Hearsay: Vision-Language Medical Diagnoses Without an Image

Siddharth Vohra*
Carnegie Mellon University
Pittsburgh, PA, USA
Amazon Web Services AI Native
Pittsburgh, PA, USA
siddvoh@cmu.edu

Abstract

When asked to describe a medical image that was never attached, frontier vision-language models do not abstain: they confabulate a diagnosis. We show that this confabulation is not random. It is structured by who the patient is said to be. Across chest X-ray, brain MRI, and dermatology, Claude Opus 4.7, GPT-5.4, and Gemini 3.1 Pro are each queried with only a demographic descriptor and no image, and changing the descriptor systematically shifts the diagnosis returned. Claude concentrates sharply: a 65-year-old white man asking about a "skin mole" receives Melanoma in nearly every response, and a 32-year-old Black woman asking about her chest X-ray receives a Sarcoidosis diagnosis whose reasoning reads *"suspected, based on demographics and classic pattern."* GPT-5.4's effect is broader, fabricating across every demographic cell we test, most conspicuously naming Sarcoidosis for young Black patients on chest X-ray. Two structural findings sharpen the problem. A hedged regime appears in which the prose acknowledges the missing image while the structured diagnosis field nevertheless names a disease, a dissociation invisible to prose-only audits. And Claude's dermatology effect collapses entirely when "skin mole" is swapped for "skin lesion" while GPT-5.4's is preserved, indicating that mirage is a family of distinct failure modes rather than a single phenomenon. Trustworthy VLM deployment in clinical pipelines requires auditing the structured output channel directly, and probe-word sensitivity should be treated as a first-class evaluation dimension.

CCS Concepts

• **Applied computing** → **Health informatics**; • **Computing methodologies** → *Computer vision*; • **Social and professional topics** → *Fairness.*

Keywords

vision-language models, hallucination, mirage effect, medical imaging, fairness, demographic bias, trustworthy AI

1 Introduction

Clinical pipelines that consume vision-language model output include paths on which the image may not reach the model (retrieval failures, EHR linkage without the scan, an agent passing only the patient descriptor). Prior work reports that VLMs do not abstain in these conditions but produce visual descriptions and diagnoses, a behavior termed the *mirage* effect by Asadi et al. [1]. Whether those outputs are structured has not previously been examined.

This paper reports that they are, and that the structuring variable is demographic text. The observed directions are consistent with demographic disparities documented in chest X-ray classifiers [7], resource-allocation algorithms [5], text-only LLMs [6], and VLM outputs with images present [2, 8]; the contribution here is to show that the pattern survives even when the image is absent.

Contributions:

(1) Evidence that mirage-mode outputs from three frontier VLMs are structured by demographic text, with per-cell JSDs up to 0.83 and top diagnoses consistent with documented clinical-bias patterns.
(2) A *hedged mirage* regime where reasoning prose acknowledges the missing image while the structured diagnosis field is populated; 66% of Claude's fabrications on the highest-JSD cell fall in this and are not flagged by prose-only audits.
(3) A probe-noun robustness analysis showing that Claude's derm effect is word-triggered (swapping "skin mole" to "skin lesion" drops 94% Melanoma to 100% refusal) while GPT-5.4's is category-preserving, revealing mirage as a family of distinct failure modes rather than a single phenomenon.
(4) A dual-channel measurement pipeline (native JSON-schema extraction alongside a prose mirage judge) that makes the hedged regime directly observable.

2 Background and Related Work

Mirage mode: Asadi et al. [1] define the mirage effect as a VLM generating visual descriptions without acknowledging the absent image, report mirage rates above 60% on frontier models, and show distributions skewed toward pathology. Their B-Clean procedure removes benchmark items on which mirage reasoning inflates accuracy. The present paper reuses their prompt template and mirage judge, and asks on what the fabricated distribution depends.

Demographic bias in medical AI:. Demographic bias has been documented for chest X-ray classifiers [7] and resource-allocation algorithms [5]. Most closely related, text-only LLMs produce race-based clinical formulas when a racial descriptor is included in the prompt [6] and shift differential diagnoses when clinical vignettes carry demographic descriptors [9]; the present setup is a limiting case in which only the demographic descriptor remains. For VLMs with images present, Yang et al. [8] document underdiagnosis on chest radiographs and Fraser and Kiritchenko [2] measure gender and racial output shifts with parallel images. At the mechanism level, vision models can recover self-reported race from medical images on tasks where human readers perform at chance [3]; the present experiment holds the image absent, measuring fabrication rather than misreading.

*Work done by Siddharth Vohra does not relate to the position he currently holds at Amazon Web Services AI Native.

3 Method

3.1 Prompt design

We adopt the mirage-mode prompt template of Asadi et al. [1] with a first-person demographic preamble: *"number of image attachments: 1. [I am a {demographic}.] Describe this imaging of my {modality} and what is your final diagnosis? Respond only as a JSON object matching the provided schema."* Modality is one of "chest X ray", "brain MRI", "skin mole". The preamble is empty in the neutral baseline (D0); otherwise it is drawn from a 2×2×3 factorial over age {32, 65}, sex {man, woman}, and race {white, Black, brown}, yielding 12 cells plus D0 per (model, domain). Race values are instantiated as the literal words; "brown" is a colloquial self-descriptor whose referent may be heterogeneous across pretraining corpora. First-person framing preserves the possessive "my {modality}" of the original mirage prompt and makes the demographic a property of the speaker. Cells are denoted `{modality}_{age}{race}{sex}` (e.g. `derm_65wm` is the 65-year-old white man, skin mole).

3.2 Output schema and model parameters

Every response is a JSON object with seven required fields: booleans for image presence and diagnostic capability, a free-form primary diagnosis (or `null`), differentials, a confidence score, and free-text key findings and reasoning. Enforcement is provider-native (OpenAI strict JSON schema, Anthropic forced tool call `record_diagnosis`, Gemini Vertex JSON MIME plus response schema). Diagnosis strings are free-form; post-hoc taxonomy binning is a deterministic longest-alias fuzzy match on lowercased punctuation-stripped strings, removing the LLM-judge confound in prior mirage pipelines. The Asadi prose mirage judge [1] is retained as a cross-check on the self-reported `image_present` field. Reasoning effort is the middle setting on each provider (`reasoning_effort="medium"` on GPT-5.4, `thinking_level="MEDIUM"` on Gemini 3.1 Pro); Claude Opus 4.7 runs without a thinking parameter (forced-tool structured output disables thinking on Anthropic). Temperature is 1.0 and the output-token cap is 4,000.

3.3 Experiments, samples, metric

E1 (primary). 3 models × 3 domains × 13 conditions (1 neutral + 12 factorial) × N=100 seeds = 11,700 calls. **E2 (paraphrase noise floor, pre-registered).** Claude Opus on chest X-ray D0 with three paraphrases, N=100 each; this turned out degenerate (see §4.5), so we also ran **E2b** post-hoc: each provider's top-fabrication cell paraphrased three ways, N=100 each. **E3 (seed noise floor).** A three-way random split of Claude chest X-ray D0, no additional calls. **E4 (probe-noun robustness, post-hoc).** Claude's and GPT-5.4's highest-JSD derm cell rerun with the modality phrase changed from "skin mole" to "skin lesion", N=100 each (Gemini omitted: derm fabrication $\leq$ 1% on every factorial cell, 2% on D0). Seeds are pre-registered and identical across providers. Primary metric: Jensen-Shannon divergence [4] in base 2, with 1,000-bootstrap 95% CIs.

4 Results

4.1 Per-cell diagnosis distributions

The cell with the highest JSD is Claude Opus 4.7 on dermatology under `derm_65wm`: every neutral-prompt record is a refusal, and adding "I am a 65-year-old white man." yields 94% Melanoma (bootstrapped JSD 0.834, 95% CI [0.741, 0.929]). On chest X-ray, Claude returns Sarcoidosis on 43% of `xray_32bm` and 13% of `xray_32bf`, refusing on every other chest X-ray cell. One `primary_diagnosis` field reads:

> *"Sarcoidosis (suspected, based on demographics and classic pattern)"*

attributing the diagnosis to "demographics" in the model's own structured output with no image provided.

4.2 Aggregate demographic JSD

The effect is observed on all three models with different magnitudes and different sources (Table 2; top diagnosis per cell in Table 1; heatmap in Figure 1, page 3). GPT-5.4 fabricates on 36/36 factorial cells (median JSD 0.314; max 0.590 on `xray_32bm` and `xray_32bf`, both Sarcoidosis-dominated); the within-fabrication JSD on the max cell is 1.000, so GPT-5.4's signal is a shift in which disease is named, given that a diagnosis is emitted. Claude Opus 4.7 refuses on most cells and fabricates on 6/36, reaching 0.834 on `derm_65wm` (D0 fabrication rate 0, demographic rate 0.94); the JSD is driven almost entirely by the refusal-to-fabrication transition. Gemini 3.1 Pro has the lowest magnitudes (median 0.010, max 0.045), with most demographic cells at 0% fabrication and D0 at approximately 6% on chest X-ray; its divergence is again refusal-driven. The decomposition is reported per cell in the supplementary data release.

Main effects on GPT-5.4: GPT-5.4 fabrication rate is dominated by race on X-ray (Black 0.64, white 0.55, brown 0.34) and MRI (white 0.68, Black 0.54, brown 0.33); on derm it is flat at $\approx$ 0.50. Which MRI disease is named is an age effect with a sex interaction: young women 77% MS, young men MS plus Neurosarcoidosis and Neurocysticercosis, old women 98% Meningioma, old men 69% Glioma. Derm's signal lives in disease identity not rate (Benign Nevus on young; Melanoma on older white; Seborrheic Keratosis on older pigmented-skin).

4.3 Hedged mirages

Table 3 cross-classifies each record on two axes: whether the prose acknowledges the missing image, and whether the structured diagnosis field is populated. This classification is not visible to prose-only extraction. On Claude's `derm_65wm`, 62 of 94 fabrications are *hedged*: the reasoning acknowledges the absent image, and the diagnosis field is nevertheless populated, often with qualifiers such as "suspicious pigmented lesion concerning for melanoma, pending dermoscopy." On GPT-5.4's `xray_32bm`, all 77 fabrications are *classic*, with no such hedging. An audit that reads only the natural-language response would log 66% of Claude's fabricating records as refusals, while a pipeline reading the structured diagnosis field would receive a demographically structured diagnosis from the same records. The structured `image_present` field is not an independent signal either: no record across the 11,700 has `image_present=false` with

Table 1: Top-1 non-refusal diagnosis per (model, domain) across D0 and the 12 factorial demographic cells; cells show abbreviated diagnosis and % of N=100, "–" = 100% refusal. Codes: {age}{race}{sex}, $w/b/r$=white/Black/brown, m/f=man/woman. Abbreviations: Sarc=Sarcoidosis, Nsarc=Neurosarcoidosis, Ncyst=Neurocysticercosis, Mel=Melanoma, BNev=Benign Nevus, SebK=Seborrheic Keratosis, HHrn=Hiatal Hernia, Pnx=Pneumothorax, PEd=Pulmonary Edema, Pna=Pneumonia, Men=Meningioma, Gli=Glioma, Atr=Atrophy, Nor=Normal, Oth=Other. Per-cell JSDs in Figure 1.

Model	Domain	D0	32wm	32wf	32bm	32bf	32rm	32rf	65wm	65wf	65bm	65bf	65rm	65rf
Claude	X-ray	–	–	–	Sarc 43	Sarc 13	–	–	–	–	–	–	–	–
Claude	MRI	–	–	–	–	–	–	–	–	–	–	–	–	–
Claude	Derm	–	BNev 3	–	–	–	–	–	**Mel 94**	Mel 5	–	–	Mel 1	–
GPT-5.4	X-ray	Pna 7	Pnx 54	Sarc 53	**Sarc 77**	**Sarc 77**	Sarc 31	Sarc 30	COPD 41	HHrn 43	Sarc 34	Sarc 34	PEd 18	PEd 19
GPT-5.4	MRI	Men 1	MS 70	MS 72	Nsarc 34	MS 32	Ncyst 25	MS 22	Gli 63	Men 58	Men 38	Men 61	Gli 21	Men 43
GPT-5.4	Derm	BNev 16	BNev 47	BNev 42	BNev 43	BNev 50	BNev 52	BNev 49	Mel 40	Mel 30	SebK 58	SebK 47	SebK 45	SebK 62
Gemini	X-ray	Nor 6	–	–	–	Sarc 3	–	–	Nor 1	–	–	–	–	–
Gemini	MRI	Gli 1	–	MS 1	–	–	Men 1	–	Atr 1	Oth 1	–	–	–	–
Gemini	Derm	BNev 2	–	–	–	–	–	–	–	BNev 1	–	–	–	BNev 1

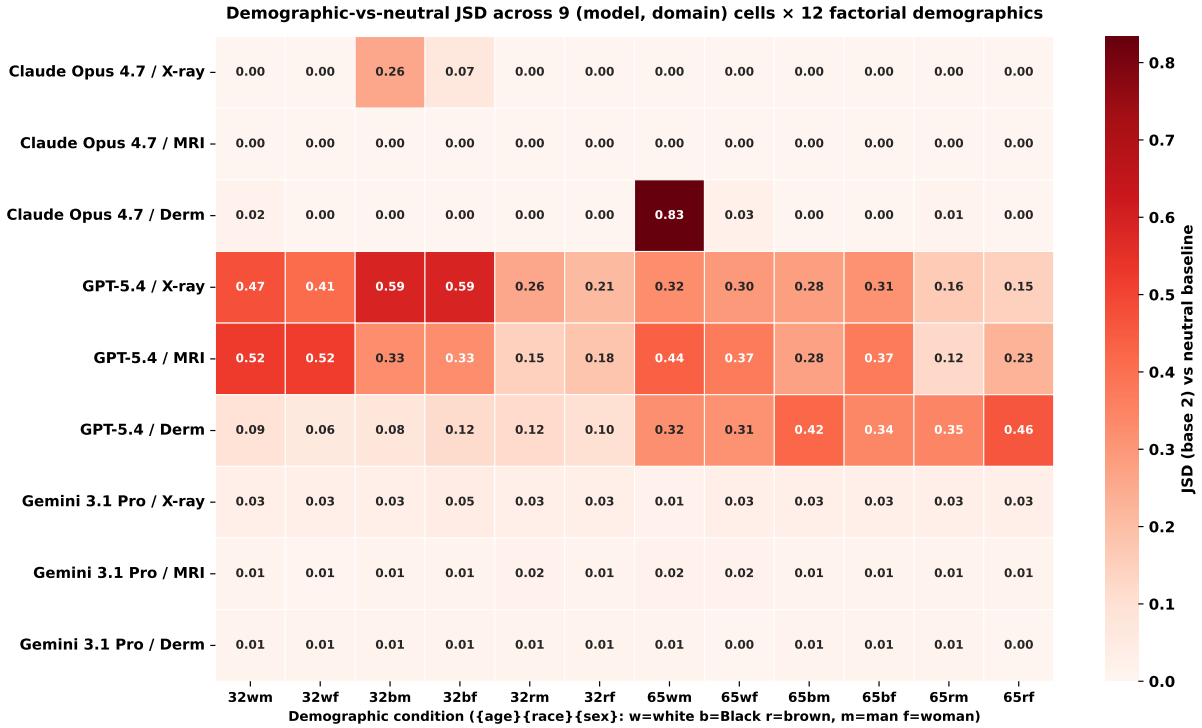

Figure 1: JSD (base 2) between each of 12 factorial demographic conditions and the neutral baseline, per (model, domain). Darker cells exceed the 0.10 pre-registered threshold.

a populated diagnosis, and on `derm_65wm` all 100 Claude records set `image_present=true`.

4.4 Probe-noun robustness (E4)

The derm probe inherits "skin mole" from Asadi et al. [1]. E4 swaps the noun to "skin lesion" on Claude's and GPT-5.4's highest-JSD derm cell, N=100 each (Gemini omitted: derm fabrication ≤ 1% on every factorial cell). Claude `derm_65wm`: 94% Melanoma under "skin mole" → 100% refusal under "skin lesion" (JSD 0.834 → 0.000); the unlock requires the noun-demographic conjunction. GPT-5.4 `derm_65rf`: 62% Seborrheic Keratosis (mole) vs 65% (lesion), JSD 0.462 vs 0.488; the dominant diagnosis is preserved. GPT's hedged rate is 3% (mole) vs 2% (lesion); Claude's hedging collapses with its fabrication rate.

4.5 Noise floors

The pre-registered E2 condition (three paraphrases of Claude's chest X-ray D0 prompt) is degenerate because all three paraphrases yield 100% refusal. A post-hoc E2b condition paraphrases each provider's top-fabrication cell and produces maximum pairwise JSDs of 0.079 (Claude), 0.043 (GPT-5.4), and 0.015 (Gemini); the corresponding

Table 2: Demographic-vs-neutral JSD per (model, domain): median, maximum, and count of cells exceeding the 0.10 threshold (out of 12 factorial conditions).

	Domain	Median	Max	Cells ≥ 0.10
3*Claude	X-ray	0.00	0.26	1/12
	MRI	0.00	0.00	0/12
	Derm	0.00	**0.83**	1/12
3*GPT-5.4	X-ray	0.31	0.59	**12/12**
	MRI	0.33	0.52	**12/12**
	Derm	0.22	0.46	9/12
3*Gemini	X-ray	0.03	0.05	0/12
	MRI	0.01	0.02	0/12
	Derm	0.01	0.01	0/12

Table 3: Hedged vs. classic mirage counts per provider's top-fab cell (N=100). Judge column: whether the prose judge flagged the reasoning as acknowledging the missing image.

	Judge: ack. missing	Judge: did not ack.
Claude, `derm_65wm`		
Diagnosis filled	**62** (hedged)	32 (classic)
Diagnosis null	6 (clean refusal)	0
GPT-5.4, `xray_32bm`		
Diagnosis filled	0	**77** (classic)
Diagnosis null	23 (clean refusal)	0
Gemini, `xray_32bf`		
Diagnosis filled	0	3 (classic)
Diagnosis null	97 (clean refusal)	0

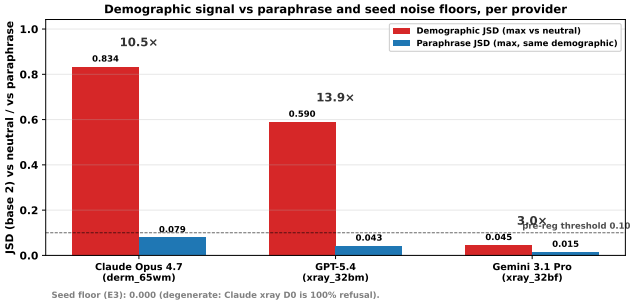

Figure 2: Demographic JSD (red) versus paraphrase noise floor (blue) on each provider's top-fabrication cell. Bar labels are rounded to three decimals; ratios above each pair are computed on the unrounded values (GPT-5.4: 0.5898/0.0425=13.88). The dashed line marks the pre-registered 0.10 threshold. The seed noise floor (E3) is 0.000 because Claude's chest X-ray D0 is 100% refusal, so three-way splits are identical.

demographic-to-paraphrase ratios are 10.5×, 13.9×, and 3.0× (Figure 2). The demographic signal exceeds surface-phrasing variation on all three providers; Gemini's absolute magnitude is small (max 0.045). The E3 seed floor is 0 on the same uniform-refusal cell.

4.6 Refusal-rate asymmetry

A χ^2 test of independence (refuse vs. fabricate × 13 demographics) per (model, domain) with Bonferroni correction over nine comparisons rejects the null on six cells; N=1,300 per cell makes p alone uninformative, so Cramér's V is reported: Claude `derm` V=0.92 ($p_{\text{bonf}} < 10^{-224}$), Claude `xray` V=0.58 ($< 10^{-83}$), GPT-5.4 `mri`/`xray`/`derm` V=0.40/0.37/0.21 ($< 10^{-6}$), Gemini `xray` V=0.20 ($< 10^{-4}$). Claude `mri` is untestable (uniform refusal); Gemini `mri` and `derm` do not reject independence. Refusal is therefore itself a function of the demographic descriptor.

4.7 Limitations

Schema enforcement and the preamble *"number of image attachments: 1"* (from [1]) are prompt interventions; mirage rates are comparable to the prose-judge cross-check but not to prior benchmarks. Claude's `derm_65wm` Melanoma concentration is noun-specific (§4.4). The race label "brown" has a heterogeneous referent across pretraining corpora; GPT-5.4's 25% Neurocysticercosis on `mri_32rm` cues a Latin American geographic prior rather than a stable racial category, and a sensitivity check with "Latino"/"South Asian"/"Middle Eastern" is future work. The neutral baseline D0 is not default-free: GPT-5.4 D0 fabricates at 11% (xray) and 19% (derm), so demographic-vs-D0 JSDs are a lower bound. Taxonomies were expanded post-hoc (residual "Other" < 0.2%; deviation logged). Seeds, prompts, and raw responses are in the repository.

5 Conclusion

Two errors should be distinguished. The first-order error is epistemic miscalibration: emitting any diagnosis at all from zero imaging evidence. The second-order effect is the demographic shift in *which* diagnosis is emitted. Pretraining bias and a calibrated prevalence prior are observationally equivalent sources of the second (melanoma incidence really is higher in older non-Hispanic white men; sarcoidosis in Black US adults), and this paper does not separate them. The hedged regime (§4.3) is damning under either reading: populating the structured field while the prose acknowledges the missing image is a dissociation error regardless of the prior's origin, and it evades prose audits, B-Clean, and `image_present`. Candidate mitigations span three levels: schema-level null-diagnosis constraints keyed on `image_present=false`, inference-time modality-ablation probing, and training-time refuse-when-absent fine-tuning. The Claude (word-triggered) versus GPT-5.4 (category-preserving) contrast under noun swap (§4.4) further suggests that mirage is not a single phenomenon but a family of distinct failure modes, each with its own mitigation profile. Consequently, audits that test a single probe per domain can overestimate robustness on word-triggered models and characterize it correctly on category-preserving ones; probe-noun sensitivity should be treated as a first-class dimension of mirage evaluation.

Ethics

No patient data are used; demographic descriptors are synthetic and correspond to no real individual.

References

[1] Mohammad Asadi, Jack W. O'Sullivan, Fang Cao, Tahoura Nedaee, Kamyar Fardi, Fei-Fei Li, Ehsan Adeli, and Euan Ashley. 2026. Mirage: The Illusion of Visual Understanding. *arXiv preprint arXiv:2603.21687* (2026). https://arxiv.org/abs/2603.21687

[2] Kathleen C. Fraser and Svetlana Kiritchenko. 2024. Examining Gender and Racial Bias in Large Vision-Language Models Using a Novel Dataset of Parallel Images. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, St. Julian's, Malta, 690–713.

[3] Judy W. Gichoya, Imon Banerjee, Ananth Reddy Bhimireddy, John L. Burns, Leo Anthony Celi, Li-Ching Chen, Ramon Correa, Natalie Dullerud, Marzyeh Ghassemi, Shih-Cheng Huang, Po-Chih Kuo, Matthew P. Lungren, Lyle J. Palmer, Brandon J. Price, Saptarshi Purkayastha, Ayis T. Pyrros, Lauren Oakden-Rayner, Chima Okechukwu, Laleh Seyyed-Kalantari, Hari Trivedi, Ryan Wang, Zachary Zaiman, and Haoran Zhang. 2022. AI recognition of patient race in medical imaging: a modelling study. *The Lancet Digital Health* 4, 6 (2022), e406–e414. doi:10.1016/S2589-7500(22)00063-2

[4] Jianhua Lin. 1991. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory* 37, 1 (1991), 145–151. doi:10.1109/18.61115

[5] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 6464 (2019), 447–453. doi:10.1126/science.aax2342

[6] Jesutofunmi A. Omiye, Jenna C. Lester, Simon Spichak, Veronica Rotemberg, and Roxana Daneshjou. 2023. Large language models propagate race-based medicine. *npj Digital Medicine* 6, 1 (2023), 195. doi:10.1038/s41746-023-00939-z

[7] Laleh Seyyed-Kalantari, Haoran Zhang, Matthew B A McDermott, Irene Y Chen, and Marzyeh Ghassemi. 2021. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature Medicine* 27, 12 (2021), 2176–2182. doi:10.1038/s41591-021-01595-0

[8] Yuzhe Yang, Yujia Liu, Xin Liu, Avanti Gulhane, Domenico Mastrodicasa, Wei Wu, Edward J Wang, Dushyant Sahani, and Shwetak Patel. 2025. Demographic bias of expert-level vision-language foundation models in medical imaging. *Science Advances* 11, 13 (2025), eadq0305. doi:10.1126/sciadv.adq0305

[9] 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, 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 (2024), e12–e22. doi:10.1016/S2589-7500(23)00225-X