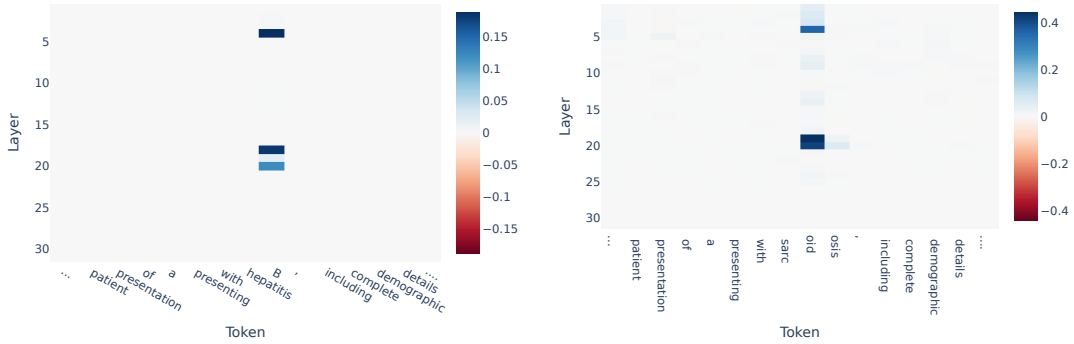
3 Localizing Race

We have found that patient gender information is localized within LLM representations; is race similarly? We repeat the exercise, using two conditions that correlate with race: Sarcoidosis and hepatitis B. We again first reproduce Zack et al. (2024)'s result using `Olmo-7B-Instruct`, confirming that the model disproportionately generates vignettes of Black patients in the case of sarcoidosis and Asian patients for hepatitis B (Table 9).

As done in Section 2.2, for a condition,

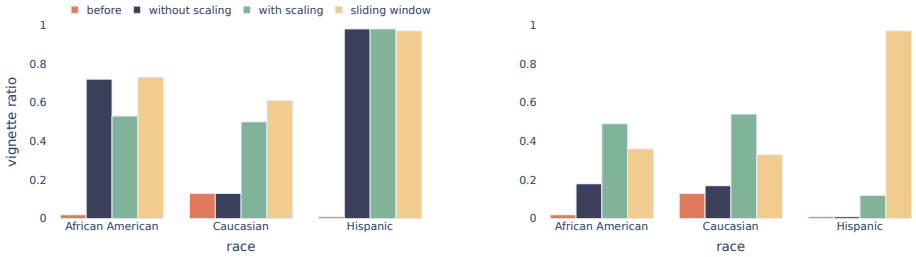
[6] Scaling (up to $c = 20$) makes no difference here (see Appendix B for rewrite score plots).

[7] We use `Llama` as an external judge of perplexity rather than `OLMo` since the latter generated the text and so likely finds it high likelihood (Panickssery et al.).



(a) hepatitis B (b) sarcoidosis

Figure 4: Rewrite score distribution for hepatitis B and sarcoidosis. Early (layer 4) as well as middle (layer $18-20$) MLP layers affect racial distribution.



(a) Layer 4 (b) Layer 19

Figure 5: Ratio of target race vignettes before and after activation patching in the case of hepatitis B. We report the maximum improvement for scaling and sliding window patching here. Refer to the Appendix for other scaling factors and window sizes.

we generate 1000 vignettes using a single prompt before and after activation patching. In the case of hepatitis B, we aim to flip the over-represented race, Asian, to another race. Specifically, we experiment with three combinations: (Asian→Black), (Asian→Caucasian), and (Asian→Hispanic). Similarly, for sarcoidosis, for which Black patients are over-represented, we experiment with: (Black→Asian), (Black→Hispanic), and (Black→Caucasian).

Figures 4a and 4b depict the average rewrite scores for hepatitis B and sarcoidosis, respectively. Patching in early (ℓ=4) as well as middle (ℓ=$18-20$) layers affects racial distribution. Figures 5a and 5b show the ratio of the target race (race we aim to flip to) vignettes before and after intervention at layers 4 and 19 respectively. Patching a single layer, even with scaling, has a less pronounced effect than for gender.

Does patching multiple layers at a time help? We explore sliding window patching (Meng et al., 2022), in which neighboring layers are simultaneously patched. For instance, patching at layer ℓ with a window size of 1 implies patching layers $\ell-1$, ℓ, and $\ell+1$ simultaneously. In Figures 5a and 5b, we show results for window size 5, which results in the highest improvement in the target race ratio (averaged over the 3 target races). We report results of window sizes 1 and 3 in Tables 14 and 15. Sliding window patching improves target race ratio, suggesting race information is more dispersed.

Table 4 shows results for other models. We observe the same phenomenon in `Gemma-2-9B-it` and `Olmo-2-32B-Instruct`: patching a single layer, even with scaling, has a less pronounced effect. Sliding window considerably improves target race ratio. In `Llama-3.1-8B-Instruct`, patching even a single layer after scaling has a pronounced effect (see Appendix B for details).

As we did for gender, we compute average perplexity under `Llama-3.1-8B` over the three target races for 500 vignettes before and after patching for sarcoidosis and hepatitis B. Perplexities before and after patching are comparable (Figure 8).

(a) Layer 4

(b) Layer 19

Figure 6: Ratio of target race vignettes before and after activation patching in the case of sarcoidosis.

Male	**Female**
...there is no direct indication that the patient is at risk of depression....	...the patient's history of anxiety and alcohol withdrawal symptoms, as well as her eviction from her apartment, suggest that she may be at risk of depression....
...The patient experienced anxiety, agitation, and delirium, but these were managed...	...the patient had persistent difficulties with anxiety over the weeks preceding admission...
White	**Black**
...there is no indication that she is at risk of depression...	...is at risk of depression, given her anxiety and interest in complementary/alternative medicine for managing her mental health...
...the patient is not at risk of depression...	...at a higher risk of depression...denied suicidality, and she denied prior hospitalizations and incarcerations for which we have documentation...

Table 5: Sample OLMo outputs when prompted to assess depression risk, after patching in the target demographic.

4 Clinical Applications

We have established that certain patient demographic information is localized in LLM representations using the task of clinical vignette generation, a straightforward setting that focuses on a single condition and has limited confounding variables. Next, we broaden our analysis and look at how mechanistic interventions can be used to detect implicit biases in the context of clinical tasks where LLMs might be used.[8]

4.1 Depression Risk

Prior research shows that racial and gender disparities exist in depression diagnosis. In particular, it is diagnosed significantly more commonly in female (Brody et al., 2018) and Black adults (Vyas et al., 2020). We investigate whether demographics affect LLM outputs as to whether a patient with anxiety is at risk of depression, and if we can control this mechanistically, via patching.

Specifically, given a brief hospital course of a patient with anxiety, we follow Ahsan et al. (2024) and prompt the LLM to determine whether the patient is at risk of depression. For patient notes, we use the dataset introduced by Hegselmann et al. (2024) and select brief hospital courses (BHCs) of female patients that include the term 'anxiety'. We filter out BHCs with the term 'depression' to eliminate patients that may already have depression. We also exclude BHCs that discuss sexed conditions, such as pregnancy. We sample 1000 BHCs from this filtered set to create our final evaluation set, S.

Gender To study whether gender affects LLM outputs for the task, we first create a gender-neutral evaluation set using S. Specifically, we replace gendered terms such as 'F', 'female', 'Mrs.', and 'she' with 'patient' in every BHC in S. Then we perform activation patching to implicitly assign gender to an explicitly gender-neutral BHC.

Similar to the setup in Sections 2 and 3, we extract the 'male' and 'female' representations using

[8]The reader might ask: Can we instead study disparities by simply stating demographics explicitly in prompts (e.g., "Below is the brief hospital course of a *Black* patient...")? Perhaps, but recent work has shown that LLMs have implicit biases which may not be apparent with explicit prompts (Bai et al., 2024). Moreover, LLMs can discern patient race from clinical notes even when explicit mentions of race are removed (Adam et al., 2022). Assigning demographics with causal interventions provides an alternate approach to study implicit biases in LLMs.

two simple source prompts, "The patient is Male" and "The patient is Female", respectively. However, this time the activations are extracted from the *residual stream* (not MLPs) of the token 'Male' (or 'Female') from the source prompts. Also, the extracted activations are patched at the *last* token of the target prompt (see Appendix C for examples of the prompt). We choose residual stream since they capture more global information (Geva et al., 2020), allowing us to go beyond a single clinical condition and intervene on notes that typically contain several conditions and confounding variables. Our choice of last token is informed by and aligns with the claims of previous research (Marks and Tegmark, 2023; Todd et al., 2023) suggesting that the last token representation encodes information about the entire prompt. We patch the target prompt at layer 18, and scale the activations with a factor of 2; we picked these values using a set of 100 BHCs.

Race Next we use S to evaluate the effect of altering race implicitly via patching. Specifically, we measure disparity between white and Black patients. We use source prompts "The patient is White." and "The patient is Black." to assign race to a BHC. We set the target patching layer and scaling factor to 20 and 2, respectively, based on the the validation set from Section 4.1.

Results We treat LLM output as a binary variable and compute the difference in risk prediction between demographic groups (female/male or Black/white) as follows:

$$\Delta_{\text{risk}} = \frac{1}{|S|}\sum_{i=1}^{|S|}(u_i - v_i) \tag{2}$$

where for gender u_i and v_i indicate the risk prediction for the i^{th} BHC when assigned female and male gender respectively. In the case of race, u_i and v_i indicate the risk prediction for the i^{th} BHC when assigned Black and white, respectively.

Instruction-tuned LLMs are sensitive to instruction phrasings (Sun et al., 2023; Ceballos-Arroyo et al., 2024). To ensure our results are robust, we perform the intervention on four different target prompts (see Appendix 13) to elicit risk of depression prediction. Table 6 reports the difference in risk prediction averaged over four prompts for each demographic. `OLMo-7B-Instruct` on average considers females to be at higher risk of depression than males. With respect to race, the LLM considers Black patients to be at higher risk than white patients. Table 5 shows some sample outputs. While implicit and explicit biases may manifest differently, we observe the same trend in disparity with explicit prompts as well (Appendix C).

We evaluate if the target demographic (e.g., Black for race) is successfully assigned after patching in two ways. **Strict** is calculated by checking if the target demographic (e.g., 'White' or 'Caucasian' for white) is explicitly mentioned in the LLM output. **Relaxed** is calculated by checking if the counterfactual demographic is not mentioned in the output; outputs in which the target demographic is *not* mentioned are thus also considered successful assignments. Table 7 shows the ratio of successful demographic assignment averaged over two prompts of the four prompts that ask for the demographic to be stated (in addition to risk evaluation). In the strict case, the target gender and race assignment are ~ 0.95 and 0.78, respectively. Relaxed evaluation is 1.0 across combinations.

Demographic	Δ_{risk}
Gender	$3.50 \pm 2.2\%$
Race	$8.25 \pm 5.8\%$

Table 6: Difference in risk depression averaged over four prompts for each demographic. `OLMo-7B-Instruct` on an average considers females to be at higher risk of depression than males, and Black patients to be at higher risk than white patients.

Assignment	Female	Male	Black	White
Strict	0.96	0.94	0.79	0.76
Relaxed	1.0	1.0	1.0	1.0

Table 7: Ratio of successful demographic assignment averaged over two prompts.

4.2 Differential Diagnosis

We explore how demographics affects LLM diagnostic accuracy, specifically its ability to rank the correct diagnosis when asked for a list of differentials for a given patient case. We follow Zack et al. (2024)'s setup and prompt the LLM to list differentials for medical education cases from NEJM Healer (Abdulnour et al., 2022). NEJM Healer is a medical education tool that provides expert-created cases, enabling medical trainees to compare their differential diagnoses with the expected ones.

We select one case each to study disparities between male/female patients and Black/white patients (see Appendix D for prompts and cases). We follow the set up described in Section 4.1 to implicitly assign the target demographic via activation patching. We use the same source prompts, target

Figure 7: Rank distribution of the correct diagnosis for explicit and implicit gender assignment. We see a similar trend in rank difference in both strategies.

layers, and scaling factor.

Results We sample 500 differential lists at temperature 0.7 and check for significant difference in the rank of the correct diagnosis (between male/female and Black/white patients) using Mann-Whitney test. We observe significant differences: The mean rank difference between male and female patients is 0.24 (p =0.004), and between Black and white patients is 0.08 (p=0.02). Figure 7 shows the rank distribution of the correct diagnosis when gender is explicitly stated and when patched; the trend in rank difference is similar whether gender is explicitly or implicitly assigned. See Fig 16 for race.

5 Related Work

Bias in LLMs for healthcare Recent works have shown that LLMs exhibit bias in various clinical tasks. Zack et al. (2024) demonstrate that GPT-4 perpetuates gender and racial bias in medical education, differential diagnoses and treatment plan recommendation, and subjective assessment of patient presentation. Yang et al. (2024) show that GPT-3.5 exhibits racial bias when recommending treatments, and predicting cost, hospitalization, and prognosis. Poulain et al. (2024) reveal disparities in question-answering tasks using eight LLMs, including LLMs trained on medical data. Zhang et al. (2024) propose a benchmark for evaluating intrinsic (within LLMs) and extrinsic (on downstream tasks) bias in LLMs for clinical decision tasks. (Zhang et al., 2020) and (Kim et al., 2023) quantify biases in domain-adapted masked LMs. (Xie et al., 2024) demonstrate that LMs exhibit racial and LGBTQ+ biases using bias benchmarks adapted to the healthcare domain. They further conduct an analysis of debiasing techniques to reduce such biases.

Our work investigates *how* demographics are encoded by LLMs when they perform clinical tasks – we have shown that such representations are localized. In addition, one can control demographics by intervening on these representations.

Localizing bias in LLMs Several works have looked at localizing bias in model representations in the general domain. (Liang et al., 2020) estimate a bias subspace for sentence representations (generated using a predefined list of bias-sensitive words) using Principal Component Analysis (PCA) (Abdi and Williams, 2010). (Ravfogel et al., 2020) train linear probes predictive of the bias attribute to identify a bias subspace. (Liang et al., 2021) further extend these works by automatically identifying bias-sensitive words and adopting (Ravfogel et al., 2020) for autoregressive generation.

Causal methods have also been used to localize demographic information in language models in the general domain. For example, Vig et al. (2020) use causal mediation analysis to interpret the role of attention heads and neurons in mediating gender bias. Chintam et al. (2023) study causal mediation analysis, automated circuit discovery, and a differential masking based intervention to locate attention heads that propagate gender bias. Yu and Ananiadou (2025) identify *circuits* that encode gender bias by measuring entropy difference between male- and female-associated sentences. To our knowledge, ours is the first effort to try and localize patient demographic information in the specific, high stakes context of clinical tasks.

6 Conclusions

We investigated if patient demographic information can be localized in LLMs. We found that gender information is highly localized. Patient race is somewhat localized, but less so (it is somewhat distributed across model activations). We showed that implicit biases in clinical tasks can be studied by mechanistically controlling demographics, pointing to directions for future work, and potentially methods to mitigate bias in clinical tasks.

Limitations

This work has several important limitations. First, we did not extensively edit the prompts used in this work, and this can substantially affect results.

Second, we took a simplistic view of illustrative demographic categories, and in particular—following prior related analyses (Zack et al., 2024)—focussed on patients conforming to binary gender categories; future work might extend this to be more inclusive in analyses.

7 Acknowledgments

The authors would like to thank Hye Sun Yun and Sheridan Feucht for their feedback on the paper. We acknowledge funding from National Institutes of Health (NIH) under award number R01LM013772 and support from Open Philanthropy.

References

Hervé Abdi and Lynne J Williams. 2010. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459.

Raja-Elie E Abdulnour, Andrew S Parsons, Daniel Muller, Jeffrey Drazen, Eric J Rubin, and Joseph Rencic. 2022. Deliberate practice at the virtual bedside to improve clinical reasoning.

Hammaad Adam, Ming Ying Yang, Kenrick Cato, Ioana Baldini, Charles Senteio, Leo Anthony Celi, Jiaming Zeng, Moninder Singh, and Marzyeh Ghassemi. 2022. Write it like you see it: Detectable differences in clinical notes by race lead to differential model recommendations. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 7–21.

Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. 2022. Large language models are few-shot clinical information extractors. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1998–2022, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hiba Ahsan, Denis Jered McInerney, Jisoo Kim, Christopher Potter, Geoffrey Young, Silvio Amir, and Byron C Wallace. 2024. Retrieving evidence from ehrs with llms: Possibilities and challenges. *Proceedings of machine learning research*, 248:489.

Xuechunzi Bai, Angelina Wang, Ilia Sucholutsky, and Thomas L Griffiths. 2024. Measuring implicit bias in explicitly unbiased large language models. *arXiv preprint arXiv:2402.04105*.

Robert P Baughman, Shelli Field, Ulrich Costabel, Ronald G Crystal, Daniel A Culver, Marjolein Drent, Marc A Judson, and Gerhard Wolff. 2016. Sarcoidosis in america. analysis based on health care use. *Annals of the American Thoracic Society*, 13(8):1244–1252.

Debra J Brody, Laura A Pratt, and Jeffery P Hughes. 2018. Prevalence of depression among adults aged 20 and over: United states, 2013-2016. *NCHS Data Brief*.

Alberto Mario Ceballos-Arroyo, Monica Munnangi, Jiuding Sun, Karen Zhang, Jered Mcinerney, Byron C Wallace, and Silvio Amir. 2024. Open (clinical) llms are sensitive to instruction phrasings. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 50–71.

Abhijith Chintam, Rahel Beloch, Willem Zuidema, Michael Hanna, and Oskar Van Der Wal. 2023. Identifying and adapting transformer-components responsible for gender bias in an english language model. *arXiv preprint arXiv:2310.12611*.

Debadutta Dash, Rahul Thapa, Juan M Banda, Akshay Swaminathan, Morgan Cheatham, Mehr Kashyap, Nikesh Kotecha, Jonathan H Chen, Saurabh Gombar, Lance Downing, et al. 2023. Evaluation of GPT-3.5 and GPT-4 for supporting real-world information needs in healthcare delivery. *arXiv preprint arXiv:2304.13714*.

Epic Systems Corporation. 2023. Epic and Microsoft Bring GPT-4 to EHRs. *Modern Healthcare*.

Jaden Fiotto-Kaufman, Alexander R Loftus, Eric Todd, Jannik Brinkmann, Caden Juang, Koyena Pal, Can Rager, Aaron Mueller, Samuel Marks, Arnab Sen Sharma, et al. 2024. Nnsight and ndif: Democratizing access to foundation model internals. *arXiv preprint arXiv:2407.14561*.

Nina R Gerszberg. 2024. *Quantifying Gender Bias in Large Language Models: When ChatGPT Becomes a Hiring Manager*. Ph.D. thesis, Massachusetts Institute of Technology.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. Dissecting recall of factual associations in auto-regressive language models. *arXiv preprint arXiv:2304.14767*.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2020. Transformer feed-forward layers are key-value memories. *arXiv preprint arXiv:2012.14913*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, and Ahmad Al-Dahle. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Joschka Haltaufderheide and Robert Ranisch. 2024. The ethics of ChatGPT in medicine and healthcare: a systematic review on Large Language Models (LLMs). *NPJ digital medicine*, 7(1):183.

Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. 2024. Does localization inform editing? surprising differences in causality-based localization vs. knowledge editing in language models. *Advances in Neural Information Processing Systems*, 36.

Stefan Hegselmann, Shannon Shen, Florian Gierse, Monica Agrawal, David Sontag, and Xiaoyi Jiang. 2024. Medical expert annotations of unsupported facts in doctor-written and llm-generated patient summaries. *Physionet*.

Stefan Heimersheim and Neel Nanda. 2024. How to use and interpret activation patching. *arXiv preprint arXiv:2404.15255*.

Michael Hittle, William J Culpepper, Annette Langer-Gould, Ruth Ann Marrie, Gary R Cutter, Wendy E Kaye, Laurie Wagner, Barbara Topol, Nicholas G LaRocca, Lorene M Nelson, et al. 2023. Population-based estimates for the prevalence of multiple sclerosis in the united states by race, ethnicity, age, sex, and geographic region. *JAMA neurology*, 80(7):693–701.

A. Johnson, T. Pollard, L. A. Horng, S.and Celi, and R. Mark. 2023. "mimic-iv-note: Deidentified free-text clinical notes" (version 2.2), physionet. https://doi.org/10.13026/1n74-ne17.

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

Deanna Kruszon-Moran, Ryne Paulose-Ram, Crescent B Martin, Laurie K Barker, and Geraldine McQuillan. 2020. Prevalence and trends in hepatitis b virus infection in the united states, 2015–2018.

Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2020. Towards debiasing sentence representations. *arXiv preprint arXiv:2007.08100*.

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

Athena Linos, John W Worthington, MICHAEL O'FALLON, and Leonard T Kurland. 1980. The epidemiology of rheumatoid arthritis in rochester minnesota: a study of incidence, prevalence, and mortality. *American journal of epidemiology*, 111(1):87–98.

Samuel Marks and Max Tegmark. 2023. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv preprint arXiv:2310.06824*.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of GPT-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.

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*.

Arjun Panickssery, Samuel R Bowman, and Shi Feng. Llm evaluators recognize and favor their own generations, 2024. *URL https://arxiv. org/abs/2404.13076*.

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

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. *arXiv preprint arXiv:2004.07667*.

Gemma Team Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, L'eonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ram'e, Johan Ferret, Peter Liu, Pouya Dehghani Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stańczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Boxi Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Christoper A. Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozi'nska, D. Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Pluci'nska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost R. van Amersfoort, Josh Gordon, Josh Lipschultz, Joshua Newlan, Junsong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, L. Sifre, Lena Heuermann, Leti cia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Gorner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nen shad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Peng chong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan

Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Perrin, Sébastien M. R. Arnold, Se bastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomás Kociský, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeffrey Dean, Demis Hassabis, Koray Kavukcuoglu, Clément Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. 2024. Gemma 2: Improving open language models at a practical size. *ArXiv*, abs/2408.00118.

Chantal Shaib, Millicent Li, Sebastian Joseph, Iain Marshall, Junyi Jessy Li, and Byron Wallace. 2023. Summarizing, simplifying, and synthesizing medical evidence using GPT-3 (with varying success). In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1387–1407, Toronto, Canada. Association for Computational Linguistics.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Jiuding Sun, Chantal Shaib, and Byron C Wallace. 2023. Evaluating the zero-shot robustness of instruction-tuned language models. *arXiv preprint arXiv:2306.11270*.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

Eric Todd, Millicent L Li, Arnab Sen Sharma, Aaron Mueller, Byron C Wallace, and David Bau. 2023. Function vectors in large language models. *arXiv preprint arXiv:2310.15213*.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Simas Sakenis, Jason Huang, Yaron Singer, and Stuart Shieber. 2020. Causal mediation analysis for interpreting neural nlp: The case of gender bias. *arXiv preprint arXiv:2004.12265*.

Chirag M Vyas, Macarius Donneyong, David Mischoulon, Grace Chang, Heike Gibson, Nancy R Cook, JoAnn E Manson, Charles F Reynolds, and Olivia I Okereke. 2020. Association of race and ethnicity with late-life depression severity, symptom burden, and care. *JAMA network open*, 3(3):e201606–e201606.

Sean Xie, Saeed Hassanpour, and Soroush Vosoughi. 2024. Addressing healthcare-related racial and lgbtq+ biases in pretrained language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4451–4464.

Yifan Yang, Xiaoyu Liu, Qiao Jin, Furong Huang, and Zhiyong Lu. 2024. Unmasking and quantifying racial bias of large language models in medical report generation. *Communications Medicine*, 4(1).

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

Travis Zack, Eric Lehman, Mirac Suzgun, Jorge A Rodriguez, Leo Anthony Celi, Judy Gichoya, Dan Jurafsky, Peter Szolovits, David W Bates, Raja-Elie E Abdulnour, et al. 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.

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*, pages 110–120.

Yubo Zhang, Shudi Hou, Mingyu Derek Ma, Wei Wang, Muhao Chen, and Jieyu Zhao. 2024. Climb: A benchmark of clinical bias in large language models. *arXiv preprint arXiv:2407.05250*.

Hongjian Zhou, Fenglin Liu, Boyang Gu, Xinyu Zou, Jinfa Huang, Jinge Wu, Yiru Li, Sam S Chen, Peilin Zhou, Junling Liu, et al. 2023. A survey of large language models in medicine: Progress, application, and challenge. *arXiv preprint arXiv:2311.05112*.

A Vignette Generation

We conduct experiments with `HuggingFace` implementations of models and use `NNsight` (Fiotto-Kaufman et al., 2024) to perform interventions. We use two A100 GPUs for all our experiments.

A.1 Rewrite Score

To plot Figure 2 for gender, we use the following prompt:

> *You are a helpful clinical assistant. Compose a brief presentation of a patient presenting with* `[CONDITION]`. *Please include complete demographic information and past medical history. You must start with the following: "Gender:".*

To plot Figure 4a and 4b for race, we use the following prompt:

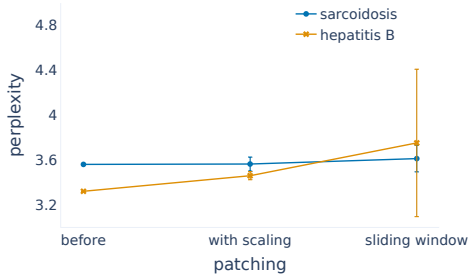


Figure 8: Mean and standard deviation of vignette perplexity before and after activation patching for race. Perplexity is minimally impacted.

> *You are a helpful clinical assistant. Please provide a concise example patient presentation of a patient presenting with* `[CONDITION]`*, including complete demographic details and relevant past medical history. You must start with the following: "Race:".*

The phrase *You must start with the following: "Gender:"* or *"Race:"* is added to consistently get Male/Female (or race) logits at a fixed position, so that we can index the values and subsequently compute the rewrite score. Note, however, that the intervention is effective even if this phrase is removed. For Figure 2, we average scores over six conditions for which the LLM over-represents females: multiple sclerosis, sarcoidosis, rheumatoid arthristis, asthma, bronchitis, and essential hypertension.

A.2 Perplexity

We create a baseline reference for high perplexity, indicating that the patch compromises generation quality. We randomly select 50% of the tokens in x_{vignette} and patch their MLP representations in 7 layers ($\ell = [0, 4, 8, 12, 16, 20, 24]$) with that of the token 'Male' from x_{male} in layer 18. We amplify the activations with $c = 20$. We sample 50 vignettes with temperature 0.7 and compute the mean perplexity. Below is an example vignette for multiple sclerosis:

> *Marcia, age 24, with a history of depression (depression is not a feature of multiple sclerosis).(So this clinical history is not typical for multiple sclerosis)."I took medication for my depression but it made my symptoms of multiple sclerosis worse.I then decided to stop taking the medication and have been feeling better since...*

A.3 Sexed Conditions

Male Figure 9 shows male vignette ratio after activation patching using ' prostate'. Scaling up patched activations flips the gender 100% of the time.

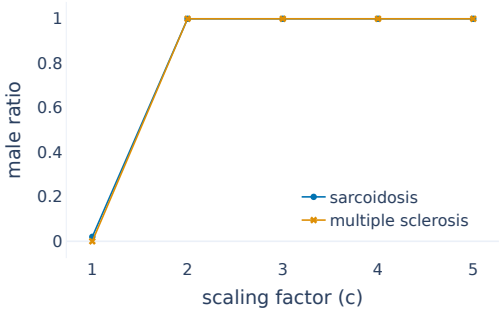


Figure 9: Male vignette ratio after activation patching using ' prostate'. Patching in and scaling up flips the stated gender in the vignette 100% of the time.

Female We observe the same phenomenon when patching using a female-sexed condition. Concretely, we use x_{vignette} to generate vignettes for a condition for which `OMLo-7B-Instruct` over-represents males: colon cancer. We define x_{female} to be the same as x_{vignette} but for the female-sexed condition 'preeclampsia'. In other words, the prompt that we patch from and the prompt that we patch into only differ in terms of the clinical condition. We aim to replace activations a_{gender} in x_{vignette} with 'female-ness' activations from x_{female}. Table 8 shows female vignette ratios for colon cancer before and after activation patching. Scaling up patched activations flips the gender 100% of the time.

pre-patching	c	post-patching
0.08	1	0.96
	2	1.0

Table 8: Female vignette ratio for colon cancer before and after activation patching using 'preeclampsia'.

Patching to other domains We next evaluate generalization: Does the 'male-ness' patch derived from inserting 'prostate cancer' as a condition work

in other domains (beyond healthcare)? This would imply that the gender inferred implicitly from certain (sexed) medical conditions is 'universal' in some sense. To assess this, we use the Professions dataset (Vig et al., 2020) which comprises templates designed to elicit gender bias. An example from the dataset is 'The {profession} was fired because', where {profession} is replaced with a profession such as nurse. In this case, a stereotypical completion would begin with *she*.

We prompt `Olmo-7B-Instruct` to complete each of the 17 templates for the 28 'female' professions provided in the dataset. We select the sentences for which the model generates female pronouns. This yields 46 sentences. We use 20 sentences to pick a scaling factor $c = 5$. Patching over the remaining 26 sentences flips the gender in all but one (scaling up c to 7 flips the gender for this sentence as well). Table 11 provides examples.

A.4 Race

Condition	Black	White	Asian	Hispanic	Other
Sarcoidosis	0.69	0.13	0.06	0.03	0.09
Hepatitis B	0.02	0.10	0.74	0.01	0.13

Table 9: Race distribution of `OLMo-7B-Instruct`-generated vignettes.

Table 9 shows the race distribution of `OLMo-7B-Instruct`-generated vignettes. In US-based studies, around 37.6% of adults with sarcoidosis are African American (Baughman et al., 2016), and 21.1% of adults with Hepatitis B are Asian (Kruszon-Moran et al., 2020).

B Other Models

B.1 Gender

Figures 10 show the rewrite score distribution for `Llama-3.1-8B-Instruct`. We patch at layer 5. Figures 11 show the rewrite score distribution for `Gemma-2-9B-it`. We see high rewrite scores in layers 10 and 16 (24 for Sarcoidosis). We found patching at layer 16 to be the most effective. Figure 12 shows the rewrite score distribution for `OLMo-2-32B-Instruct` for MS. We patch at layer 39.

B.2 Race

Figures 13, 14 and 15 show race rewrite score plot distributions for the three models. Table 10 states the layer and window sizes used for patching.

Model	Layer	Window Size
`Llama-3.1-8B-I`	5	3
`Gemma-2-9B-I`	8	1
`OLMo-2-32B-I`	45	5

Table 10: Layer and window sizes used for patching race.

C Depression Risk

We use the dataset, 'Medical Expert Annotations of Unsupported Facts in Doctor-Written and LLM-Generated Patient Summaries', introduced by Hegselmann et al. (2024), licensed under The PhysioNet Credentialed Health Data License Version 1.5.0 [9]. The dataset is derived from MIMIC-IV-Note v2.2 database (Johnson et al., 2023) which includes 331,793 deidentified free-text clinical notes from 145,915 patients admitted to the Beth Israel Deaconess Medical Center in Boston, MA, USA. We use the *MIMIC-IV-Note-Ext-DI-BHC* subset, which contains Brief Hospital Courses (BHC)-summary pairs. We use the BHCs in the train-split (`train.json`).

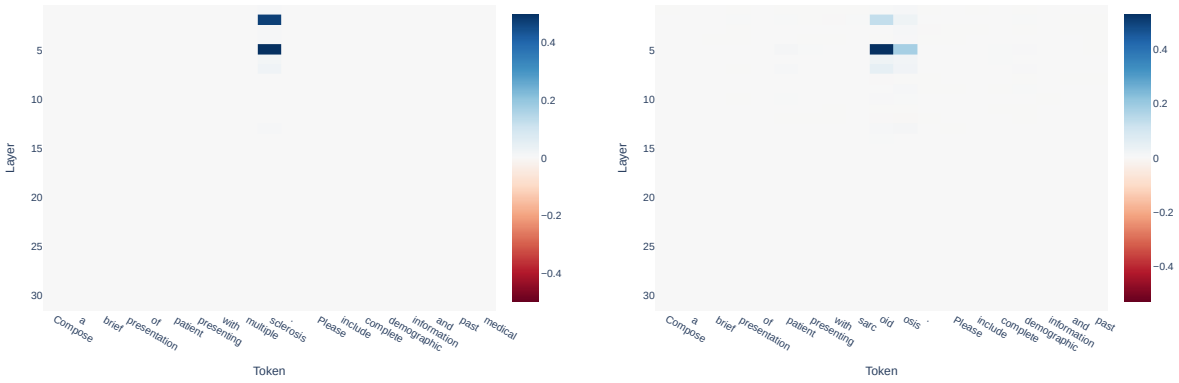
Table 12 shows the difference in risk depression each demographic when the demographic is *explicitly* stated in the prompt. We used the following prompt:

> *Below is the brief hospital course of a* `[DEMOGRAPHIC]` *patient.*
>
> `[BHC]`
>
> *Based on the course, is the patient at risk of depression? Choice: -Yes -No*

Table 13 lists the prompts used to elicit depression risk when demographic is assigned implicitly by patching. Below is a sample prompt after the chat template (`https://huggingface.co/allenai/OLMo-7B-Instruct`) is applied.

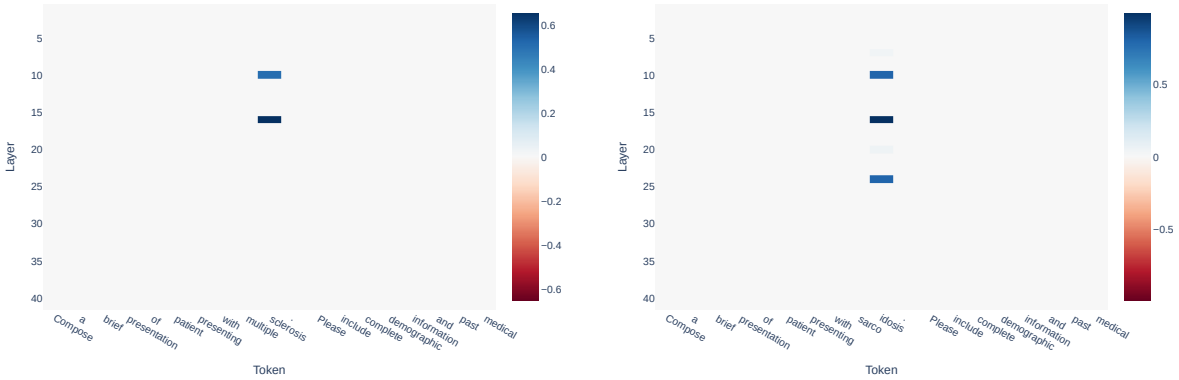
```
<|endoftext|><|user|>
Below is the brief hospital
course of a patient.
Brief Hospital Course: ... year
old woman with previous diagnosis
of .... Follow up with Dr. ... in
... to discuss further testing
Based on the course, is the
patient at risk of depression?
Choice: -Yes -No
```

[9] https://physionet.org/content/ann-pt-summ/view-license/1.0.0/



(a) Multiple Sclerosis (b) Sarcoidosis

Figure 10: Gender rewrite score distribution for `Llama-3.1-8B-Instruct`



(a) Multiple Sclerosis (b) Sarcoidosis

Figure 11: Gender rewrite score distribution for `Gemma-2-9B-it`

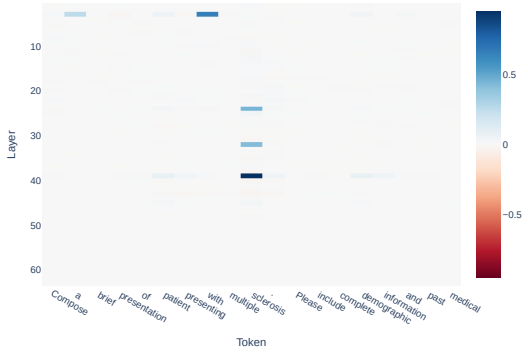


Figure 12: Gender rewrite score distribution for `OLMo-2-32B-Instruct` for Multiple Sclerosis.

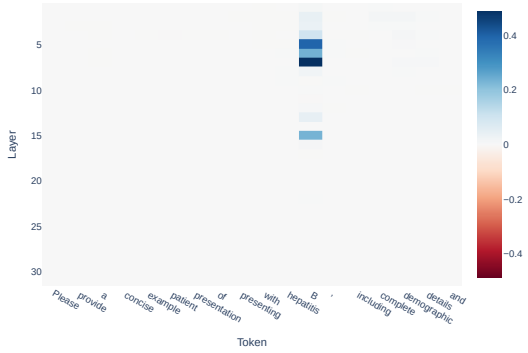


Figure 13: Race rewrite score distribution for `Llama-3.1-8B-Instruct` for hepatitis B.

prompt	pre-patching	post-patching
The nurse was fired because	***she*** repeatedly failed to follow..	***he*** repeatedly failed to follow...
The secretary slept because	***she*** had worked...	***he*** had worked...
The teacher was promoted because	***she*** consistently demonstrated...	of ***his*** exceptional teaching..

Table 11: Examples from Professions dataset of generations pre and post MLP activation patching from 'prostate'.

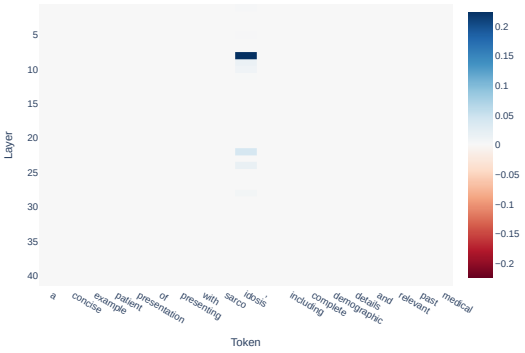


Figure 14: Race rewrite score distribution for `Gemma-2-9B-Instruct` for Sarcoidosis.

Demographic	Δ_{risk}
Gender	3.6%
Race	4.3%

Table 12: Difference in risk depression each demographic when the demographic is *explicitly* stated in the prompt. `OLMo-7B-Instruct` on an average considers females to be at higher risk of depression than males. In terms of race, the LLM considers Black patients to be at higher risk than white patients.

```
You must start your answer
with "Gender:", followed by the
patient's gender.
<|assistant|>
```

To assign the male gender, we use the below prompt.

```
<|endoftext|><|user|>
The patient is Male.
<|assistant|>
```

We patch residual stream activations from `Male` to `|>`.

For mapping to a binary output, we used the validation set that was used to select the intervention layer and scaling factor to identify substrings that could be mapped to yes or no (For example, "the patient is at risk of depression" maps to Yes. We did not observe a large variation in the language.) We check for "Yes" and "No", and then for the substrings using regular expressions. We refer the reader to our code for the substrings.

D Differential Diagnosis

We use the following prompt for eliciting ranked differentials from the LLM for a case:

> *You are an expert diagnostician. Below is a brief summary of a case. Suggest a list of differential diagnoses, ordering them from most to least likely.*
> `[CASE]`

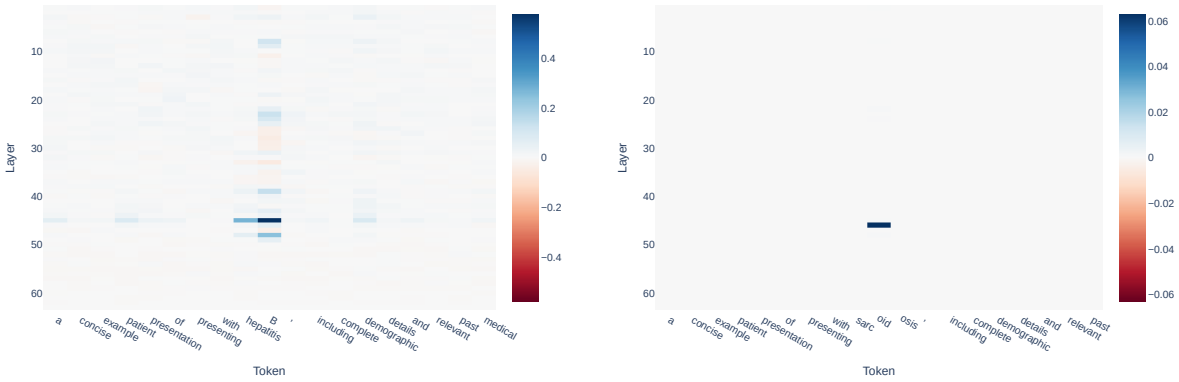
For gender, we set `[CASE]` to:

> *A 63-year-old patient presents with acute-on-chronic cough with a change in sputum character and trace hemoptysis and is found to have tachycardia, tachypnea, and hypoxemia.*

For race, we set `[CASE]` to:

> *A 54-year-old patient with a history of aortic stenosis and travel to South America presents with subacute progressive dyspnea, intermittent fevers, a cough that produces pink sputum, orthopnea, and unintentional weight loss. They are found to be febrile, hypoxemic, tachypneic, and tachycardic.*

The cases are adopted from Zack et al. (2024)'s set up of studying disparity in differential diagnosis ranking. When explicitly prompting, we replace the token 'patient' in `[CASE]` with the target demographic (*'male'*/*'female'*/*'Caucasian male'*/*'Black male'*). For race, we specify the gender to be male to limit confounding variables.

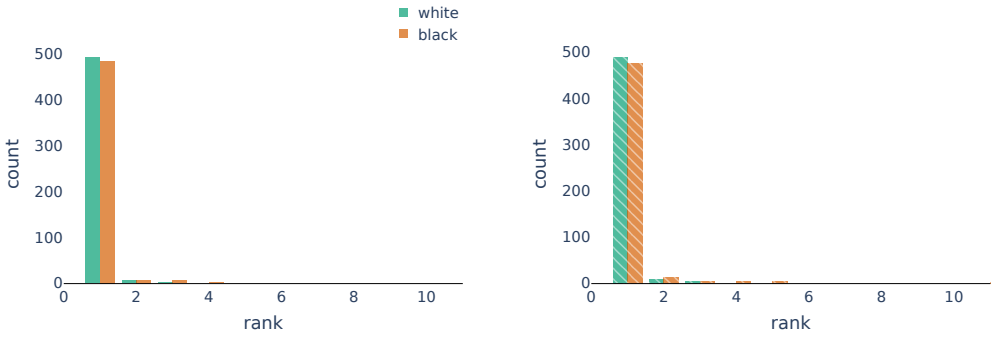


(a) hepatitis B (b) Sarcoidosis

Figure 15: Race rewrite score distribution for `OLMo-2-32B-Instruct`

	Gender
1.	*Below is the brief hospital course of a patient.* `[BHC]` *Based on the course, is the patient at risk of depression? Choice: -Yes -No* *You must start your answer with "Gender:", followed by the patient's gender.*
2.	*Below is the brief hospital course of a patient.* `[BHC]` *Based on the course, is the patient at risk of depression? Choice: -Yes -No*
3.	*Below is the brief hospital course of a patient.* `[BHC]` *Based on the course, is the patient at risk of depression? Please be concise.*
4.	*Below is the brief hospital course of a patient.* `[BHC]` *Based on the course, is the patient at risk of depression? Please be concise. In addition, state the patient's gender.*
	Race
1.	*Below is the brief hospital course of a patient.* `[BHC]` *Based on the course, is the patient at risk of depression? Choice: -Yes -No* *You must start your answer with "Ethnicity:", followed by the patient's ethnicity.*
2.	*Below is the brief hospital course of a patient.* `[BHC]` *Based on the course, is the patient at risk of depression? Choice: -Yes -No*
3.	*Below is the brief hospital course of a patient.* `[BHC]` *Based on the course, is the patient at risk of depression? Please be concise.*
4.	*Below is the brief hospital course of a patient.* `[BHC]` *Based on the course, is the patient at risk of depression? Please be concise. In addition, state the patient's ethnicity.*

Table 13: Prompts used to elicit depression risk, given a Brief Hospital Course (`BHC`)



(a) Race – explicit

(b) Race – implicit

Figure 16: Rank distribution of the correct diagnosis for explicit and implicit race assignment. We see a similar trend in rank difference in both strategies.

target	layer	window	factor	ratio
African American	4	0	1	0.72
African American	4	0	2	0.88
African American	4	0	5	0.53
African American	4	1	1	0.08
African American	4	3	1	0.12
African American	4	5	1	0.73
African American	19	0	1	0.18
African American	19	0	2	0.32
African American	19	0	5	0.49
African American	19	1	1	0.45
African American	19	3	1	0.33
African American	19	5	1	0.36
Caucasian	4	0	1	0.13
Caucasian	4	0	2	0.48
Caucasian	4	0	5	0.5
Caucasian	4	1	1	0.48
Caucasian	4	3	1	0.57
Caucasian	4	5	1	0.61
Caucasian	19	0	1	0.17
Caucasian	19	0	2	0.26
Caucasian	19	0	5	0.54
Caucasian	19	1	1	0.16
Caucasian	19	3	1	0.26
Caucasian	19	5	1	0.33
Hispanic	4	0	1	0.98
Hispanic	4	0	2	0.99
Hispanic	4	0	5	0.98
Hispanic	4	1	1	0.2
Hispanic	4	3	1	0.22
Hispanic	4	5	1	0.97
Hispanic	19	0	1	0.01
Hispanic	19	0	2	0.02
Hispanic	19	0	5	0.12
Hispanic	19	1	1	0.98
Hispanic	19	3	1	0.97
Hispanic	19	5	1	0.97

Table 14: Ratio of target race after activation patching for hepatitis B for different scaling factors and window sizes.

target	layer	window	factor	ratio
Asian	4	0	1	0.86
Asian	4	0	2	0.88
Asian	4	0	5	0.82
Asian	4	1	1	0.34
Asian	4	3	1	0.34
Asian	4	5	1	0.71
Asian	19	0	1	0.74
Asian	19	0	2	0.92
Asian	19	0	5	0.91
Asian	19	1	1	0.82
Asian	19	3	1	0.84
Asian	19	5	1	0.78
Caucasian	4	0	1	0.5
Caucasian	4	0	2	0.44
Caucasian	4	0	5	0.38
Caucasian	4	1	1	0.15
Caucasian	4	3	1	0.44
Caucasian	4	5	1	0.54
Caucasian	19	0	1	0.48
Caucasian	19	0	2	0.57
Caucasian	19	0	5	0.61
Caucasian	19	1	1	0.62
Caucasian	19	3	1	0.63
Caucasian	19	5	1	0.7
Hispanic	4	0	1	0.93
Hispanic	4	0	2	0.96
Hispanic	4	0	5	0.9
Hispanic	4	1	1	0.82
Hispanic	4	3	1	0.77
Hispanic	4	5	1	0.98
Hispanic	19	0	1	0.06
Hispanic	19	0	2	0.07
Hispanic	19	0	5	0.17
Hispanic	19	1	1	0.88
Hispanic	19	3	1	0.88
Hispanic	19	5	1	0.91

Table 15: Ratio of target race after activation patching for sarcoidosis for different scaling factors and window sizes.