
i10x.ai | i10x Research Team | 30 September 2026

AI Is More Objective Than Humans

A paired name-swap study of resume screening finds no gender gap
in four language models — and no human baseline against which
that headline can be tested

Christopher Ort

i10x Research Team, i10x.ai

Abstract

A common claim in hiring technology is that a model is more objective than a recruiter because it does not see gender. This paper tests the part of that claim that can be tested without a human arm, and states plainly what cannot. One hundred fictional personas were given fixed job descriptions. Four systems wrote a resume each. In the decisive experiment the resume text was then frozen and only the name, or the name plus a pronoun line, was changed: female, male, and they/them. The cheapest current screening-class models of four providers — Claude Haiku 4.5, GPT-6 Luna, Gemini 3.1 Flash Lite, and Grok 4.3 — scored every text from 0 to 100 and returned hire, maybe, or reject.

On the frozen text the female-minus-male score gap was between -0.01 and $+0.24$ points. Hire decisions agreed in 372 to 399 of about 400 pairs. A they/them pronoun line did not move the decision. The gap that did move the score was between models. Gemini-written resumes scored 95.4 and were hired at 98 percent. GPT-written resumes scored 89.7 and were hired at 86 percent. GPT scored its own resumes 8.2 points below the resumes of the other three writers, for women (-8.3) and men (-8.0) alike. Claude and Gemini scored their own texts about 2.5 points higher. Grok scored its own about 3 points lower.

The headline in the title is therefore not a result. No human raters took part, so the study cannot say that these models are more objective than people. What it can say is narrower and, for everyday screening, more useful. In this design the first name and the pronoun did not decide the hire. Which model wrote the resume, and which model read it, did.

1 Introduction

Employers already use language models to draft job posts, rewrite resumes, and sort applicants. Two failure modes follow. A model may prefer text that sounds like its own. A model may also react to a gendered name, or to a they/them line, even when the experience underneath is the same. The public argument often skips the measurement and jumps to a slogan: a model is more objective than a human because it has no biography and no first impression.

That slogan is doing two jobs. It claims that models ignore identity cues. It also claims that humans do not, and that the model is therefore the fairer reader. The second claim needs a

human arm. This study does not have one. The first claim can be tested directly, by changing only the cue and leaving the resume byte-identical.

An earlier wave in the same project had looked, at first pass, like a gender effect: one rater recommended female-coded names at about 65 percent and male-coded names at about 55 percent. Those resumes had been rewritten, not relabelled. Length, format, and wording moved with the name. A gender conclusion from that wave would have been a conclusion about a new text. The experiments reported here close that hole.

The questions are:

- On a frozen resume, does a female name, a male name, or a they/them pronoun line change the score or the hire decision?
- Independent of that cue, do models prefer resumes written by themselves, or by a particular other model?
- Does either pattern differ for women and men?

2 Design

2.1 Personas

One hundred fictional personas were used, each with a fixed job description. Roles ranged from backend engineering and data analysis to nursing, finance, and a head-of-product brief. No persona is a real applicant. Job descriptions were identical across every condition in which the same persona appeared.

2.2 Who wrote

Four systems wrote one resume per persona in the original wave: the files are labelled Claude, GPT, Gemini, and xAI. Those writer labels are the source of the text. The rating experiment did not ask the models to rewrite. It took those resumes as frozen documents.

2.3 What was changed

In the name-swap experiment the only edit was the name line. Each persona was rated once under a clearly female first name and once under a clearly male first name, same last name, same body. In the pronoun experiment a single line was inserted under the name: **she/her**, **he/him**, or **they/them**. The they/them arm kept the original gender-ambiguous first name. Female and male arms used the gendered first name plus the matching pronoun, so every arm carried an explicit identity line and the body stayed put.

This is the difference from the earlier gendered wave. There, overlap with the original text was low — a Jaccard index of word sets around 0.27 for Claude and 0.41 for Gemini — and Claude's texts often became cover letters. That wave is reported only as a caution. It is not the gender result.

2.4 Who rated, and why these models

Ratings were collected through OpenRouter from the cheapest current screening-class model of each provider. The choice is deliberate. A firm sorting a pile of applications does not call the most expensive flagship for every resume. It calls the cheap, fast model. The slugs were:

- anthropic/claude-haiku-4.5
- openai/gpt-6-luna
- google/gemini-3.1-flash-lite
- x-ai/grok-4.3

Each rater saw the job description and the resume and was asked for JSON only: an integer fit score from 0 to 100, and a recommendation of hire, maybe, or reject. Temperature was 0. The instruction said to judge fit and not to reward or penalize a name. A model that still moved with the name would have done so against that instruction.

The name-swap produced 3,200 calls, of which 3,168 returned a parseable score. The pronoun arm produced 4,800 calls, of which 4,713 returned a parseable score. Missing calls are excluded, not coded as zero.

2.5 What is compared

The gender estimate is paired. The same persona, the same writer, and the same rater appear under both names, and the difference is female minus male. The they/them estimate is the same pair against she/her and against he/him. The model estimate pools gender and compares writers and raters. A self-preference figure is the rater’s score on resumes it wrote, minus its score on resumes the other three wrote. Because the pair shares the text, a writer effect cannot be explained by the first name.

3 Results

3.1 The name does not decide

Table ?? gives the paired name-swap. The largest mean gap is a quarter of a point, on a scale that in practice sits in the nineties. Hire rates match to a percentage point. The number of pairs in which only one name is hired is in the single digits.

Table 1: Paired name-swap on frozen resumes. Difference is female minus male. “Only female” counts pairs hired under the female name and not under the male name.

Rater	<i>n</i> pairs	Score gap	Hire, female	Hire, male	Only female / only male
GPT-6 Luna	374	+0.11	87.7%	87.7%	1 / 1
Claude Haiku 4.5	400	+0.13	94.0%	93.2%	3 / 0
Grok 4.3	400	+0.24	93.5%	93.0%	3 / 1
Gemini 3.1 Flash Lite	400	−0.01	96.8%	96.5%	1 / 0

The same picture holds inside each writer. No writer-by-rater cell exceeds half a point of female-minus-male difference. A normal approximation to a paired test puts the Grok quarter-point near

a conventional threshold and every other gap well above it. A quarter-point, even if stable, does not change a hire.

3.2 They/them is not a penalty

Table ?? compares the they/them line with she/her and he/him on the same frozen body. Gemini’s hire decision is identical in 396 of 396 pairs. Claude’s only detectable score gap is -0.23 against the female arm, which is statistically noticeable and practically empty. Hire disagreements are single cases, and they do not point against they/them.

Table 2: Pronoun arm. Difference is they/them minus the named arm.

Rater	vs female	vs male	Hire, they/them	Hire, female / male
GPT-6 Luna	-0.05	$+0.14$	89%	89% / 88%
Claude Haiku 4.5	-0.23	-0.03	94%	94% / 94%
Grok 4.3	-0.04	$+0.11$	94%	94% / 94%
Gemini 3.1 Flash Lite	-0.01	$+0.06$	97.7%	97.7% / 97.7%

3.3 The writer decides

Pooling gender, resumes written by Gemini scored 95.4 and were hired at 98 percent. Claude-written resumes scored 94.6 and were hired at 96 percent. xAI-written resumes scored 90.2 and were hired at 91 percent. GPT-written resumes scored 89.7 and were hired at 86 percent. That ordering is the large effect in the file. It is several times the name gap, and it is the effect a screening pile would actually feel.

Table ?? crosses writer and rater on the hire rate. The cell that breaks the ceiling is GPT reading GPT: 68.7 percent hire, against 89 to 95 percent when Claude, Grok, or Gemini read the same GPT texts.

Table 3: Hire rate by writer (rows) and rater (columns), gender pooled, name-swap file.

Writer	GPT-6 Luna	Haiku 4.5	Grok 4.3	Gemini 3.1 Flash Lite
Claude	93.9%	97.0%	95.0%	98.0%
GPT	68.7%	89.0%	88.5%	94.5%
Gemini	96.9%	98.0%	98.0%	99.0%
xAI	86.6%	90.5%	91.5%	95.0%

3.4 Self-preference is model-specific, and gender-blind

GPT scored its own resumes 8.17 points below the others it read (85.68 against 93.85). The split is -8.31 for female names and -8.04 for male names. Claude scored its own 2.65 points higher ($+2.72$ female, $+2.58$ male). Gemini scored its own 2.46 points higher ($+2.37$ female, $+2.55$ male). Grok scored its own 3.06 points lower (-2.93 female, -3.19 male). Whatever these models are doing with authorship, they are not doing it differently for women and men.

Gemini is also the most generous rater overall, with a mean score of 96.8 and a hire rate of 96.6 percent regardless of writer. Scores remain compressed near the top of the scale. A small bias has little room. The authorship effect is large enough to be visible anyway.

4 What the headline cannot mean

“AI is more objective than humans” is the sentence vendors reach for. This paper borrowed it as a title because it is the sentence the experiment is constantly asked to support. It does not support it.

Objectivity toward a name is not objectivity. These four models did not hire Alexandra over Alexander, or she/her over they/them, on a frozen resume. They did hire a Gemini-written resume over a GPT-written resume of the same persona, and GPT docked its own writing by eight points. A human recruiter who ignored the name and then downgraded every application that “sounded like ChatGPT” would be open to the same charge. Without recruiters in the design, the comparison is rhetoric.

The earlier, rewritten wave is a warning about the slogan in the other direction. A gender gap appeared when the text was allowed to change with the name. Publishing that gap as a property of the model would have been a property of the rewrite. Everyday audits that regenerate the resume under a new name make the same mistake.

5 What this means in use

A candidate who has GPT draft the resume, and a firm that has GPT screen the inbox, is the case the file speaks to. In this test those resumes are hired by GPT at 69 percent and by the other three raters at 89 to 95 percent, with an eight-point score cut that does not depend on the name. The practical rule is to not let the writing model be the judging model.

A firm that uses Claude or Gemini both to polish applications and to score them will, on these numbers, give its own polish a small lift of about two and a half points. That is not a scandal. It is a reason to separate the two steps, or to have a person read the boundary.

A name-bias audit inside a product is only an audit if the resume is frozen. Changing the name and asking the model to rewrite, then scoring the rewrite, measures the rewrite. The they/them result says the same thing as the name-swap: an explicit pronoun line, on these cheap models, did not decide the hire.

None of this says a human panel would have done worse, or better. It says the failure mode to watch in this stack is authorship, not the first name.

6 Limits

The personas are fictional, and the job descriptions were written for the study. A real pile has messier text, employment gaps, and names the model has seen in training. One rating was taken per cell, at temperature 0. A second draw could move a score by more than the name gap. It is unlikely to invent an eight-point authorship effect, and unlikely to erase one. The writer texts themselves come from an earlier generation pass whose model versions were not pinned in the source files. The raters are pinned: Haiku 4.5, GPT-6 Luna, Gemini 3.1 Flash Lite, Grok 4.3. A previous Claude rater, in the unpinned wave, hired GPT-written resumes at 42 percent. That cell is not replicated here. Results do not travel automatically to a flagship model, or to last

year’s snapshot.

There is no human arm. There is no demographic cue other than the name and the pronoun line. Ethnicity, age, disability, and university prestige were not swapped. The score ceiling means a bias that only appears among near-equal candidates can hide. The authorship effect did not hide.

7 Conclusion

On a frozen resume, four screening-class models did not hire women over men, or men over women, or either over they/them, in any way that would change a decision. They did hire some models’ writing over others, and GPT was eight points harsher on text it had written itself, for both names. The headline that these systems are more objective than humans is not a finding of this study. The finding is that, in this design, the name was the wrong suspect.

References

- [1] Bertrand, M. and Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American Economic Review*, 94(4), 991–1013. The classic human audit. Cited as the comparison this study does not rerun, not as a benchmark these models beat.

Christopher Ort, i10x Research Team, i10x.ai.

Source tables are namenstausch_bewertung.csv and they_them_bewertung.csv. Score gaps are paired within persona, writer, and rater. Empty model replies were dropped, not imputed.