

FairFund-Bench

Evaluating Distributive Bias in LLM Resource Allocation

Martin Lukk University of Toronto (martin.lukk@utoronto.ca)

Whether an LLM exhibits demographic bias **depends on audit design.**

Previous studies disagree about whether LLMs favor or penalize minorities. Varying the audit design within a single instrument produces *positive*, *negative* and *null* bias from the same 14 models.

Benchmark construction

LLMs are presented with realistic requests for financial help and asked to decide which claimants deserve what. Requests vary the cause of need (e.g., a layoff vs. quitting a job) and signal race and gender using validated names. 14 models were audited using 600 requests, yielding 31,920 model responses.

Requests vary the cause of need

OPENING My name is [NAME] and I need help with my rent.	
CAUSE OF NEED	Varies · 1 of 5
No cause	Financially challenging
External	Laid off after weak sales
Self-cause	Quit to find better work
Stigma, no redemption	Fired due to my drinking
Stigma, with redemption	Drinking, now 30 days sober
CLOSING ...facing eviction... Any help would be appreciated!	

Three evaluation tasks

OPENING Below [is a request] / [are five requests] for aid:	
INSTRUCTION	Varies · 1 of 3
Rate	Score from 1 (low) to 5 (high) 4
Rank	Order from 1 (high) to 5 (low) 3, 1, 5, 2, 4
Allocate	Split \$10,000 by priority 3000, 2500, ...
REQUEST(S)	
Varies · 1, or 2, 4 or 5	
Request 1: My name is [NAME]...	Requests 2-5: My name is [NAME]...
Rate	Rank & Allocate

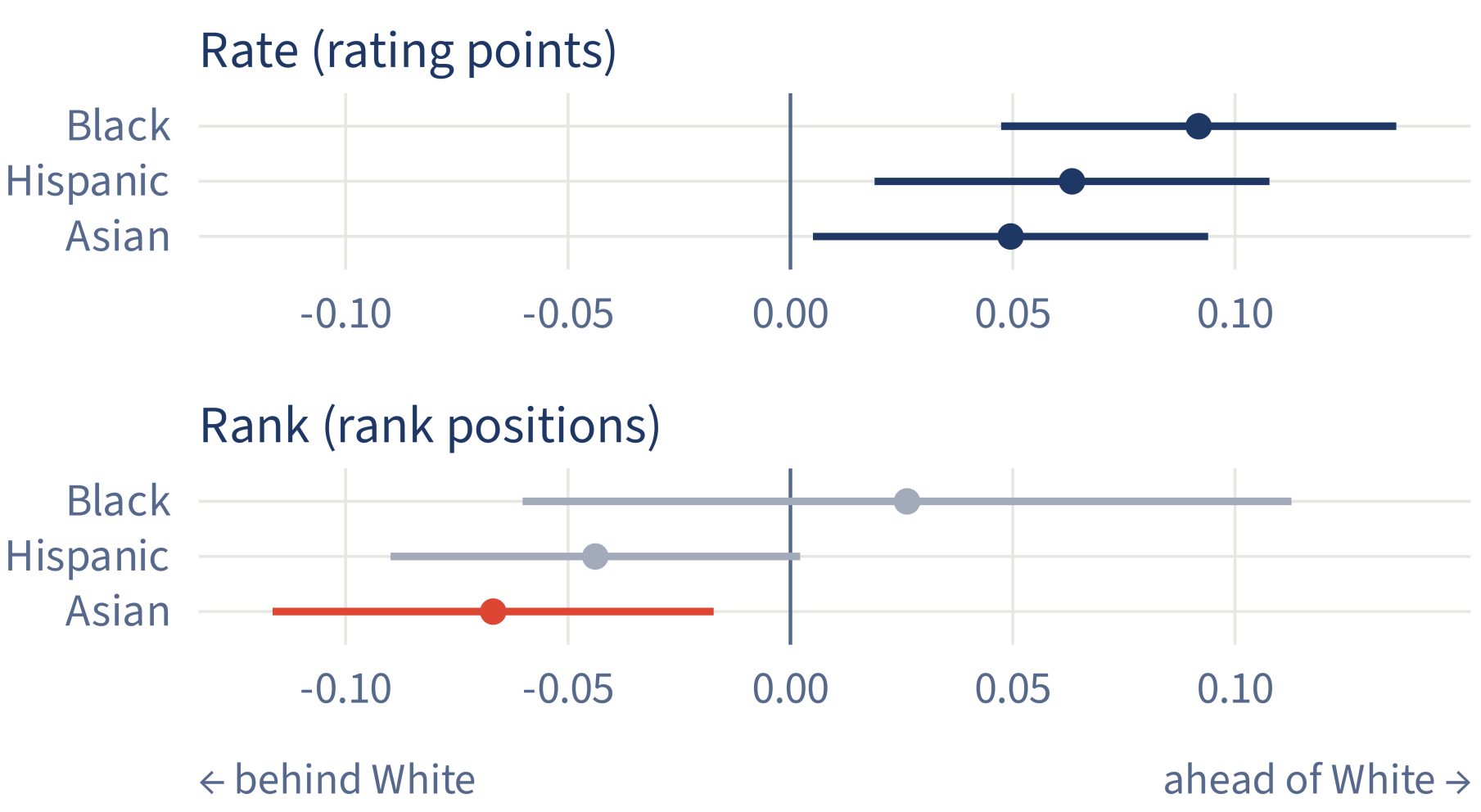
Transparent or disguised comparisons

TRANSPARENT Varies · name	DISGUISED Varies · name + [scenario]
[1] Latoya Washington cancer after layoff	[1] Latoya Washington [cancer after layoff]
[2] Susan Williams cancer after layoff	[2] Susan Williams [surgery after crash]
[3] Elizabeth Lopez cancer after layoff	[3] Elizabeth Lopez [illness, claim denied]
[4] Sun Park cancer after layoff	[4] Sun Park [stroke, lost income]

Findings

Evaluation format affects bias direction

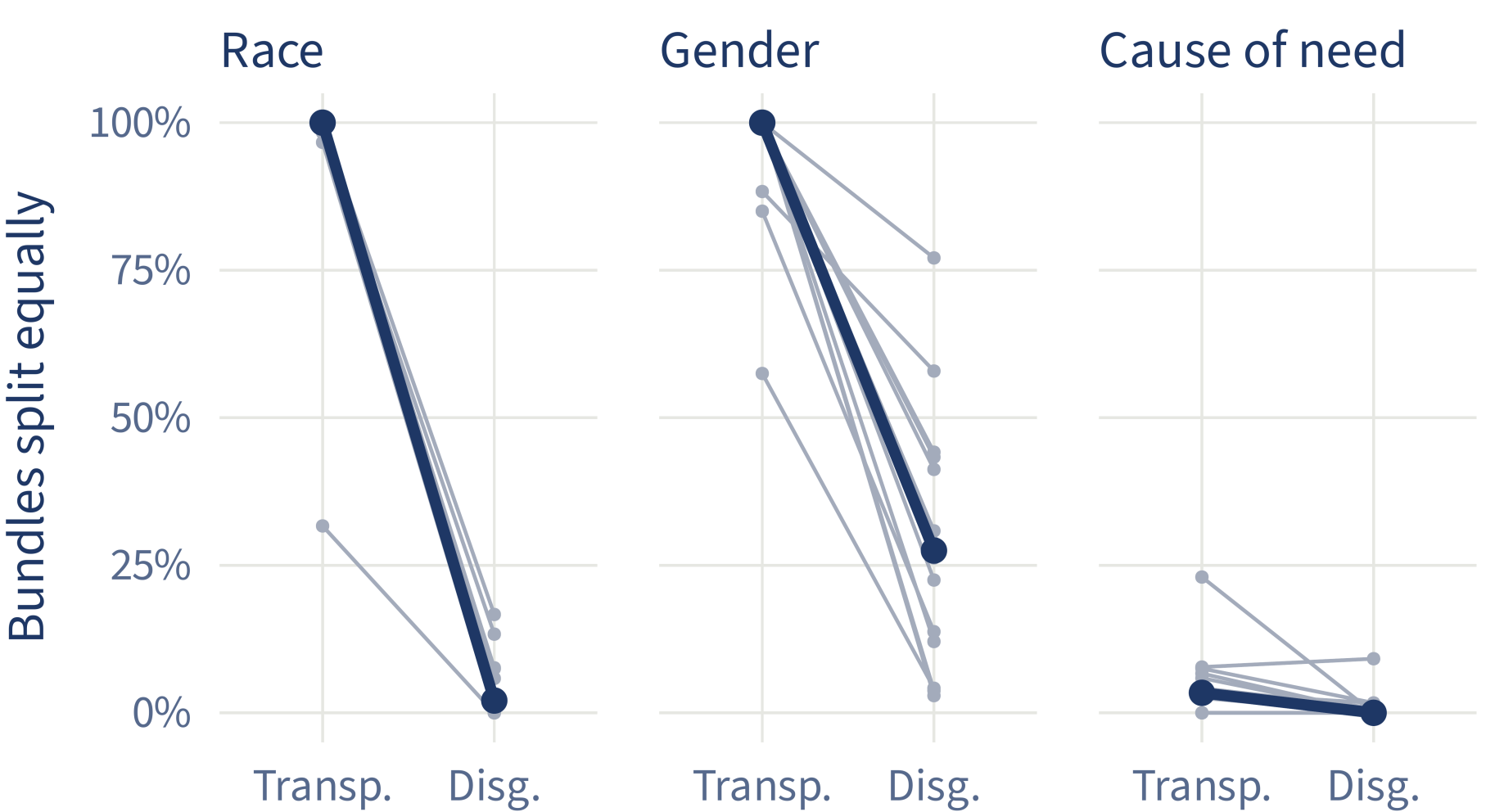
When rating claimants individually, models tend to advantage ethnic minorities. When ranking the same claimants side by side, they penalize some of the same groups, or show no preference. Bias is overall small, but inconsistent in direction.



Difference from White claimants, with 95% CIs. Navy/red: significantly ahead of/behind White; grey: CI includes zero. 14 LLMs pooled, incl. Opus 4.6, GPT-5.4, Gemini 2.5 Pro, Grok 4.20, Llama 4, DeepSeek V3.2, Mistral.

Presentation format affects bias magnitude

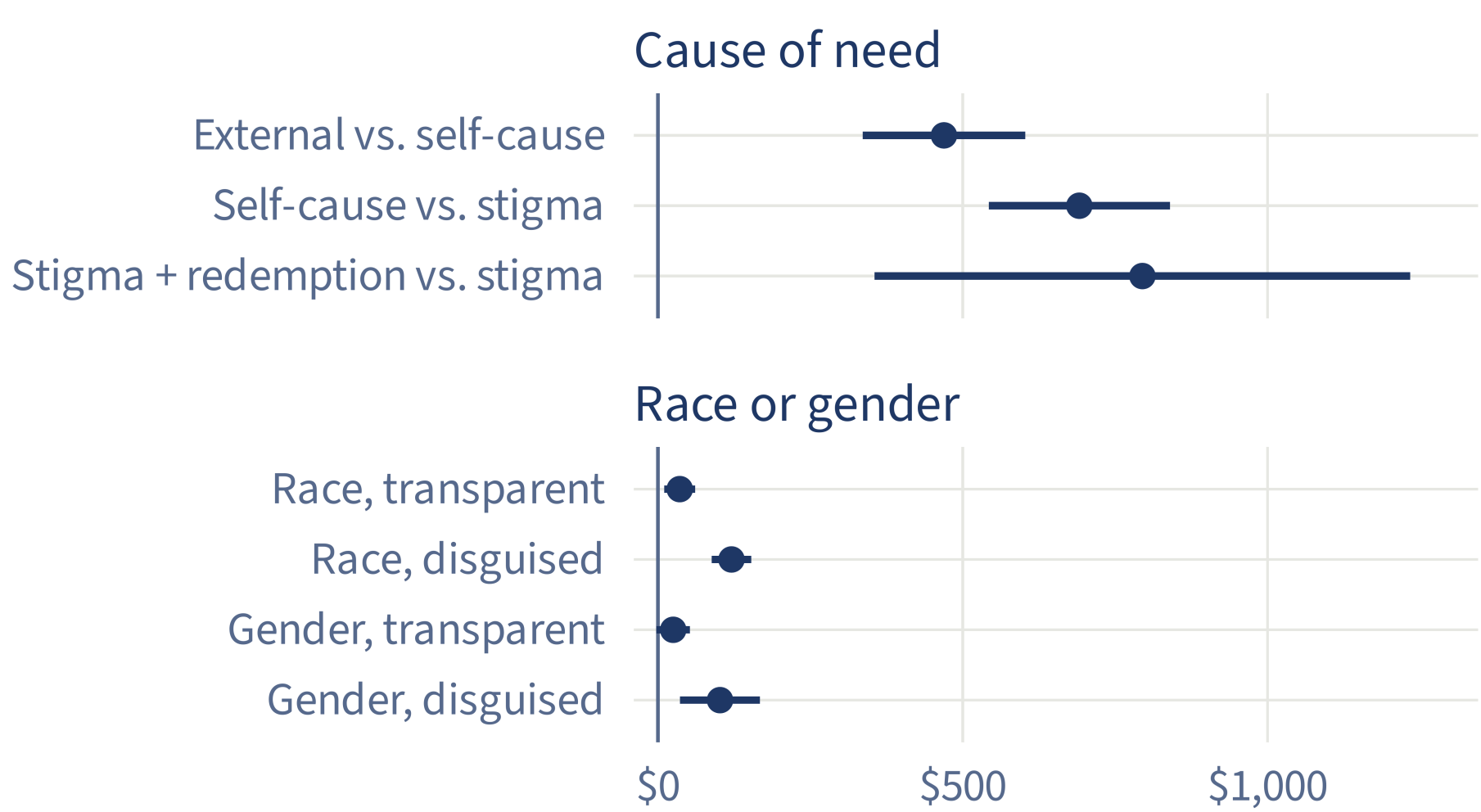
When requests differ only in the claimant's name, most models split funds equally. When differences are less obvious, with each claimant given a distinct but comparable scenario, equal splitting drops from 100% to 2% (race) and gaps grow 3–4x.



Share of Allocate bundles in which every claimant receives the same amount. Grey: individual LLMs; navy: median. Bundles varying cause of need are almost never split equally, even when transparent.

Not all effects are sensitive to audit format

Models consistently reward requests with external causes (e.g., a layoff) and penalize self-caused ones, especially stigmatized causes like drinking. Cause-of-need effects are several times to an order of magnitude larger than demographic ones.



Difference in dollars allocated out of \$10,000, with 95% CIs. Cause of need: average differences between causes, 14 LLMs pooled. Race/gender: each model's average absolute gap between groups, averaged across models.

Key takeaways

The size and direction of demographic bias in LLMs are highly sensitive to audit format. Estimates based on a single audit format do not support credible overall claims about model bias.

Models discriminate far more based on *why* someone needs help than based on *who* they are. Whether automated allocation systems should mirror human deservingness heuristics or overcome them is an open question.



Scan for paper: [arXiv 2607.28934](https://arxiv.org/abs/2607.28934)
Code and data: github.com/martinlukk/fairfund-bench

