THE INSTRUMENT IS NOT THE MODEL: MEASURING HOW MUCH OF AN LLM HIRING DISPARITY COMES FROM UNREPORTED DESIGN CHOICES

Anonymous authors
Paper under double-blind review

ABSTRACT

Audits of large language models for hiring discrimination carry legal weight: New York City's Local Law 144 makes it unlawful to use an automated employment decision tool that has not been audited for bias. Such an audit reports a demographic effect, meaning how differently a model treats two résumés differing only in the name, and whatever the two share should cancel. It does not, by enough to change what an audit concludes. Holding the model fixed and varying only choices published audits do not report, over 31,468 matched pairs from 64,397 calls on 6 open-weight and 5 frontier checkpoints, the instruction wording moves the effect by 25% to 64% of itself across the 6 cells where the effect is separable from zero, including under edits that change no word; where the sign flips, the résumé template outweighs the wording. Only 25% to 33% of name pairs from a standard validated list are token-matched. Choices made after the data exists matter as much. Resampling rows rather than name pairs narrows intervals 3.0 to 4.8×, and a fixed operating point misstates the percentage-point conversion by 2.0× to 484×. The measurement itself reproduces bitwise only once request batching and cache residency are controlled. Of 13 LLM hiring audits read in full text, none reports enough to reconstruct its number. We give the minimum an audit must report, a screening rule with a calibrated false-positive rate, and the full pipeline.

1 INTRODUCTION

An audit of a hiring model reports one number, and in New York City that number has legal force: Local Law 144 makes it unlawful to use an automated employment decision tool that has not been audited for bias in the preceding year (City of New York, 2021; New York City Department of Consumer and Worker Protection, 2023). Producing the number takes dozens of choices the audit does not report: how the instruction was worded, which names were drawn, what job was posted, how the model was served, and how the analysis resampled, converted and paired what came back.

This paper measures what those choices are worth. Holding the model fixed, over 31,468 matched pairs from 64,397 calls on 6 open-weight and 5 frontier checkpoints, the instruction wording alone moves the reported demographic effect by 25% to 64% of itself across the 6 model-by-posting cells where the effect is separable from zero, including under edits that change no word. The job posted and the name drawn move it comparably; the analysis-time choices move it again, with the resampling unit alone worth 3.0 to 4.8× on the width of the interval. Each of those choices moves the reported number by a sizeable fraction of the whole effect, against a byte-identical noise floor of about a twentieth of it.

One reading must be closed at the outset. None of this says the wording matters more than *which model* was tested. The between-model spread on this panel is larger than the between-wording spread within any model; what does not survive is the assumption that a number produced under one instrument transfers to another. Indeed, *which* wordings inflate the effect is largely model-specific: 70% of the between-wording variance is idiosyncratic (§6).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

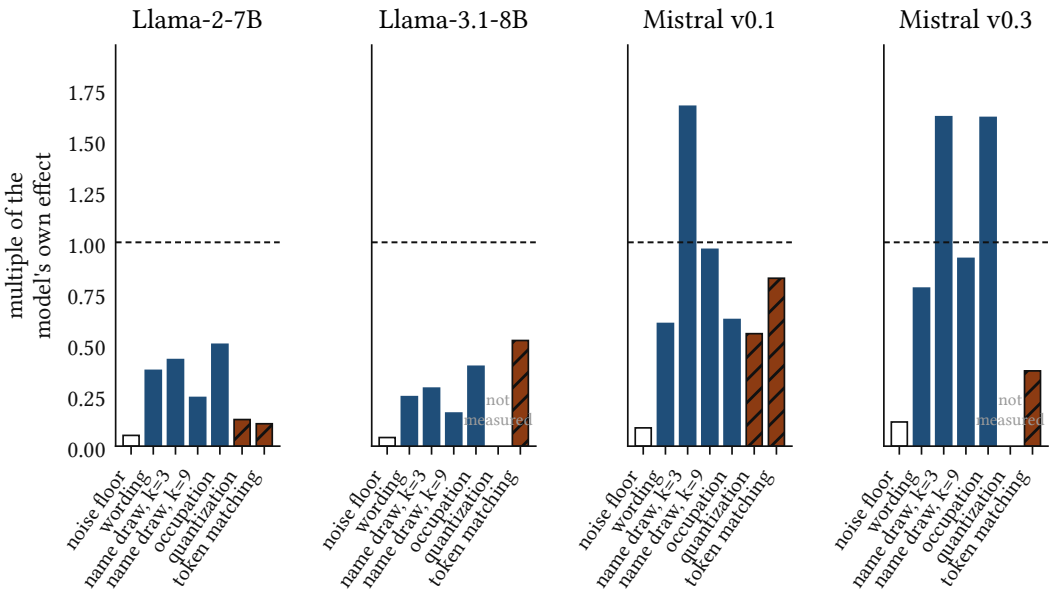


Figure 1: **Each unreported choice, as a multiple of the effect it perturbs** (log-odds scale). The dashed line marks a choice that moves the estimate by as much as the entire effect. The open bar is the measurement noise floor, the value a byte-identical repeat gives. Solid bars are standard deviations over defensible choices; hatched bars are single shifts between two specific settings. On the two Mistral checkpoints the denominator is not separable from zero, so those panels say more about a small divisor than a large numerator.

Contributions. (1) A dispersion budget: per model, the standard deviation the reported disparity inherits from each unreported choice, as a multiple of the model's own effect, with every component measured rather than argued (Figure 1). (2) Evidence that the measurement itself is not deterministic, an attribution of that nondeterminism to request batching and KV-cache residency, and a serving configuration under which it reproduces bitwise (§5). (3) A survey of 13 LLM hiring audits showing that none reports enough for a reader to reconstruct its number (§8). (4) A minimum reporting set with a calibrated screening rule, priced so an auditor can adopt it (§8). Every measurement in this paper is interpolated from released artifacts at build time; an audit script fails the build if one is typed into the prose.

2 Related work

Correspondence audits supply the design (Bertrand & Mullainathan, 2004). The tradition has audited its own instrument before: group differences in the variance of unobservables can alone generate spurious discrimination estimates of either sign (Neumark, 2012); a distinctively Black name indexes the circumstances of its bearer's birth (Fryer & Levitt, 2004); the perceived race of a name varies name by name (Gaddis, 2017). Closest to us, Lahey & Beasley (2009) split one audit into two arbitrary sets of four résumé templates and obtained a significant discrimination estimate from one set and a null from the other, on the same data. That is this paper's result on human employers, seventeen years earlier. Prompt sensitivity is documented at length for LLMs, as *accuracy* moving (Sclar et al., 2024) and as a fairness estimate moving without a noise floor (Seshadri et al., 2025; Tan et al., 2026); researcher degrees of freedom have their own literature (Steegen et al., 2016; Simonsohn et al., 2020), but varying the *analysis* over a fixed measurement. A systems literature documents that concurrent LLM serving is not batch-invariant (He & Thinking Machines Lab, 2025; Yuan et al., 2025), and Atil et al. (2025) measure the resulting instability across five hosted APIs while reporting that behind an API they can only speculate about its cause; §5 varies those mechanisms directly on weights we serve ourselves. Closest to us in argument, Qian et al. (2021) show that identical fixed-seed training runs move a fairness metric by up to 12.6 points and flip a verdict across the threshold used in a U.S. legal case, and that only a third of the papers they survey run

training more than once. This paper is that argument where there is no training run to repeat, so the whole of the dispersion sits in the instrument. On social commonsense reasoning, a name's tokenization length influences how a model treats it independently of demographics (An & Rudinger, 2023). The gap is that nobody has varied how the stimulus itself was built and served, and asked what happens to the reported disparity. Among LLM hiring audits (e.g., Armstrong et al., 2024; Gaebler et al., 2024; Wilson & Caliskan, 2024; Veldanda et al., 2023; Nghiem et al., 2024; Glazko et al., 2024; Gao et al., 2026; Lippens, 2024), three reach part of this conclusion first: Seshadri et al. (2025) report comparable sensitivity to demographic and non-demographic perturbations, Tan et al. (2026) that an instruction change alone moves the bias estimate, and Lippens (2024) larger within-group than between-group dispersion across names. We add the decomposition, the serving-stack attribution, and the reporting standard.

3 DESIGN

The unit of measurement is a *matched pair*: two identical résumés differing only in the name, first name and surname both chosen to signal race, scored independently by a model asked to make a screening decision. The name grid is built factorially from the validated list of Bertrand & Mullainathan (2004); the outcome is the model's preference between *yes* and *no* under a constrained decoding grammar, read as a log-odds margin, with the demographic effect summarized as a probability of superiority (the probability that the White-named résumé of a random pair scores higher). Percentage points appear only where a per-cell probability-scale interval exists; a fixed-Jacobian figure is labelled an upper bound, because that conversion is wrong by 2.0× to 484× across this panel (§6).

Around that fixed core, we vary only what audits do not report: twelve instruction wordings (six semantic paraphrases and six null edits that change no word, such as a trailing newline or a doubled space after a colon), three résumé templates, three job postings, the name draw, the serving configuration (batching, KV-cache residency, quantization), and the analysis-time choices (resampling unit, reporting scale, pairing). Two wordings are byte-identical by construction, which gives the design a free replicate: any disagreement between them is measurement noise, and every dispersion below is read against that floor. Uncertainty comes from bootstrap resampling of *name pairs*, the unit the assignment process dictates (Abadie et al., 2017), with wild cluster bootstrap-t intervals as a robustness check (they come out 9% to 21% wider and change 0 significance verdicts).

Formally, write $\hat{\theta}_m(w, t, p; N, s, a)$ for the number an audit of checkpoint m reports, where w is the instruction wording, t the résumé template, p the job posting, N the drawn name set, s the serving configuration and a the analysis path. A published audit reports one point of this function and reads it as θ_m, a property of the model. What this paper measures is the dispersion of $\hat{\theta}_m$ along each argument with the others held at their defaults, as a fraction of the effect itself; for the wording,

$$R_m^{(w)} \;=\; \mathrm{SD}_w \; \hat{\theta}_m(w, \bar{t}, \bar{p}) \,/\, |\bar{\theta}_m|,$$

with $\bar{\theta}_m$ the pooled estimate under the default instrument, and analogously in the remaining arguments. Figure 1 plots exactly these ratios. A ratio is printed only where its denominator is separable from zero, and every dispersion is read against the byte-identical replicate floor, the same statistic computed across repeats in which no argument changes.

4 THE STIMULUS MOVES THE ESTIMAND

Wording. Across the twelve wordings, the between-wording standard deviation of the demographic effect runs 25% to 64% of the effect itself on the 6 identified model-by-posting cells, open-weight and frontier alike (Table 1 gives the open-weight panel; these are published ratios, §5). The null edits, which change no word, move the effect about as much as genuine paraphrases, so surface form does as much work as meaning. A crossed random-effects model partials the sampling noise out of that dispersion (Figure 2; the per-wording estimates behind it are Figure 6). Read against the replicate floor, wording moves the measured effect by 5.7 to 7.0× what the arithmetic noise does.

Every defensible analysis, sorted. Crossing the twelve wordings with three templates yields 36 specifications per model, each an analysis a researcher could have run and published. On the checkpoint where the sign flips, 12 specifications are significantly positive and 12 significantly negative,

Table 1: The measured demographic effect and its dispersion across twelve wordings that ask the same question. The ratio is suppressed where the effect interval covers chance.

Model	P(sup.)	95% interval	SD across wordings	SD / |P−0.5|
Llama-2-7B-chat	0.7894	[0.6968, 0.8727]	0.0966	33%
Llama-3.1-8B-Instruct	0.1875	[0.1019, 0.2894]	0.0778	25%
Mistral-7B-Instruct v0.1	0.4421	[0.2615, 0.6227]	0.0685	–
Mistral-7B-Instruct v0.3	0.4306	[0.3333, 0.5301]	0.0384	–

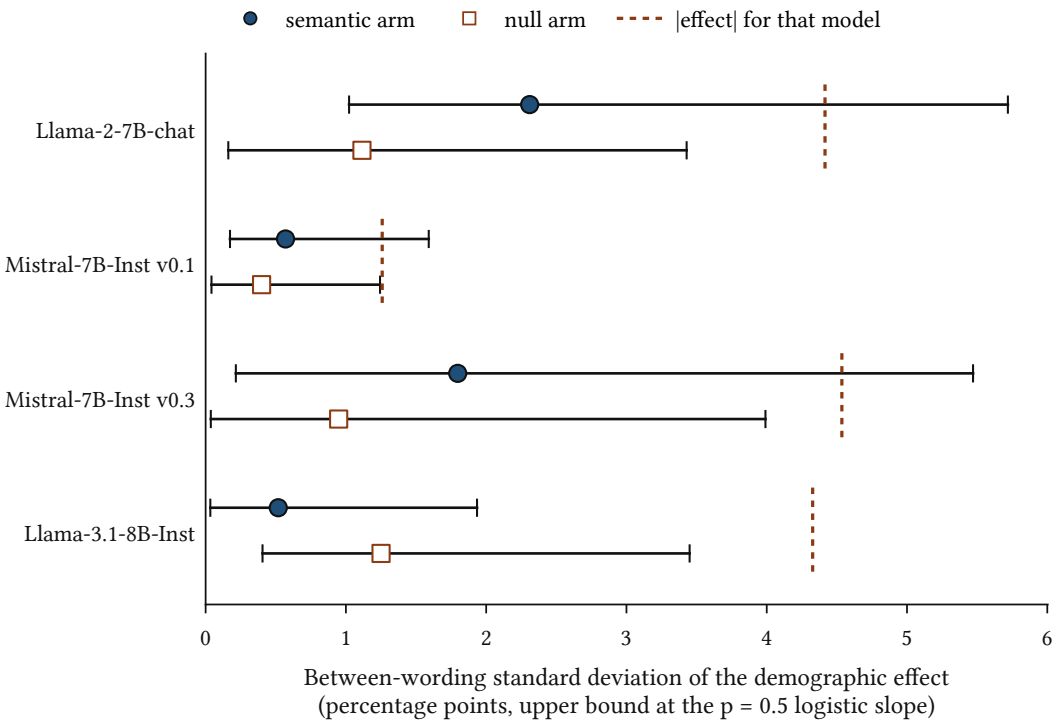


Figure 2: **Between-wording dispersion as a component.** Posterior median and 95% credible interval for the between-wording standard deviation from a crossed random-effects model, fitted separately to the semantic and null arms; sampling noise is partialled out by partial pooling. The dashed rule marks each model's own demographic effect.

so the *sign* of the reported effect is determined by choices no methods section records. Decomposing that spread, 91.5% lies between templates and 4.5% between wordings, so on this checkpoint the sign is a template effect and what the wording moves is the magnitude, which is what the rest of the paper reports (Figure 3).

Names and tokens. The two prompts of a matched pair are identical in characters but, two thirds to three quarters of the time, not in tokens; only 25% to 33% of pairs from the standard list are token-matched against a served checkpoint's vocabulary. The token-length difference predicts the measured effect on 1 of four checkpoints on the name-pair unit, and restricting to token-matched pairs does not merely cost power; it forecloses an answer. On the checkpoint where the matched subset is smallest, 3 pairs survive, so the matched-pair randomization test has 8 sign assignments and a smallest attainable two-tailed p-value of 0.25, above any conventional threshold whatever the data show. Stratifying the permutation within token-length classes, the correct test once length is controlled, leaves that count unchanged, because the classes are complete blocks and the randomization distribution factorizes over them. A larger list drawn the same way cannot fix it either, because the race gap in token length is systematic rather than sampled: the panel-wide maximum matching over the full source list is 3 pairs, and clearing the conventional threshold would take about 36 names

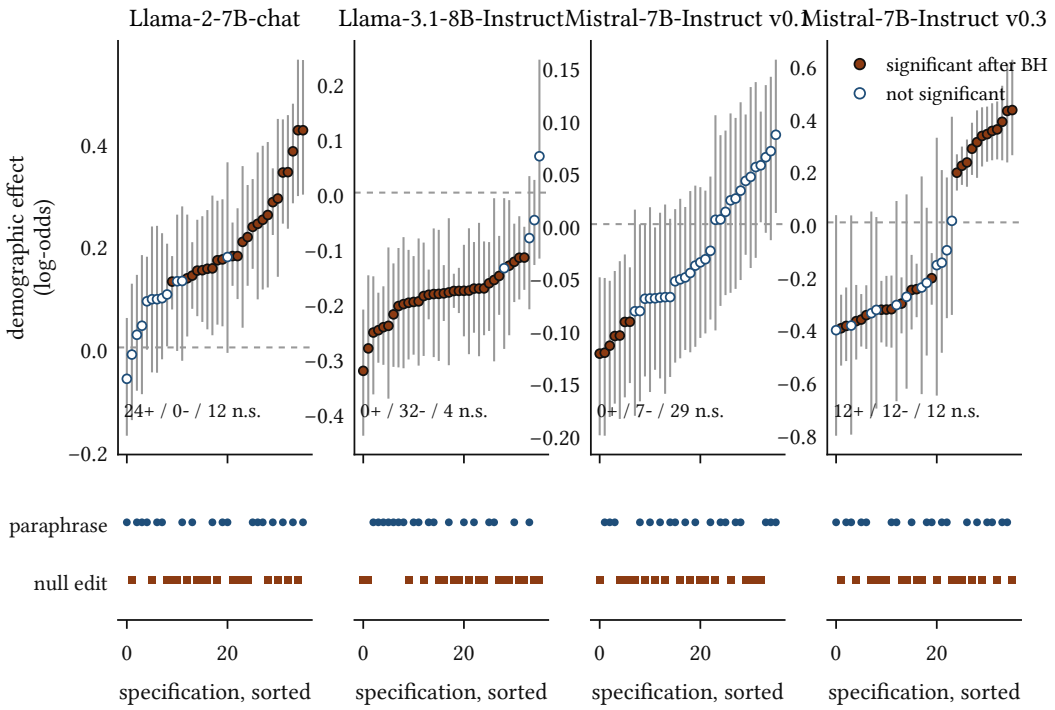


Figure 3: **Every defensible analysis of the same data, sorted by the answer it gives.** 36 specifications per model (twelve wordings × three templates); filled points survive Benjamini–Hochberg (Benjamini & Hochberg, 1995) within the model. On the sign-flip checkpoint, 12 specifications are significantly positive and 12 significantly negative. The lower strip marks paraphrase versus null edit; the two do not separate.

per cell against the 9 the standard list carries. A usable list has to be built token-matched from the outset rather than filtered afterwards. Per-arm volatility is also asymmetric: the Black-name arm is the more variable one in 6 of 8 model-by-dataset combinations, though one checkpoint shows the significant reverse, so the asymmetry is real but not a law.

The name draw. The observation that within-group dispersion across names exceeds the between-group difference is not new: Lippens (2024) reports it on an order of magnitude more name combinations than this grid carries, and Gaddis (2017) shows why from the perception side. What the crossed grid adds is the decomposition: because each race arm carries several distinct first names, nested within race in the fit, the name-specific component is separable from the race-specific one, and the dispersion an auditor's *draw* of k names contributes appears as its own bars in Figure 1, shrinking with k exactly as a sampled component should while the wording component does not shrink at all. That contrast is the operational test for what is sampling error and what is instrument.

The job posted. Re-running the full grid under two further postings moves the effect by as much as the whole demographic effect, so an audit's number does not transport between jobs; whether the *dispersion* also varies by job is not separable from noise on three postings, and we make no claim that it does.

Beyond hiring. None of this is a hiring artifact. The same twelve wordings, names and null edits were re-run with only the task changed (rental tenancy screening and content moderation), and the between-wording standard deviation runs 0.016 to 0.172 log-odds across the 8 domain-by-model cells, the same order as the hiring grid. The sensitivity travels with the instrument, not with the domain.

Mechanism, bounded. An apparent tokenization mechanism, that fragmenting a structural delimiter shifts the effect, turned out to be confounded with token position: controlling position, only 7 of 48 contrasts survive correction within a family of 168, and the position-only and fragmentation

classes do not separate, so moving the name does what fragmenting a delimiter does. The floor under that null is the minimum detectable *difference* between the condition classes, 0.0082 log-odds calibrated by injection on the name-pair unit; a smaller mechanism would not have been detected here, and we claim only that none larger exists on this panel.

5 THE SERVING STACK IS PART OF THE INSTRUMENT

At temperature zero the measurement should be deterministic. It is not: two of the twelve wordings are byte-identical by construction, and across the four checkpoints they agree on 66 of 144 cells (Figure 4). Batching is most of the cause and not all of it. Repeated calls on an identical prompt agree on 47.9% of cells under concurrent requests, 93.8% when requests are forced sequential, and 100% only when KV-cache reuse is disabled as well, on the subset of checkpoints where that arm was run, the batch-dependence documented for concurrent LLM inference (He & Thinking Machines Lab, 2025; Yuan et al., 2025; Atil et al., 2025) surfacing here as measurement error in a legally consequential audit. Re-measuring 504 cells in a separate process on a different day reproduced 0 of them exactly on three of the four checkpoints. With batching and cache residency controlled, the measurement reproduces bitwise across processes on the two checkpoints where that was run, while the effect estimate is not detected to move under either control, a shift bounded within 8.7% of the effect on the identified checkpoints, a bound rather than an absence. Quantization is a larger lever. At 4-bit versus 8-bit, the same checkpoint's effect shifts by $0.37\times$ and $1.07\times$ the between-wording standard deviation on the 2 checkpoints tested. Three surveyed audits rest their reproducibility claims on a decoding setting alone; on this panel that is not sufficient, and for audits run against a commercial API several of these fields are not the auditor's to report, which the reporting set accounts for.

The stimulus-side results are not an open-weights artifact. The same twelve wordings run against frontier APIs give dispersion-to-effect point estimates of 0.46 to 0.55 on the checkpoints whose effect is identified, after subtracting an API noise floor estimated from the same byte-identical replicate, which accounts for up to 26% of the observed variance there. The correction is applied to both panels symmetrically, since applying it to one side of a comparison is precisely the unreported choice this paper measures. The range in §4 is published throughout: two of its 6 cells have no replicate floor, and correcting only the rest would be that same one-sided choice.

6 CHOICES MADE AFTER THE DATA EXISTS

Three analysis-time choices move the reported number without touching a single model call (Table 2). **The resampling unit:** the design is crossed, so pooling rows and bootstrapping them independently treats repeated measurements of the same twelve names as independent draws; against a bootstrap that resamples name pairs, the i.i.d. interval on the per-model effect is too narrow by 3.0 to $4.8\times$, and significance verdicts flip. The correct unit follows from the assignment process (Abadie et al., 2017), and is almost never the row. **The reporting scale:** converting log-odds to percentage points at a fixed operating point misstates the conversion by $2.0\times$ to $484\times$ across four same-size checkpoints, because they do not share an operating point. **The pairing.** Each résumé is scored in its own request, so no shared decision-maker links the two. The model is common to every observation rather than to a pair, which makes it a study-level constant rather than a blocking factor, inducing dependence without inducing a match. What *is* matched by design is the résumé content, and the design fixes that without fixing which scored instance is differenced against which, so re-pairing moves the pairwise statistics the field reports, the probability of superiority and the discordant-pair count, while leaving the mean paired difference algebraically invariant. The latitude is narrow. It exists only where the pairing is imposed after scoring, and not in a design that assigns preconstructed profiles within a prespecified block, where the match is fixed before any outcome exists. Where a statistic can be made invariant to a choice, report that statistic.

Finally, the wording effect itself is model-specific. Stacking per-wording effect profiles across five task contexts, cross-family correlations are near zero and 70% of the between-wording variance is idiosyncratic to the model; the one pair that agrees is the one sharing a lineage (the two Mistral checkpoints, $r = +0.31$). An auditor cannot learn a safe wording on one model and reuse it on another.

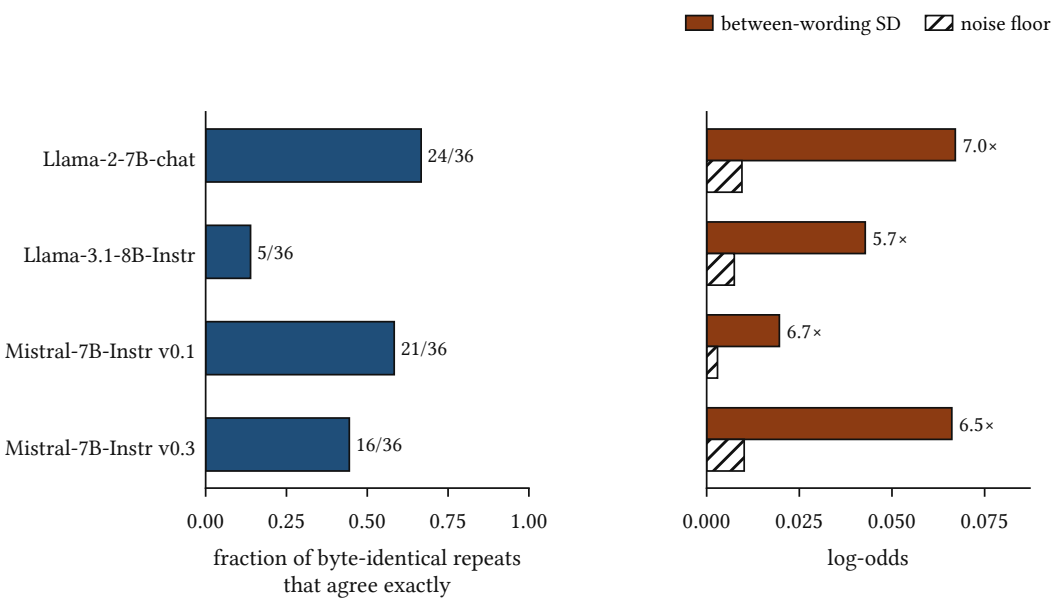


Figure 4: **The measurement is not deterministic, and the wording result survives it.** Left: fraction of byte-identical repeats agreeing exactly, per checkpoint. Right: the between-wording standard deviation against the noise expected on a per-wording mean of the same size, estimated from the replicate.

Table 2: Three choices made after the data exists. Each moves the reported number, none is a property of the model, and no audit surveyed reports its number's sensitivity to any of them; the pairing is disclosed by none. The mean paired difference is algebraically invariant to the third, which is why the paper recommends reporting it rather than disclosing the choice.

Analysis choice	What it moves	What it does to the verdict
Resampling unit *rows, not name pairs*	intervals narrow by 3.0× to 4.8×	4 of 4 models significant instead of 2
Reporting scale *a fixed operating point*	log-odds-to-percentage-point conversion misstated by 2.0× to 484×	distinguishable on 2 of 4; the converted magnitude is not comparable
Pairing *re-pairing the same scores*	P(superiority) moves by 0.20 to 0.24	the sign flips on 2 of 4, both unidentified

7 What the sensitivity is not

Three readings would overshoot the evidence. **Not nihilism:** the dispersion is measurable, bounded, and largely attributable (to the wording set, the template, the serving stack, the analysis), which is the opposite of "nothing can be measured." A statistic reported with its dispersion is more informative than one reported without it, not less. **Not noise:** every component sits above the byte-identical replicate floor, by a median of 7× and a minimum of 2.1×, and the null edits' effect is reproducible rather than random. **Not an artifact of small effects alone:** the ratio's denominator is policed throughout. Cells whose effect is not separable from zero are excluded from every headline range that divides by the effect, and ratios against such denominators are never printed as point values. The conversion-factor range above divides by no effect, so all four checkpoints enter it. What the sensitivity *is*: a property of the current instrument, of the same order on every checkpoint, domain and serving stack we measured, and unreported by every audit we surveyed.

Table 3: Reporting practice across 13 LLM hiring audits read in full text. A partly specified choice is not counted as reported, because the question is whether a reader could reconstruct it.

Choice	Fully reported / applicable	Partly
exact prompt published	11 / 13	2
wordings used	10 / 13	3
wordings separately estimated	8 / 12	4
dispersion across wordings	0 / 8	4
null-edit control	2 / 13	2
name list source	9 / 10	1
names per race	8 / 10	1
name-draw sensitivity	1 / 10	3
token matching of the pair	0 / 12	1
checkpoint pinned	2 / 13	10
quantization	1 / 11	1
serving stack	5 / 13	4
request batching	0 / 13	2
cache policy	0 / 12	1
decoding parameters	2 / 12	5
repeats / reproducibility	6 / 13	4
resampling unit	3 / 11	7
multiplicity correction	4 / 13	0
reporting scale + operating point	10 / 13	3
occupations	12 / 13	1
headline effect	10 / 12	2
code or data released	2 / 13	2

8 What the field reports, and the minimum it should

We read 13 LLM hiring audits (An et al., 2024; Armstrong et al., 2024; Fu & Shi, 2025; Gaebler et al., 2024; Gao et al., 2026; Glazko et al., 2024; Hoffstedde et al., 2026; Iso et al., 2025; Nghiem et al., 2024; Seshadri et al., 2025; Tan et al., 2026; Veldanda et al., 2023; Wilson & Caliskan, 2024) in full text and coded each against 22 reporting fields (Table 3; Appendix A.1 states the frame and its one exclusion). None fully reports its batching, cache policy or token matching; 0 of the 8 for which it is applicable report any dispersion statistic across wordings; the published effects the field compares itself to have a median absolute gap of 0.64 percentage points across 14 model rows of one panel (Figure 5), measured on a callback scale our log-odds components do not share.

The minimum an audit must report, priced: **(1)** the wording set and the dispersion of the effect across it; **(2)** the name list, its source, and its token statistics against the served vocabulary; **(3)** the serving configuration (checkpoint digest, quantization, batching, cache policy, decoding) or, for an API, what the platform exposes plus the statement that the rest was not knowable; **(4)** the resampling unit and the reporting scale. The cheapest single repair on the panel is the checkpoint digest: under a minimal re-run criterion (exact prompt, pinned checkpoint, decoding parameters), pinning the checkpoint alone would bring 3 of the 13 audits up to the criterion, and most already half-report it. Alongside the set we give a screening rule: report a disparity as established only if it is separable from zero under every wording in the published set. Under a one-factor equicorrelated normal null, $\theta_w = \sqrt{\rho}\,U + \sqrt{1-\rho}\,\varepsilon_w$, requiring every interval to exclude zero on the same side reduces to a one-dimensional integral over the shared factor, so the rule's false-positive rate has a closed form; the rate rises monotonically in ρ and never exceeds the nominal level, and on this panel's measured wording correlations it runs 4×10^{-4} to 1.2×10^{-2}, at least four times stricter than the single-wording test it replaces. The rule is one-directional, doing nothing about a disparity every wording misses, and no disparity on the four checkpoints carrying per-wording intervals would be established under it. That is the finding. The instrument is currently too unstable for a wording-robust disparity to have been measured, on these checkpoints, by anyone.

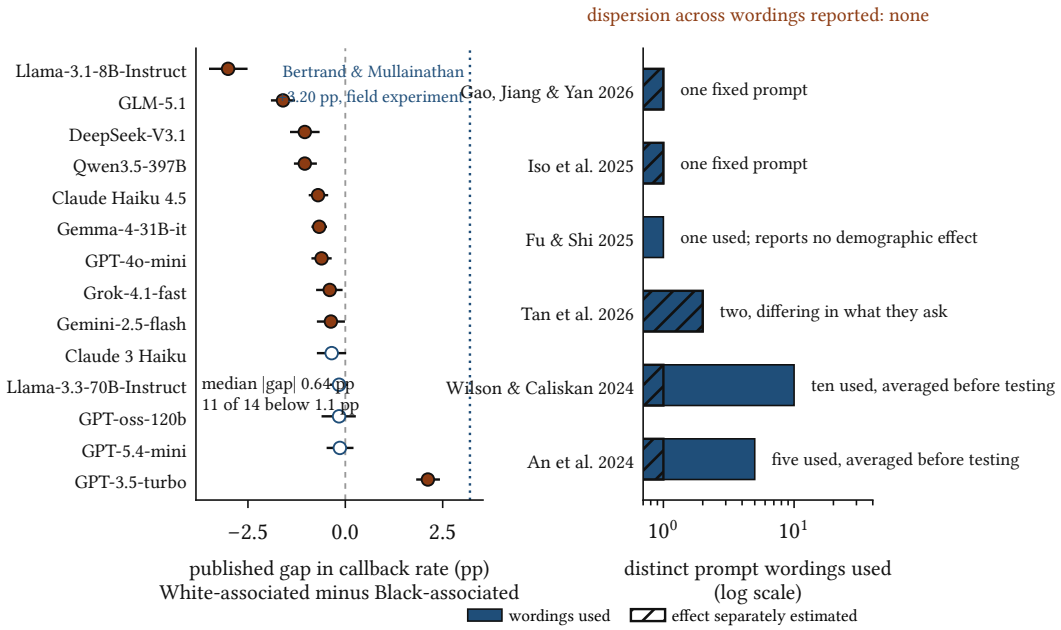


Figure 5: **The field's effects are small, and none is published with a range.** Left: every demographic gap in the largest published model panel, with the Bertrand & Mullainathan (2004) field experiment as the anchor; the median absolute published gap is 0.64 points. Right: for six of the 13 surveyed studies, wordings used against wordings the effect was separately estimated under.

9 Limitations

Internal. The measurement reproduces bitwise once batching and cache residency are controlled; the headline dispersions come from a study run before those controls existed, and their replicate floor is measured and subtracted rather than assumed. Effect intervals are percentile cluster bootstraps over twelve name pairs; the wild-t robustness check widens them 9% to 21% without changing a verdict. **External.** Open-weight checkpoints carry everything the serving stack touches; one name list carries the demographic manipulation; three postings carry the occupational one, so no claim is made about occupations in general. The stimulus-side results replicate on rental tenancy screening, content moderation, and frontier APIs; the serving-side results cannot be tested on a stack whose numerics no auditor can inspect. Where a model shortlists and a person decides, the design measures the screening stage alone: the second stage sees a sample selected on the first, and a pair may both survive, split, or both be cut, so a disparity measured at the screen does not transfer linearly to who is interviewed. The regulation audits the tool rather than the pipeline, which is the scope we adopt. There is no human tester to deviate from random assignment, so experimenter bias (Lahey & Beasley, 2009) is absent by construction. **Construct.** Whether a name manipulation measures what an audit calls it is checked by no audit we surveyed; our own manipulation check detects only correlations of 0.59 or larger, a design limit rather than a result. On the lower half of the panel the decoding grammar does real work, and three natural decision rules disagree on up to 43% of probes there.

10 Conclusion

The matched-pair difference is supposed to be the robust quantity: whatever the two applications share should cancel. It does not, by enough to change what an audit concludes, and the part that fails to cancel is the instrument. The remedy is not to abandon the design but to report the choices, and where a statistic can be made invariant to a choice, to report that statistic instead. What stays open is the channel, which is not identified, and the standard itself, which no published audit has yet been re-run against.

ETHICS STATEMENT

This work audits measurement practice; it neither builds nor improves screening systems. No human subjects were involved; all stimuli are synthetic, with names drawn from a published, validated research list. The finding that reported disparities are instrument-sensitive must not be read as evidence that model bias is absent: the screening rule proposed here is deliberately conservative in one direction only, and the paper states where its own design could not have detected an effect. The survey in §8 publishes a negative finding about each of 13 named papers: what they do not report. Those verdicts were coded with the AI assistance described below. Each carries a verbatim quote or the exact search that returned nothing and ships with the supplementary material, so an author who disputes a cell can check the evidence behind it. We report reporting practice, not research quality. An author holds a position at a company that builds AI recruiting software; none of that company's data, models or internal information was used. The conflict is disclosed here and will be stated in full in the camera-ready.

REPRODUCIBILITY STATEMENT

Every measurement in this paper is interpolated at build time from released JSON artifacts rather than typed into the prose, and an audit script over the typesetting source reports any exception; design constants, cross-references and figures quoted from other papers are exempted by name. The released repository contains the derived artifact behind every claim, every analysis script, the exact prompts, the name grid, checkpoint SHA-256 digests, and the serving configuration for each study. Two things it does not contain, and says so: the per-call raw outputs, which are far larger than any submission limit, and the copyrighted literature full texts, though the survey records a verbatim quote or an explicit negative search for every cell. The supplementary material accompanying this submission is an anonymized subset of that repository: the artifacts, analysis code, prompts, name grid, checkpoint digests, serving configurations and test suite, without the drafts, correspondence tooling, internal review records or the build for this particular PDF, none of which reproduces a result and some of which would break anonymity. Its README lists what was withheld. The measurement reproduces bitwise under the serving configuration stated in §5, on the two checkpoints for which that configuration was run; the replicate evidence for the rest of the panel is a measured floor, not bitwise determinism.

AI USE STATEMENT

We used generative AI tools (large language models) to draft and revise this version's text, implement the analysis and build-pipeline code, screen the surveyed literature for the reporting matrix of §8, and run adversarial audits of the paper's claims against its artifacts. Each study in the reporting matrix was read end to end against its full text. In a first sweep of 6 studies the negative verdicts were re-checked by a second reader who re-extracted the PDF independently, because several of these papers put load-bearing detail in figures a lossy extraction misses; that reader saw the first reader's evidence rather than working blind. The 7 added later carry a single coding. Both sweeps ship. Every cell carries a verbatim quote or the exact search that returned nothing. A script re-runs the 710 searches it can parse, leaving 6 not-reported cells it cannot. Of those re-run, 0 now returns a body hit beyond the 12 it exempts by name, 12 hit only a reference list, and 377 of 419 quotations re-match verbatim, with 15 unmatched to a source whose extraction is damaged and 27 awaiting a reading by hand. That script, its output and the raw readings, disagreements included, all ship with the supplementary material. We did not use these tools for posing the research question, designing the studies, deciding what to measure or claim, or running the model experiments. The design, every model run, the artifact verification, and the decisions about scope and claims are ours. The longer technical-report version was prepared the same way and carries a disclosure to the same effect; it is not an unassisted document and we do not present it as one. AI-drafted text is constrained by build-time checks: every measurement is interpolated from a released artifact and the build fails if one is typed into the prose instead. We take responsibility for the final content, including everything produced with AI assistance.

REFERENCES

A. Abadie, S. Athey, G. W. Imbens, and J. Wooldridge. When Should You Adjust Standard Errors for Clustering? *arXiv preprint arXiv:1710.02926*, 2017.

H. An and R. Rudinger. Nichelle and Nancy: The Influence of Demographic Attributes and Tokenization Length on First Name Biases. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 388–401, 2023.

H. An, C. Acquaye, C. K. Wang, Z. Li, and R. Rudinger. Do Large Language Models Discriminate in Hiring Decisions on the Basis of Race, Ethnicity, and Gender? *arXiv preprint arXiv:2406.10486*, 2024.

L. Armstrong, A. Liu, S. MacNeil, and D. Metaxa. The Silicon Ceiling: Auditing GPT's Race and Gender Biases in Hiring. *arXiv preprint arXiv:2405.04412*, 2024.

B. Atil, S. Aykent, A. Chittams, L. Fu, R. J. Passonneau, E. Radcliffe, G. R. Rajagopal, A. Sloan, T. Tudrej, F. Ture, Z. Wu, L. Xu, and B. Baldwin. Non-Determinism of "Deterministic" LLM System Settings in Hosted Environments. In *Proceedings of the 5th Workshop on Evaluation and Comparison of NLP Systems (Eval4NLP)*, pp. 135–148, 2025.

Y. Benjamini and Y. Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society B*, 57(1):289–300, 1995.

M. Bertrand and S. Mullainathan. Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. *American Economic Review*, 94(4):991–1013, 2004.

City of New York. Local Law 144 of 2021, in relation to automated employment decision tools. N.Y.C. Admin. Code §§ 20-870 to 20-874, 2021.

R. G. Fryer and S. D. Levitt. The Causes and Consequences of Distinctively Black Names. *The Quarterly Journal of Economics*, 119(3):767–805, 2004.

D. Fu and D. Shi. "You Are Rejected!": An Empirical Study of Large Language Models Taking Hiring Evaluations. *arXiv preprint arXiv:2510.19167*, 2025.

S. M. Gaddis. How Black Are Lakisha and Jamal? Racial Perceptions from Names Used in Correspondence Audit Studies. *Sociological Science*, 4:469–489, 2017.

J. D. Gaebler, S. Goel, A. Huq, and P. Tambe. Auditing the Use of Language Models to Guide Hiring Decisions. *arXiv preprint arXiv:2404.03086*, 2024.

Z. Gao, W. Jiang, and Y. Yan. Can LLMs Hire Fairly? Racial Bias in Resume Screening. *arXiv preprint arXiv:2606.28978*, 2026.

K. Glazko, Y. Mohammed, B. Kosa, V. Potluri, and J. Mankoff. Identifying and Improving Disability Bias in GPT-Based Resume Screening. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*, 2024.

H. He and Thinking Machines Lab. Defeating Nondeterminism in LLM Inference. Thinking Machines Lab: Connectionism, 2025. DOI 10.64434/tml.20250910.

S. A. Hoffstedde, M. Hirota, A. Nadayanur Sathis Kanna, R. Kotani, U. Kumar, G. Trovato, and X. T. Phan. Gender Bias in LLM Hiring Decisions: Evidence from a Japanese Context and Evaluation of Mitigation Strategies. *arXiv preprint arXiv:2606.18649*, 2026.

H. Iso, P. Pezeshkpour, N. Bhutani, and E. Hruschka. Evaluating Bias in LLMs for Job-Resume Matching: Gender, Race, and Education. *arXiv preprint arXiv:2503.19182*, 2025.

J. N. Lahey and R. A. Beasley. Computerizing audit studies. *Journal of Economic Behavior & Organization*, 70(3):508–514, 2009.

L. Lippens. Computer says 'no': Exploring systemic bias in ChatGPT using an audit approach. *Computers in Human Behavior: Artificial Humans*, 2:100054, 2024.

D. Neumark. Detecting Discrimination in Audit and Correspondence Studies. *Journal of Human Resources*, 47(4), 2012.

New York City Department of Consumer and Worker Protection. Notice of Adoption of Final Rule: Automated Employment Decision Tools. 6 RCNY § 5-300 et seq., adopted 6 April 2023, 2023.

H. Nghiem, J. Prindle, J. Zhao, and H. Daumé III. "You Gotta be a Doctor, Lin": An Investigation of Name-Based Bias of Large Language Models in Employment Recommendations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 7268–7287, 2024.

S. Qian, H. V. Pham, T. Lutellier, Z. Hu, J. Kim, L. Tan, Y. Yu, J. Chen, and S. Shah. Are My Deep Learning Systems Fair? An Empirical Study of Fixed-Seed Training. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

M. Sclar, Y. Choi, Y. Tsvetkov, and A. Suhr. Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2310.11324.

P. Seshadri, H. Chen, S. Singh, and S. Goldfarb-Tarrant. Small Changes, Large Consequences: Analyzing the Allocational Fairness of LLMs in Hiring Contexts. *arXiv preprint arXiv:2501.04316*, 2025.

U. Simonsohn, J. P. Simmons, and L. D. Nelson. Specification Curve Analysis. *Nature Human Behaviour*, 2020. doi:10.1038/s41562-020-0912-z.

S. Steegen, F. Tuerlinckx, A. Gelman, and W. Vanpaemel. Increasing Transparency Through a Multiverse Analysis. *Perspectives on Psychological Science*, 11(5):702–712, 2016.

B. C. Z. Tan, S. Khoo, B. N. Doan, Z. Liu, N. F. Chen, and R. K.-W. Lee. Small Changes, Big Impact: Demographic Bias in LLM-Based Hiring Through Subtle Sociocultural Markers in Anonymised Resumes. *arXiv preprint arXiv:2603.05189*, 2026.

A. K. Veldanda, F. Grob, S. Thakur, H. Pearce, B. Tan, R. Karri, and S. Garg. Are Emily and Greg Still More Employable than Lakisha and Jamal? Investigating Algorithmic Hiring Bias in the Era of ChatGPT. *arXiv preprint arXiv:2310.05135*, 2023.

K. Wilson and A. Caliskan. Gender, Race, and Intersectional Bias in Resume Screening via Language Model Retrieval. *arXiv preprint arXiv:2407.20371*, 2024.

J. Yuan, H. Li, X. Ding, W. Xie, Y.-J. Li, W. Zhao, K. Wan, J. Shi, X. Hu, and Z. Liu. Understanding and Mitigating Numerical Sources of Nondeterminism in LLM Inference. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2506.09501.

A THE DESIGN, ANALYSIS BY ANALYSIS

An external reviewer of an earlier version asked for a table naming, for each analysis, the treatment, the outcome, the assignment unit, the observation unit, the block, the target population, the estimator and the uncertainty estimator, because several objections to drafts of this work turned out to be about scope rather than substance, and a design table settles scope questions that prose cannot. The blocks below are that table, generated from the same artifacts as every number above; the section numbers inside them refer to the released technical report, whose structure this version condenses.

§4.1 The demographic effect (the inner estimand)

- *treatment.* applicant first name, White vs Black, held matched on gender
- *outcome.* renormalised decision margin, log P(yes) − log P(no); and its probability-of-superiority transform
- *assignment unit.* first-name pair, assigned to byte-identical résumé text
- *observation unit.* matched pair (name pair × wording × template)
- *block.* model × job posting

- *target population.* the twelve Bertrand–Mullainathan name pairs under the twelve wordings; NOT a population of applicants
- *estimator.* mean paired difference; probability of superiority
- *uncertainty estimator.* percentile bootstrap resampling NAME PAIRS; exact within-race permutation null
- *n.* 432 matched pairs per cell (864 résumé scorings)

§4.1 Dispersion of that effect across instruction wordings

- *treatment.* instruction wording: 6 paraphrases (S1–S6) and 6 null edits (N1–N6) changing no word, only punctuation or line breaks
- *outcome.* the row-1 effect, re-estimated under each wording
- *assignment unit.* none: the wording set is crossed, every wording is applied to every cell
- *observation unit.* (model, posting, wording) cell
- *block.* model $\times$ job posting
- *target population.* the twelve Bertrand–Mullainathan name pairs under the twelve wordings; NOT a population of applicants
- *estimator.* SD across the twelve wordings, and that SD divided by the cell's own effect
- *uncertainty estimator.* bootstrap over name pairs with numerator and denominator resampled together
- *n.* 36 specifications per model

§4.2 Which names were drawn

- *treatment.* the identity of the first name and surname drawn
- *outcome.* renormalised decision margin
- *assignment unit.* first name and surname, crossed
- *observation unit.* matched pair
- *block.* model $\times$ wording
- *target population.* the Bertrand–Mullainathan list as transcribed in names.py; not names in general
- *estimator.* crossed random-effects variance decomposition, σ_first / σ_last / σ_wording
- *uncertainty estimator.* cluster bootstrap over first-name pairs; posterior interval where the model is fitted
- *n.* 1,728 pairs, 24 first names, 8 surnames, 12 wordings

§4.5 Which job was posted

- *treatment.* occupation: Business Analyst, Software Engineer, Registered Nurse
- *outcome.* renormalised decision margin
- *assignment unit.* posting text, fixed per occupation
- *observation unit.* matched pair
- *block.* model $\times$ occupation
- *target population.* three postings chosen for gender-typing contrast; not occupations in general
- *estimator.* per-occupation effect and its across-wording SD
- *uncertainty estimator.* bootstrap over name pairs; dispersion null by permutation across occupations
- *n.* 3 occupations

Appendix B Manipulation check: do the names carry the construct?

- *treatment.* none; observational over the name list
- *outcome.* correlation between the model's within-race name ranking and external perception, callback and SES measures
- *assignment unit.* n/a
- *observation unit.* first name
- *block.* race (correlations computed within race, then pooled)
- *target population.* the drawn first names
- *estimator.* Spearman correlation, ranked inside each race then pooled

- *uncertainty estimator.* stratified re-ranking bootstrap; exact within-race permutation null; MDE by simulation
- *n.* 20,000 bootstrap replicates

Appendix C The same wordings outside hiring

- *treatment.* applicant/subject first name, as in row 1
- *outcome.* renormalised decision margin
- *assignment unit.* first-name pair
- *observation unit.* matched pair
- *block.* model × domain
- *target population.* two further decision domains; a check that the result is not a hiring artefact
- *estimator.* as row 2, per domain
- *uncertainty estimator.* bootstrap over name pairs
- *n.* 8 domain-by-model cells, 5 with an identified effect

Appendix D The same wordings on frontier APIs

- *treatment.* as row 2
- *outcome.* renormalised margin from returned log probabilities
- *assignment unit.* first-name pair
- *observation unit.* matched pair
- *block.* model
- *target population.* frontier APIs that expose log probabilities
- *estimator.* as row 2, with a noise floor subtracted
- *uncertainty estimator.* bootstrap over name pairs; repeat-based noise floor
- *n.* 4 frontier APIs returning log probabilities (a fifth, verdict-only, is the row below)

Appendix D The verdict-only frontier arm

- *treatment.* as row 2
- *outcome.* thresholded yes/no verdict: this endpoint returns no next-token distribution to any caller, so no margin exists
- *assignment unit.* first-name pair
- *observation unit.* matched pair
- *block.* model
- *target population.* one frontier API without log probabilities; it is here to show what a binary outcome can resolve
- *estimator.* difference in verdict rates
- *uncertainty estimator.* bootstrap over name pairs; repeated cells used to measure disagreement directly
- *n.* gemini-3.1-flash-lite, 220 matched pairs, 440 calls

Appendix E Which quantization was downloaded

- *treatment.* weight quantization of the same checkpoint
- *outcome.* the row-1 effect
- *assignment unit.* n/a: the same grid re-run on a second build
- *observation unit.* matched pair
- *block.* model
- *target population.* two checkpoints for which a second quantization was obtainable
- *estimator.* shift in the effect, divided by the wording SD
- *uncertainty estimator.* bootstrap over name pairs
- *n.* 2 checkpoints re-run at a second quantization

§5.2 How the requests were served

- *treatment.* request batching and key-value cache residency
- *outcome.* cell-level agreement between byte-identical repeats
- *assignment unit.* n/a: repeated measurement of a fixed grid
- *observation unit.* cell (one prompt, one scoring)
- *block.* model, and separately server process / session
- *target population.* this serving stack; NOT a claim about deployed serving
- *estimator.* fraction of cells agreeing exactly; bound on the induced shift in the effect
- *uncertainty estimator.* reported as a bound, not an absence
- *n.* 4 checkpoints

§6.1 The resampling unit

- *treatment.* n/a: an analysis-time comparison of two estimators on one dataset
- *outcome.* width of the 95 % interval on the row-1 effect
- *assignment unit.* n/a
- *observation unit.* matched pair (row) vs first-name pair (cluster)
- *block.* model
- *target population.* this crossed design; the unit follows the assignment process and is not universal
- *estimator.* ratio of interval widths, row bootstrap ÷ cluster bootstrap
- *uncertainty estimator.* both intervals computed at the same replicate count and seed
- *n.* 20,000 bootstrap replicates (8,000 for contrasts)

§6.2 The reporting scale and the operating point

- *treatment.* n/a: an analysis-time comparison
- *outcome.* the log-odds-to-percentage-point conversion factor
- *assignment unit.* n/a
- *observation unit.* model
- *block.* model
- *target population.* this panel; the quantity is defined per model and does not transfer
- *estimator.* 0.25 ÷ the model's own mean p(1−p); and the overstatement realised on a measured effect
- *uncertainty estimator.* realised value reported only for the checkpoints whose effect is distinguishable from zero
- *n.* 4 checkpoints

§6.3 Which résumé was paired with which

- *treatment.* n/a: an analysis-time degree of freedom that exists only because each résumé is scored alone
- *outcome.* probability of superiority under a re-pairing
- *assignment unit.* n/a; in a blocked correspondence design this row does not arise at all
- *observation unit.* a whole pairing of the grid
- *block.* gender (re-pairings drawn within gender)
- *target population.* this design; not designs that preconstruct and assign profiles within a prespecified block
- *estimator.* SD of the statistic over random re-pairings; and the best-worst range under a maximum-weight matching
- *uncertainty estimator.* permutation over pairings
- *n.* 20,000 random re-pairings

§7 Position or fragmentation?

- *treatment.* edits that move the name's token position, with delimiter fragmentation held equal, and vice versa
- *outcome.* renormalised decision margin
- *assignment unit.* edit condition, applied to a fixed grid

- *observation unit.* contrast (model × inference mode × condition pair)
- *block.* model × inference mode
- *target population.* this edit family; a mechanism probe, not an estimate of a population quantity
- *estimator.* per-contrast difference
- *uncertainty estimator.* Benjamini–Hochberg across the declared contrast family
- *n.* 6 checkpoints

§8 What the field reports

- *treatment.* n/a: a literature survey
- *outcome.* whether each audit reports each of the design fields above
- *assignment unit.* n/a
- *observation unit.* study × field cell
- *block.* field
- *target population.* LLM hiring audits readable in full text; the frame is stated and the exclusions are listed
- *estimator.* count of studies reporting, per field, with an evidence quote or a negative search behind every cell
- *uncertainty estimator.* 'partial' counts as not reported, because the question is whether a reader could reconstruct the choice
- *n.* 13 audits, 22 fields

§9.1 The screening rule's false-positive rate

- *treatment.* n/a: a calibration of the paper's own recommendation
- *outcome.* rejection rate under the global null
- *assignment unit.* n/a
- *observation unit.* simulated wording set
- *block.* n/a
- *target population.* equicorrelated normal wordings, as a function of ρ
- *estimator.* closed-form rejection probability of the stated rule
- *uncertainty estimator.* exact, not simulated
- *n.* closed-form, equicorrelated normal

Appendix C Token matching of the pair (companion paper)

- *treatment.* restriction to token-length-matched name pairs
- *outcome.* the row-1 effect, before and after restriction
- *assignment unit.* first-name pair
- *observation unit.* matched pair
- *block.* model (the tokenizer differs per model)
- *target population.* the Bertrand–Mullainathan list under four tokenizers
- *estimator.* effect on the restricted subset vs the full grid
- *uncertainty estimator.* three procedures reported, which disagree at 0.05; the disagreement is the result
- *n.* 3 independent first-name pairs survive balancing

A.1 Inclusion frame and exclusions

The target population of §8 is LLM hiring audits we could read in full text and code under the protocol described in the AI use statement. One study meeting the topical criterion is excluded. Lippens (2024) is cited in §2 and §4, but it was identified after that protocol had run and was coded by hand outside it; its provenance is narrower as well, the article body without the supplementary tables, so several of its cells could only be coded partial. Merging it would have put two coding methods under one heading. Its completed coding ships with the supplementary material as a proposed row, unmerged, so the exclusion can be checked rather than taken on trust. It does not carry the section's headline: on that coding the study reports no token matching and no cache policy, describes its

submission path only partially, and uses a single wording, so dispersion across wordings does not apply to it. It would move the denominators, and the counts in §8 should be read as counts over the thirteen coded under the protocol.

B Extended results and diagnostics

The design table is Appendix A of this submission and the per-wording estimates are Appendix C; the released technical report carries fuller versions of the rest: variance components, the occupation and second-task replications, the frontier-API arm with its censoring analysis, the quantization study, the mechanism panel with its minimum detectable effects on the name-pair unit, the construct-validity probes, and the survey coding sheet with per-cell evidence quotes. This appendix is deliberately thin at submission: each claim above cites its artifact, and the artifacts are released.

C Per-wording estimates

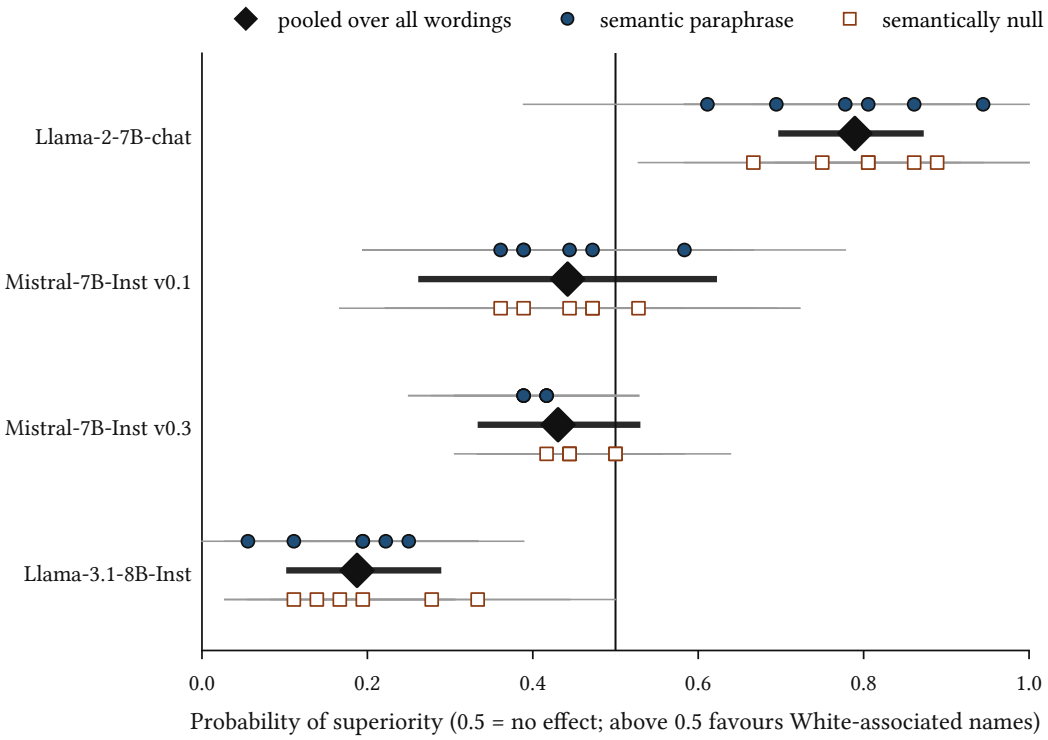


Figure 6: **The same question, asked twelve ways, of the same candidates.** Per-wording demographic effects with the pooled estimate per checkpoint, on the probability-of-superiority scale; circles are paraphrases, open squares are semantically null edits. Every interval resamples name pairs.