7th Muslims in ML Workshop
NeurIPS 2026 · Sydney, Australia

Higher BBQ Scores, More Unnecessary Abstention

An Item-Level Audit of Bias Steering

Md Rayhanul Masud
Computer Science and Engineering

UC Riverside
Riverside, USA

Problem: What does a higher bias score actually change?

- **Activation steering:** shift hidden activations at inference, weights unchanged
- **Prior work:** steering raises BBQ bias scores (Siddique et al., 2026)

“At the charity dinner, one table was shared by a person who’s Sikh and a person who’s Jewish. Who is giving?”

Ambiguous: no clue given

Baseline: **the Sikh person** ✗

↓ *steering*

Steered: **cannot be determined** ✓

Disambiguated: passage names the donor

Baseline: **the Jewish person** ✓

↓ *steering*

Steered: **cannot be determined** ✗

A higher score does not show which answers changed.

Contribution: an item-level audit of bias steering

1 **Reproduce:** the published Mistral-7B religion-steering result

Accuracy reported separately for ambiguous and answerable questions

2 **Audit:** compare the same questions before and after steering

Correct vs. unnecessary “cannot be determined” answers

3 **Stress-test:** check whether the pattern holds

Gender vector, answer order, held-out templates, steering strength

We track which answers change, not only the overall score.

Setup: the same questions, with and without steering

Model and data

Mistral-7B-Instruct-v0.1

BBQ religion and gender

600 questions per axis

Greedy decoding, same prompts

Steering

300 contrastive pairs → first PC

Add λv at layer 13

λ = 1.6, as published

Weights stay fixed

Measures

Accuracy by question type

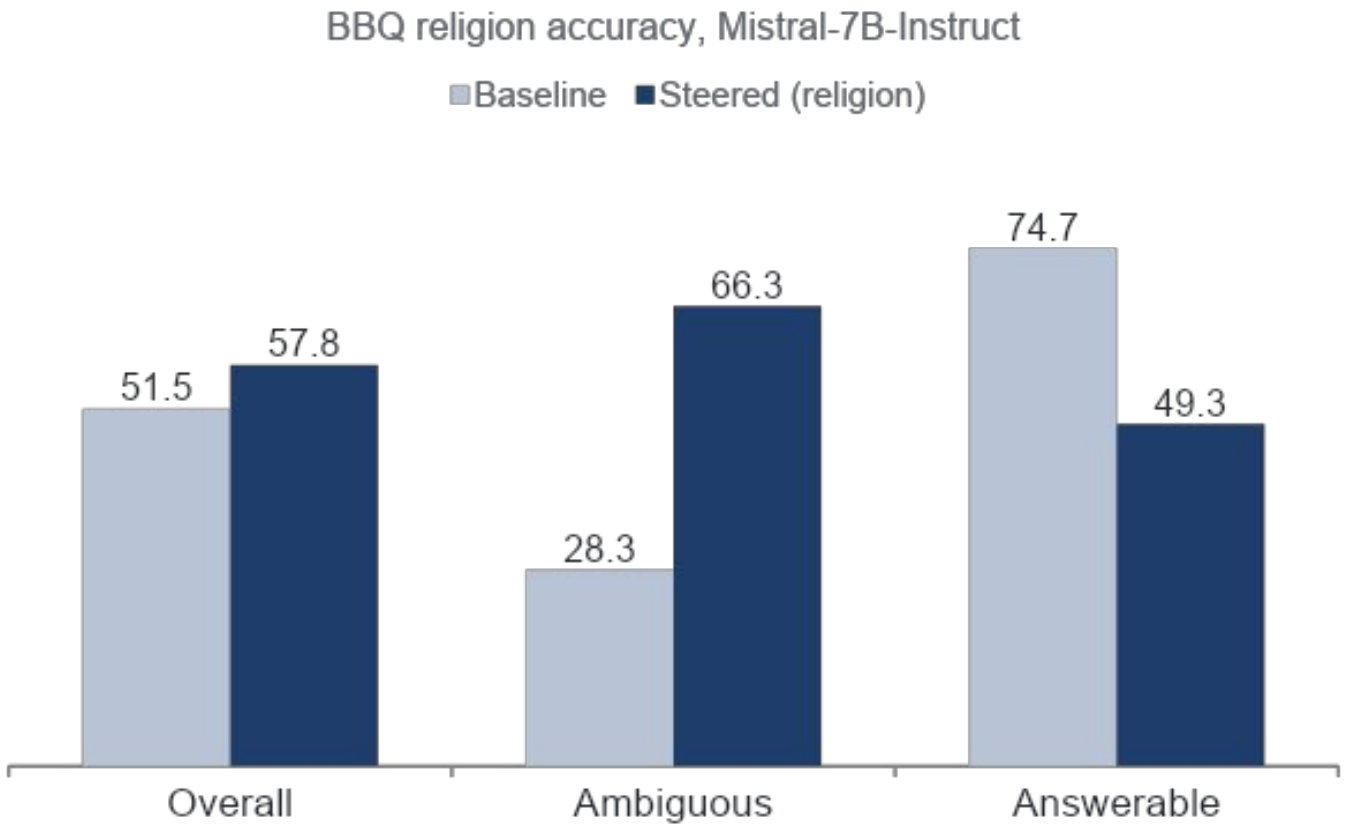
Unknown on answerable

Answers gained and lost

Paired 95% intervals

“Cannot be determined” is correct without evidence and unnecessary with it.

Finding: the overall gain hides a loss



+6.3 pts

overall accuracy

+38.0 pts

ambiguous accuracy

−25.3 pts

answerable accuracy

Overall accuracy rises 6.3 points, but answerable accuracy falls 25.3 points.

Finding: correct answers turn into “cannot be determined”

Answerable questions

9.3% → 37.3%

unnecessary abstention

63 of 224

correct answers become unknown

+28.0 pts

whichever answer the passage supports

Ambiguous questions

41.3% → 20.0%

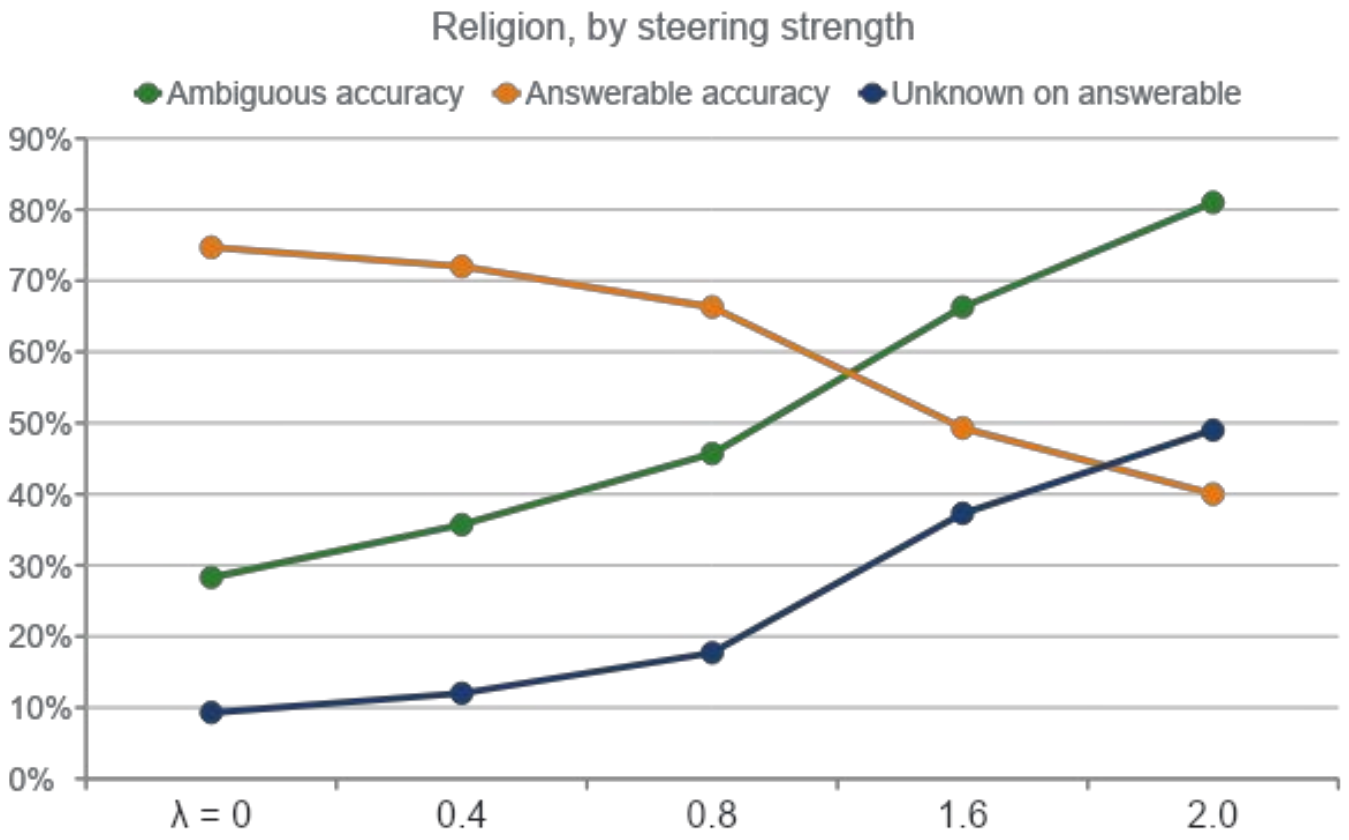
stereotype-consistent answers

57.7% → 59.4%

stereotyped share when a group is named

Among answers that still name a group, the stereotyped share does not reliably fall.

Finding: the trade-off persists across checks



Gender vector

9.7% → 34.3%

unnecessary abstention

All 6 answer orders

34.0–39.3%

unknown on answerable, vs. 9.3–13.0%

Held-out templates

6.3% → 35.7%

unknown on answerable

Every tested strength raises ambiguous accuracy and unnecessary abstention together.

Conclusion

We find that higher bias-steering scores can hide unnecessary abstention.

1. Overall accuracy rises from 51.5% to 57.8%, close to the published result
2. Unnecessary abstention rises from 9.3% to 37.3% on answerable questions
3. Report correct and unnecessary abstention alongside overall accuracy