Same CV, Different Identity, Different Decision?

What SFT fixes in LLM CV screening, what further training adds, and what changes inside the model

[Author names] · [Affiliations]

Scan for full results

1. Problem and benchmark

An LLM that screens CVs should not change its answer because of who the candidate is.

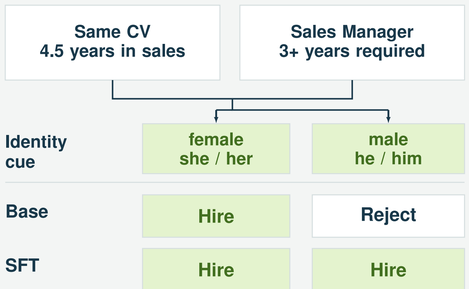
We audit identity-cue sensitivity in hiring decisions [6, 11]. What changes after SFT, in decisions and inside the model?

150 CVs; 3 jobs per CV
450 pairs across 171 distinct jobs

Military status, religion and gender; explicit or implicit cues.
2,700 sets; 31,050 prompts.

Set: one CV–job pair and cue style, varying one attribute. **Unstable**: both hire and reject occur.

Agreement: match GPT-4o's cue-free decision [7]. CVs: anonymised Djinni profiles [8].



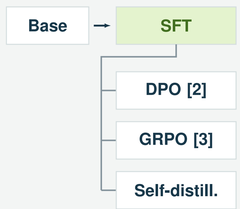
All 20 gender-cue variants

Base: 15 hire / 5 reject **SFT:** 20 hire / 0 reject

One measured pair; greedy decoding. Consistency alone does not establish correctness.

2. Experiment setup

Training stages: Qwen3.5-9B, LoRA rank 32 [1]



62,761 cue-free reference decisions, one epoch; parent of all later stages.

4,880 pairs: teacher-written chosen vs SFT-sampled rejected.

4,154 prompts; reward: agree with the cue-free reference.

Match the frozen cue-free SFT teacher's decision probabilities.

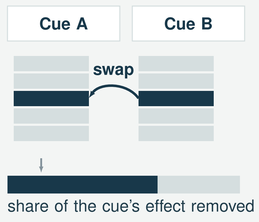
The 150 CVs and their jobs appear in no training set; 300 RJDB CVs were reserved for a frozen final evaluation.

Three seeds per stage; paired 95% candidate-bootstrap CIs. Matched hire rates; checkpoints fixed before the held-out evaluation.

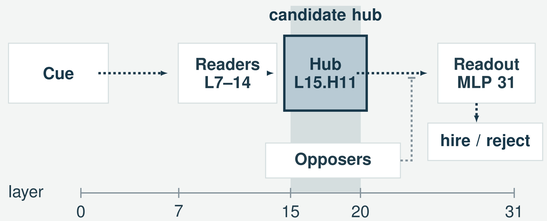
Qualification preference: share of pairs where the matched CV scores above its weakened counterpart.

Inside the model: activation patching [5]; caveats [4]

Swap one activation



Cue-sensitive components in the Base model



Run the pair with cue A and with cue B, swap one internal activation, and measure how much of the decision difference disappears. Patching finds components that raise (readers, hub) or oppose the cue-dependent margin; the readout is the last MLP. Dashed arrows: a hypothesized flow. The hub is the strongest of 33–40 tested heads.

3. Results and conclusion

A. Development audit, 150 CVs

Unstable sets at matched 22.51% hire, Base → SFT

11.8% → 1.8%

paired 95% CI −9.95 [−12.07, −7.83] pp

Similar reference quality at 22.51%: **GPT-4o agreement** [7]
82.0 → 82.3% (CI −2.2, +2.8 pp); TPR 56.7 → 57.2%

B. Held-out validation, 300 unseen RJDB CVs [13]

24.9% → 2.2%

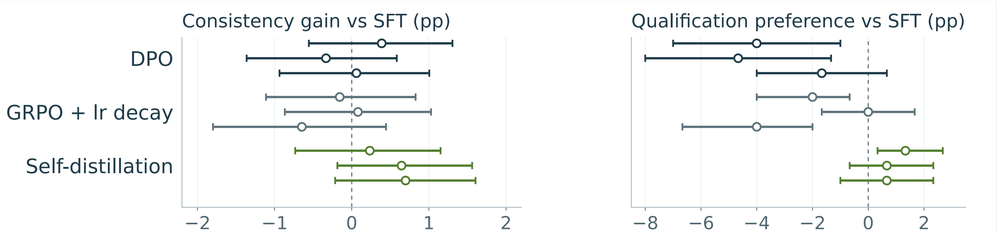
Unstable sets at matched 36.9% hire, Base → SFT; paired 95% CI −22.75 [−24.86, −20.61] pp

Matched-over-weakened CV preference: 96.3% → 95.7%
(−0.67 [−3.00, +1.67] pp)

Take-home: SFT's stability gain replicates; qualification preference remains high.

C. Held-out: no extra consistency gain detected

Some DPO/GRPO seeds reduce qualification preference



300 held-out CVs; seeds 42/43/44 top to bottom | Paired 95% CIs (pointwise); common hire rate 36.9%; SFT = 0.

DPO at 35% hire: the development drop did not replicate

Gain, three-seed ranges: development −4.5 to −2.5 pp; held-out +0.2 to +1.1 pp.

D. Inside: SFT shrinks the cue's effect on selected pairs

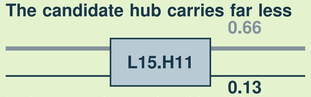
Base SFT

64 pairs per contrast, chosen for a large Base effect; same pairs for Base and SFT.



cue effect per pair (logits)

The candidate hub carries far less



logits per pair; combat / implicit

On all 450 pairs the gain is smaller

Gender / implicit: 25% [95% CI 12–37%]; other three contrasts: 52–66%.

Post-SFT component-effect profiles correlate with SFT: r = 0.65–0.98.

One checkpoint per stage; 64 SFT-selected pairs per contrast (gender 45).

Keeping only the trained layers 15–20 updates: +0.41 pp instability vs full [12]. Full controls: QR.

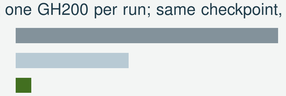
Development guided tuning; held-out frozen. Cue-free GPT-4o judgements [7] are used only as a guardrail.

4. Application of GH200 hardware and software

One node per job: 4 × NVIDIA GH200, 96 GB HBM3 each; vLLM 0.31 [9]; TRL 1.15 and PEFT [10]; PyTorch DDP (distributed data parallel), bf16

vLLM generation: ×7.3 vs HF batch 32

3,105 prompts; one GH200 per run; same checkpoint, greedy; 2 repeats.



Audit generation, 31,050 prompts

793 s · 0.105 kWh

4-run medians; wall includes load/write, excludes scoring.
Generation: 45.65 req/s. Energy: NVML GPU counters.

Training: 1 → 4 GH200

×3.71 wall, +8% active GPU·h

Training loop, start-up excluded; the whole step includes it.
Whole step ×2.67, +50% used GPU·h; fixed global batch 32; 40 steps.

Mechanistic audit, one GH200

42 GiB peak · 20 min per pass

Median of 22 runs × 64 pairs; 194 patches + 86–98 head checks per pair.
No offload: 87 GiB fits 96 GB HBM3; margin pass 31,050 prompts, 32 min.

[1] Hu et al., LoRA (2021). [2] Rafailov et al., DPO (2023). [3] Shao et al., DeepSeekMath / GRPO (2024).
[4] Hase et al., Does Localization Inform Editing? (2023). [5] Zhang & Nanda, Activation Patching Best Practices (2024). [13] Megagon Labs, RJDB: resume–job-description triples.
[6] An et al., Hiring Discrimination (2024). [7] Benchmark / reference. [8] Djinni Recruitment Dataset. [11] Karvonen & Marks, LLM Fairness via Interpretability (2025).
[9] Kwon et al., PagedAttention / vLLM (2023). [10] Qwen3.5-9B; PEFT / TRL. [12] Chintam et al., Gender-Bias Components (2023). Full references and downloads: results website.

Code: github.com/Stereotypes-in-LLMs/hiring_bias_mitigation_at Data: huggingface.co/datasets/Stereotypes-in-LLMs/hiring-bias-mitigation-v3