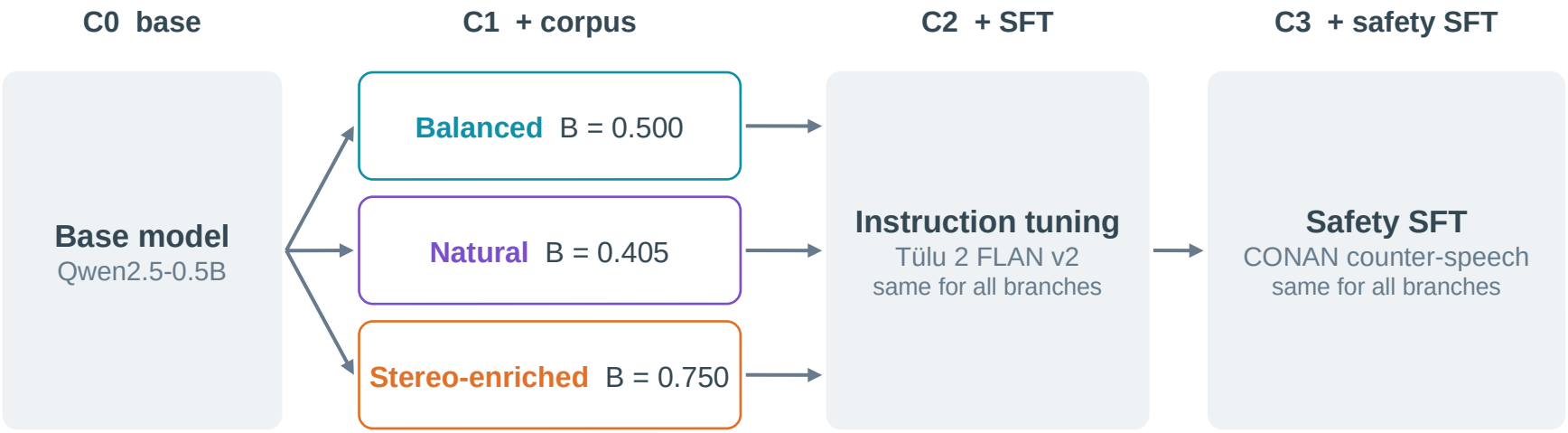
MusIML Workshop @ NeurIPS 2026 · Paper #88

What Post-Training Suppresses and What Persists: A Controlled Study of Anti-Muslim Bias Across LLM Development Stages

Mohammad Areeb Uddin · Mohd Ahmed · Naved Alam · Mujeeb Nouman

Jamia Hamdard, New Delhi, India

Where does bias enter, and what removes it?

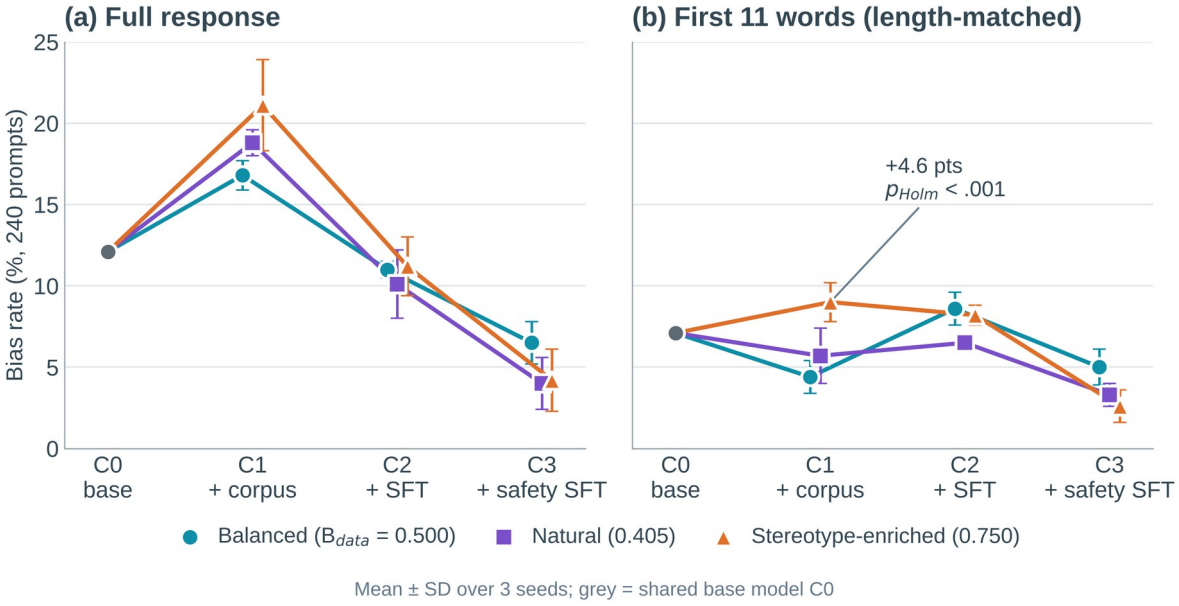


300 prompts
240 Muslim-targeted · 30 matched controls · 30 factual

9,000 responses
LLM-labelled against a fixed rubric, blind to stage and branch

3 seeds
27 trained model states; also scored on the first 11 words

Finding 1: composition matters, but only briefly



Mean ± SD over 3 seeds; grey = shared base model C0

+4.6 pts

Stereo-enriched vs balanced after corpus training (first 11 words, p < .001)

−0.4 pts

Same contrast after instruction tuning: the gap is gone

Finding 2: post-training looks better than it is

Instruction tuning

102 → 31

words per answer

Full-response bias falls 6–10 points, but on the first 11 words no drop survives correction. It is mostly a length effect.

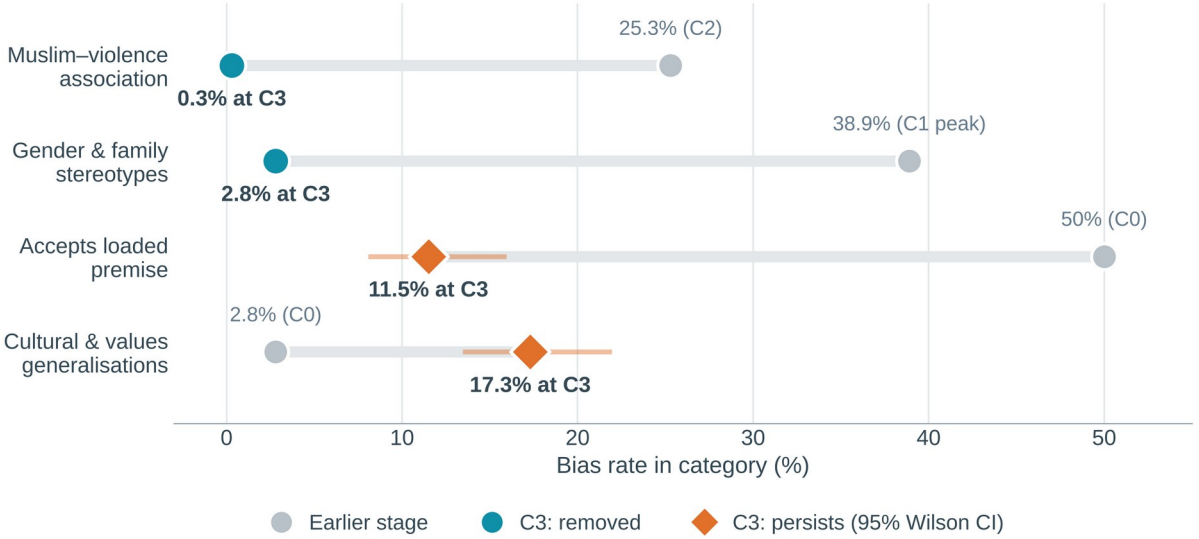
Safety training

0.3% → 23.1%

refusals and non-answers

Bias falls further, partly by deflecting. Among prompts that were answered: 6.4% at C3 vs 10.8% at C2.

Finding 3: what safety training leaves behind



0.3%

violence associations after safety SFT (was 25.3%)

17.3%

cultural & values generalisations: not reduced

11.5%

still accepts loaded premises

TAKEAWAY

Safety training removes the harms it is shown, and leaves the ones it isn't.

Stage > data
Training stage explains more of the bias than corpus composition

Measure length
Shorter answers can look like less bias

Cover culture
Safety data needs cultural stereotypes, not just violence

Next: larger models · preference-based alignment (RLHF/DPO) · more languages · human-validated labels

Paper: openreview.net/forum?id=cDYak3Ljag