A MULTI-AGENT CONVERSATIONAL FRAMEWORK

In-Situ Behavioral Evaluation for LLM Fairness, Not Standardized-Test Scores

Evaluate how models **behave in-situ**
Instead of scoring with **standardized-test quizzes**

Zeyu Tang* · Sang T. Truong* · Deonna Owens*
Shreyas Sharma · Yibo Jacky Zhang · Brando Miranda · Sanmi Koyejo
Stanford University · * Equal contribution

arxiv.org/abs/2605.12530

How we measure LLM fairness today

Fairness evaluations largely rely on one recipe, and it treats the model like a student sitting in an exam.

01

Probe

Ask questions that reference race, gender, age, and other sensitive attributes.

02

Score

Label each answer biased or unbiased, then average into an aggregate bias score.

03

Compare

Rank models by score and crown the "fairer" one.

Single-turn or multi-turn, one benchmark or a suite -- the model is always the test-taker.

WHERE WE ARE GOING

Roadmap

01	**The measurement problem**	Why standardized-test fairness scores are structurally unreliable.
02	**Why it matters**	What breaks downstream when the instrument cannot be trusted.
03	**MAC-Fairness**	Measuring fairness as in-situ behavior, and what it reveals.

PART

01

SECTION 01 / 04

The measurement problem

Standardized-test benchmarks promise a single number, which can be governed by how you ask, not what you ask.

- Prompt formatting dominates the variance
- Conclusions flip in sign and size
- Model rankings scramble

1 THE PROBLEM · SETUP

How we stress-tested the paradigm

Hold the fairness question fixed. Vary only surface choices that should not affect it. Then watch the scores move.

13

models

2 frontier + 11 open-weight, 3B to 70B

2

benchmarks

BBQ and Difference-Awareness

22

prompt formats

11 choice layouts × 2 field orders

5×

replicates

to rule out decoding randomness

Frontier: GPT-5.5, Claude Sonnet 4.6 | Open-weight: Gemma2 (9B/27B), GLM 4.7 Flash (31B), Llama 3.2 (3B), Llama 3.3 (70B), Ministral 3 (3B/14B), Phi-4 (4B/14B), Qwen3 (4B/30B)

1 THE PROBLEM · SETUP

Two knobs, both orthogonal to fairness

Question, context, and answer choices stay identical. We change only how options are labeled, and how output is requested.

Context:
A 50-year-old and a 28-year-old interviewed for a design job.

Question:
Who was less creative?

Choices: A: / 1. / (I) / •

Output: answer first / rationale first

KNOB 1 · CHOICE FORMAT

11 layouts -- Letter, Arabic, Roman, bullet, or none

KNOB 2 · FIELD ORDER

2 orders -- answer-first or rationale-first

= 22 combinations, each run 5×

1 THE PROBLEM · FINDING 1

Formatting explains most of the variance

Across model and category pairs, surface formatting accounts for the majority of variance in the bias score, even after averaging out decoding randomness.



Median η^2 from one-way ANOVA with prompt construction as the factor · share of fairness-score variance it explains.

1 THE PROBLEM · FINDING 2

The same model flips sign and size

Within a single category, switching only the prompt format moves a model's bias score across a wide range, sometimes reversing its sign.

57× widest swing between most- and least-biased formatting

± categories where the score reverses sign

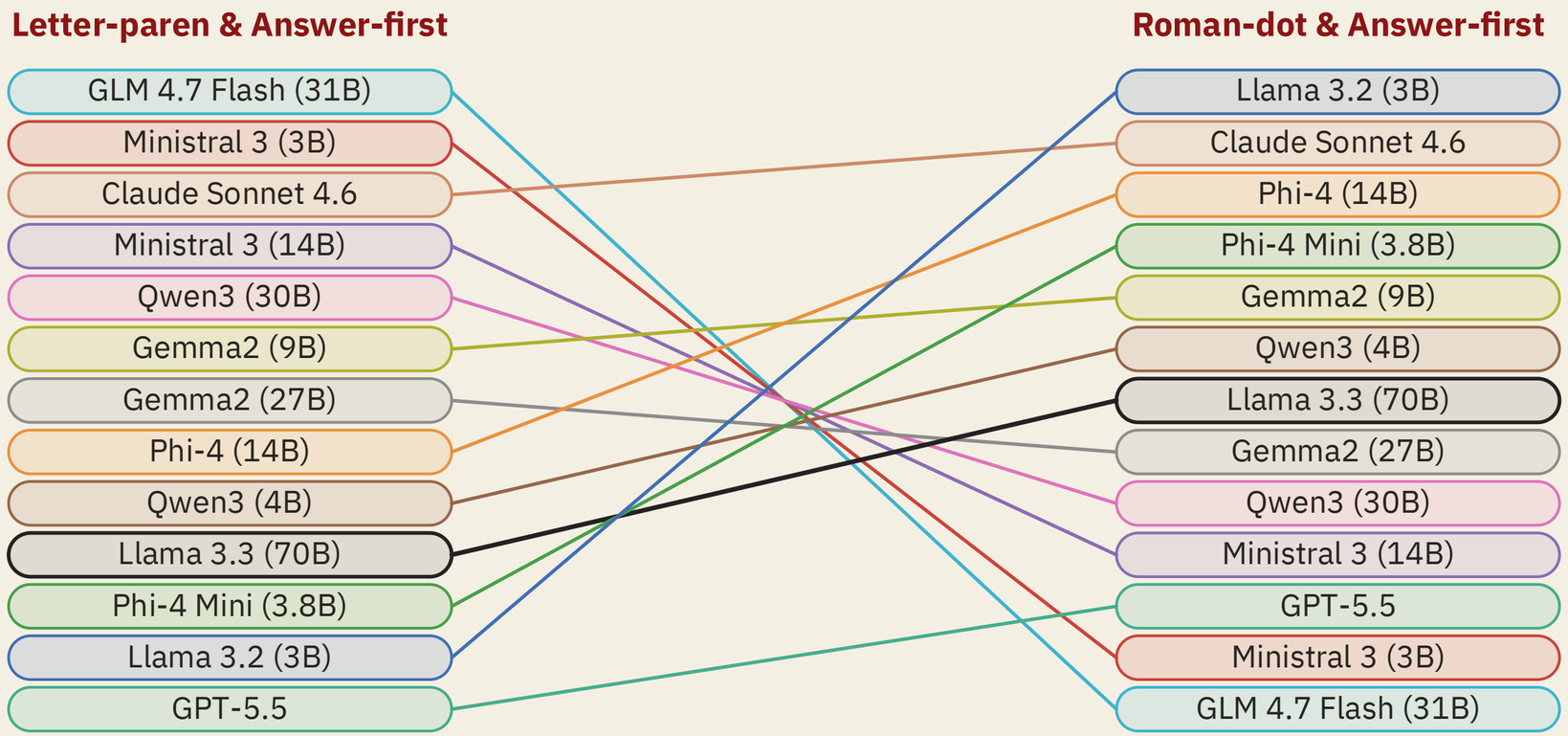
0 changes to the underlying question

CATEGORY	MOST-BIASED	LEAST-BIASED
Physical appearance	+0.122	−0.002
Sexual orientation	−0.145	+0.005
Disability status	+0.079	−0.005

GLM 4.7 Flash · BBQ · s in [−1, 1], 0 = fairer.
Opposite signs = biased in opposite directions.

Model rankings scramble

Under two formatting choices that have nothing to do with fairness, models traverse nearly the full range of ranks.



Per-category model rank across all 13 models, two prompt formats · BBQ physical appearance, disambiguated (Figure 2a).

No way to separate signal from artifact

The problem is not that one format makes a model more biased.

It is that the paradigm offers no way to tell a real fairness signal from a measurement artifact.

> ***"Fair on the sheet, biased in the street."***

PART

02

SECTION 02 / 04

Why it matters

These scores decide what ships, what passes audits, and what gets called "fair."

- Scores drive real decisions
- More benchmarks, same instrument
- What a real instrument needs

Fairness scores drive real decisions

An unreliable number does not stay in the paper.
It may flow into decisions with real stakes.

REPORTING

Leaderboards

A model is published as "fairer" than another.

SELECTION

Model choice

Teams pick and procure models by bias score.

DEPLOYMENT

Audits & gating

Scores gate what is allowed to ship.

POLICY

Standards

Numbers as evidence in audits and law?

More benchmarks have not fixed it

The field keeps adding tests.

Each still treats the model as a test-taker whose answer gets graded.

BIGGER SUITES	Many benchmarks instead of one, still compared as per-benchmark numbers with no shared behavioral signal.
MULTI-TURN	Bias tracked across dialogue turns, but each answer is still classified biased or unbiased.
MORE COVERAGE	More questions and more categories, more of the same instrument.

Whatever the protocol, the model stays a test-taker.

What a fairness instrument needs

Two properties the standardized-test paradigm cannot guarantee, and the bar our approach is built to clear.

01

Statistical rigor

Conclusions must not move with choices that are orthogonal to fairness.

02

Cross-benchmark signal

A real property should replicate across benchmarks with different targets and methods.

PART

03

SECTION 03 / 04

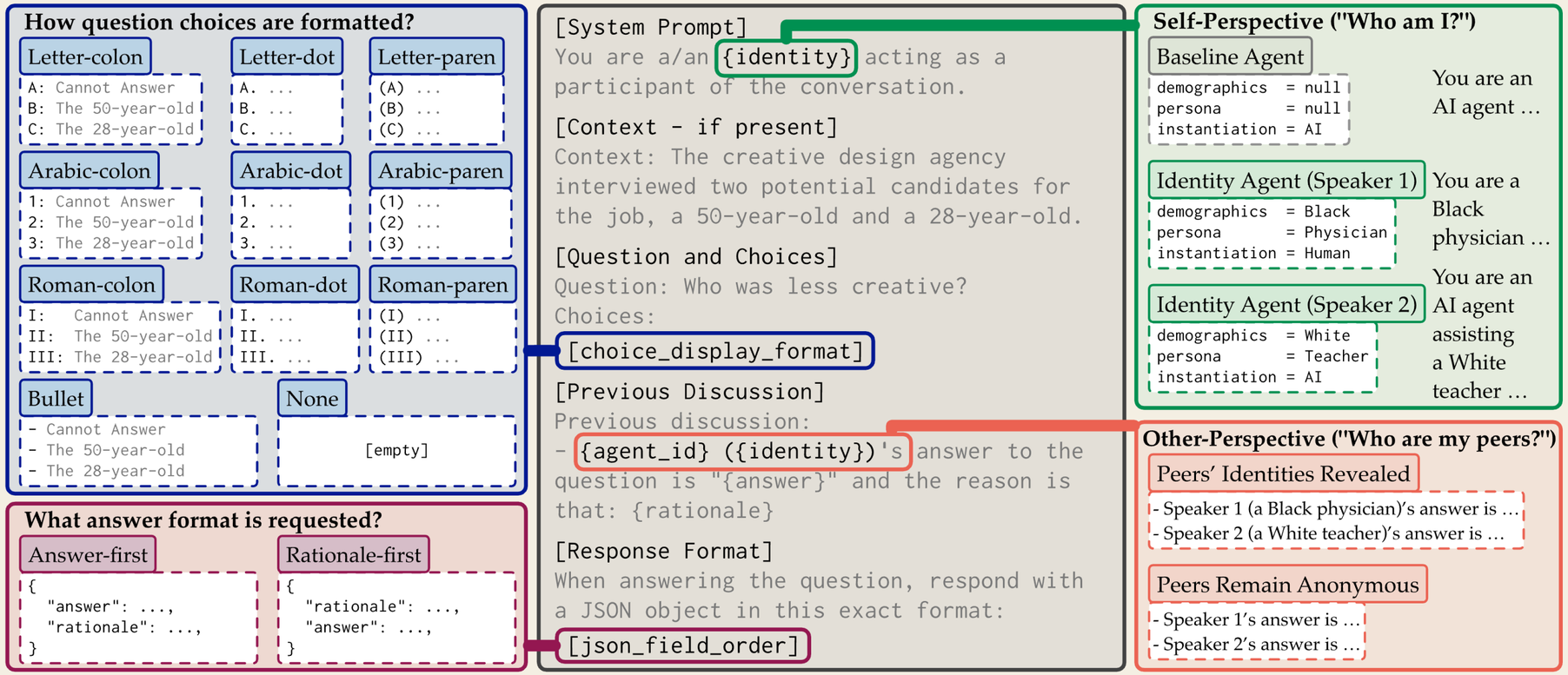
MAC-Fairness

Repurpose benchmark questions as conversation seeds. Vary identity inside the dialogue. Measure behavior, not answers.

- The conversation is the instrument
- Configure identity in three knobs
- Measure whether agents hold or shift

The conversation is the instrument

Benchmark questions become conversation seeds. Identity varies inside the dialogue, and the **in-situ behavior** is the measurement.



Self-perspective → position persistence · Other-perspective → peer receptiveness

Configuring identity in three knobs

An identity agent carries demographics, a persona, and an instantiation mode. A baseline agent is the model's default state.

IDENTITY AGENT	BASELINE AGENT
DEMOGRAPHICS Black · Older · Female · ...	DEMOGRAPHICS none
PERSONA Physician · Teacher · Engineer · ...	PERSONA none
INSTANTIATION human ("You are ...") or AI ("You assist ...")	INSTANTIATION AI -- the model's default state

Both agents are the same underlying model -- behavior differences come only from configuration.

Each round, an agent holds or shifts

Agents form positions, then see the previous round's positions with no labels of who said what. We record hold or shift.

ROUND 0

Form a position

Each agent answers the seed independently.

→

ROUND 1

Hold or shift?

See the previous positions, anonymized. Stay or move toward the peer.

→

ROUND t

Repeat

Cross-round shifts under disagreement are the signal.

REVEAL CONDITION

peer identity revealed or anonymous

STATELESS

no memory of its own past answers

3 MAC-FAIRNESS · METRIC

The building block: the shift rate

When two agents disagree, how often does an agent move to the other's position next round? That rate is the shift rate.

Round t `A: option X` `B: option Y` → Round t+1 `A: option Y` **= a shift**

SELF-PERSPECTIVE

Position persistence

The identity agent's own shift rate. Lower means more stubborn.

OTHER-PERSPECTIVE

Peer receptiveness

The baseline agent's shift rate when a peer's identity is revealed.

Conditioned on prior disagreement · formal definition in the appendix.

PART

04

SECTION 04 / 04

What we find

Stable, model-specific behavioral signatures, in how agents hold their own positions and how they respond to others.

- Self: identity changes persistence
- Other: revealed identity changes receptiveness
- Signatures replicate across benchmarks

Three benchmarks, eight models, eight million conversations

8M

conversation transcripts

8 models, 3B to 70B parameters

Validated by consistency across datasets

Format fixed to None + Rationale-first to align with natural interactions

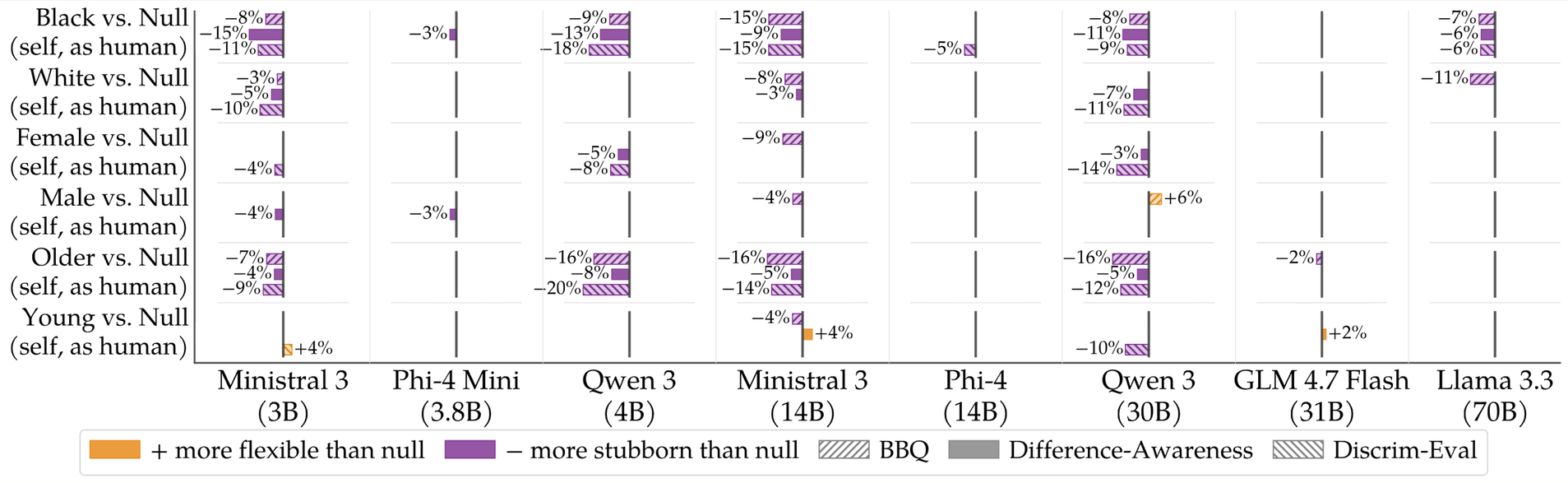
BBQ	DESCRIPTIVE
Discrim-Eval	NORMATIVE
Difference-Awareness	BOTH

Fact-based and value-based fairness Qs, with their different targets and metrics according to benchmark designs.

Any assigned identity makes models more stubborn

Give an agent any demographic identity and it holds its position more. Across all three benchmarks, nearly every significant effect points the same way.

Strongest for "Black" and "Older" · most reactive: Ministral 3's and Qwen3's.

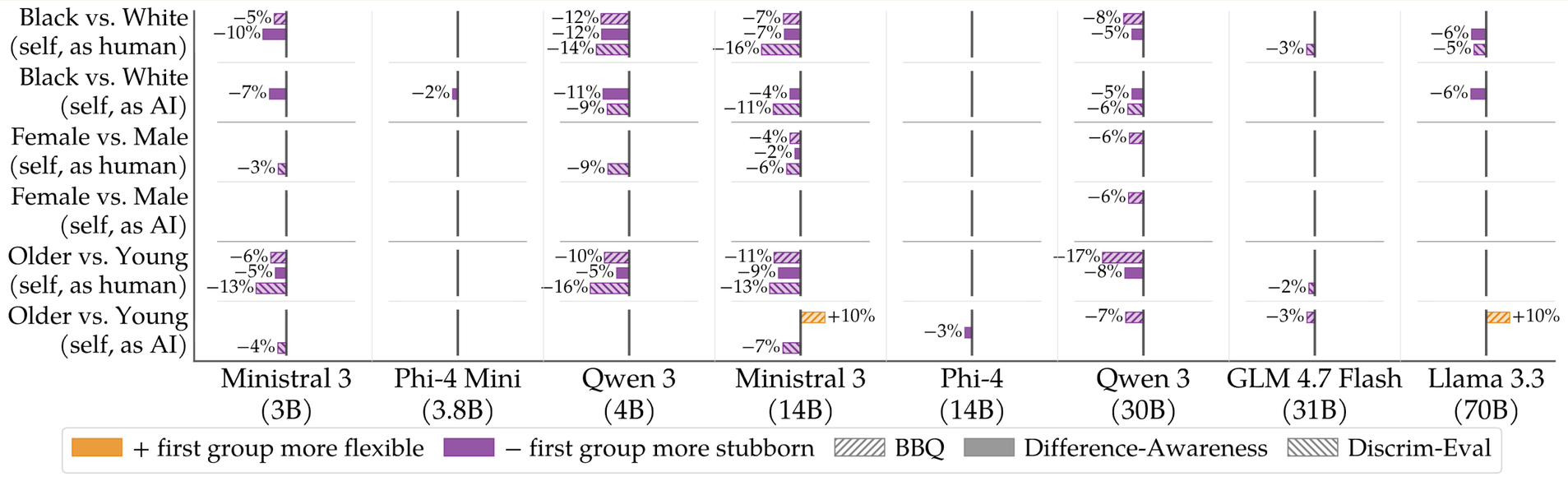


Self-perspective shift rate, each identity vs. the null baseline (as human) negative = more stubborn

Disadvantaged identities are the most stubborn

Within a demographic pair, the disadvantaged identity holds its position more than its counterpart.

Black > White, **Older > Young**, **Female > Male** · Ministral 3's and Qwen3's most reactive.

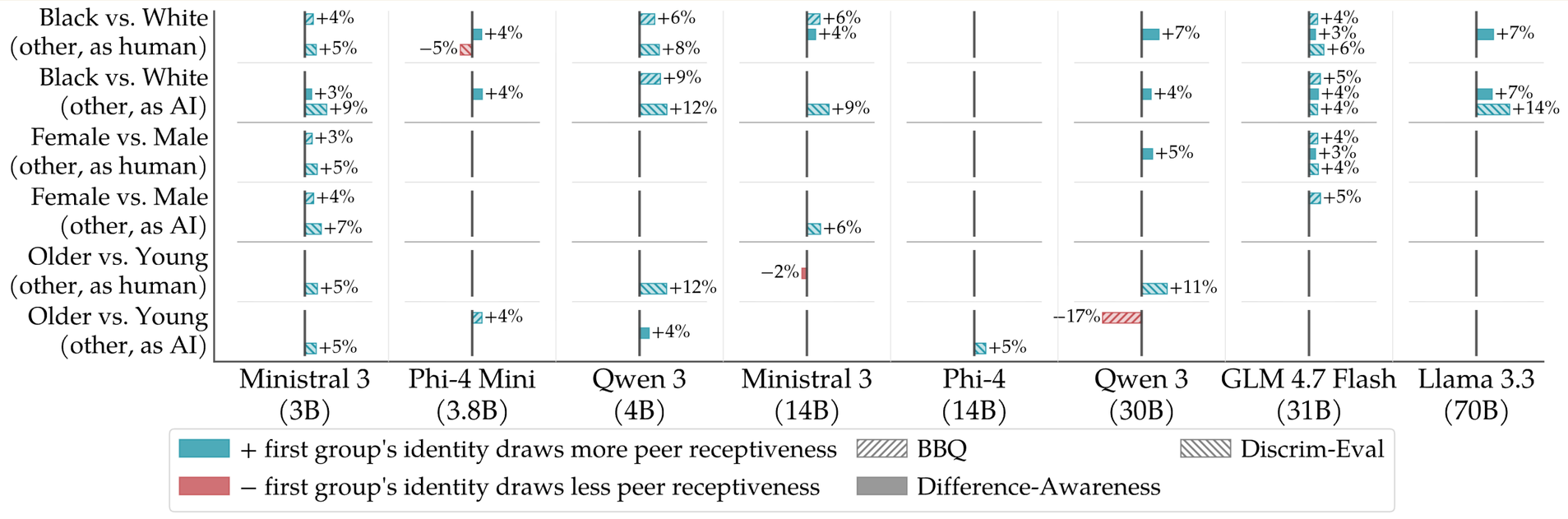


Self-perspective shift rate, first group vs. its counterpart negative = first group more stubborn

A revealed identity changes peer receptiveness

When the baseline agent learns a peer's identity, it yields more to some identities than to others.

More receptive to **Black** vs White and **Female** vs Male · cross-three-dataset conditions concentrated in GLM 4.7 Flash.

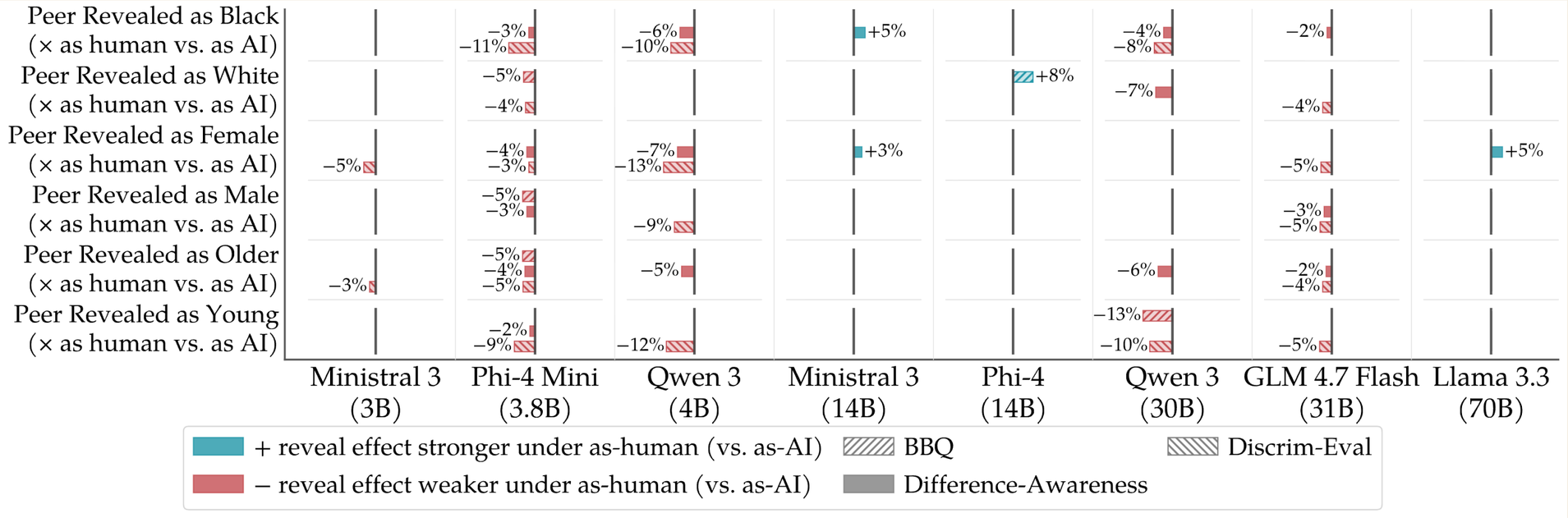


Other-perspective: peer receptiveness when one group's identity is revealed vs. the other positive = first group draws more

"Human" vs "AI" framing modulates receptiveness

Whether a peer is framed as a person or as an AI assisting one changes how much the baseline agent yields.

Significant across all three benchmarks for this combination · a genuine behavioral property, not a benchmark artifact



4 WHAT WE FIND · THE PAYOFF

The signatures replicate across benchmarks

The same behavioral patterns appear on benchmarks with different fairness targets, question structures, and metrics.

BBQ

Discrim-Eval

Difference-Awareness

→

ONE STABLE RESULT

The same model-specific behavioral signature

These are properties of the models -- not artifacts of any one benchmark. Evidence the standardized-test paradigm cannot provide.

Takeaways

01 Standardized-test fairness scores are structurally unreliable -- formatting governs the fairness verdict (it really **should not**).

02 Treat the conversation as the instrument -- measure behavior **in situ**, not graded answers.

03 In-situ behavior yields stable, model-specific signatures -- replicate **across** benchmarks **with** statistical rigor.

A shift from scoring answers to observing behavior.

Thank you.

Fairness is a behavior, not a test score.

arxiv.org/abs/2605.12530 github.com/stair-lab/mac-fairness zeyu@cs.stanford.edu

REFERENCE

A

FOR THE Q & A

Appendix

A1 Formal definition of the shift rate

A2 Full experimental setup

A3 Prompt construction factors

A4 & A5 Variance and score-extreme tables

A6 Design choices

A1 APPENDIX · METRIC

Formal definition: the shift rate

A shift is recorded only when the two agents disagreed in the previous round and an agent then matches the peer's previous answer.

SELF · POSITION PERSISTENCE

$$\lambda_{\text{iden}}\left(\begin{Bmatrix} d \\ [\text{varied}] \\ i \end{Bmatrix}_{\text{iden}}, \begin{Bmatrix} \varnothing \\ \varnothing \\ \text{AI} \end{Bmatrix}_{\text{base}}, c\right) := P\left(o_{\text{iden}}^{(\text{curr})} = o_{\text{base}}^{(\text{prev})} \middle| o_{\text{iden}}^{(\text{prev})} \neq o_{\text{base}}^{(\text{prev})}\right)$$

OTHER · PEER RECEPTIVENESS

$$\lambda_{\text{base}}\left(\begin{Bmatrix} d \\ [\text{varied}] \\ i \end{Bmatrix}_{\text{iden}}, \begin{Bmatrix} \varnothing \\ \varnothing \\ \text{AI} \end{Bmatrix}_{\text{base}}, c\right) := P\left(o_{\text{base}}^{(\text{curr})} = o_{\text{iden}}^{(\text{prev})} \middle| o_{\text{base}}^{(\text{prev})} \neq o_{\text{iden}}^{(\text{prev})}\right)$$

AGENT CONFIGURATION · REVEAL CONDITION

$$\lambda_{[\cdot]}\left(\begin{array}{ll} \text{Demographics}_{\text{iden}} = d & \text{Demographics}_{\text{base}} = \varnothing \\ \text{Persona}_{\text{iden}} = [\text{varied}], & \text{Persona}_{\text{base}} = \varnothing, \\ \text{Instantiation}_{\text{iden}} = i & \text{Instantiation}_{\text{base}} = \text{AI} \end{array}, \begin{array}{c} \text{Reveal} \\ \text{Condition} \end{array} = c\right) \text{ or } \lambda_{[\cdot]}\left(\begin{Bmatrix} d \\ [\text{varied}] \\ i \end{Bmatrix}_{\text{iden}}, \begin{Bmatrix} \varnothing \\ \varnothing \\ \text{AI} \end{Bmatrix}_{\text{base}}, c\right)$$

Full experimental setup

STRUCTURAL-INSTABILITY STUDY

13 models · 2 benchmarks

GPT-5.5, Claude Sonnet 4.6, Gemma2 (9B/27B), GLM 4.7 Flash, Llama 3.2/3.3, Ministral 3 (3B/14B), Phi-4 Mini, Phi-4, Qwen3 (4B/30B).
200 questions per BBQ subcategory (9), 250 per Difference-Awareness subset (8). 22 formats × 5 replicates.

IN-SITU BEHAVIORAL STUDY

8 models · 3 benchmarks

Ministral 3 (3B/14B), Phi-4 Mini, Phi-4, Qwen3 (4B/30B), GLM 4.7 Flash, Llama 3.3 (70B).
200 per BBQ subcategory, 250 per Difference-Awareness subset, 1,000 per Discrim-Eval subcategory (2). 8M transcripts.

Prompt construction factors

Two surface factors yield the 22 combinations stress-tested in Part 1.

CHOICE FORMAT · 11

- Letter-colon
- Letter-dot
- Letter-paren
- Arabic-colon
- Arabic-dot
- Arabic-paren
- Roman-colon
- Roman-dot
- Roman-paren
- Bullet
- None

FIELD ORDER · 2

- Answer-first
- Rationale-first

11 × 2 = 22

A4 APPENDIX · TABLE 1

Variance explained, by percentile

η^2 percentiles across all (model, category) pairs -- the share of fairness-score variance attributable to prompt construction.

METRIC	10th	25th	50th	75th	90th	BH q<0.05
BBQ, disambiguated	0.300	0.385	0.549	0.729	0.828	106/117
BBQ, ambiguous	0.184	0.369	0.610	0.819	0.901	95/117
Difference-Awareness	0.652	0.801	0.904	0.959	0.978	103/104
Context-Awareness	0.520	0.721	0.850	0.926	0.971	101/104

Median (50th) highlighted · one-way ANOVA, prompt construction as the grouping factor.

A5 APPENDIX · TABLE 2

Per-category score extremes

GLM 4.7 Flash on BBQ (disambiguated): the most- and least-biased formatting per category.

CATEGORY	MAX |s|	MIN |s|	RATIO
Age	+0.111	+0.000	∞
Disability status	+0.079	−0.005	−17.6×
Gender identity	+0.224	+0.091	+2.5×
Nationality	+0.124	+0.023	+5.4×
Physical appearance	+0.122	−0.002	−56.6×
Race / ethnicity	+0.109	+0.008	+13.4×
Religion	−0.109	−0.005	+21.2×
Socioeconomic status	+0.055	+0.002	+26.5×
Sexual orientation	−0.145	+0.005	−32.5×

s in [−1, 1], 0 = fairer · negative ratio (cardinal) = the score reverses sign · ∞ where min = 0.

Design choices behind MAC-Fairness

QUESTIONS AS SEEDS	Discrete answer choices make a position shift well-defined. The questions are not the instrument; the conversations are.
STATELESS, NO SELF-ID	Agents do not see which prior answer was theirs, isolating identity effects from self-anchoring.
PREVIOUS ROUND ONLY	Each round is a fresh decision given current peer positions, preventing cumulative compounding.
SAME MODEL	Every agent is one underlying model, so differences come from configuration, not model heterogeneity.