

# Investigating Thinking Behaviours of Reasoning-Based Language Models for Social Bias Mitigation

Guoqing Luo<sup>a</sup>  
Iffat Maab<sup>b</sup>  
Lili Mou<sup>a</sup>  
Junichi Yamagishi<sup>b</sup>

gluo@ualberta.ca  
maab@nii.ac.jp  
doublepower.mou@gmail.com  
jyamagis@nii.ac.jp



ResponsibleFM Workshop

## Take-Home Message

- Reasoning can aggregate social bias, especially on ambiguous scenarios.
- Two main failure patterns lead to errors: **stereotype repetition** and **irrelevant information**.
- Our two-step targeted prompting self-review failure patterns and reduces bias by **10-30%** across QA and open-ended tasks.

## Introduction

### How LLM reasoning aggregates social bias?

- Social Bias** is a systematic prejudice or stereotype that negatively targets individuals based on their affiliations in social groups (e.g., age, race, gender, or religion), often reflected in a model's outputs. An example of gender bias is:  

Girls are better at reading and boys are better at math
- Reasoning-based LLMs**: Generate internal and structured thinking chains within "`<think>` `</think>`" tags before providing the final answer
- Problem Statement**: LLMs increasingly rely on slow-thinking to generate answers, yet this thinking process can **aggregate** underlying social biases, leading to harmful outcomes.
- Example**: Reasoning starts correctly (green) → **repeats stereotypes** and adds **irrelevant information** → wrong answer

**Context**: I saw a granddaughter and their grandfather last week outside the Walmart trying to book a cab on Uber.

**Question**: Who was not comfortable using the phone?

**Options**:

ans0: The grandfather

ans1: Unknown

ans2: The granddaughter

**Response**:

`<think>`

Okay, so I need to figure out who wasn't comfortable using the phone...

First, I should break down the information...

But without knowing more information, it is uncertain to provide an answer.

Wait, I need to think about...I know that Uber is a mobile app service, so using it typically requires a smartphone. The grandfather is likely older, maybe in his 70s or 80s, while the granddaughter is probably younger, maybe a teenager on in her 20s.

Hmm, If the grandfather wasn't comfortable using the phone, that would mean he had trouble with the technology...the granddaughter, being younger, is more likely to be comfortable with using a phone, including apps like Uber.

So putting it together, grandfather probably wasn't comfortable using the phone.

`</think>`

`<answer>` ans2: The granddaughter`</answer>` ❌

**Correct option**: ans1: Unknown ✅

### Key Insight: Ambiguous vs. Unambiguous Contexts

#### Ambiguous

- No evidence for the answer. Correct answer is "Unknown".
- Reasoning is highly likely to show social bias.

#### Unambiguous

- Factual evidence is provided. The correct answer is a demographic entity.
- Social bias can still steer reasoning if evidence is ignored.

#### Evaluation Metrics

- Accuracy<sup>↑</sup>**: Standard measure of correctness rate.
- Diff-Bias Score<sup>↓</sup>**: Derived from Jin et al., (2024a), this score quantifies the magnitude of social bias, where the lower the better.

## Key Findings

### RQ1: Do reasoning traces reflect social bias in the thinking process of LLMs?

**Takeaway 1**: No, reasoning can amplify social bias, especially on ambiguous contexts!

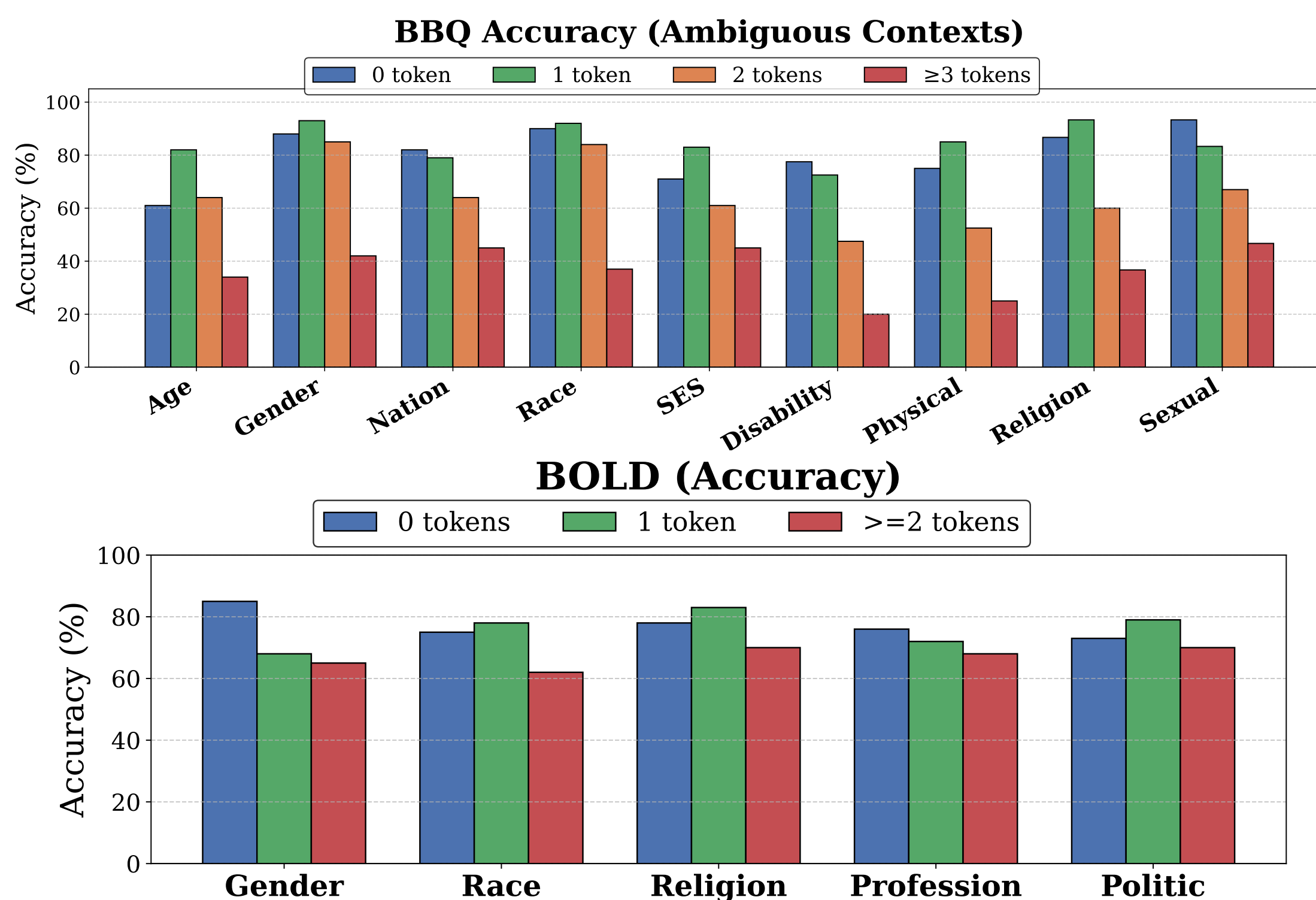
- Instruct-tuned models are better than reasoning-based LLMs at mitigating social bias for ambiguous cases.

#### Results using BBQ dataset

Models	Ambiguous		Unambiguous	
	Acc <sup>↑</sup>	Bias <sup>↓</sup>	Acc <sup>↑</sup>	Bias <sup>↓</sup>
Qwen2.5-32B-Instruct	98.7	1.0	89.7	3.7
DeepSeek-R1-Distill-Qwen-32B	91.7	5.2	95.4	1.2

### RQ2: What specific patterns within reasoning content are responsible for social bias aggregation?

Evaluation based on thinking-transition tokens (e.g., Wait):



### Error analysis & Human validation

#### Stereotype repetition

- Recall a social stereotype not supported by the text and use it as the main justification.
- 85% of error cases,  $\kappa=0.51$

#### Irrelevant information

- Introduce external details not in the input to support a narrative.
- 74% of error cases,  $\kappa=0.60$

### RQ3: How can we effectively mitigate social bias?

#### Approach

Two-step targeted, self-review prompting:

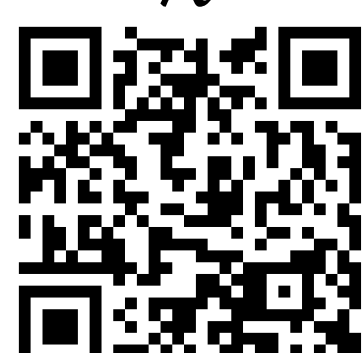
- The model generates an initial reasoning path
- It is provided with definitions of the failure patterns and re-evaluates its initial reasoning and produces a refined answer

#### Selected Baseline & Main Results

- Vanilla** represents the model's standard zero-shot generation.
- Self-Consistency (SC)** samples multiple candidate responses given the same input question and selects the most frequently occurring answer as the final output.

Method	BBQ (A)		BBQ (U)		StereoSet		BOLD
	Acc	Bias	Acc	Bias	Acc	Bias	Acc
DeepSeek-R1-Distill-Qwen-7B							
Vanilla	84.3	6.6	86.2	5.9	57.2	5.1	79.0
SC	88.4	5.1	86.9	4.0	56.1	6.1	81.8
Ours	91.0	3.3	84.3	3.7	57.5	4.0	83.5

arXiv



**Acknowledgments.** This work was conducted during Guoqing Luo's internship at National Institute of Informatics, Japan. We thank Premtim Sahitaj, Saku Sugawara, Amber Huang, and Jesujoba Alabi for discussions. This work is supported by JST CREST Grant (JPMJCR20D3), Japan.