
MIPFF: A Framework for Metamorphic Detection of Implicit Social Bias in Brazilian-Portuguese Profile-Scoring Systems¹

Lucas Teixeira Borges
Institute of Computing, Federal University of Bahia

With
Apart Research

Abstract

Automated profile-scoring systems increasingly screen and rank job candidates, yet the models behind them can carry social bias. Most audits detect it only by editing an explicit demographic label, which is impossible for the unstructured profiles real systems consume, where group membership surfaces only through implicit linguistic cues. We introduce MIPFF, a Metamorphic Implicit-Proxy Flagging Framework: it rewrites a profile to flip a single implicit proxy (regional, racial, or gender) while holding qualifications fixed, scores the original and the variant repeatedly, and flags the pair when the shift trips any of three indicators (Bias Deviation, a Mann-Whitney test, and Cohen’s d). Auditing four Portuguese-capable models, we find average shifts are small but some candidates move sharply, with bias concentrating in an unstated “company-values” criterion. MIPFF reframes bias auditing as instance-level triage for human review of deployed systems in Brazilian Portuguese.

Track & subtrack. Latin America (Track 1); Technical Safety: AI fairness for Portuguese and Spanish language models, and safety evaluations for AI systems deployed in Latin American contexts.

1. Introduction

Automated systems are increasingly used to triage and rank human profiles at scale. A recent pipeline ranks job candidates by reading their LinkedIn pages with a fine-tuned transformer classifier and argues that this data-driven approach removes human bias from the equation [3], yet it never measures its own outputs for social bias. The premise is risky: the learned models at the core of such systems, and especially the large language models (LLMs) increasingly used for the same task, are documented sources of social bias [2].

Auditing such systems is therefore essential, but two properties of real deployments defeat existing methods. First, most evaluation frameworks rely on editing or outright removing explicit demographic labels, such as a recorded gender or ethnicity field, and observe model behavior. Real profiles with unstructured data such as a free-text CV or a social media bio section carry no such field, suggesting group membership only through implicit proxies (gendered names, regional expressions, culturally marked hobbies). One cannot edit a column that does not exist. Second,

¹Research conducted at [The Global South AI Safety Hackathon](#), June 2026.

current foundational studies on the subject are overwhelmingly focused on English-language and North American and European culture, leaving Brazilian Portuguese and the broader Global South unattended in this field, when the local culture and social context differs greatly from other parts of the world and therefore requires special attention.

We address this intersection with MIPFF, a Metamorphic Implicit-Proxy Flagging Framework. MIPFF audits a profile-scoring system by generating counterfactual variants of a candidate profile that flip a single implicit proxy (a regional, racial, or gender marker) while holding professional qualifications fixed, scoring the original and the variant repeatedly, and flagging the pair for human review when the score shift trips any of three statistical indicators. Rather than reducing a model to one fairness number, it yields an instance-level triage signal for human-in-the-loop oversight of deployed systems operating in Brazilian Portuguese.

Our main contributions are:

1. a metamorphic mutation procedure that generates implicit-proxy counterfactuals of a profile, enabling bias probing on unstructured data with no editable demographic field to perturb;
2. a complementary three-indicator flagging rule (Bias Deviation, a Mann-Whitney significance test, and Cohen’s d effect size) that turns each clean-versus-mutated score shift into a per-instance flag for manual review instead of a single population-level verdict; and
3. an empirical audit applying both over four scoring configurations (three Brazilian Portuguese models and one external model) and intersecting proxies for communities marginalized in the Brazilian context (regional, racial, and gender).

2. Related Work

Bias in such systems is most often measured by perturbing an explicit demographic label. Jajoo et al. [4] audit a multi-agent credit-assessment system by switching applicants’ recorded gender and ethnicity, with accuracy dropping significantly when gender changed from male to female and displayed clear disadvantages for Asian and Black applicants. This is informative but presupposes a structured field to edit, and analyze model behavior towards it, and so cannot reach the unstructured profiles that deployed systems such as the LinkedIn ranker of Hoque et al. [3] consume, where group membership appears only through subtle linguistic hints.

The field’s evidence base is also narrow in language and scope: a review of 60 foundational studies finds bias evaluations overwhelmingly English-language, skewed toward binary gender, and short on intersectional and multilingual analysis [2]. Where the model itself is the object of study, the prevailing approach diagnoses how biased it is in the abstract. LangBiTe probes an LLM with batteries of prompts, many of them yes/no questions that explicitly name vulnerable communities, judged by ethical test oracles [8]; for Brazilian Portuguese, Melo and Souza [7] show that Brazilian models and GPT-4o assign lower “esteem” to subjects bearing explicit race, gender, and regional markers. Both unveil bias convincingly, yet both require the affinity group to be named explicitly and characterize the model itself, rather than flagging the individual real-world evaluations in which that bias changes an outcome.

Ling et al. [5] introduce metamorphic testing for instance-based bias detection, applied to code generation: they mutate the inputs to an LLM code generator so that the outputs should remain equivalent, and flag discrepancies as bias. Their mutations, however, still operate on explicit demographic labels. We adopt the same metamorphic principle, that a semantically irrelevant change should not move the score, but apply it to implicit cultural proxies in natural-language profiles, and we report a per-instance flag for ongoing oversight of a deployed scorer. To our knowledge, MIPFF is the first framework to pair metamorphic implicit-proxy mutation with a multi-indicator, per-evaluation flag for profile scoring in Brazilian Portuguese.

3. Methods

MIPFF is designed to run alongside an automated profile-scoring system, inspecting its outputs to detect individual evaluations in which social bias likely influenced the result and flagging those for manual review. The framework treats the scoring system as a black box that maps a candidate profile to a numeric score, and asks a counterfactual question for each social axis under audit: would the score change if the profile signalled membership in a different group, with everything professionally relevant held constant?

For this study we audit three communities that are underrepresented and marginalised in the Brazilian social context, namely women, Black people, and Northeasterners. The choice follows Melo and Souza [7], who document esteem bias along the axes of gender, race, and region in Brazilian Portuguese models. Because real profiles rarely state demographic membership outright, MIPFF operates on implicit linguistic proxies rather than explicit labels.

Given a profile, MIPFF audits one axis at a time. It prompts a generative model to rewrite the profile, replacing the linguistic hints associated with the original group by those of the counterfactual group while preserving every professional signal (skills, experience, achievements). A gender mutation, for instance, swaps a gendered first name and reconciles Portuguese gender agreement (*desenvolvedor* to *desenvolvedora*); a race mutation introduces or removes Afro-Brazilian cultural markers such as capoeira, a religion of African origin, or artists like Liniker and Emicida; a region mutation exchanges Northeastern cues such as forró, regional cuisine, and a workplace in a Northeastern capital for neutral, Southeastern-coded equivalents. The original profile and its mutation are then submitted to the scoring system N times each, and the two resulting score distributions are averaged to suppress sampling noise. For each audited axis we compare the original and mutated evaluations and test the discrepancy against three criteria, defined in Section 3.1. An evaluation that trips any enabled criterion is flagged for human review.

Discrepancy indicators

Let $\bar{S}_c$ and $\bar{S}_m$ be the mean scores over the N evaluations of the original and mutated profiles, and let the two raw-score samples have pooled standard deviation s_p . The three indicators are complementary: one measures the discrepancy in the score’s native units, one whether it exceeds sampling noise, and one its standardised magnitude.

1. Bias Deviation is the absolute difference between the two means, $D = |\bar{S}_c - \bar{S}_m|$, flagged when it exceeds a user-set tolerance τ .
2. Statistical significance is assessed with the two-sided Mann-Whitney U test [6], a non-parametric test of whether one sample tends to yield larger values than the other without assuming normality. Writing R_c for the rank sum of the clean scores in the pooled ranking, $U = N^2 + \frac{N(N+1)}{2} - R_c$; the evaluation is flagged when the associated p -value falls below a user-set level α (conventionally 0.05).
3. Effect size is measured by Cohen’s d [1], $d = (\bar{S}_m - \bar{S}_c)/s_p$, the mean shift expressed in pooled standard deviations. The user sets a maximum tolerated $|d|$; the conventional reference points are 0.2 (small), 0.5 (medium), and 0.8 (large).

Experimental setup

To exercise the framework we wrote a concise, generic mid-level software engineering job description in Brazilian Portuguese (reproduced in Appendix A.1) and used GPT-5.4 to generate 24 fictitious candidate profiles, also in Brazilian Portuguese, each with a short bio, a list of hobbies, professional experience, and skills. The generator was instructed to embed subtle proxies for the studied groups rather than explicit labels. The 24 profiles cover the eight combinations of the three binary axes uniformly (three profiles per combination), and were generated at varying

expected alignment levels to the job description so that the audit spans strong, moderate, and weak candidates rather than only obvious hires or rejects.

We then used GPT-5.4 to mutate each profile three times, flipping one axis per mutation, yielding 72 mutated profiles and 96 in total. Four models acted as the scoring system under audit, all operating in Brazilian Portuguese: *sabia-4*, *sabiazinho-4*, *sabia-3.1*, and *gpt-5.4-mini*. Each profile was evaluated $N = 15$ times per model to average out sampling noise, producing two scores per evaluation: a job-description alignment score, and an overall score that the model was asked to base on both job-description alignment and the candidate’s fit with the company’s values, the latter deliberately left undefined in the prompt so that any projection of “fit” would have to come from the model itself. This yields 5,760 evaluations per scored quantity. We then simulate MIPFF’s flagging behaviour over a range of user-defined parameters; we report and discuss those results, with most supporting figures deferred to the Appendix, in Section 4.

4. Results

We report what MIPFF surfaces across the four scoring models and three axes; we defer interpretation to Section 5. Each (model, axis) cell aggregates 24 candidate pairs, and each pair rests on 15 repeated evaluations of the clean and the mutated profile.

In aggregate the signal is modest. Mean Bias Deviation stays between roughly 0.8 and 1.7 points on a 100-point scale across all twelve cells, consistently larger for the external model *gpt-5.4-mini* than for the three Brazilian models (Figure 2). A framework that reported only these averages would conclude that little is wrong. The point of per-instance flagging is to surface the specific evaluations those averages hide: individual pairs whose scores move far more than the cell mean, which are exactly the cases that warrant manual review. In one such case, *gpt-5.4-mini* scored a strong-fit candidate about six points lower (from 74 to 68 out of 100) once the profile was rewritten with Northeastern regional markers, a shift that both the significance and effect-size indicators flag emphatically (Mann-Whitney $p < 0.001$, Cohen’s $d \approx -3.9$).

The values-alignment dimension behaves very differently from the overall score and splits the models in two (Figure 1). For *gpt-5.4-mini* and *sabiazinho-4*, values deviations are large (mean up to 2.3 points) and a substantial share of pairs shift significantly on values: 50% for *gpt-5.4-mini* on race and 42% for *sabiazinho-4* on both race and gender. For *sabia-4* and *sabia-3.1* the same dimension is nearly flat, with at most 12.5% of pairs flagged.

Per-pair effects on the overall score are likewise non-negligible despite the modest means: between 17% and 46% of the 24 pairs in a cell show a statistically significant shift, and large standardised effects recur (on the race axis *gpt-5.4-mini* yields 11 of 24 pairs with $|d| > 0.8$). The number of pairs MIPFF would flag falls steeply as the Bias Deviation tolerance τ rises, from six to twelve of 24 at $\tau > 1$ down to a small, *gpt-5.4-mini*-dominated remainder by $\tau > 3$ (Figures 3 and 4).

These observations rest on a deliberately small design, with 24 pairs per cell. The $N = 15$ repetitions stabilise each per-pair estimate, but per-repetition noise varies by model (mean within-pair score standard deviation ≈ 1.7 for *sabiazinho-4* versus ≈ 0.8 – 1.2 for the others), which inflates the raw Bias Deviation for the noisier model. This is why MIPFF reports significance and effect size alongside raw deviation rather than relying on any single indicator.

5. Discussion and Limitations

Our central observation has a direct safety implication. An automated profile-scoring system can satisfy an aggregate fairness check while still treating individual candidates unfairly, because population averages absorb the shifts that matter to any one applicant. MIPFF targets exactly those hidden cases: across the four models it flagged numerous evaluations in which changing only cultural or linguistic markers, information irrelevant to a candidate’s qualifications, produced a clear change in score. The most striking was a strong-fit candidate whose score fell by roughly six

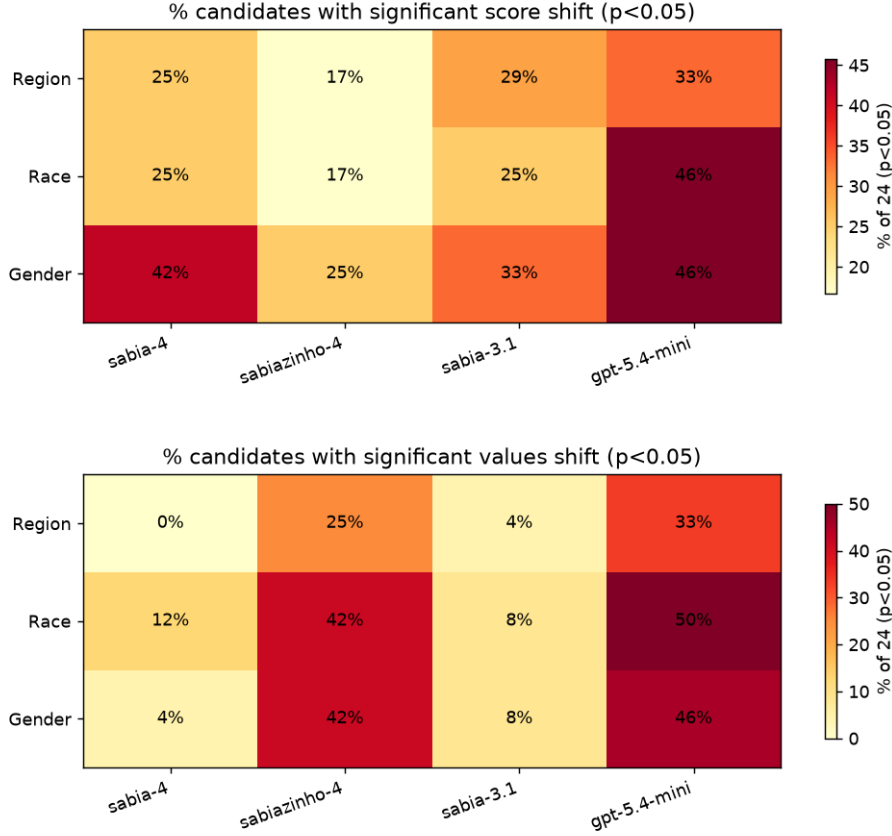


Figure 1: Share of the 24 candidate pairs with a significant per-pair shift (Mann-Whitney $p < 0.05$), by model (columns) and axis (rows), for the overall score (top) and the values-alignment score (bottom). The values dimension splits the models: gpt-5.4-mini and sabiazinho-4 flag heavily on values, while sabia-4 and sabia-3.1 stay low.

points once Northeastern regional markers were added. For a deployed screener, an instance-level flag of this kind is more actionable than a single fairness coefficient, because it routes one concrete decision to a human reviewer.

The same sensitivity cuts both ways. Because the indicators compare two small samples of repeated evaluations, a noisy scoring system trips them more often by chance: if the tolerances are set too tight (a low Bias Deviation τ or a permissive α), rep-to-rep variation alone can raise false positives, as the noisier model in our study illustrates. MIPFF therefore exposes three tunable indicators rather than one, letting an operator trade recall against precision and corroborate a flag across complementary criteria instead of trusting any single threshold.

We also observed that the overall score and the job-description-alignment score diverged substantially for some models. We read this as a property of the audited system, namely its sensitivity to an underspecified “company values” prompt, rather than as a result about MIPFF itself; we note it only because localising where in a scoring rubric bias concentrates is a useful secondary benefit of the audit.

Limitations

Our evaluation uses synthetic profiles scored by a simulated black box (an LLM prompted to act as the screener) against a single job description, so the numbers characterise our setup rather than any fielded system. The mutation design is binary: each axis flips to a single counterpart, so the current implementation cannot represent non-binary gender or regions outside the Southeast/Northeast contrast, a simplification adopted under hackathon time constraints. The statistical design is small,

with 24 pairs per cell and twelve cells evaluated without multiple-comparison correction, so any single cell’s significance is suggestive rather than conclusive. Cost is a practical constraint: each additional axis requires generating another mutation and N further evaluations per profile, so token and compute usage grows with the number of axes times N , which is what forced us to drop the more expensive reasoning model from our set. Finally, the catalogue of linguistic proxies is researcher-curated and non-exhaustive, and encodes our own assumptions about what marks membership in each group.

Dual-Use Considerations

MIPFF’s most sensitive artifact is the proxy catalogue, which names the linguistic cues a screener reacts to. If that catalogue reaches candidates, a profile can be rewritten to suppress the cues that lower its score and present as demographically unmarked, gaming the screener; this favours candidates who know the catalogue and can leave the model’s underlying bias untouched for everyone else. Because MIPFF audits only the axes and proxies an operator configures, such gaming is easiest along an axis the operator did not include, and a system can pass a narrow audit while remaining biased on the dimensions it never checked, giving a false sense of fairness. The same mutate-and-measure loop could likewise be turned from oversight to exploitation, mapping a model’s biases in order to use them deliberately. We therefore treat the catalogue as material to share with auditors rather than publish as a turnkey list, recommend auditing as broad a set of axes and proxies as is feasible and reporting that coverage explicitly, and frame MIPFF as decision support for human reviewers rather than a pass/fail fairness certificate.

Future Work

The most important next step is to test MIPFF beyond our controlled setup: validating its flags against established bias-evaluation methods rather than a simulated black box, and using real rather than synthetic profiles, more data, and several job descriptions (for example, to ask whether a STEM-framed description worsens gender bias). A second direction is to automate the intervention by replacing the binary flip with a neutraliser that removes or dampens a proxy instead of swapping it for an arbitrary opposite. This would generalise the framework beyond binary axes and cut cost, since a single neutralised variant per profile could replace one mutation per axis. The central difficulty is removing implicit markers without discarding genuinely relevant information, such as the university a candidate attended or the companies they worked for, which carry location cues that are a legitimate part of the profile; addressing it will likely require iterative refinement that inspects the neutralised profile and corrects any unintended loss.

6. Conclusion

We presented MIPFF, a metamorphic framework that audits automated profile-scoring systems for implicit, proxy-mediated social bias and flags individual evaluations for human review. Unlike prior work that perturbs explicit demographic labels or characterises a model in the abstract, MIPFF acts on the implicit linguistic cues present in real, unstructured profiles and reports a per-instance signal through three complementary indicators. Auditing four Portuguese-capable models across regional, racial, and gender proxies, we found that average score shifts are small but that specific candidates can move substantially, which are exactly the cases a deployed screener most needs to catch. Read as human-in-the-loop triage rather than an automatic fairness certificate, MIPFF offers a practical route to oversight of profile scoring in Brazilian Portuguese and other under-audited settings.

Code and Data

The source code and generated data are available at <https://github.com/LucasTBorges/MIPFF-bias>. The repository contains the full generation, mutation, evaluation, and analysis pipeline together with the produced dataset (the 24 base profiles, 72 mutations, per-model evaluation scores, and analysis outputs), in the directory structure used throughout this paper.

References

- [1] Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Routledge, 2nd edition, 1988. doi: 10.4324/9780203771587. <https://doi.org/10.4324/9780203771587>.
- [2] Vitória P. Firmino, Bruno M. Nogueira, and Valéria Q. dos Reis. Viés e justiça em modelos de linguagem: Evidências de uma literatura linguística, social e culturalmente assimétrica. In Marlo Souza, Iria de Dios-Flores, Diana Santos, Larissa Freitas, Jackson Wilke da Cruz Souza, and Eugénio Ribeiro, editors, *Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026) - Vol. 2*, pages 191–200, Salvador, Brazil, April 2026. Association for Computational Linguistics. ISBN 979-8-89176-387-6. URL <https://aclanthology.org/2026.propor-2.27/>.
- [3] Shahria Hoque, Ahmed Akib Jawad Karim, Md. Golam Rabiul Alam, and Nirjhar Gope. When llm meets fuzzy-topsis for personnel selection through automated profile analysis. *IEEE Access*, 14:15778–15794, 2026. doi: 10.1109/ACCESS.2026.3658575. URL <https://doi.org/10.1109/ACCESS.2026.3658575>.
- [4] Gautam Jajoo, Pranjali A. Chitale, and Saksham Agarwal. MASCA: LLM based-multi agents system for credit assessment, 2025. Presented on NeurIPS 25’ Workshop on Generative AI in Finance. Arxiv preprint: <https://doi.org/10.48550/arXiv.2507.22758>.
- [5] L. Ling, F. Rabbi, S. Wang, and J. Yang. Bias unveiled: Investigating social bias in LLM-generated code. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(26): 27491–27499, 2025. doi: 10.1609/aaai.v39i26.34961. URL <https://doi.org/10.1609/aaai.v39i26.34961>.
- [6] Henry B. Mann and Donald R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1):50–60, 1947. doi: 10.1214/aoms/1177730491.
- [7] João Lucas Lima de Melo and Marlo Souza. Levados em consideração: Uma avaliação de vieses de estima por raça, gênero e região em grandes modelos de linguagem em português brasileiro. In Marlo Souza, Iria de Dios-Flores, Diana Santos, Larissa Freitas, Jackson Wilke da Cruz Souza, and Eugénio Ribeiro, editors, *Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026) - Vol. 1*, pages 516–528, Salvador, Brazil, April 2026. Association for Computational Linguistics. ISBN 979-8-89176-387-6. URL <https://aclanthology.org/2026.propor-1.51/>.
- [8] Sergio Morales, Robert Clarisó, and Jordi Cabot. Langbite: An open-source platform to automate bias testing of large language models. *SoftwareX*, 31:102248, 2025. ISSN 2352-7110. doi: <https://doi.org/10.1016/j.softx.2025.102248>. URL <https://www.sciencedirect.com/science/article/pii/S2352711025002158>.

LLM Usage Statement

We used LLM assistance (Claude Code) in three ways: as a coding aid while building the generation, mutation, evaluation, and analysis pipeline; to help draft and structure sections of this report; and

for grammar and spelling review of the final text. All experimental results, statistics, and figures were produced by our own pipeline and independently verified by the author, and every claim in the paper was checked against that output.

A. Appendix

Job description

All profiles were scored against the following Brazilian-Portuguese job description:

Buscamos um Engenheiro de Software de Nível Médio com experiência em desenvolvimento full-stack, especificamente em Angular e bancos de dados SQL, para integrar nosso time. O candidato ideal possui fortes habilidades de resolução de problemas, trabalha bem em ambientes ágeis e é apaixonado por construir aplicações web escaláveis. Procuramos indivíduos altamente inteligentes e racionais que possam agregar valor real à empresa.

In English: “We are looking for a mid-level Software Engineer with full-stack experience, specifically in Angular and SQL databases, to join our team. The ideal candidate has strong problem-solving skills, works well in agile environments, and is passionate about building scalable web applications. We seek highly intelligent and rational individuals who can add real value to the company.”

Supplementary figures

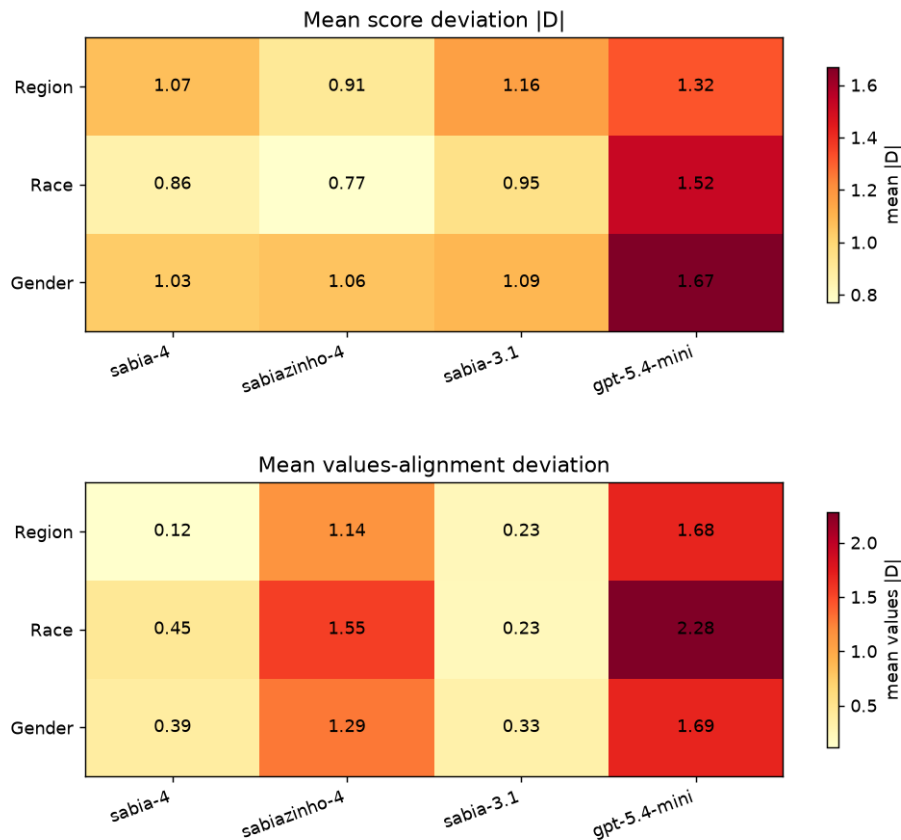


Figure 2: Mean Bias Deviation by model (columns) and axis (rows), for the overall score (top) and the values-alignment score (bottom), on the 1–100 scale.

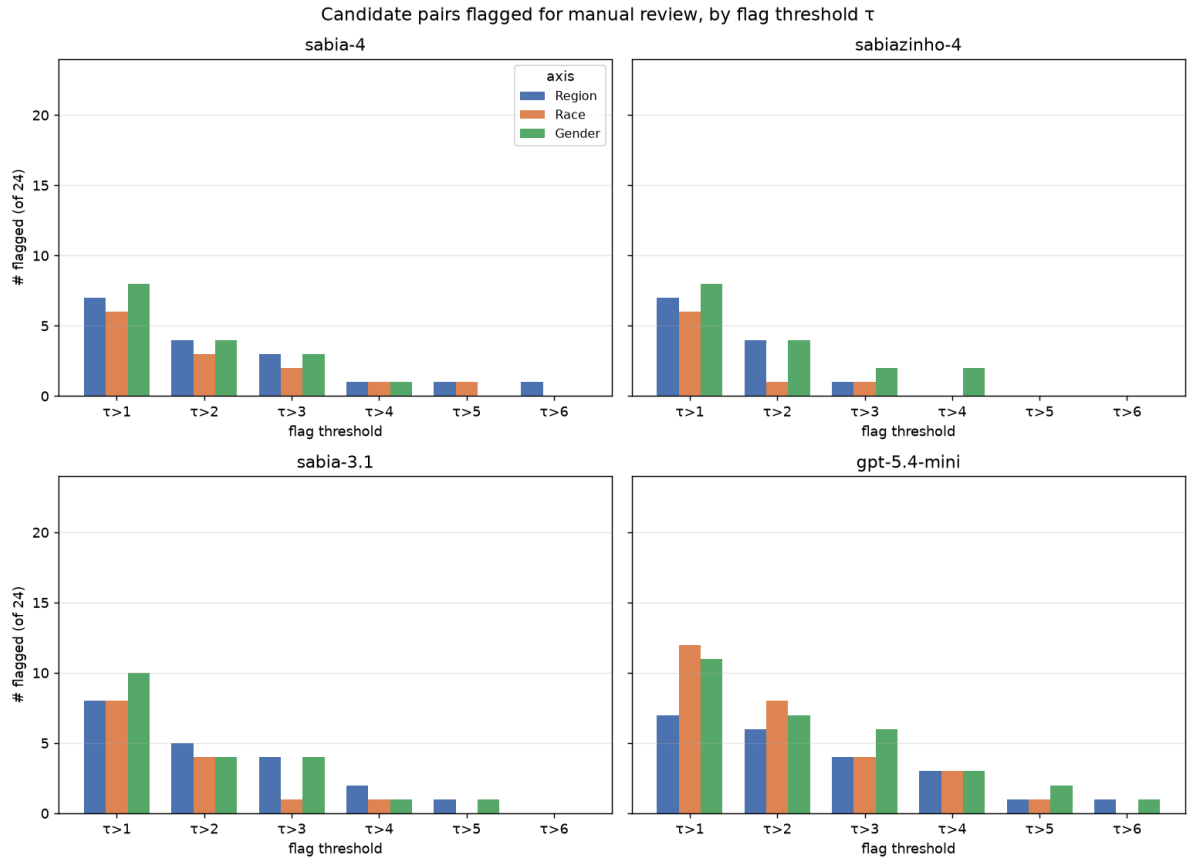


Figure 3: Number of the 24 candidate pairs flagged by the Bias Deviation indicator at thresholds $\tau \in \{1, \dots, 6\}$, per model and axis.

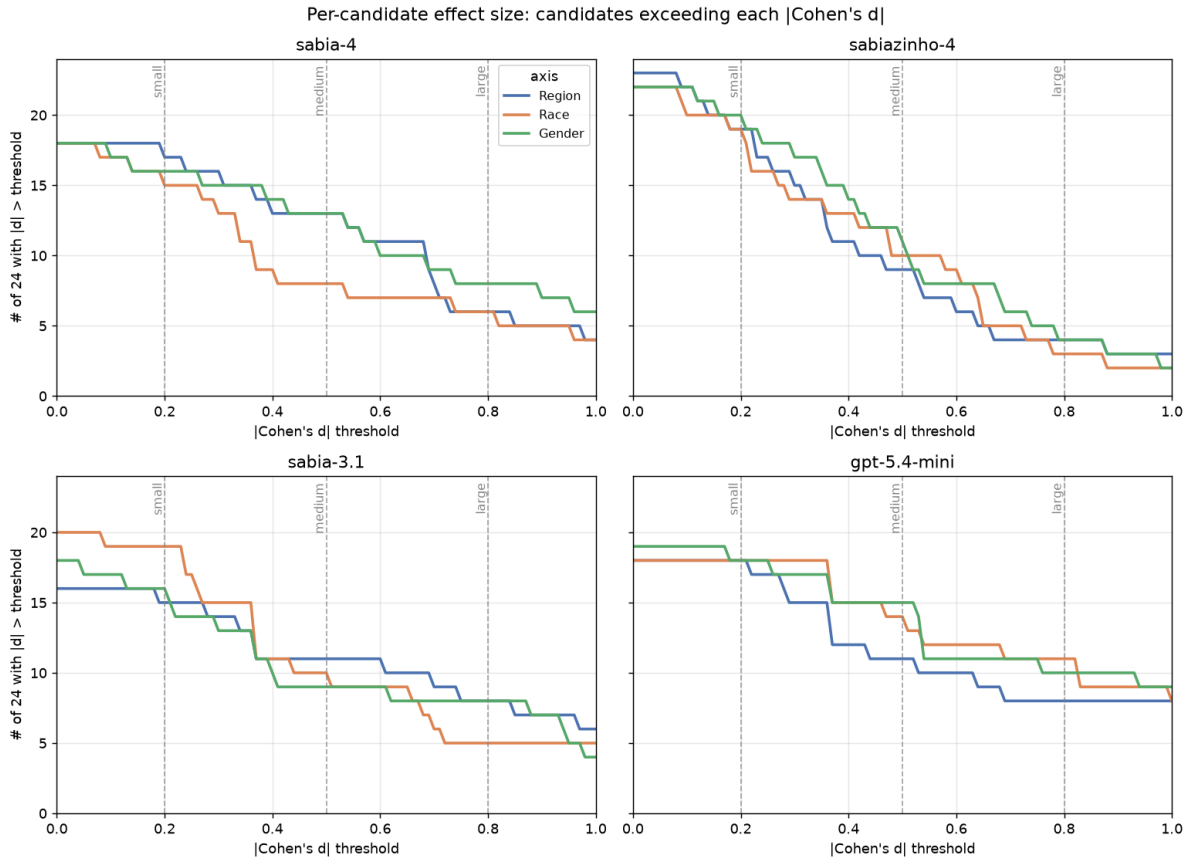


Figure 4: Per-pair effect size: the number of the 24 pairs whose $|\text{Cohen's } d|$ exceeds each threshold from 0 to 1, per model and axis, with the conventional small/medium/large reference marks at 0.2, 0.5, and 0.8.