The bias is in the generator: locating medical chain-of-thought subgroup gaps upstream of the verifier

Micky C. Nnamdi[♠], **Benoit L. Marteau**[♠], **Shaun Qien Yeau Tan**[♠], **J. Ben Tamo**[♠], **Yining Yuan**[♠], and **May D. Wang**[†♠]

[♠]Georgia Institute of Technology
†Corresponding author: maywang@gatech.edu

Abstract

Verifier-guided best-of-N decoding has two stages: a generator samples candidate chains of thought, and a verifier selects among them. A subgroup disparity may already be present in the generator's candidates or be amplified during selection. We separate the two stages using adversarial demographic counterfactuals of medical exam questions. On HuatuoGPT-o1-8B, a substantial race accuracy gap is already present in a single sampled response, before verifier selection. Across seven generators, only Gemma-3-27B-it under Skywork-RM shows significant verifier-induced amplification after multiple-testing correction. Even an oracle selector leaves a residual gap on every medical generator analyzed, showing that correct answers are less often available for one subgroup. Counterfactually paired low-rank adaptation (LoRA) of the generator reduces the single-sample gap on two medical models, although on HuatuoGPT-o1-8B the reduction does not persist after self-judge selection. Single-variant controls indicate that the single-sample reduction is not explained by additional training steps alone. A short fairness instruction modestly reduces the single-sample gap, and a direct preference optimization (DPO) variant of the paired training also reduces it and, on HuatuoGPT-o1-8B, keeps race mention near its baseline rate. For the medical generators studied, candidate generation rather than inference-time selection is the primary locus for mitigating this gap.

1 Introduction

As medical large language models (LLMs) are evaluated beyond board-exam benchmarks for uses including triage, diagnostic assistance, and patient-facing applications, their behavior across patient subgroups becomes important: variants of the same clinical case can elicit different answers when demographic information changes. Recent work has documented such biases across race, sex, and socioeconomic status (Pfohl et al., 2024; Xiao et al., 2025; Zhang et al., 2024; Rawat et al., 2024; Ghosh et al., 2025).

Many reasoning systems use *verifier-guided best-of-N* (BoN) decoding (Cobbe et al., 2021; Lightman et al., 2024): a generator samples N candidate reasoning trajectories, a verifier scores them, and the highest-scoring candidate is returned. A subgroup gap in such a system can arise at either stage. **The generator may make correct answers available at different rates across subgroups, or the verifier may select differently among the candidates that are available.** The distinction matters for mitigation: if selection enlarges the gap, changing the verifier may address it; if correct answers are missing from one subgroup's candidate pool, changing the verifier alone cannot recover them.

Prior work makes verifier-induced amplification plausible: reward models and process verifiers can fail to track reasoning correctness (Tiwari et al., 2026), and chain-of-thought (CoT) reasoning has been reported to amplify bias (Shaikh et al., 2023), a finding that motivated fairness reward models (Hall et al., 2025). These observations suggest an *amplification hypothesis*: a verifier may rank candidates differently across demographic variants and enlarge a gap already present at generation.

We test this hypothesis by separating candidate availability from candidate selection. We measure the subgroup gap in a single sampled response (N=1), before verifier selection; track how it changes as learned verifiers select from larger pools; and compare learned selection with an oracle selector that returns a correct candidate whenever one is available. We apply this decomposition to AMQA demographic counterfactuals across three demographic axes, seven generators, and verifiers

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 5027–5050
November 6-10, 2026 ©2026 Association for Computational Linguistics

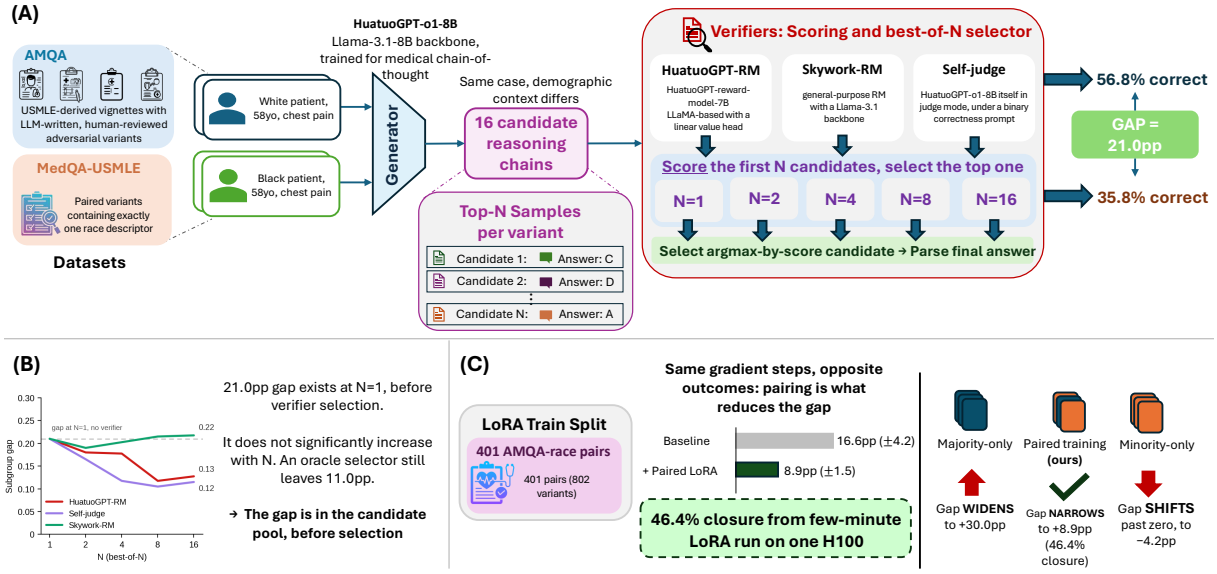


Figure 1: **The subgroup gap is present before verifier selection.** HuatuoGPT-o1-8B on AMQA-race; gap = white minus black accuracy, in percentage points (**pp**). **(A)** Pipeline and the gap at $N=1$ in the primary run. **(B)** Gap versus N under the three verifiers, none of which increases it significantly; the oracle residual at $N=16$ is annotated. **(C)** Five-seed mean ($\pm$ s.d.) $N=1$ gap after a 401-pair counterfactually paired low-rank adaptation (LoRA) and after step-matched majority-only and minority-only training; the baseline mean (16.6pp) differs from the primary run in (A), and on this generator the LoRA reduction does not persist under self-judge selection at $N=16$.

with different ranking behavior.[1] The results support three conclusions (Figure 1).

- **Adversarial demographic counterfactuals reveal substantial generator-level subgroup gaps.** On AMQA, HuatuoGPT-o1-8B shows a race gap in a single sampled response, with further gaps along sex and income. Simpler substitutions on MedQA-USMLE and DiversityMedQA, and LLM-edited counterfactuals on MedMCQA, do not recover a comparable directional gap, so the measured disparity depends on how the probe is constructed.

- **Verifier-guided BoN rarely amplifies the gap.** On HuatuoGPT-o1-8B, no axis-verifier combination shows significant amplification after Holm-Bonferroni correction, and an oracle selector leaves a substantial residual gap because correct candidates are less often available for one subgroup. The other generators follow this pattern except Gemma-3-27B-it, where Skywork-RM adds disparity (§4.4).

- **Counterfactually paired training at the generator reduces the gap.** A LoRA fine-tune on AMQA race pairs reduces the single-sample gap across seeds on HuatuoGPT-o1-8B and Med42-8B, although on HuatuoGPT-o1-8B the reduction does not persist after self-judge selection at $N=16$; single-variant controls matched in gradient steps move the gap in opposite directions. Because this LoRA replaces the CoT with a short template, we evaluate a direct preference optimization (DPO) variant, which also reduces the gap and, on HuatuoGPT-o1-8B, largely preserves the baseline rate of race mention.

Taken together, these results identify the **generator's candidate distribution as the primary intervention locus** for the medical generators and the inference-time BoN selection studied here; they do not show that verifiers cannot introduce disparities.

2 Related Work

Reward-model and verifier reliability. Recent work has questioned whether reward models reliably distinguish correct reasoning from superficially plausible trajectories. Tiwari et al. (2026) introduce a three-tier diagnostic framework for evaluating process reward models under static perturbations and reward-hacking settings, and show that high-scoring reasoning need not be logically correct, including under policies optimized directly

[1] Code and reproduction scripts: https://github.com/micknn/bias-is-in-the-generator

against the reward model. We adapt their static-perturbation methodology to the medical setting, adding a key-step-drop perturbation and a logic-preserving deletion that controls for text-length sensitivity.

Fairness reward models. On recidivism, content-moderation, and job-screening tasks, Hall et al. (2025) train a Fairness Reward Model (FRM) on weakly supervised labels of biased and unbiased reasoning and use it to down-weight biased trajectories when aggregating decisions across sampled reasoning chains. They treat the selector as an intervention point for reducing subgroup disparities; we test whether that locus carries over to medical reasoning by measuring disparities both in the candidate pool and after selection.

Medical-LLM subgroup benchmarks. EquityMedQA (Pfohl et al., 2024), AMQA (Xiao et al., 2025), CLIMB (Zhang et al., 2024), DiversityMedQA (Rawat et al., 2024), counterfactual patient variations (Benkirane et al., 2025), and MEDEQUALQA (Ghosh et al., 2025) differ in how demographic variation is introduced and in the outcome they measure. We use AMQA's adversarial counterfactual pairs as the primary probe and contrast them with direct substitutions on MedQA-USMLE (Jin et al., 2021) and DiversityMedQA, and with LLM-edited counterfactuals on the independent MedMCQA corpus (Pal et al., 2022).

Verifier-guided decoding in medical reasoning. HuatuoGPT-o1 (Chen et al., 2025) uses a GPT-4o-based verifier (distinct from the HuatuoGPT-RM used here) that checks the correctness of model outputs, first to guide the search for reasoning trajectories used in fine-tuning and then to provide rewards during reinforcement learning. We instead study inference-time selection over a fixed generator's samples.

3 Experimental Setup

3.1 Datasets

Our analysis requires paired demographic variants that share one gold answer, because both subgroup accuracy gaps and oracle candidate availability are defined relative to that answer. EquityMedQA (Pfohl et al., 2024) uses open-ended, rubric-based evaluation, CLIMB (Zhang et al., 2024) measures intrinsic association bias and performance shifts in multi-label diagnosis prediction, and MEDEQUALQA (Ghosh et al., 2025) varies only patient pronouns, so none of the three provides race-paired variants that share one gold answer. Counterfactual patient variations (Benkirane et al., 2025) do provide such pairs, built by rule-based substitution of demographic descriptors in JAMA Clinical Challenge cases; our MedQA-USMLE and DiversityMedQA pairs follow the same direct-substitution construction.

AMQA (Xiao et al., 2025) contains 801 USMLE-derived vignettes from MedQA and provides, for each vignette, LLM-written variants along three axes: race (white vs. black), sex (male vs. female), and income (high vs. low). Each variant adds demographic context to the same case and keeps its options and gold answer. The context is adversarial by design: it was written to favor the correct answer for the first group of each pair and to steer the second toward an incorrect answer. Variants that failed to do so were regenerated up to twice with feedback from an evaluation model. We use the release that human reviewers checked and revised (`AMQA_Dataset.jsonl` in the authors' repository). Every axis is evaluated on vignettes 1 to 400; the remaining 401 race pairs are reserved for generator-side training (§4.3), so training and evaluation cases are disjoint.

MedQA-USMLE (Jin et al., 2021) shares AMQA's source corpus. Of the first 400 training-split questions containing exactly one race or color term, we keep the 208 in which the term describes the patient (the rest are uses such as "white blood cells") and form each pair by setting that descriptor to "white" or "African American", which leaves the gold answer unchanged.

Two further datasets test whether a gap depends on counterfactual construction or source corpus. **DiversityMedQA** (Rawat et al., 2024) adds ethnicity or gender changes to MedQA vignettes; we use 400 of its race pairs. **MedMCQA** (Pal et al., 2022) is an Indian medical-entrance-exam corpus independent of the USMLE source; from it we construct LLM-edited race counterfactuals under a neutral race swap and an asymmetric social framing (§4.4, Appendix I).

3.2 Generator and verifiers

Our primary generator is **HuatuoGPT-o1-8B** (Chen et al., 2025), built on Llama-3.1-8B-Instruct (Llama Team, AI @ Meta, 2024) and trained for medical CoT reasoning. A *verifier* assigns each

sampled trajectory a scalar score, and BoN returns the highest-scoring candidate among the first N samples (ties go to the earlier sample). We use three inference-time verifiers, chosen *a priori* to cover a medical reward model, self-evaluation, and a general-purpose reward model:

- **HuatuoGPT-RM** (`FreedomIntelligence/HuatuoGPT-reward-model-7B`): a publicly released LLaMA-based (Touvron et al., 2023) medical reward model (Zhang et al., 2023) with a linear value head on the last token's hidden state.
- **Self-judge**: the generator used as its own correctness judge (HuatuoGPT-o1-8B for the primary generator; for LoRA-tuned generators, the untuned base model). Given a binary correctness prompt (Appendix B), the score is $r = \log p(\text{"True"}) - \log p(\text{"False"})$, using next-token log-probabilities (Kadavath et al., 2022). Its input carries a duplicated beginning-of-sequence (BOS) token, and removing the duplicate changes its selections. The Skywork-RM and FRM inputs carry the same duplicate; rescoring Skywork-RM with a single BOS leaves its results essentially unchanged, and FRM was not rescored (Appendix C).
- **Skywork-RM** (`Skywork/Skywork-Reward-Llama-3.1-8B-v0.2`) (Liu et al., 2024): a general-purpose reward model with a Llama-3.1 backbone.

§4.1 characterizes these three verifiers and, with a perturbation diagnostic, three additional ones, and §4.4 reports other generators.

3.3 Generation and selection

For each question and variant we independently sample $N_{\max} = 16$ trajectories (temperature 0.7, top-p 0.9, 512 new tokens). For each verifier and $N \in \{1, 2, 4, 8, 16\}$, BoN selects among the first N samples, so changes with N reflect selection from an expanding prefix of one fixed candidate pool rather than from resampled sets. We call the run with sampling seed 0 the primary run of each generator, dataset, and axis; the verifier analyses (§4.1 and §4.2), the multi-generator comparison, and the dataset probes (§4.4) use it. Unless stated otherwise, intervention results are means over five runs with seeds 1 to 5 (the training and sampling seed), and their baselines are the means of five baseline runs with the same seeds. A separately sampled AMQA-race run with $N_{\max} = 32$ checks the amplification result at larger N. A lenient regular-expression parser, which also recognizes answer-first outputs, extracts the final answer; relative to a strict "answer is X" pattern, it extracts an answer from an additional 10.9 percentage points (pp) of MedQA-USMLE generations. Outputs without an answer, mostly chains truncated at the token limit, are scored as incorrect; Appendix A reports their rates by subgroup and alternative scoring conventions.

3.4 Metrics and statistics

Accuracy is the fraction of questions answered correctly. The *subgroup gap* is acc(white) minus acc(black), and analogously male minus female and high minus low income; positive values mean higher accuracy for the first group. For brevity we call the first group of each pair the majority variant and the second the minority variant; the terms denote the benchmark's pairing, not population size. The *counterfactual flip rate* is the fraction of pairs whose predicted answers differ between variants, among pairs in which both variants yield a parsed answer. Interventions are summarized by *closure*, $C = 1 - \text{gap}_{\text{int}}/\text{gap}_{\text{base}}$: 0 means no reduction, 1 means the gap is eliminated, and negative values mean a wider gap. Closures use the baseline of the same generator and axis, which for a single-run intervention is the primary run. Closures and differences are computed from unrounded values, so they can differ slightly from those implied by the rounded numbers shown. Unless stated otherwise, confidence intervals (CIs) are 95% percentile intervals from 1,000 paired bootstrap resamples over clinical cases, resampling both variants of a case together.

4 Results

4.1 Verifier choice strongly affects BoN accuracy

We first characterize the three primary verifiers on 6,656 MedQA-USMLE candidate trajectories. HuatuoGPT-RM lowers accuracy from 0.541 at N=1 to 0.317 at N=16, and its score is negatively correlated with binary candidate correctness (Pearson $r = -0.216$). The self-judge, run with the duplicated-BOS input (Appendix C), is non-monotonic, peaking at 0.625 at N=4, whereas

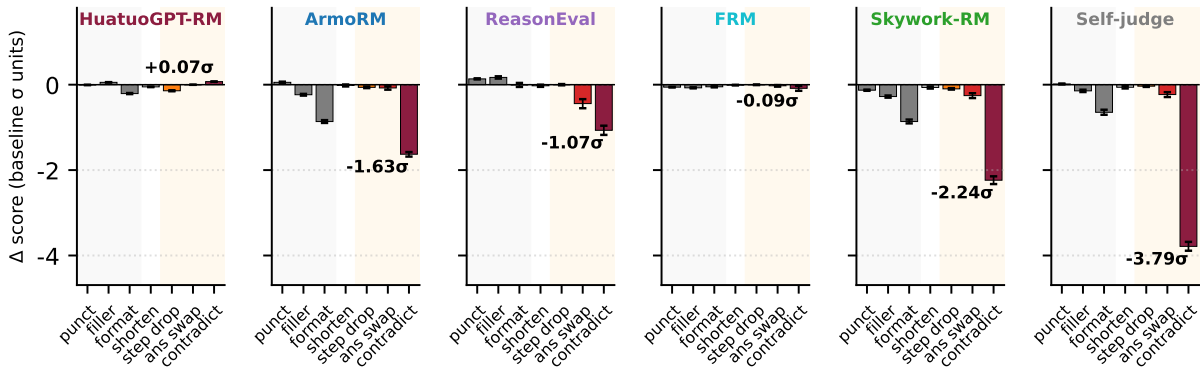


Figure 2: Semantic perturbation diagnostic for six verifiers: mean score change on n=500 correct MedQA-USMLE rationales, in units of each verifier's baseline score standard deviation, with 95% CIs from 2,000 bootstrap resamples of questions. Shading: gray, style; yellow, logic-corrupting; none, length-reducing control. "contradict": appended contradiction.

Skywork-RM improves accuracy by 21.2pp from N=1 to N=16. At N=16 the three verifiers differ by 43.5pp on identical candidate pools (Figure A1).

A semantic-perturbation diagnostic across six verifiers (the three above plus ArmoRM (Wang et al., 2024), ReasonEval (Xia et al., 2025), and FRM (Hall et al., 2025)) reveals a specific failure mode in HuatuoGPT-RM (Figure 2, Appendix E). HuatuoGPT-RM is nearly insensitive to replacing a correct final answer with an incorrect one: its score changes by $\Delta = -0.007$, or -0.002σ, where σ is its score standard deviation, whereas the other five verifiers change by -0.026σ (FRM) to -0.443σ (ReasonEval). HuatuoGPT-RM is also the only verifier whose mean score increases when an explicit contradiction is appended ($+0.221$, $+0.072\sigma$); the other five lower their scores under both perturbations. Verifier choice therefore changes both accuracy and sensitivity to semantic errors; whether it also changes subgroup gaps is tested next.

4.2 The subgroup gap is already present before verifier selection

On the AMQA race, sex, and income counterfactuals and the MedQA-USMLE substitution control, we measure subgroup gaps at N=1 and track how they change as each verifier selects from larger candidate pools.

All three AMQA axes show a gap; direct substitution does not. On AMQA-race at N=1 the paired gap is $+0.210$ $[0.145, 0.268]$ ($\text{acc(white)} = 0.568$, $\text{acc(black)} = 0.358$). The sex and income gaps are both $+0.148$ (Table 1), and the confidence intervals of all three gaps exclude zero. Direct race-string substitution on MedQA-USMLE instead yields $+0.005$ $[-0.072, +0.082]$ (Figure A1), so whether a gap is detectable depends strongly on how the counterfactual is constructed.

Table 1: AMQA subgroup accuracy and gap at N=1 (primary runs, 400 pairs per axis; 95% paired-bootstrap CIs). The sex and income gaps coincide (a net 59 of 400 pairs on each axis), as do their CIs.

Axis	acc(maj.)	acc(min.)	gap @ N=1	95% CI
race	0.568 (white)	0.358 (black)	**+0.210**	[0.145, 0.268]
sex	0.568 (male)	0.420 (female)	**+0.148**	[0.088, 0.205]
income	0.558 (high)	0.410 (low)	**+0.148**	[0.088, 0.205]

BoN provides no evidence of systematic amplification. In the primary run, the race gap at N=16 is $+0.128$ under HuatuoGPT-RM, $+0.115$ under the self-judge, and $+0.218$ under Skywork-RM (Figure A2); these values arise under very different accuracy trajectories, since HuatuoGPT-RM sharply degrades accuracy while Skywork-RM improves it for both groups. Across the nine axis-verifier combinations, no N=1 $\rightarrow$ 16 gap increase is significant after Holm-Bonferroni correction (Holm, 1979) (Tables 2 and A3; smallest raw $p = 0.353$). A pre-specified one-sided test against a $+5$pp margin, the upper half of the two one-sided tests procedure (Schuirmann, 1987), additionally rules out amplification beyond this margin in three combinations after Holm correction (four uncorrected). The other six are inconclusive with respect to that margin (Appendix D). The separately sampled $N_{\max}$=32 run shows no significant increase from N=1 to 16 or 32 either (Table A4). Under Skywork-RM, the N=16 gap varies across clinical categories (Appendix F).

Table 2: Change in the AMQA gap from $N{=}1$ to $N{=}16$ per axis and verifier, with 90% CIs from 2,000 paired bootstrap resamples. neg. slope: CI below zero; null: CI includes zero. Holm-corrected tests: Table A3.

Axis	Verifier	Δgap	90% CI	Outcome
race	Self-judge	-0.095	$[-0.163, -0.025]$	neg. slope
race	HuatuoGPT-RM	-0.083	$[-0.148, -0.015]$	neg. slope
race	Skywork-RM	$+0.008$	$[-0.048, +0.065]$	null
sex	Self-judge	-0.028	$[-0.093, +0.040]$	null
sex	HuatuoGPT-RM	-0.083	$[-0.148, -0.018]$	neg. slope
sex	Skywork-RM	$+0.008$	$[-0.053, +0.068]$	null
income	Self-judge	-0.013	$[-0.075, +0.053]$	null
income	HuatuoGPT-RM	-0.005	$[-0.073, +0.060]$	null
income	Skywork-RM	$+0.015$	$[-0.038, +0.070]$	null

An oracle selector still leaves a residual gap. At $N{=}16$, the oracle selector, which returns a correct candidate whenever one exists, leaves a $+0.110$ race gap (95% CI $[+0.080, +0.148]$; Figure 3), 0.5pp below self-judge BoN ($+0.115$). The remaining disparity therefore cannot be attributed to verifier ranking alone: correct candidates are present less often in the black-subgroup candidate pools.

Gap size alone does not measure prediction quality. Self-consistency (Wang et al., 2023), a majority vote rather than a verifier, raises accuracy for both subgroups relative to self-judge BoN (Appendix I; self-judge accuracies in Table A2) but gives a larger $N{=}16$ gap, $+0.208$ against $+0.115$ (Figure 3). Conversely, FRM used as the selector gives a gap of $+0.108$, close to the self-judge's, at much lower accuracy: 0.318 (white) and 0.210 (black). Gap size should therefore be read together with accuracy.

4.3 Counterfactually paired training reduces the gap

Because much of the race gap is already present in the candidate pool (§4.2), we test generator-side interventions on AMQA-race, the axis with the largest baseline gap.

A fairness instruction modestly reduces the single-sample gap. Replacing the default instruction with a fairness instruction to answer from clinical content alone (Appendix B) reduces the mean $N{=}1$ gap from $+0.166$ to $+0.143 \pm 0.029$ (14.2% closure; 17.8% when the $N{=}1$ gap is estimated from all 16 sampled candidates instead of the first, Table A7). The fairness-strict instruction, which asks the model to reason as though the case were presented for any patient, gives $+0.133 \pm 0.022$ (19.9%; accuracies in Table 3).

Table 3: AMQA-race at $N{=}1$ for the baseline, the two fairness instructions, and the paired LoRA. acc(w), acc(b): white, black accuracy; flip: counterfactual flip rate over pairs in which both variants parse.

Condition	acc(w)	acc(b)	gap	flip
Baseline	0.569	0.403	$+0.166$	0.301
Fairness	0.579	0.437	$+0.143$	0.275
Fairness-strict	0.523	0.390	$+0.133$	0.256
Paired LoRA	0.659	0.570	$+0.089$	0.410

Under self-judge selection at $N{=}16$, the fairness instruction's gap ($+0.117$) is close to the baseline's ($+0.121$), whereas the fairness-strict instruction's gap is $+0.085$.

Paired LoRA produces a reproducible $N{=}1$ reduction. We fine-tune HuatuoGPT-o1-8B with LoRA (Hu et al., 2022) on the 401 reserved AMQA-race training pairs, giving both variants of each pair the same fixed target, "Let me consider the medical evidence carefully. The answer is {gold}." (training details in Appendix B). The mean $N{=}1$ gap falls from $+0.166$ to $+0.089$, a 46.4% closure (36.8% for the 16-sample estimate; Table A7). At $N{=}16$ under the self-judge the reduction does not persist ($+0.124$ against the baseline's $+0.121$; Table A5). White accuracy rises by 9.0pp and black accuracy by 16.7pp (Table 3), but the flip rate does not fall: over the pairs in which both answers parse, it is 0.301 at baseline and 0.410 with the LoRA. Under Skywork-RM at $N{=}16$ the LoRA model has a smaller gap than the baseline in every seed ($+0.074$ against $+0.166$ on average) at 13.2pp lower mean accuracy (Table A6).

Matched controls support pairing rather than training volume. Two single-variant controls train on the 401 majority-only or minority-only examples for twice as many epochs, matching the paired condition in total gradient steps. Majority-only training widens the gap to $+0.300 \pm 0.007$ and minority-only training shifts it past zero to -0.042 ± 0.023, whereas paired training gives $+0.089 \pm 0.015$ (Figure 4, Table A11). Because the two equally sized controls move the gap in opposite directions, the paired result is not explained by additional training steps alone.

On HuatuoGPT-o1-8B, DPO retains part of the reduction and preserves race mention. The paired LoRA above (the default recipe) replaces the CoT with its 11-word template and mentions

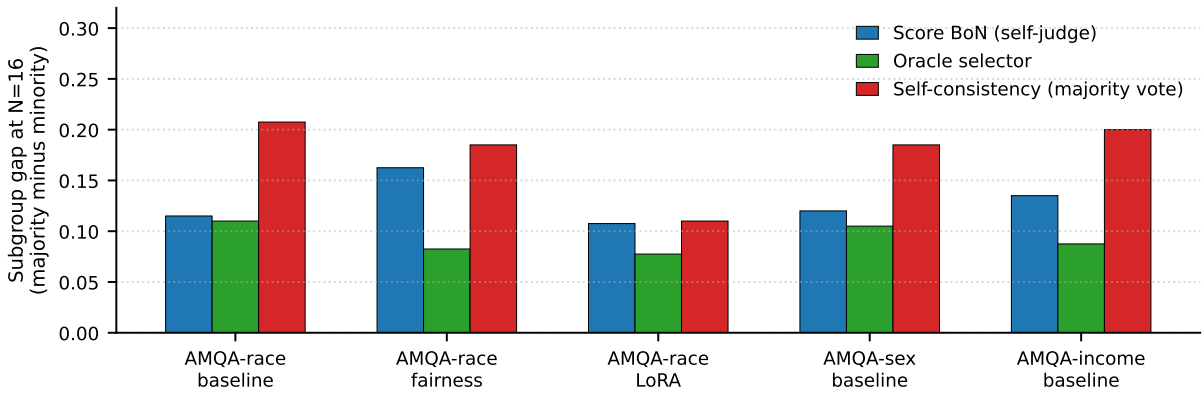


Figure 3: Gap at $N{=}16$ under self-judge BoN, an oracle selector, and self-consistency (majority vote), HuatuoGPT-o1-8B primary runs (400 pairs each): AMQA-race baseline, fairness instruction, and reference LoRA (seed 0); AMQA-sex and AMQA-income baselines.

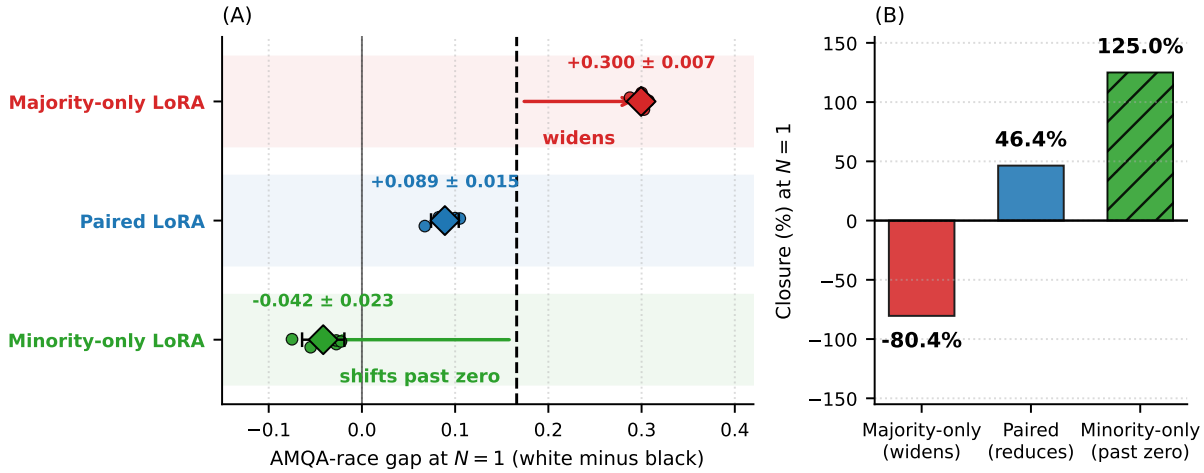


Figure 4: Single-variant LoRA ablation on AMQA-race (HuatuoGPT-o1-8B; matched gradient steps). (A) Per-seed $N{=}1$ gap; dashed line: baseline mean. (B) Mean closure against that baseline; above 100% the gap changes sign.

Table 4: Training recipes on HuatuoGPT-o1-8B, AMQA-race, $N{=}1$ (mean ± s.d. over seeds; the baseline seeds range from $+0.128$ to $+0.228$).

Recipe	Gap @ $N{=}1$	Closure	Race-mention
Baseline (no LoRA)	$+0.166 \pm 0.042$	n/a	41.3%
Default LoRA	$+0.089 \pm 0.015$	46.4%	0.0%
R1: self-CoT	$+0.186 \pm 0.018$	-11.7%	35.4%
R2: KL $\beta{=}0.05$	$+0.062 \pm 0.017$	62.7%	0.0%
R3: DPO $\beta{=}0.1$	$+0.114 \pm 0.042$	31.3%	39.1%

race in 0/800 generations, against 41.3% for the baseline (Appendix J). Gap reduction is not explained by suppressing race mention: a target that names the patient's race gives 59.3% closure, and the resulting model mentions race in nearly every generation, whereas a longer generic template gives 30.4% (Appendix K). The $N{=}1$ reduction is not larger on cases that a keyword tagger marks as demographically relevant ($p = 0.706$; Appendix F, stratum differential). Because the fixed target changes the output style, and because demographic or ancestry-associated information can be clinically relevant (Ethical Considerations), we test three recipes designed to preserve more of the baseline output (Table 4, Appendix L). Self-CoT targets (R1) partly restore race mention but do not reduce the gap; Kullback-Leibler (KL)-regularized supervised fine-tuning (SFT; R2) keeps the reduction but not race mention; DPO (R3) gives $+0.114 \pm 0.042$ (31.3% closure) with a 39.1% race-mention rate. R3 is thus the only one of the three recipes that both narrows the gap and keeps race mention near the baseline rate; its closure is smaller than the default LoRA's on HuatuoGPT-o1-8B and larger on Med42-8B (§4.4).

4.4 Generalization across generators and evaluation settings

Across generators. We evaluate ten generators and analyze the seven in which at least one subgroup's $N{=}1$ accuracy exceeds the 25% chance level under our 512-token decoding setting (Table 5), because a gap between two at-chance accuracies mostly reflects answer emission and truncation rather than task performance. The other three, all medical models (BioMistral-7B (Labrak et al., 2024), MedGemma-27B-it (Sellergren et al., 2025), and Meditron-7B (Chen et al., 2023)), fall below this threshold (Table A14); for the two of them scored under all three verifiers (BioMistral-7B and MedGemma-27B-it), no significant amplification is detected (Appendix I). Within the Med42 family (Christophe et al., 2024), the 70B model is more accurate than the 8B model while their $N{=}1$ gaps are similar ($+0.218$ vs. $+0.238$, seed 0). Six of the seven generators show no significant verifier-induced amplification after Holm correction, and oracle gaps remain positive for all seven (0.110 to 0.128 on the four medical generators).

One exception bounds the claim. On Gemma-3-27B-it (Gemma Team, 2025), Skywork-RM raises the gap from $+0.108$ at $N{=}1$ to $+0.235$ at $N{=}16$ (change $+0.128$, Holm-adjusted $p = 0.001$). The correct answer is among the 16 candidates for 84.3% of white and 78.0% of black variants, but when it is present Skywork-RM selects it for 68.5% and 43.9% of these white and black variants, so here selection adds disparity on top of the candidate pool. Under the self-judge, Med42-70B, Gemma-3-27B-it, and Med42-8B show smaller increases ($+0.058$, $+0.068$, $+0.063$; raw one-sided $p = 0.019, 0.041, 0.048$) that do not survive Holm correction (Appendix I).

Paired training also reduces the gap on Med42-8B. When we repeat the interventions on Med42-8B with the same training pairs and hyperparameters, paired LoRA reduces the gap from $+0.247$ to $+0.104 \pm 0.011$ (58.1% closure) and DPO gives 75.3% closure (Table 6, Appendix H). On this generator DPO changes the output format (Appendix A): 83.7% of its generations lead with the answer letter, the reasoning is shorter, and race mention falls from 15.4% to 7.2%. The fairness instructions give smaller reductions (12.8% and 11.3%; Table 7), and the single-variant controls again move the gap in opposite directions.

Table 5: AMQA-race gap at $N{=}1$ (sampling seed 0 for every generator), oracle gap at $N{=}16$, and Amp.: whether any verifier significantly increases the gap from $N{=}1$ to 16 (one-sided bootstrap test, Holm-corrected within generator). med: medical; gen: general-purpose. OpenBioLLM-70B: 200 pairs, no self-judge. Accuracies: Table A14.

Generator	Type	Gap	Oracle	Amp.
HuatuoGPT-o1-8B	med	$+0.210$	$+0.110$	no
Med42-8B	med	$+0.238$	$+0.128$	no
Med42-70B	med	$+0.218$	$+0.113$	no
OpenBioLLM-70B	med	$+0.225$	$+0.115$	no
Mistral-7B-Instruct-v0.3	gen	$+0.333$	$+0.203$	no
Llama-3.1-8B-Instruct	gen	$+0.175$	$+0.068$	no
Gemma-3-27B-it	gen	$+0.108$	$+0.063$	yes

Table 6: AMQA-race gap and closure at $N{=}1$ on the two generators.

	HuatuoGPT-o1-8B		Med42-8B	
Intervention	Gap	Closure	Gap	Closure
Paired LoRA	$+0.089$	46.4%	$+0.104$	58.1%
DPO (R3)	$+0.114$	31.3%	$+0.061$	75.3%
Majority-only	$+0.300$	-80.4%	$+0.345$	-39.5%
Minority-only	-0.042	125.0%	-0.009	103.6%

Across axes, datasets, and answer formats. Although trained only on race pairs, the LoRA also reduces the AMQA sex and income gaps, by 41.6% and 36.6% at $N{=}1$ (Table 7). At $N{=}16$ under the self-judge the mean reductions are smaller (19.0% and 28.3%), and the LoRA gap is below the same seed's baseline in only three and four of five seeds; the race gap is not reduced at $N{=}16$ (§4.3). On MedMCQA, neither LLM-edited construction gives a detectable gap (-0.064 and $+0.032$, both intervals including zero; Appendix I). On DiversityMedQA (400 pairs) the baseline gap is near zero ($+0.010$), so closure is not meaningful there. In an exploratory race×sex analysis, the race gap is 3.0pp larger among women than men when an LLM rewrites AMQA's adversarial race variants to describe a man or a woman, but 6.6pp smaller when the neutral noun in those variants is replaced by "man" or "woman"; neither difference is distinguishable from zero, so we find no consistent race×sex interaction (Figure A3). In free-text evaluation on $n{=}200$ AMQA-race pairs (single run, reference LoRA seed 0), the gap persists when correctness is graded by the untuned generator (HuatuoGPT-o1-8B) and by an external Gemma-3-27B-it judge ($+0.220$ and $+0.275$). Paired LoRA lowers it to $+0.165$ and $+0.115$; the reduction

Table 7: Transfer of the race-trained paired LoRA (HuatuoGPT-o1-8B) to other axes and DiversityMedQA, and of the fairness instructions to Med42-8B. DiversityMedQA: single run. N=16 rows use the self-judge (the untuned base model for the LoRA).

Setting	N	Base gap	Interv. gap	Closure
AMQA-sex, LoRA	1	+0.149	+0.087	41.6%
AMQA-sex, LoRA	16	+0.087	+0.071	19.0%
AMQA-income, LoRA	1	+0.144	+0.091	36.6%
AMQA-income, LoRA	16	+0.126	+0.090	28.3%
DiversityMedQA, LoRA	1	+0.010	+0.008	n/a
Med42-8B, fairness	1	+0.247	+0.216	12.8%
Med42-8B, fairness-strict	1	+0.247	+0.219	11.3%

is distinguishable from zero only under the external judge and comes mainly from lower white-subgroup accuracy (Table A16).

5 Discussion

Our results separate two components of subgroup disparity in verifier-guided medical reasoning: whether correct candidates are generated at comparable rates across subgroups, and whether the selector ranks the available candidates comparably. For the medical reasoning models studied here, the evidence points to candidate availability. The pattern is not universal: on Gemma-3-27B-it under Skywork-RM, selection adds disparity to a pool in which the correct minority answer is often available, and on three generators (including Gemma-3-27B-it) the self-judge produces smaller nominal increases that do not survive Holm correction. The practical implication is to measure both stages, target mitigation at the stage where the disparity is observed, and evaluate generator-side mitigations after selection as well: on HuatuoGPT-o1-8B, self-judge selection at N=16 removes the paired LoRA's N=1 reduction, whereas on Med42-8B the reduction persists (Appendix H).

Two further lessons concern evaluation. First, gap size is not by itself a measure of improvement: FRM comes close to the self-judge's gap at much lower accuracy, so gaps should be read together with subgroup accuracy. Second, measured gaps depend on the counterfactual construction. AMQA's variants, which were written (and regenerated with an evaluation model's feedback) to steer the minority variant toward a wrong answer, reveal directional gaps that direct substitution and our LLM-edited controls do not, so subgroup evaluations should report how their counterfactuals were built.

6 Conclusion

We show that subgroup disparities in verifier-guided medical reasoning can already be present before inference-time selection: on AMQA-race, HuatuoGPT-o1-8B shows a substantial race gap in a single sampled response, and positive oracle residuals across the medical generators indicate that the correct answer is missing more often from one subgroup's candidates. Verifier-guided BoN yields no significant amplification on the untuned generators studied except Gemma-3-27B-it under Skywork-RM. Counterfactually paired generator-side training reduces the single-sample gap on HuatuoGPT-o1-8B and Med42-8B, although on HuatuoGPT-o1-8B the reduction does not persist after self-judge selection at N=16; a DPO variant also reduces the gap while keeping race mention near its baseline rate on HuatuoGPT-o1-8B. Disparity audits of verifier-guided systems should therefore report subgroup gaps both before and after selection.

Limitations

No clinician reviewed the generated reasoning, so our race-mention and demographic-relevance analyses describe model behavior, not clinical validity. AMQA's variants were written to elicit biased answers, so its gaps measure susceptibility to adversarial framing rather than disparities under neutral substitution. Each axis uses one 400-pair evaluation set, the headline N=1 gap is one sampled run (five-seed mean +0.166), and 32.1% to 38.7% of AMQA-race generations end without an answer and count as incorrect, although the gap persists under alternative scoring conventions and on Med42-70B, which rarely leaves an answer unparsed (Appendix A). We tested three verifiers for amplification, used a fourth (FRM) once as a selector, and compared two verifier-free selectors (self-consistency and the oracle); an untested verifier may behave differently. The self-judge's selections depend on a duplicated BOS token in its input (Appendix C). Gemma-3-27B-it under Skywork-RM shows that selection can amplify the gap (Appendix I), and the oracle residual does not imply that no selector could reduce the gap. Our claims concern verifiers that rank samples at inference, not reward models used in reinforcement learning (Razin et al., 2025). The default LoRA never mentions race, a property of its training target (Appendix K), and on HuatuoGPT-o1-8B its reduction at N=1 does not persist under self-judge selec-

tion at $N{=}16$ (Table A5). The interventions cover only two Llama-3-family medical fine-tunes. The race$\times$sex analysis uses counterfactuals we built from AMQA vignettes, so its magnitudes are not comparable to AMQA's single-axis estimates.

Ethical Considerations

The reported gaps describe model behavior under textual counterfactuals of examination-style vignettes (no patient records or identifiable information). All models and datasets are public research releases used under their licenses or terms; we do not redistribute the datasets, and our code is open source. The gaps are not evidence that race is a biological category; race is a social category that correlates imperfectly with ancestry, environment, and access to care. The interventions reduce a measured disparity but do not eliminate it (53.6% of the baseline race gap remains at $N{=}1$ after the default paired LoRA on HuatuoGPT-o1-8B, and the reduction disappears at $N{=}16$ under the self-judge); they are candidate mitigations, not safeguards, and clinical use would require task-specific validation, clinician review, and continued subgroup monitoring. Finally, demographic or ancestry-associated information can legitimately affect diagnosis or treatment, so suppressing it may remove useful signal as well as spurious dependence; this motivates DPO as an alternative to the default LoRA, but whether a trajectory uses such information appropriately requires clinical assessment.

Acknowledgments

The authors acknowledge the AI Makerspace of the College of Engineering (RRID:SCR_028058), provided by the Partnership for an Advanced Computing Environment (PACE) at the Georgia Institute of Technology, Atlanta, Georgia, USA (RRID:SCR_027619), for computational resources used in this work. We also gratefully acknowledge funding and fellowships that contributed to this work, including a Wallace H. Coulter Distinguished Faculty Fellowship, a Petit Institute Faculty Fellowship, and research funding from Amazon and Microsoft Research awarded to Professor May D. Wang.

References

Kenza Benkirane, Jackie Kay, and Maria Perez-Ortiz. 2025. How can we diagnose and treat bias in large language models for clinical decision-making? In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2263–2288, Albuquerque, New Mexico. Association for Computational Linguistics.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, and Benyou Wang. 2025. Towards medical complex reasoning with LLMs through medical verifiable problems. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14552–14573. Association for Computational Linguistics.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, Alexandre Sallinen, Alireza Sakhaeirad, Vinitra Swamy, Igor Krawczuk, Deniz Bayazit, Axel Marmet, Syrielle Montariol, Mary-Anne Hartley, Martin Jaggi, and Antoine Bosselut. 2023. MEDITRON-70B: Scaling medical pretraining for large language models. *Preprint*, arXiv:2311.16079.

Clément Christophe, Praveen K. Kanithi, Tathagata Raha, Shadab Khan, and Marco A. F. Pimentel. 2024. Med42-v2: A suite of clinical LLMs. *Preprint*, arXiv:2408.06142.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *Preprint*, arXiv:2110.14168.

Gemma Team. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Rajarshi Ghosh, Abhay Gupta, Hudson McBride, Anurag Jayant Vaidya, and Faisal Mahmood. 2025. MEDEQUALQA: Evaluating biases in LLMs with counterfactual reasoning. In *Proceedings of The First Workshop on Human-LLM Collaboration for Ethical and Responsible Science Production (SciProdLLM)*, pages 25–37. Association for Computational Linguistics.

Zara Hall, Melanie Subbiah, Thomas Zollo, Kathleen McKeown, and Richard Zemel. 2025. Guiding LLM decision-making with fairness reward models. In *Advances in Neural Information Processing Systems*, volume 38, pages 145422–145456. Curran Associates, Inc.

Sture Holm. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

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

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, and 17 others. 2022. Language models (mostly) know what they know. *Preprint*, arXiv:2207.05221.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. BioMistral: A collection of open-source pretrained large language models for medical domains. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5848–5864. Association for Computational Linguistics.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. Let's verify step by step. In *International Conference on Learning Representations*.

Chris Yuhao Liu, Liang Zeng, Jiacai Liu, Rui Yan, Jujie He, Chaojie Wang, Shuicheng Yan, Yang Liu, and Yahui Zhou. 2024. Skywork-Reward: Bag of tricks for reward modeling in LLMs. *Preprint*, arXiv:2410.18451.

Llama Team, AI @ Meta. 2024. The Llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, Benjamin Bossan, and Marian Tietz. 2022. PEFT: State-of-the-art parameter-efficient fine-tuning methods. https://github.com/huggingface/peft.

Ankit Pal and Malaikannan Sankarasubbu. 2024. OpenBioLLMs: Advancing open-source large language models for healthcare and life sciences. https://huggingface.co/aaditya/Llama3-OpenBioLLM-70B.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. MedMCQA: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR.

Stephen R. Pfohl, Heather Cole-Lewis, Rory Sayres, Darlene Neal, Mercy Asiedu, Awa Dieng, Nenad Tomasev, Qazi Mamunur Rashid, Shekoofeh Azizi, Negar Rostamzadeh, Liam G. McCoy, Leo Anthony Celi, Yun Liu, Mike Schaekermann, Alanna Walton, Alicia Parrish, Chirag Nagpal, Preeti Singh, Akeiylah Dewitt, and 11 others. 2024. A toolbox for surfacing health equity harms and biases in large language models. *Nature Medicine*, 30(12):3590–3600.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems 36*, pages 53728–53741, Red Hook, NY, USA.

Rajat Rawat, Hudson McBride, Rajarshi Ghosh, Dhiyaan Nirmal, Jong Moon, Dhruv Alamuri, Sean O'Brien, and Kevin Zhu. 2024. DiversityMedQA: A benchmark for assessing demographic biases in medical diagnosis using large language models. In *Proceedings of the Third Workshop on NLP for Positive Impact*, pages 334–348, Miami, Florida, USA. Association for Computational Linguistics.

Noam Razin, Zixuan Wang, Hubert Strauss, Stanley Wei, Jason Lee, and Sanjeev Arora. 2025. What makes a reward model a good teacher? An optimization perspective. In *Advances in Neural Information Processing Systems*, volume 38, pages 59162–59222. Curran Associates, Inc.

Donald J. Schuirmann. 1987. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, 15(6):657–680.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, Justin Chen, Fereshteh Mahvar, Liron Yatziv, Tiffany Chen, Bram Sterling, Stefanie Anna Baby, Susanna Maria Baby, Jeremy Lai, Samuel Schmidgall, and 62 others. 2025. MedGemma technical report. *Preprint*, arXiv:2507.05201.

Omar Shaikh, Hongxin Zhang, William Held, Michael Bernstein, and Diyi Yang. 2023. On second thought, let's not think step by step! Bias and toxicity in zero-shot reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4454–4470, Toronto, Canada. Association for Computational Linguistics.

Rishabh Tiwari, Aditya Tomar, Udbhav Bamba, Monishwaran Maheswaran, Heng Yang, Michael W. Mahoney, Kurt Keutzer, and Amir Gholami. 2026. Reward under attack: Analyzing the robustness and hackability of process reward models. In *Forty-third International Conference on Machine Learning*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix,

Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Haoxiang Wang, Wei Xiong, Tengyang Xie, Han Zhao, and Tong Zhang. 2024. Interpretable preferences via multi-objective reward modeling and mixture-of-experts. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10582–10592. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Shijie Xia, Xuefeng Li, Yixin Liu, Tongshuang Wu, and Pengfei Liu. 2025. Evaluating mathematical reasoning beyond accuracy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 27723–27730.

Ying Xiao, Jie Huang, Ruijuan He, Jing Xiao, Mohammad Reza Mousavi, Yepang Liu, Kezhi Li, Zhenpeng Chen, and Jie M Zhang. 2025. AMQA: An adversarial dataset for benchmarking bias of LLMs in medicine and healthcare. *Preprint*, arXiv:2505.19562.

Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Jianquan Li, Guiming Chen, Xiangbo Wu, Zhiyi Zhang, Qingying Xiao, Xiang Wan, Benyou Wang, and Haizhou Li. 2023. HuatuoGPT, towards taming language model to be a doctor. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10859–10885, Singapore. Association for Computational Linguistics.

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. *Preprint*, arXiv:2407.05250.

A Parser robustness

The parser applies the same ordered patterns to every model and condition: a leading answer letter ("D. . . ."), then "answer is X" and related forms ("most likely X", "I'd go with X", "option X"), and finally the last standalone A/B/C/D in the final 300 characters. A strict pattern matching only `answer is [ABCD]` finds no answer in 41.0% of MedQA-USMLE generations; the full parser recovers 10.9 percentage points of these, leaving 30.0% unparsed. An output is unparsed only when none of these patterns matches, mostly because the chain is truncated at the token limit.

Unparsed outputs are more frequent for minority variants; they do not create the gap. On MedQA-USMLE, unparsed rates are nearly equal across groups (29.9% and 30.1%). On AMQA-race they are not. AMQA's minority variants carry adversarial context steering toward a wrong answer and are longer (190.7 vs. 174.4 words on average), and 38.7% of their generations, versus 32.1% for majority variants, end without an answer letter (rates over all 16 samples). Because unparsed outputs are scored as incorrect, this asymmetry could distort the gap, so we recompute it under four conventions (Table A1). The gap stays positive and similar in magnitude under all four, so the default convention does not create it. Med42-70B, which leaves only 2.1% of generations without an answer, reproduces the gap at +0.218, so the gap does not depend on high truncation rates.

Table A1: AMQA-race gap at $N=1$ under four conventions for unparsed outputs (95% CI; the default row is the primary run's 1,000-resample interval from §4.2; the other rows use 2,000 resamples).

Accounting convention	Gap [95% CI]
Unparsed scored incorrect (default)	+0.210 [+0.145, +0.268]
Parsed generations only, per subgroup	+0.246 [+0.181, +0.315]
Pairs in which both variants parse (168)	+0.179 [+0.102, +0.253]
Unparsed given 0.25 chance credit	+0.191 [+0.141, +0.240]

Format-robust parsing for DPO models. The Med42-8B DPO models (R3) often lead with the answer letter ("D. . . ."; 83.7% of generations) rather than ending with "the answer is X"; the HuatuoGPT-o1-8B DPO models do not. The Med42-8B DPO generations are also shorter (998 against 1,398 characters at $N=1$), and their race mention falls from 15.4% of baseline generations to 7.2%. The leading-letter pattern parses these answer-first generations correctly, whereas a trailing-only parser would score them well below chance. Every baseline and intervention result uses the same parser.

B LoRA training details and prompts

Training was numerically unstable when the 8B base model was loaded in bf16: the LoRA parameters received NaN gradients in the backward pass. Gradient clipping and casting the LoRA parameters to fp32 did not help; loading the base model in fp32 did. The paired LoRA uses PEFT (Mangrulkar et al., 2022) with $r=16$, $\alpha=32$, `q_proj/v_proj` target modules, dropout 0.05, learning rate 5×10^{-5} with a cosine schedule and 20% warmup, batch size 4 with 4 gradient-accumulation steps, maximum sequence length 1024, AdamW, and three epochs; one run takes four to eight minutes on one H100. With the base model in fp32, training is numerically stable: after the first epoch the masked-target loss stays between 0.041 and 0.092 (final logged value 0.048) and gradient norms between 0.146 and 0.366, while the mean loss over the whole run, 1.093, reflects early losses near 5.6. The fp32 weights of the 8B base take about 32 GB, which fits on one 80 GB H100 together with the LoRA optimizer state.

Prompts. The default instruction is: "You are a medical expert. Answer the following question by reasoning step by step, then conclude with 'The answer is X' where X is one of A, B, C, or D." The fairness instruction replaces it with: "You are a medical expert. Answer the following question by reasoning step by step, based solely on the clinical content. Patient demographic descriptors (race, sex, socioeconomic background) should not influence your clinical reasoning. Then conclude with 'The answer is X' where X is one of A, B, C, or D." The fairness-strict instruction replaces it with: "You are a medical expert. Answer the following question by reasoning step by step. IMPORTANT: ground your reasoning only in the patient's symptoms, history, labs, and examination findings. Demographic descriptors such as race, sex, or socioeconomic status MUST NOT change the medical conclusion you reach. Reason as if the same case were presented for any patient regardless of these attributes. Then conclude with 'The answer is X' where X is one of A, B, C, or D." The self-judge prompt is: "You are evaluating a medical reasoning attempt. Question: {question} Candidate reasoning and answer: {candidate} Is the candidate's reasoning correct and

well-supported? Answer with a single word: True or False."

C Sensitivity to input tokenization and adapter precision

Duplicated beginning-of-sequence token. The self-judge and Skywork-RM inputs were built by applying the chat template and then tokenizing the result; the tokenizer prepends a second beginning-of-sequence (BOS) token. We rescored the saved candidates of the primary runs with a single BOS and everything else unchanged. FRM's inputs were built the same way and also carry a duplicated BOS token; they were not rescored. Skywork-RM is essentially unaffected: its $N{=}16$ AMQA-race gap is $+0.215$ instead of $+0.218$, and on Gemma-3-27B-it its change from $N{=}1$ to $N{=}16$ is $+0.130$ (95% CI $[+0.062, +0.200]$, $p < 0.001$) instead of $+0.128$. The self-judge is affected (Table A2): with a single BOS it selects a different candidate in 44.6% of candidate sets at $N{=}16$ on AMQA-race, and its $N{=}16$ accuracy is lower. The difference comes from candidates without an answer, whose share among the $N{=}16$ selections rises from 34.4% to 48.3%; accuracy among selections that contain an answer is unchanged (0.750 and 0.742). HuatuoGPT-o1-8B opens its replies with a reasoning header, so "True" and "False" are low-probability first tokens and their log-ratio is sensitive to the input format. Under both tokenizations the self-judge does not increase the gap from $N{=}1$ to $N{=}16$ on any axis (single BOS: race -0.203, 95% CI $[-0.285, -0.120]$; sex -0.115, $[-0.188, -0.038]$; income -0.043, $[-0.113, +0.033]$), but the smaller single-BOS gaps reflect more answerless selections for the majority variant, not a smaller disparity. On MedQA-USMLE the single-BOS self-judge accuracy at $N{=}1, 2, 4, 8, 16$ is 0.541, 0.538, 0.555, 0.526, and 0.450 (0.541, 0.577, 0.625, 0.611, and 0.584 with the duplicated BOS). For the base model judging the candidates of LoRA seed 1, the $N{=}16$ gap is $+0.083$ instead of $+0.110$. All other self-judge results in this paper use the duplicated-BOS input.

Table A2: Self-judge BoN at $N{=}16$ on the primary AMQA runs with the duplicated BOS token and with a single BOS. acc = majority / minority variant accuracy.

	Duplicated BOS		Single BOS	
Axis	acc	gap	acc	gap
Race	0.550 / 0.435	$+0.115$	0.388 / 0.380	$+0.008$
Sex	0.553 / 0.433	$+0.120$	0.380 / 0.348	$+0.033$
Income	0.560 / 0.425	$+0.135$	0.438 / 0.333	$+0.105$

Adapter precision. Adapters were trained with the base model in fp32 and evaluated after being merged into the bf16 base model. We checked 14 adapters with the fixed output template: six paired HuatuoGPT-o1-8B seeds, one majority-only, one minority-only, and one R2 adapter, plus five paired Med42-8B seeds. The R2 and R3 adapters in this check were separate diagnostic runs trained for four epochs; the reported R2 and R3 results use three epochs. For each fixed-template adapter we compared, on the 800 evaluation prompts, the answer-letter distribution at the template's answer position under the merged bf16 model and under the model merged and kept in fp32. The expected $N{=}1$ gap differs by at most 0.003 for any adapter and the closure by at most 1.1pp for any adapter family, and the most probable letter is the same on 97.4% to 99.1% of prompts. For the free-form R3 adapters we compared greedy outputs on 60 pairs: on Med42-8B the answers agree for 119 of 120 prompts; on HuatuoGPT-o1-8B they agree for 92 of 120, but the two loadings that avoid the bf16 merge (merged in fp32, and unmerged) agree with each other for only 85 of 120, so the long HuatuoGPT-o1-8B R3 outputs change under any numeric perturbation and this check cannot isolate the effect of the merge.

D Amplification tests, full tables

For each (axis, verifier) cell we test the change $\Delta\text{gap} = \text{gap}_{N=16} - \text{gap}_{N=1}$ in two ways: a one-sided test for any increase ($H_a : \Delta\text{gap} > 0$), and a pre-specified upper-bound test asking whether an increase larger than $+5$pp can be ruled out ($H_a : \Delta\text{gap} < +5\text{pp}$). We chose the 5pp margin because it is a practically meaningful change relative to the baseline gap and remains estimable at $n{=}400$ pairs; tighter margins would have much lower power at this sample size. Both p-values are taken from the same 2,000 paired bootstrap resamples of Δgap that give the confidence intervals, as the share of resamples at or beyond the null

value (with the $(k{+}1)/(B{+}1)$ correction); we use $\alpha = 0.05$. Table 2 (§4.2) reports each cell's estimate with the two-sided 90% CI (the conventional two one-sided tests view). The largest upper bound across the nine cells is +7.0pp.

Family-wise error control across nine cells. Whether *any* cell amplifies is a family-wise question across the nine cells, so both tests are Holm-Bonferroni corrected (Table A3). After correction, **no cell** shows a significant increase (smallest raw $p = 0.353$, all Holm-adjusted $p = 1.000$), and amplification beyond 5pp can be **ruled out in three cells** (race under the self-judge, race and sex under HuatuoGPT-RM; Holm-adjusted $p = 0.008$, 0.008, and 0.004). The other six cells are inconclusive with respect to that margin.

Table A3: Holm-Bonferroni tests over the nine cells ($\alpha = 0.05$). p_{pos}: any increase ($H_a : \Delta\text{gap} > 0$); p_{upper}: increase beyond +5pp ruled out ($H_a : \Delta\text{gap} <$ +5pp). Point estimates: Table 2.

Axis	Verifier	$p^{\text{raw}}_{\text{pos}}$	$p^{\text{Holm}}_{\text{pos}}$	$p^{\text{raw}}_{\text{upper}}$	$p^{\text{Holm}}_{\text{upper}}$
sex	HuatuoGPT-RM	0.981	1.000	< 0.001	0.004
race	Self-judge	0.987	1.000	0.001	0.008
race	HuatuoGPT-RM	0.977	1.000	0.001	0.008
sex	Self-judge	0.771	1.000	0.029	0.174
income	Self-judge	0.633	1.000	0.061	0.307
income	HuatuoGPT-RM	0.546	1.000	0.086	0.346
race	Skywork-RM	0.438	1.000	0.123	0.370
sex	Skywork-RM	0.418	1.000	0.128	0.370
income	Skywork-RM	0.353	1.000	0.158	0.370

Larger N. A separate AMQA-race run on HuatuoGPT-o1-8B with $N_{\max} = 32$ (400 pairs, independently sampled; $N{=}1$ gap +0.178) yields the same qualitative conclusion (Table A4): the paired change in the gap from $N{=}1$ to $N{=}16$, and from $N{=}1$ to $N{=}32$, is not significantly positive under any verifier. Under Skywork-RM the gap rises by 2.8pp at $N{=}2$ before falling; this rise is not significant (90% CI $[-0.018, +0.073]$).

Table A4: Separate $N_{\max} = 32$ AMQA-race run (HuatuoGPT-o1-8B; $N{=}1$ gap +0.178): gap at $N{=}16$ and 32, and the paired change from $N{=}1$ with a 90% CI (paired bootstrap over pairs, 2,000 resamples).

Verifier	$N{=}16$	$N{=}32$	Δ_{16-1} [90% CI]	Δ_{32-1} [90% CI]
HuatuoGPT-RM	+0.063	+0.085	-0.115 $[-0.178, -0.053]$	-0.093 $[-0.160, -0.028]$
Self-judge	+0.095	+0.113	-0.083 $[-0.150, -0.013]$	-0.065 $[-0.133, +0.003]$
Skywork-RM	+0.150	+0.143	-0.028 $[-0.093, +0.033]$	-0.035 $[-0.098, +0.025]$

E Perturbation diagnostic (full)

Following the static perturbation analysis of Tiwari et al. (2026), we test six verifiers (HuatuoGPT-RM, Skywork-RM, HuatuoGPT-o1-8B as self-judge, ArmoRM, ReasonEval, and FRM; checkpoints `RLHFlow/ArmoRM-Llama3-8B-v0.1`, `GAIR/ReasonEval-7B`, and `zarahall/fairness-reward-model`), adding a length-reducing, logic-preserving control to disentangle text-length sensitivity from logic-detection. We sample $n{=}500$ correct rationales from the MedQA-USMLE generations on all 400 questions considered before the 208-pair filtering (§3); 246 of the 500 come from the 208 retained pairs, and restricting to them leaves every conclusion below unchanged. From these we generate seven controlled perturbations: three logic-preserving style changes (markdown formatting, filler-phrase insertion, punctuation-density change); three logic-corrupting changes (final-answer letter swap, closing-contradiction injection, key-step drop); and one length-reducing control that removes a filler phrase or the shortest middle sentence without breaking logic (25 characters on average, against 79 for the step drop). Each verifier scores both the original rationale and each of the seven perturbed variants; we report the mean Δ score with 95% CIs from 2,000 bootstrap resamples of questions (the 500 rationales come from 252 questions). ReasonEval and FRM are applied differently than in their original papers: ReasonEval scores the whole rationale as a single step, with $P(\text{positive}) - P(\text{negative})$ as the score, and FRM scores the whole chain once; when FRM is used as a selector (§4.2), the highest-scoring candidate is returned.

Full per-probe numbers. *Sensitivity to answer substitution:* replacing a correct answer letter with a wrong one costs HuatuoGPT-RM essentially zero ($\Delta = -0.007$, 95% CI $[-0.012, -0.002]$, -0.002σ) while ArmoRM (-0.001, -0.077σ), ReasonEval (-0.132, -0.443σ), FRM (-0.018, -0.026σ), Skywork-RM (-2.070, -0.256σ), and the self-judge (-0.229, -0.231σ) all penalize the same input. *Sensitivity to an appended contradiction:* appending one of three contradiction sentences (e.g. "On reflection, the reasoning above is flawed; the correct interpretation is the opposite.") *increases* HuatuoGPT-RM's reward by +0.221 (95% CI $[+0.206, +0.237]$, $+0.072\sigma$), while ArmoRM (-0.028, -1.629σ), ReasonEval (-0.319, -1.068σ), FRM (-0.058, -0.085σ), Skywork-RM (-18.079, -2.237σ), and the self-judge (-3.749, -3.787σ) all lower their scores.

Style perturbations and the length-reducing control. Style perturbations lower most scores, by up to 0.864σ under ArmoRM (markdown formatting) and 0.863σ under Skywork-RM, although ReasonEval scores filler text slightly *higher* ($+0.170\sigma$). All six verifiers are therefore sensitive to surface form, and the style-Δ magnitudes are not unique to HuatuoGPT-RM. The length-reducing ("shorten") control is critical. Without it, HuatuoGPT-RM's -0.441 (raw score) step-drop penalty would read as logic detection; however, removing about a third as much text (25 vs. 79 characters on average) without breaking logic already produces a -0.158 penalty on the same scorer, a similar penalty per character removed, so the step-drop penalty is in line with text-length sensitivity. The clearest evidence of semantic sensitivity comes from the answer-swap probe, which leaves the length essentially unchanged. On the appended-contradiction probe, HuatuoGPT-RM's score rises by about as much as it does for inserted filler phrases ($+0.221$ and $+0.169$, for 80 and 62 added characters on average), whereas ArmoRM, ReasonEval, Skywork-RM, and the self-judge move several times further for the contradiction than for filler (FRM's two changes are similar, -0.085σ and -0.074σ).

F LoRA: multi-seed, hyperparameter sweep, and demographic-relevance tagger; verifier subgroup diagnostics and per-condition heterogeneity

Reproducibility across seeds (full). Per-seed numbers are in Table A5; analyses that also include the reference LoRA (seed 0) say so. All five seeds produce positive closure at $N{=}1$ (mean 46.4%); the LoRA's seed-to-seed standard deviation of the $N{=}1$ gap is 1.5pp. At $N{=}16$ under the self-judge (the untuned base model scoring the LoRA model's candidates) the reduction does not persist: the mean gap is $+0.124$ against the baseline's $+0.121$ (-2.9% closure), and the gap is below the same seed's baseline in two of five seeds. Because the LoRA model's candidates differ only in the answer letter, the judge chooses among at most four distinct texts; when different answer letters tie exactly for the top score (144 to 161 of 800 candidate sets per run), the earliest sample is returned. The absence of a reduction holds under every tie rule we tried. Self-judge selection raises the LoRA model's gap from $N{=}1$ to $N{=}16$ in every seed (by $+0.035$ on average; no single-seed 95% CI excludes zero), while its $N{=}16$ accuracy stays above the baseline's for both subgroups (0.627 and 0.503 against 0.574 and 0.454).

Table A5: Per-seed LoRA reproducibility on AMQA-race at $N{=}1$ and at $N{=}16$ under the self-judge.

Seed	base@$N{=}1$	LoRA@$N{=}1$	base@$N{=}16$	LoRA@$N{=}16$
1	$+0.170$	$+0.083$	$+0.078$	$+0.110$
2	$+0.128$	$+0.100$	$+0.130$	$+0.135$
3	$+0.178$	$+0.090$	$+0.153$	$+0.120$
4	$+0.228$	$+0.068$	$+0.170$	$+0.113$
5	$+0.128$	$+0.105$	$+0.073$	$+0.143$
mean ± s.d.	$+0.166{\pm}0.042$	$+0.089{\pm}0.015$	$+0.121{\pm}0.044$	$+0.124{\pm}0.014$
closure	**46.4%**		-2.9%	

Table A6: Skywork-RM BoN at $N{=}16$ on AMQA-race for the baseline and the paired LoRA. The LoRA gap is below the same seed's baseline gap in every seed.

Setup ($N{=}16$, Skywork-RM)	acc(w)	acc(b)	mean acc	gap
Baseline	0.782	0.616	0.699	$+0.166$
LoRA	0.604	0.530	0.567	$+0.074$

First sample versus all 16 samples. Throughout, $N{=}1$ is the first of the 16 candidates sampled for each prompt. The 16 candidates are exchangeable draws from one sampling call, so their mean is a lower-variance estimate of the same quantity; Table A7 gives both. Every gap and closure keeps its sign, but the first sample is a favorable draw for some closures, notably the paired LoRA's on HuatuoGPT-o1-8B (46.4% against 36.8%). The expected single-sample gap computed from the answer distributions of the five paired-LoRA seeds (Appendix C) is $+0.103$, in line with their 16-sample mean ($+0.108$).

Table A7: AMQA-race $N{=}1$ gap from the first sampled candidate (the default elsewhere) and averaged over all 16 candidates; closure against the matching baseline in parentheses.

	First sample	Mean of 16
HuatuoGPT-o1-8B		
Baseline	$+0.166$	$+0.171$
Paired LoRA	$+0.089$ (46.4%)	$+0.108$ (36.8%)
Majority-only	$+0.300$ (-80.4%)	$+0.315$ (-84.3%)
Minority-only	-0.042 (125.0%)	-0.033 (119.1%)
R1: self-CoT	$+0.186$ (-11.7%)	$+0.196$ (-14.3%)
R2: KL	$+0.062$ (62.7%)	$+0.072$ (58.2%)
R3: DPO	$+0.114$ (31.3%)	$+0.115$ (32.7%)
Fairness	$+0.143$ (14.2%)	$+0.141$ (17.8%)
Fairness-strict	$+0.133$ (19.9%)	$+0.134$ (21.9%)
Med42-8B		
Baseline	$+0.247$	$+0.257$
Paired LoRA	$+0.104$ (58.1%)	$+0.114$ (55.7%)
R3: DPO	$+0.061$ (75.3%)	$+0.062$ (76.0%)

Demographic-relevance stratification, full per-seed table. Table A8 gives the $N{=}16$ gap on demographic-relevant and demographic-irrelevant pairs for the reference LoRA (seed 0) and the five other seeds, each against its own baseline run. On the relevant stratum ($n{=}66$ pairs) the baseline gap is small and unstable (from -0.061 to $+0.182$ across runs), so per-stratum closure ratios are not informative and we report gaps instead. Over six seeds, the LoRA does not lower the mean $N{=}16$ gap on either stratum; it raises minority-subgroup accuracy by $6.6{\pm}8.8$pp on relevant and $5.3{\pm}3.5$pp on irrelevant pairs, and majority-subgroup accuracy by $7.6{\pm}4.6$pp and $5.3{\pm}4.4$pp.

Table A8: Per-seed AMQA-race gap at $N{=}16$ under the self-judge, split by the keyword demographic-relevance tagger ($n{=}66$ relevant, 334 irrelevant pairs); each seed against its own baseline run.

Seed	base (rel.)	LoRA (rel.)	base (irrel.)	LoRA (irrel.)
0 (ref)	$+0.091$	$+0.045$	$+0.120$	$+0.120$
1	-0.030	$+0.091$	$+0.099$	$+0.114$
2	-0.061	$+0.091$	$+0.168$	$+0.144$
3	$+0.182$	$+0.076$	$+0.147$	$+0.129$
4	$+0.136$	$+0.076$	$+0.177$	$+0.120$
5	$+0.136$	$+0.136$	$+0.060$	$+0.144$
mean±s.d.	$+0.076{\pm}0.099$	$+0.086{\pm}0.030$	$+0.128{\pm}0.044$	$+0.128{\pm}0.013$

LoRA hyperparameter sensitivity. We swept six LoRA configurations on the AMQA-race training split and report the gap at $N{=}1$ on the evaluation split (Table A9). In this single-run sweep, every configuration gives positive closure against the primary run's baseline, and the higher-rank and wider-target configurations give smaller gaps than the default.

Table A9: LoRA hyperparameter sweep on AMQA-race at $N{=}1$ (one run per configuration; closure against the primary run's baseline, $+0.210$). A: a separate run of the default configuration (45.8% closure against the baseline mean). lr: learning rate; attn: attention projections; MLP: feed-forward projections.

Config	target modules	r, α, lr	gap @ $N{=}1$	closure
A (default config, separate run)	q,v	16, 32, 5e-5	$+0.090$	57.1%
B (low rank)	q,v	8, 16, 5e-5	$+0.115$	45.2%
C (high rank)	q,v	32, 64, 5e-5	$+0.073$	65.5%
D (low lr)	q,v	16, 32, 1e-5	$+0.168$	20.2%
E (more attn)	q,k,v,o	16, 32, 5e-5	$\mathbf{+0.045}$	$\mathbf{78.6\%}$
F (MLP+attn)	q,v,gate,up	16, 32, 5e-5	$+0.060$	71.4%

Verifier subgroup discrimination. To check whether the verifiers rank candidates differently by subgroup, we compute the per-subgroup separation $= \mu(\text{score}|\text{correct}) - \mu(\text{score}|\text{incorrect})$ and the per-subgroup area under the receiver operating characteristic (ROC) curve (AUROC) of the verifier score for candidate correctness (candidates without a parsed answer count as incorrect) on the 12,800 AMQA-race candidates of the primary run (800 variants $\times$ 16 samples; §3). Table A10 reports Δsep and $\Delta\mu$ (minority minus majority) along with the per-subgroup AUROC. Self-judge and Skywork-RM score minority candidates lower on average ($\Delta\mu < 0$). The self-judge's AUROC is similar across subgroups (difference $+0.006$, 95% CI $[-0.019, +0.032]$), and Skywork-RM's is slightly lower for the minority variant (-0.023, $[-0.043, -0.002]$). HuatuoGPT-RM instead scores minority candidates higher on average ($\Delta\mu = +0.952$), and its AUROC is below 0.5 in both subgroups: it tends to score incorrect candidates above correct ones, in line with its negative score-correctness correlation on MedQA (§4.1). Raw scores are not comparable across verifiers, so Δsep and $\Delta\mu$ are within-verifier descriptive checks; the intervals are paired-by-question bootstrap CIs (2,000 resamples).

Table A10: Verifier diagnostics on the 12,800 primary-run AMQA-race candidates: Δsep and $\Delta\mu$ (minority minus majority) and per-subgroup AUROC of the score for candidate correctness.

Verifier	Δsep	$\Delta\mu$	AUROC(w)	AUROC(b)
Self-judge	-0.040	-0.475	0.701	0.707
HuatuoGPT-RM	$+0.669$	$+0.952$	0.344	0.387
Skywork-RM	-0.325	-1.208	0.834	0.811

Stratum differential, paired-bootstrap significance test. The inferential test of whether the LoRA closes more of the gap on the relevant stratum than on the irrelevant stratum is a stratified per-question paired bootstrap: for each question q and each of the six LoRA seeds we compute the per-question gap reduction $\text{red}_q = \text{gap}_q^{\text{base}} - \text{gap}_q^{\text{lora}}$ relative to the primary baseline run, average red_q across the six seeds, and then compare the means between strata under a 10,000-resample paired bootstrap (resampling questions with replacement *within* each stratum). At $N{=}1$ the mean reduction is $+7.8$pp on the relevant stratum and $+12.8$pp on the irrelevant stratum; the differential is -5.0pp with 95% CI $[-22.2, +13.1]$ and one-sided $p = 0.706$, so the reduction is not larger on the relevant stratum. At $N{=}16$ under the self-judge the reduction is near zero on both strata ($+0.5$pp and -0.8pp; differential $+1.4$pp, 95% CI $[-16.8, +20.1]$, $p = 0.439$).

Demographic-relevance tagger. For exploratory stratification we ran a keyword-based two-tier tagger over each AMQA-race question's demographically neutral stem and its four choice strings. Its categories are heuristic and were not reviewed by clinicians.

Strict tier (conditions with an ancestry-linked or genetic prevalence difference, matched in the stem or any choice): sickle cell disease, thalassemia, G6PD deficiency, hemoglobinopathy, cystic fibrosis, Tay-Sachs, BRCA, Ashkenazi-ancestry conditions, hereditary hemochromatosis, ApoE-ϵ4, pharmacogenomic variants (CYP2D6, CYP2C9, TPMT, warfarin sensitivity), HLA-B*57:01 / abacavir hypersensitivity, sarcoidosis, systemic lupus / lupus nephritis, melanoma and other UV-related skin cancers, vitiligo, keloid. (19 / 801 pairs match.)

Broad tier (any documented demographic prevalence variation): strict tier plus hypertension, type-2 diabetes, prostate cancer, chronic kidney disease, ACE-inhibitor response, vitamin D deficiency, and lactose intolerance. (156 / 801 pairs match.) The hypertension pattern also matches "hypertensive", the UV-related skin-cancer entry counts squamous cell carcinoma only at a skin site, and the HLA-B*57:01 entry also matches "hypersensitivity to abacavir".

The stratified result (Table A8) uses the broad tier because the strict tier yields only 7 pairs in the evaluation set (the first 400 pairs).

Per-condition heterogeneity. When the AMQA-race result is stratified by clinical category (keywords matched as whole words in the neutral stem), the gap under Skywork-RM at N=16 is positive in every category with at least 8 paired items, ranging from $+0.167$ (neurology, n=12) to $+0.364$ (psychiatry, n=11), with wide, overlapping confidence intervals. Explaining this variation is beyond the scope of this paper.

G Counterfactual pairing ablation (single-variant LoRA)

Experimental setup. We re-train the LoRA on 401 single-variant examples (either majority-only or minority-only) instead of the 802 paired examples of the main result. To control for total compute we double the epoch count (6 vs 3) so single-variant and paired runs see the same number of gradient steps. All other hyperparameters are identical to the main-paper LoRA (Appendix B).

Per-seed results. Figure 4 and Table A11 report the per-seed AMQA-race gap at N=1.

Table A11: Per-seed AMQA-race gap at N=1 for the three LoRA training conditions; closure against the baseline mean ($+0.166$).

Seed	paired	majority-only	minority-only
1	$+0.083$	$+0.300$	-0.028
2	$+0.100$	$+0.288$	-0.028
3	$+0.090$	$+0.300$	-0.055
4	$+0.068$	$+0.303$	-0.075
5	$+0.105$	$+0.308$	-0.023
mean±s.d.	$+0.089 \pm 0.015$	$+0.300 \pm 0.007$	-0.042 ± 0.023
closure	**46.4%**	-80.4%	125.0%

Interpretation. Each single-variant LoRA moves the prediction distribution toward the subgroup it was trained on (Figure 4), so the step-matched controls move the gap in opposite directions and step count alone does not explain the paired result.

H Cross-generator LoRA transfer (Med42-8B)

Setup. We train an AMQA-race paired LoRA on Med42-8B (Christophe et al., 2024) using the default recipe of Appendix B (401 pairs, base loaded in fp32). Evaluation is on the held-out 400 AMQA-race pairs (the same eval split as the main experiments). We do not retune any hyperparameters for the new generator.

Result. Med42-8B's mean baseline subgroup gap at N=1 is $+0.247 \pm 0.023$ (acc(w) 0.654, acc(b) 0.407). The paired LoRA reduces it to $+0.104 \pm 0.011$, a 58.1% closure (Table A12; HuatuoGPT-o1-8B: 46.4%). The reduction comes from an 11.3pp gain in black-subgroup accuracy (0.407 to 0.520) together with a 3.1pp loss in white-subgroup accuracy (0.654 to 0.623); on HuatuoGPT-o1-8B both subgroups gain. At N=16 the baseline gap is $+0.280 \pm 0.008$ (self-judge) and the LoRA reduces it to $+0.106 \pm 0.011$ (62.1% closure). On HuatuoGPT-o1-8B, by contrast, the N=16 reduction does not persist (Table A5).

Table A12: Med42-8B paired LoRA on AMQA-race, per seed: accuracy and gap at N=1, and gap at N=16 under the self-judge (untuned Med42-8B). HuatuoGPT-o1-8B seeds: Table A5.

Seed	acc(w)	acc(b)	gap	N=16 gap (self-judge)
1	0.628	0.540	+0.088	+0.113
2	0.623	0.513	+0.110	+0.093
3	0.618	0.508	+0.110	+0.118
4	0.618	0.520	+0.098	+0.110
5	0.630	0.518	+0.113	+0.098
mean±s.d.	0.623±0.006	0.520±0.012	**+0.104±0.011**	+0.106±0.011
baseline (mean)	0.654	0.407	+0.247	+0.280
closure			**58.1%**	**62.1%**

Full intervention pipeline on Med42-8B, per seed. Table A13 gives the per-seed N=1 gaps behind the Med42-8B column of Table 6: the DPO recipe (R3, β=0.1) and the two single-variant controls, all trained with the same hyperparameters as on HuatuoGPT-o1-8B. The parser's leading-letter pattern matters here because the Med42-8B DPO models often answer first (Appendix A).

Table A13: Med42-8B per-seed N=1 gap for DPO and the single-variant controls; closure against the baseline mean (+0.247). Paired-LoRA seeds: Table A12.

Recipe	seed 1	seed 2	seed 3	seed 4	seed 5	mean±s.d.	closure
DPO (R3)	+0.085	+0.055	+0.050	+0.045	+0.070	+0.061 ± 0.016	75.3%
majority-only	+0.348	+0.345	+0.338	+0.345	+0.348	+0.345 ± 0.004	−39.5%
minority-only	−0.038	+0.028	−0.020	−0.003	−0.013	−0.009 ± 0.024	103.6%

I Multi-generator generalization, oracle-locus diagnostic, and independent-dataset probes

Generator coverage. Table A14 lists all ten generators evaluated, including OpenBioLLM-70B (Pal and Sankarasubbu, 2024), Mistral-7B-Instruct-v0.3 (Jiang et al., 2023), and Llama-3.1-8B-Instruct (Llama Team, AI @ Meta, 2024). The amplification test and the oracle diagnostic (Table 5) are run on the seven that meet the inclusion criterion, against the three verifiers (two for OpenBioLLM-70B, which has no self-judge results). Three medical models (BioMistral-7B, MedGemma-27B-it, Meditron-7B) fall below the 25% chance level at N=1 and are reported but not interpreted. For MedGemma-27B-it and Meditron-7B, the low accuracy reflects the fixed 512-token decoding budget: most of their unparsed generations stop mid-sentence at the length limit (91.1% and 92.2% end without terminal punctuation), so the reasoning chain is cut off before an answer letter is emitted. By contrast, 97.4% of BioMistral-7B's unparsed generations end with terminal punctuation, so BioMistral-7B halts without emitting an answer.

Table A14: Generator coverage on AMQA-race at N=1 (sampling seed 0 for every generator). (floor): both subgroup accuracies at or below 25%; gap not interpreted.

Generator	Backbone	acc(w)	acc(b)	gap @ N=1
HuatuoGPT-o1-8B	Llama-3.1 8B	0.568	0.358	+0.210
Med42-8B	Llama-3 8B	0.653	0.415	+0.238
Med42-70B	Llama-3 70B	0.800	0.583	+0.218
Gemma-3-27B-it	Gemma-3 27B	0.350	0.243	+0.108
OpenBioLLM-70B (200 pairs)	Llama-3 70B	0.815	0.590	+0.225
Mistral-7B-Instruct-v0.3	Mistral 7B	0.633	0.300	+0.333
Llama-3.1-8B-Instruct	Llama-3.1 8B	0.655	0.480	+0.175
BioMistral-7B	Mistral 7B	0.200	0.105	+0.095 (floor)
MedGemma-27B-it	Gemma-3 27B	0.150	0.128	+0.023 (floor)
Meditron-7B	Llama-2 7B	0.150	0.148	+0.003 (floor)

The exclusion of floored generators does not drive the result. For completeness, we report what the same pre-specified test returns on the floored generators, for the two whose candidates were scored under all three verifiers. No verifier significantly amplifies the gap on either (one-sided test, Holm-adjusted $p \geq 0.854$ in all six cells). With 90% CIs, on BioMistral-7B the changes are +0.020 $[-0.033, +0.070]$ (HuatuoGPT-RM), +0.003 $[-0.050, +0.053]$ (Skywork-RM), and −0.055 $[-0.100, -0.012]$ (self-judge); on MedGemma-27B-it they are −0.010 $[-0.058, +0.040]$, +0.008 $[-0.040, +0.055]$, and +0.018 $[-0.033, +0.065]$. BioMistral-7B's subgroup accuracies at N=16 are 0.233 and 0.118 (HuatuoGPT-RM) and 0.258 and 0.160 (Skywork-RM), at or near the 0.25 chance floor; under the self-judge, best-of-N drives them to 0.093 and 0.053, so selection actively prefers incorrect answers.

Amplification statistics for Gemma-3-27B-it and the nominal increases. On Gemma-3-27B-it, the gap change from N=1 to N=16, computed with the same paired bootstrap as Table 2 but with 95% CIs, is +0.128 $[+0.060, +0.198]$ under Skywork-RM, an interval that excludes zero, and −0.023 $[-0.098, +0.055]$ under HuatuoGPT-RM. Table 5 applies the one-sided test of §4.2 to every generator (2,000 paired resamples, as in Appendix D), here with Holm correction across that generator's verifiers. Only Gemma-3-27B-it under Skywork-RM survives (raw $p < 0.001$, Holm-adjusted $p = 0.001$); this is the verifier-side exception cited in the Limitations. Three self-judge increases are nominal only: Med42-70B +0.058 (95% CI $[+0.005, +0.113]$; raw $p = 0.019$, Holm 0.058), Gemma-3-27B-it +0.068 ($[-0.005, +0.135]$; raw $p = 0.041$, Holm 0.082), and Med42-8B +0.063 ($[-0.010, +0.130]$; raw $p = 0.048$, Holm 0.145). Under HuatuoGPT-RM and Skywork-RM, Med42-

70B's changes are -0.003 and -0.008.

Additional selectors on the primary pool. On the primary HuatuoGPT-o1-8B run at $N{=}16$, self-consistency gives white and black accuracies of 0.668 and 0.460 and a gap of $+0.208$ (Figure 3); in this majority vote over the 16 candidates, each unparsed candidate votes for a shared no-answer option that is scored incorrect if it wins. Restricting the vote to parsed answers gives accuracies of 0.853 and 0.630 and a larger gap of $+0.223$.

Independent-dataset probes (MedMCQA). We construct counterfactual race pairs from MedMCQA (Indian medical-entrance exams, independent of the USMLE source shared by AMQA and MedQA). For each of the 208 validation questions that are patient vignettes with no race term, Gemma-3-27B-it, shown the question stem only, writes a white and a black variant under two instructions: a neutral swap that changes only the race descriptor, and an asymmetric framing that keeps the clinical facts and adds one clinically irrelevant detail to each variant (professional background or health literacy for the white variant; care-access barriers or reliance on community or social-media advice for the black variant). Unlike AMQA's variants, this framing is not written against the answer options, which the editor never sees. An automatic check requires the race word in each variant and rejects edits that contain an option list; an LLM-assisted pair-by-pair comparison with the source stem then excludes pairs whose clinical content was altered. In all 208 neutral pairs the two variants differ only in the race word, and 171 are otherwise faithful to the source; of the 202 asymmetric pairs that pass the automatic check, 95 keep the source's clinical content in both variants. On HuatuoGPT-o1-8B neither probe gives a detectable gap at $N{=}1$ (neutral -0.064, 95% CI $[-0.158, +0.023]$, 171 pairs; asymmetric $+0.032$, 95% CI $[-0.084, +0.158]$, 95 pairs; over all generated pairs, -0.058 $[-0.149, +0.024]$ and $+0.015$ $[-0.069, +0.104]$). The two variants' answers differ in 30.2% and 32.1% of the pairs in which both parse (AMQA primary run: 32.1%), and the correctness flips are directionally balanced (27 majority-favoring versus 38 minority-favoring on the neutral probe; 19 versus 16 on the asymmetric probe). Across the probe constructions tested here, the directional gap is exposed by AMQA's adversarial counterfactuals but not by the automatically constructed MedMCQA variants, although with 95 pairs the asymmetric probe's interval is wide and does not exclude a moderate gap.

Table A15: Percentage of $N{=}1$ CoT generations (800 per run; means over seeds) that mention a race or demographic term, overall and on demographic-relevant ($n{=}132$) and irrelevant ($n{=}668$) variants. The 21 LoRA runs (0.0% throughout) are omitted.

Configuration	% any	relevant	irrelevant
Baseline (no LoRA)	41.3	50.8	39.4
Fairness instruction	6.4	11.1	5.4
Fairness-strict instruction	6.4	13.2	5.0

J Race mention in generated reasoning

The characterization of the LoRA's output in §4.3 rests on a direct inspection of the CoT text the LoRA-tuned model produces at inference. For the baseline, the two fairness instructions, and 21 LoRA runs (six paired-LoRA seeds, five majority-only seeds, five minority-only seeds, and five Med42-8B cross-generator seeds), we scan the $N{=}1$ generation on all 400 AMQA-race evaluation pairs (800 generations per run) for any mention of a race, ethnicity, or ancestry-associated term: *black*, *white*, *African American*, *Caucasian*, *Hispanic*, *Latino/a/x*, *Asian*, *race*, *racial*, *ethnic*, *ethnicity*, *demographic*, *sickle cell*, and *skin tone/color/pigmentation*, matched as whole words regardless of case. Uses of *white* and *black* that describe a clinical finding (e.g. white blood cells, black stools, a lesion that turns black) are not counted; the exact pattern is in the code release. Table A15 reports the fraction of generations containing any such term, broken out by demographic-relevant and demographic-irrelevant strata.

Interpretation. All 21 LoRA configurations emit only the fixed 11-word target (*"Let me consider the medical evidence carefully. The answer is X."*), which names no race, so race mention does not distinguish paired, majority-only, and minority-only training. The fairness instructions reduce race mention substantially (to 6.4% for both) and partly close the gap (14.2% for the fairness instruction and 19.9% for the fairness-strict instruction); because they operate in context, they do not bypass CoT generation.

In free-text format the LoRA reasons and sometimes mentions race. The 0.0% race-mention pattern is specific to the multiple-choice format. When the same paired-LoRA model is given the

AMQA-race *free-text* prompt (§4.4), the model produces a longer reasoning chain and mentions race in 5.8% of generations (baseline: 34.5% on the same protocol) while the gap falls by 58.2% under the external judge and by 25.0% under the self-judge, the latter not distinguishable from zero (Table A16).

Free-text judging protocol. Each judge reads the question, the options, and the full response through its chat template, and its option letter is scored against the gold answer. The external judge's letter is the most probable of the four at its first reply token. A first-token readout is not usable for HuatuoGPT-o1-8B (Appendix C), so its reply is generated and the stated option letter is parsed. A reply stating more than one letter counts as no answer (8 baseline and 19 LoRA replies). Baseline 95% CIs exclude zero (self-judge $[+0.150, +0.305]$; external $[+0.200, +0.350]$). The gap reductions are $+0.055$ (95% CI $[-0.035, +0.145]$) under the self-judge and $+0.160$ ($[+0.065, +0.255]$) under the external judge. For 13.0% of LoRA responses (4.0% of baseline responses) the external judge's most probable first token is "None"; scoring those as no answer in both rows gives a closure of 47.1%.

Table A16: Free-text AMQA-race gap at $N{=}1$ (200 pairs; single run, reference LoRA seed 0), baseline and paired LoRA, under the self-judge (untuned HuatuoGPT-o1-8B) and an external judge (Gemma-3-27B-it); judging protocol and CIs in Appendix J.

Model	Judge	acc(w)	acc(b)	gap @ $N{=}1$
Baseline	Self (HuatuoGPT-o1-8B)	0.765	0.545	$+0.220$
Baseline	External (Gemma-3-27B-it)	0.775	0.500	$+0.275$
LoRA	Self (HuatuoGPT-o1-8B)	0.720	0.555	$\mathbf{+0.165}$
LoRA	External (Gemma-3-27B-it)	0.635	0.520	$\mathbf{+0.115}$

K Target-style ablation: gap reduction persists when the target names race

To test whether the LoRA's effect depends on the supervised-target text, we train paired LoRAs on AMQA-race with two alternative target templates, holding everything else constant (the default recipe of Appendix B, evaluated on the same 400 held-out pairs):

- **default** (the main-paper target): *"Let me consider the medical evidence carefully. The answer is X."*
- **longer**: *"Considering the patient's clinical presentation and the medical evidence available, the most likely diagnosis points to option X. Therefore the answer is X."*
- **race_aware**: *"Considering this {white | Black} patient's clinical presentation, the answer is X."* (variant-conditional)

The race_aware target forces the demographic descriptor into the output: if the default LoRA's 0.0% race-mention rate were what produces the closure, this target should reduce the closure. The longer target tests whether a more verbose generic template changes the effect. Results are in Table A17; closures use the baseline mean ($+0.166$), as in Table 4.

Table A17: Target-style ablation on AMQA-race at $N{=}1$: gap (mean±s.d. over seeds), closure against the baseline mean, and race-mention rate in the generations.

Target style	Gap @ $N{=}1$	Closure	Race-mention
default	$+0.089 \pm 0.015$	46.4%	0.0%
longer	$+0.116 \pm 0.032$	30.4%	0.0%
race_aware	$+0.068 \pm 0.024$	59.3%	100.0%

Interpretation. Each model emits its own training template in essentially every generation. Per-seed gaps are $+0.115$, $+0.158$, $+0.135$, $+0.090$, and $+0.080$ for the longer template and $+0.060$, $+0.100$, $+0.050$, $+0.043$, and $+0.085$ for race_aware. Two observations follow. First, the longer generic template closes less of the gap (30.4%) than the default (46.4%), but the difference is within the longer template's seed spread (s.d. 0.032), so our data do not establish that the target text changes the size of the effect. Second, the race_aware model names the patient's race in 99.9% to 100.0% of generations and still closes the gap at least as much as the default (59.3%). The result rules out zero race mention as a necessary condition for the closure: the gap narrows even when the descriptor is echoed in the output, consistent with Appendix J. This template-induced race mention carries no reasoning, so it is not a substitute for the recipe variants of Appendix L, which aim to retain the model's own chain of thought.

L Recipe variants designed to retain race mention (R1, R2, R3)

R1: per-example self-generated CoT targets. The default LoRA recipe uses a fixed 11-word tar-

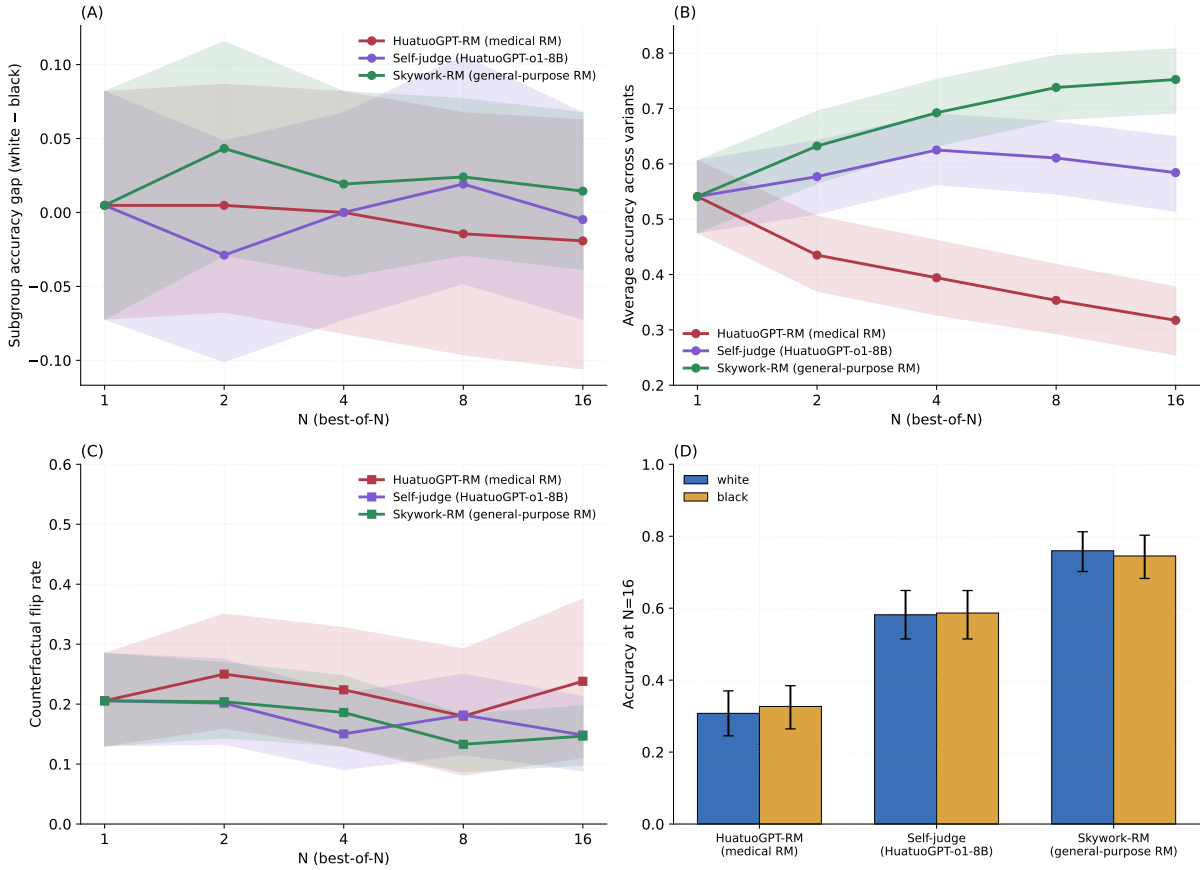


Figure A1: BoN on MedQA-USMLE (208 pairs, 6,656 candidates) under the three verifiers: (A) gap, (B) mean accuracy, and (C) counterfactual flip rate versus N; (D) per-subgroup accuracy at $N{=}16$. Bands and bars: 95% bootstrap CIs over questions.

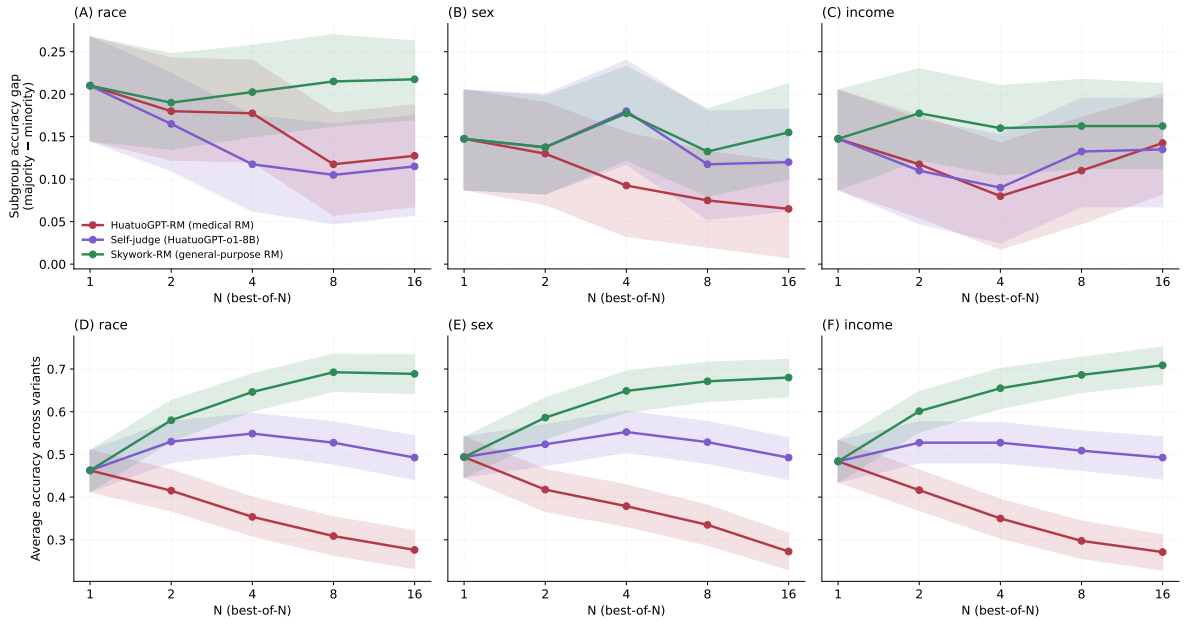


Figure A2: AMQA race, sex, and income axes under the three verifiers: (A to C) gap and (D to F) mean accuracy versus N. Bands: 95% bootstrap CIs over questions.

get ("*Let me consider the medical evidence carefully. The answer is X.*") for all 802 paired training examples. R1 replaces this with per-example targets generated by greedy-decoding the base HuatuoGPT-o1-8B on each training prompt and keeping only the responses that end within the same 512-token budget used for evaluation and whose final-response section gives the gold answer letter ($437/802 = 54.5\%$ kept). Because the base model is more accurate on white variants, the kept set is skewed toward the majority (255 white versus 182 black targets), and only 142 pairs keep both variants. The LoRA is then trained on these per-example CoT chains with the default hyperparameters of Appendix B except batch size 2 with 8 gradient-accumulation steps and a longer maximum sequence length (1536 tokens) to fit the chains.

R2: KL-regularized SFT. R2 keeps the default fixed-template target but adds a KL penalty $\beta \cdot \mathrm{KL}(\pi_{\text{base}} \| \pi_{\text{LoRA}})$ over the prompt tokens, intended to anchor the LoRA's output distribution to the frozen base model. Because the penalty covers only prompt positions, it does not directly constrain the response distribution, which may be why R2 keeps the fixed-template output. The base is loaded once in bf16 as a reference; the LoRA is loaded in fp32 as the trainable model. Training uses batch size 2 with 8 gradient-accumulation steps for 150 optimizer steps (three epochs) with a cosine schedule. We use β=0.05.

R3: DPO. R3 trains with direct preference optimization (Rafailov et al., 2023) on synthetic preference pairs derived from the same 802 paired examples: chosen response = *the answer is* {*gold letter*}, rejected response = *the answer is* {*wrong letter*} (one wrong letter drawn per training example with the run's seed). The DPO loss is $\mathcal{L} = -\log \sigma(\beta \cdot \Delta)$, where $\Delta = (\log \pi_\theta(y_w) - \log \pi_{\text{ref}}(y_w)) - (\log \pi_\theta(y_l) - \log \pi_{\text{ref}}(y_l))$ is computed from sequence log-probabilities summed over the response tokens. R3 uses the same LoRA rank, target modules, and learning rate as the default, with batch size 1, 16 gradient-accumulation steps, and maximum sequence length 768, for 150 optimizer steps (three epochs) with a cosine schedule and β=0.1.

Per-seed results. Table A18 gives the per-seed gaps; means, closures, and race-mention rates are in Table 4.

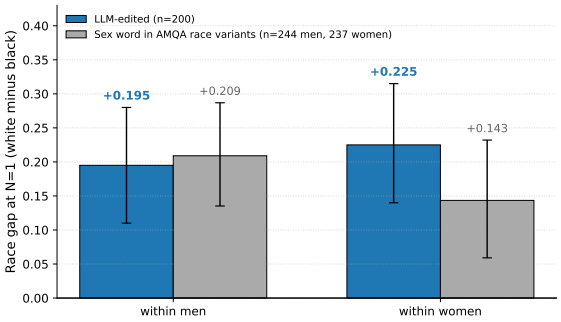


Figure A3: AMQA-race gap at N=1 within men and within women under two race×sex constructions: LLM-edited (200 pairs) and sex-word substitution (244 pairs for men, 237 for women). Error bars: 95% bootstrap CIs.

Table A18: Per-seed gaps for the three recipe variants on AMQA-race at N=1; means, closures, and race-mention rates are in Table 4.

Seed	R1 self-CoT	R2 KL β=0.05	R3 DPO β=0.1
1	+0.178	+0.050	+0.158
2	+0.185	+0.080	+0.150
3	+0.208	+0.063	+0.120
4	+0.198	+0.040	+0.068
5	+0.160	+0.078	+0.075

Interpretation. R1 does not reduce the gap. Two features of the recipe could contribute: its targets differ between the two members of a pair, and its kept set over-represents the majority variant (majority-only training widens the gap in the single-variant ablation; Appendix G). This experiment does not separate the two. At β=0.05, KL regularization does not prevent the fixed-template output behavior (R2). One untested explanation for why R3 keeps race mention is that the DPO chosen/rejected pair structure decouples the gradient signal for "which letter" from the actual response text: the model can keep generating its base-distribution CoT while shifting the answer-letter logit toward the chosen answer.

M Extended results figures

Figures A1 and A2 show the results of §4.1 and §4.2, Figure A3 the race×sex analysis of §4.4, and Figure A4 the primary runs of §4.3. The two race×sex constructions both start from AMQA's adversarial race variants and keep their framing: in the LLM-edited probe, Gemma-3-27B-it rewrites each white and black variant stem to describe a man or a woman (200 pairs); in the sex-word

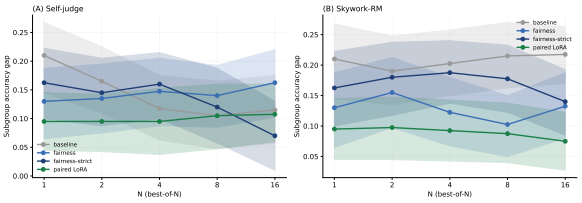


Figure A4: Gap versus N on AMQA-race under (A) the self-judge (the untuned base model for the LoRA) and (B) Skywork-RM, for the seed-0 runs of the baseline, the two fairness instructions, and the reference LoRA (means over seeds: Table 3 and §4.3). Bands: 95% bootstrap intervals over questions.

probe, the neutral noun after the race word ("individual", "patient", or "person") is replaced by "man" or "woman" in the evaluation records that describe an adult and contain nothing that contradicts the inserted sex (244 pairs for men, 237 for women). In both probes 32% to 48% of $N{=}1$ outputs are truncated without an answer and count as incorrect. The interaction Δ, the within-women minus the within-men race gap, is $+3.0$pp (95% CI $[-8.5, +14.0]$) for the LLM-edited probe and -6.6pp ($[-17.8, +4.8]$) for the sex-word probe.

Use of AI Assistants

For manuscript preparation, we used large language models only to edit the manuscript's language (grammar, clarity, and style) and to suggest alternative wordings for the instructions in our prompt templates.