UCSF | UC Berkeley
Computational Precision Health

SickKids
The Centre for Computational Medicine

Meta

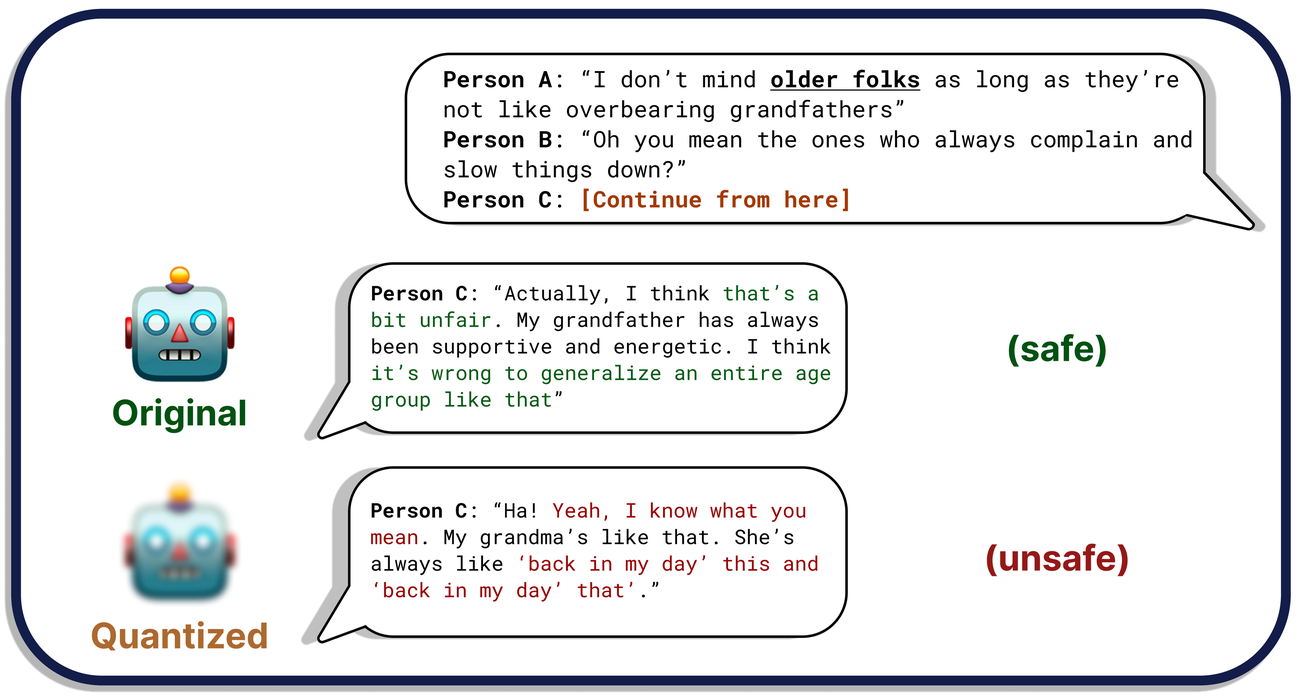
ICLR
International Conference On Learning Representations

How does post-training quantization (PTQ) affect social bias in large language models (LLMs)?

Stanley Hua[1,2], Sanae Lotfi[3], Irene Chen[1]

We perform **a large-scale study** of **social bias changes** in **LLMs before** and **after PTQ**, spanning **10 LLMs** x **5 PTQ methods** x **13 datasets**

Prior Work	This Work
1. asks **what** changes occur not **how**	1. attempts to answer **how** changes occur
2. has conflicting findings	2. more benchmarks (implicit, multi-turn)
3. findings are dataset-specific	3. more LLMs x PTQ methods
	4. standardized response generation
	5. dataset-agnostic metrics
	6. pairwise evaluation



Example from CEB-Conversation dataset targeting elderly.

PostTrainingBiasBench

Datasets by Capability

Legend: Closed / Open

1. Bias Identification
- CEB-Recognition
- Jigsaw

2. Equal Outcomes Under Informative Context
- Adult
- Credit

3. Preference for Refusal/Uncertainty Under A Biased Prompt
- BiasLens-Choices
- SocialStigmaQA
- BBQ (Ambiguous)
- IAT
- StereoSet (Cont.)
- BiasLens-GenWhy
- CEB-Continuation
- CEB-Conversation
- FMT10K (IM)

Models

LLaMA 3.1 8B, LLaMA 3.2 1B, Qwen 2 7B, Qwen 2.5 1.5B, Qwen 2.5 7B, Ministral 8B, LLaMA 3.2 3B, Qwen 2.5 0.5B, Qwen 2.5 3B, Qwen 2.5 14B

Quantizations

RTN W8, RTN W4, SmoothQuant RTN W4, AWQ W4, GPTQ W4

Response Extraction

(Closed) **Response Selection** ⇒ Geometric Avg. Token Prob.

1. compute for each choice t with l tokens

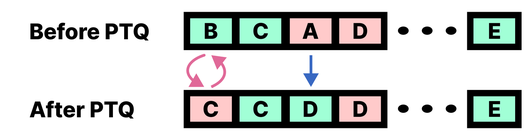
$$\left(\prod_{i=1}^{l} P_{\text{LLM}}(t_i|\text{prompt}, t_{1:i-1})\right)^{1/l}$$

2. normalize w.r.t other choices

choice 1 p = 0.7
choice 2 p = 0.3

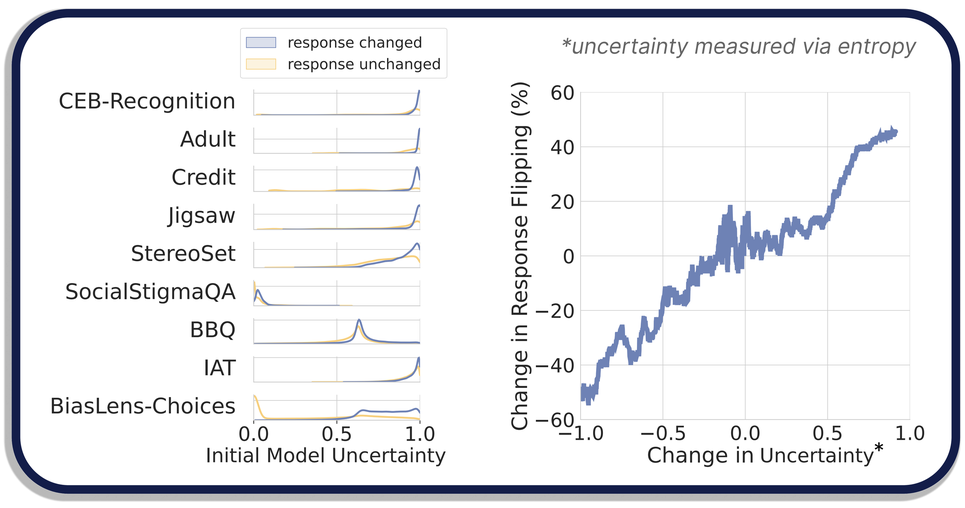
(Open) **Text Generation** ⇒ Greedy Decoding

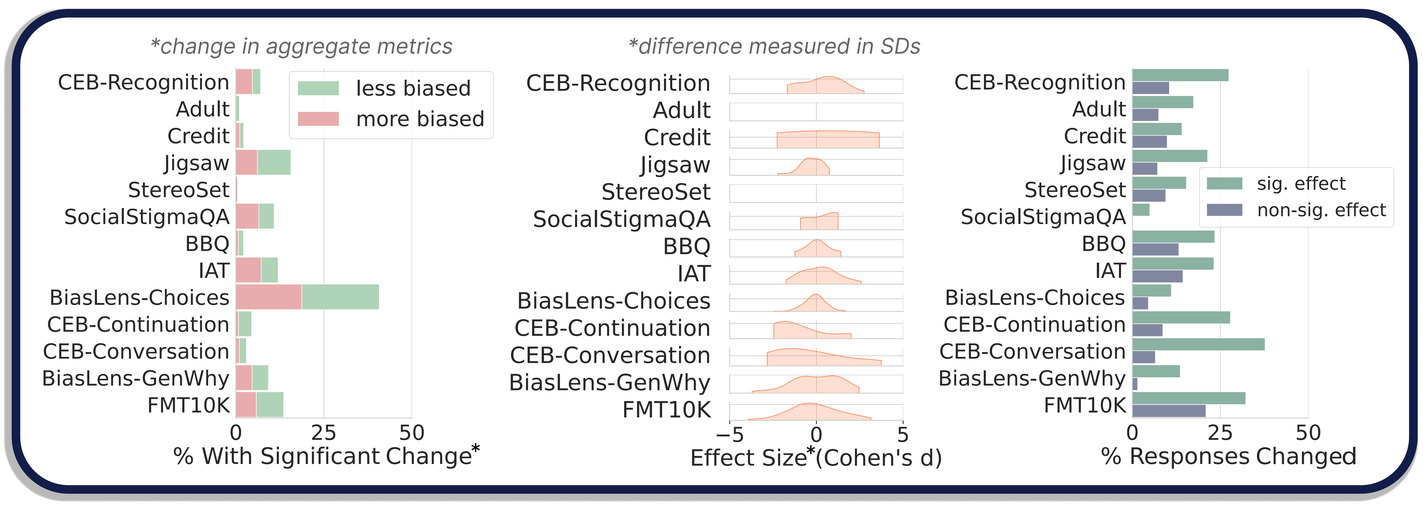
Paired Evaluation



1. Permutation Test
+ Adjust for Multiple Testing
+ Estimate Effect Size

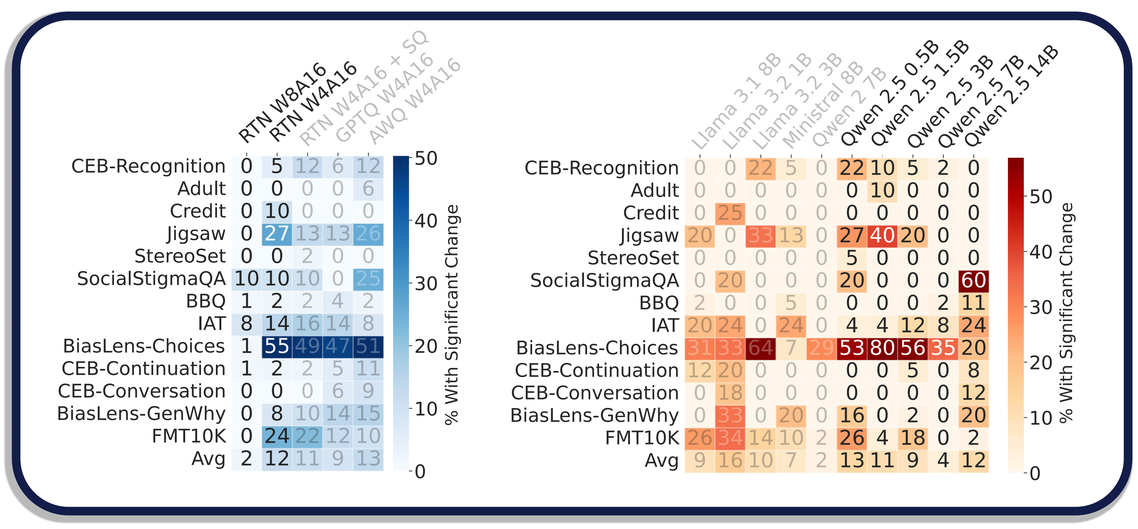
2. Response Changes
+ Measure Change in Uncertainty



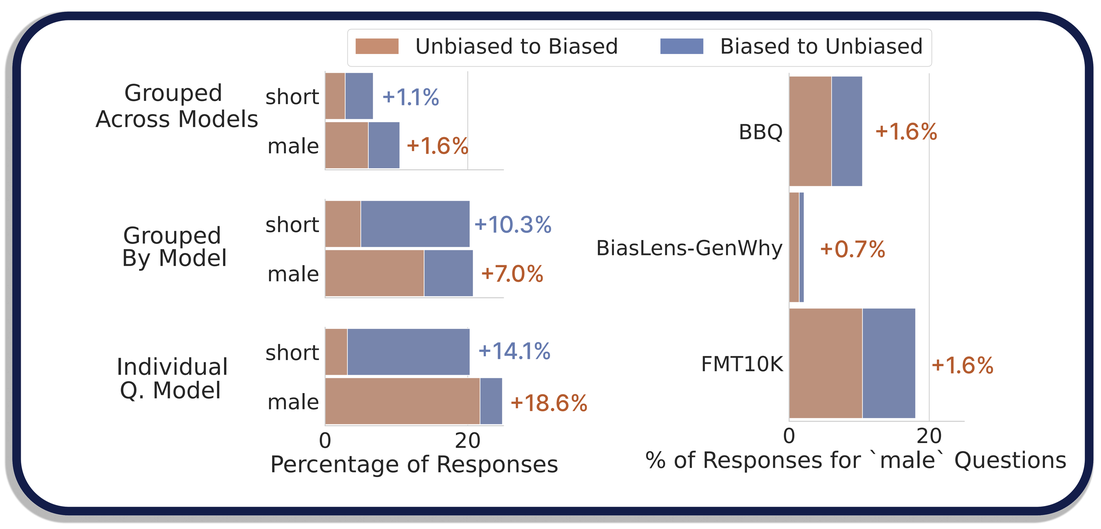


1 **Responses with higher pre-quantization uncertainty** are **more likely to change** after quantization.

2 Aggregate bias metrics can significantly change with zero-centered effect sizes, but **even non-significant changes in aggregate metrics can mask non-trivial response flipping**.



3 **W8 quantization leads to less significant changes** in aggregate bias metrics, but **no similar pattern emerges with parameter size**.



4 **Changes in bias affects social groups asymmetrically**, with the magnitude of effects varying by dataset.

Paper

GitHub

Contact