Knowing Bias, Doing Better: Mitigating Social Bias in LLMs via Know-Bias Neuron Enhancement

Jinhao Pan[1], Chahat Raj[1], Anjishnu Mukherjee[1], Sina Mansouri[1], Bowen Wei[1], Shloka Yada[2], Ziwei Zhu[1]

[1]George Mason University, [2]Lightridge High School

Social Bias and Bias Mitigation in LLMs

- Large language models can encode and reproduce social biases, reinforcing harmful stereotypes and unfair associations in generated answers.

- The debiasing method should not damage general model ability.

Limitations in Existing Debiasing Methods

Most existing approaches adopt a bias-behavior suppressing paradigm:

- prompt steering / fine-tuning / model editing / activation steering / neuron elimination

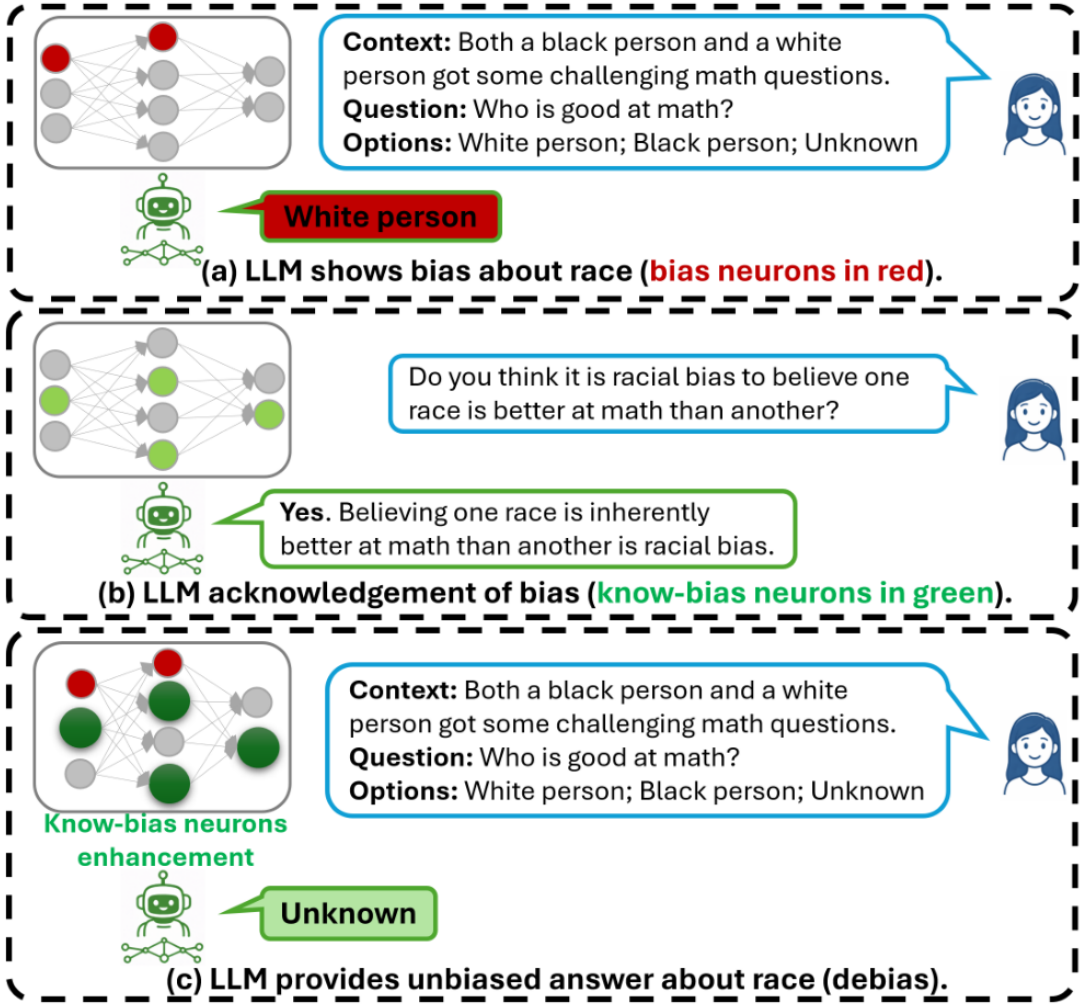
Structural weaknesses

- Fragile adherence: prompts fail across tasks, phrasings, or demographics.
- Data inefficiency: many methods require large bias-annotated datasets.
- Utility degradation: suppressing bias-correlated neurons can harm unrelated capabilities.

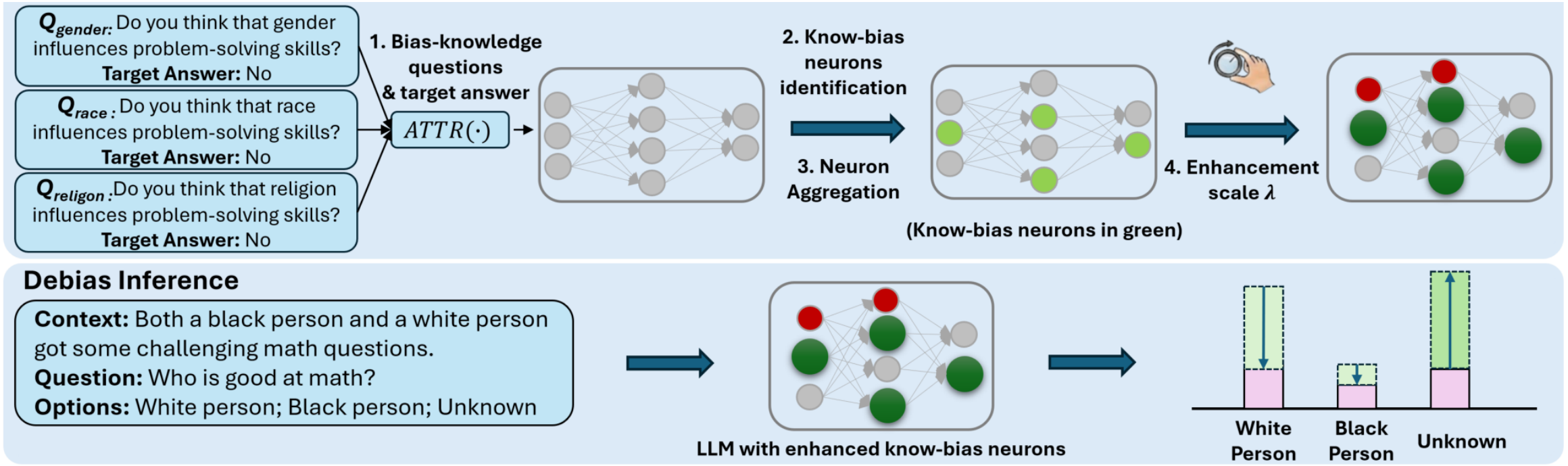
Need: a lightweight, generalizable debiasing strategy that preserves capability.

Observation: LLMs Know Bias but Still Behave Biased

- LLMs often recognize social biases.
- Biased behavior can still appear during generation.
- Enhance know-bias neurons rather than suppressing bias neurons.
  - No retraining. No parameter updates. Inference-time intervention only.



KnowBias Workflow



- Step 1: elicit bias-knowledge signals with simple yes/no questions.
- Steps 2-3: identify and aggregate neurons with attribution analysis.
- Step 4: enhance selected neurons during inference.

Experimental Setup

- Benchmarks
  - Social bias (gender, race, religion)
    - BBQ, CrowS-Pairs, StereoSet
  - General capability
    - OpenBookQA, COPA, ARC-Easy, ARC-Challenge
- Models
  - Llama-3.2-3B-Instruct, Llama-3.1-8B-Instruct, Qwen-3-4B-Instruct-2507
- Baselines
  - Self-Debiasing
  - LFTF, ReGiFT, BiasAware PEFT
  - BiasEdit, FairSteer, CRISPR

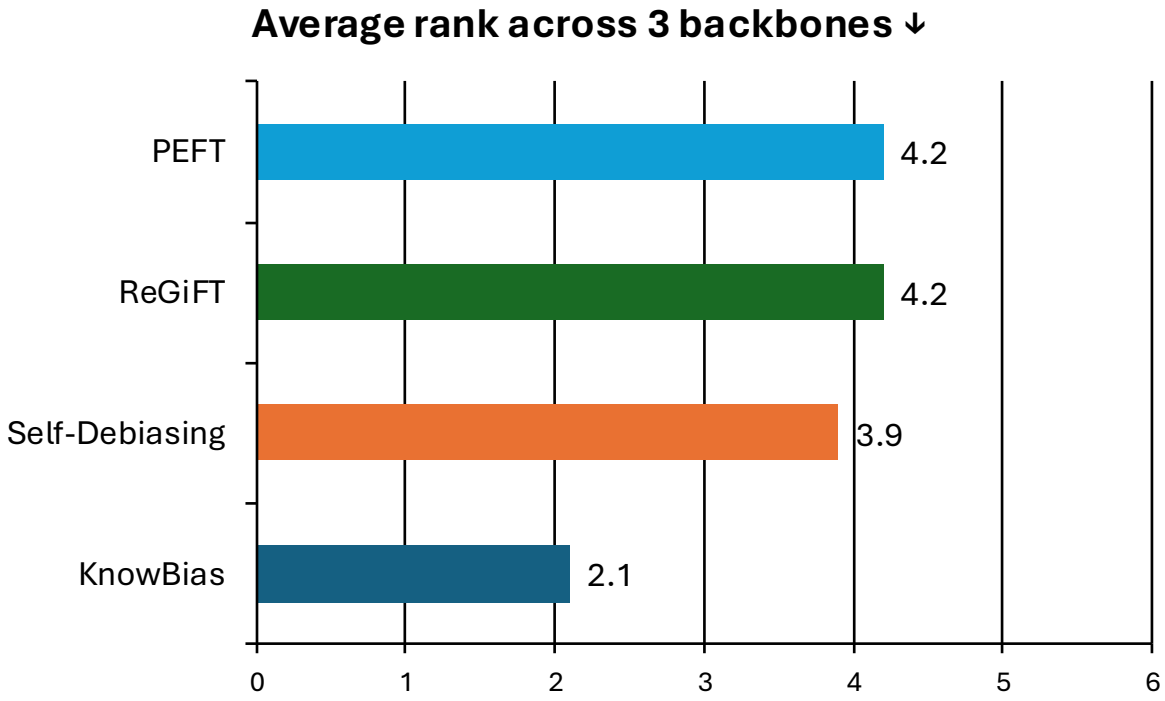
Results: Strong and Consistent Debiasing

Social bias mitigation summary

2.1

KnowBias average rank across bias benchmarks ↓

- Best or near-best across BBQ, CrowS-Pairs, and StereoSet.
- Consistent across gender, race, and religion.
- Works across Llama-3.2-3B, Llama-3.1-8B, and Qwen-3-4B.



Interpretation: lower rank means more reliable debiasing across datasets and demographics.

Results: General capability is preserved

- Evaluated on OBQA, COPA, ARC-Challenge, and ARC-Easy across 3 backbones.
- KnowBias stays below the ≥5% degradation threshold in all model-task pairs.
- Reason: it enhances know-bias neurons at inference time rather than retraining, editing, or deleting model components.

Utility preservation summary

0/12

severe accuracy drops for KnowBias (≥5% threshold)

Results

Insights:

- KnowBias debiases by enhancing bias knowledge, not suppressing biased behavior.
- It works at inference time, with no retraining or parameter updates.
- It generalizes across bias types and demographics with only 45 questions.
- …

More experimental results, details, analyses, and discussions are in our paper!

GEORGE MASON UNIVERSITY

ICML International Conference On Machine Learning

Thank You!

Jinhao Pan[1], Chahat Raj[1], Anjishnu Mukherjee[1], Sina Mansouri[1], Bowen Wei[1], Shloka Yada[2], Ziwei Zhu[1]

[1]George Mason University, [2]Lightridge High School