Equity Language Without Equity: Three Structural Markers of Persistent Gender Asymmetry in GPT Models

Anonymous ACL submission

Abstract

Safety-aligned language models report lower toxicity scores across model generations, but toxicity classifiers measure surface-form harm. We identify three structural markers of gender representation asymmetry that survive across safety alignment boundaries, and in two cases intensify, in the OpenAI GPT lineage (450,000 completions, 15 models, GPT-2 through to GPT-5). First, *definitional othering*: women-directed completions reference men at 30.5-67.2% across all five generations while men-directed completions reference women at near-zero rates (0.0-12.2%), with the gap *largest* at the GPT-4 alignment boundary (+60.3 pp). Second, *topic space contraction*: the GPT-4 alignment boundary reduces the unique topic ratio for women relative to men (W/M: 0.91 to 0.58, a 36% relative decline), narrowing representational range. Third, *equity reframing*: GPT-4 and GPT-5 apply equity-coded language to women-directed completions at twice the rate applied to men (37.3% vs. 16.9% at GPT), co-occurring with lower harm scores: surface substitution, not substantive change. None of the three markers are detectable by standard evaluation. Together they constitute a structural harm signature that accompanies safety alignment.

1 Introduction

Safety training for large language models is evaluated using surface-form classifiers. Across the OpenAI GPT lineage, published benchmarks show monotonic improvement: explicit harmful content declines from GPT-2 to GPT-5. This improvement is real but incomplete. Toxicity classifiers measure surface harm; they do not measure structural representation asymmetry. We define *structural representation asymmetry* as systematic differences in how demographic groups are framed, referenced, or allocated representational range in model outputs: differences that are relational (emerging from cross-condition comparison) and sub-lexical (not encoded in individual harm-flagged tokens).

Structural asymmetry has two defining properties. First, it is relational: it emerges from comparison across demographic conditions, not from absolute scores within a single condition. Second, it is invisible to single-axis evaluation: a completion that scores low on toxicity may nonetheless define one demographic group in terms of another, narrow that group's representational range, or apply equity-coded language asymmetrically to substitute for equitable representation. All three properties are absent from the feature space of Detoxify-class classifiers.

We present three findings that document this structural harm signature across 15 models and five GPT generations. Each finding is derived from a distinct measurement approach; each is invisible to standard toxicity evaluation; and each reveals a dimension of structural asymmetry that either persists through or is introduced by safety alignment. The three are unified by the same structural logic: safety alignment optimises against lexical harm signals, leaving relational and representational asymmetries (which have no lexical surface) undisturbed or amplified.

Contributions. (1) We identify *definitional othering* as a structural property of GPT generation: women-directed completions consistently reference men at rates far exceeding the converse, with the gap widening at the GPT-4 alignment boundary. (2) We quantify *topic space contraction*: women's representational range declines 36% relative to men's at the alignment boundary while men's expands. (3) We document *equity reframing*: safety-aligned models apply equity-coded language to women-directed completions at double the rate applied to men, and this framing correlates negatively with harm scores, functioning as surface substitution. (4) We demonstrate that all three markers

evade Detoxify-class toxicity scoring, showing that surface-form harm detection is insufficient to identify discriminatory content.

2 Related Work

Gender bias in language model output has been studied across generation (Sheng et al., 2019), coreference (Zhao et al., 2018), and semantic association (Bolukbasi et al., 2016). The dominant evaluation paradigm measures bias as deviation from parity in absolute scores within a single demographic condition (Blodgett et al., 2020).

Structural approaches to representation are less common. Noble (2018) documents how search systems produce asymmetric representational frames without explicit slurs. Collins (1990) develops the concept of definitional othering: the structural tendency to define marginalised groups in relation to dominant groups, which we operationalise as cross-demographic referencing asymmetry. Crenshaw (1989) situates such structural asymmetries within systems of differential visibility and epistemic standing.

Safety alignment evaluation typically reports toxicity score trajectories (Perez et al., 2022; Ganguli et al., 2022; Ouyang et al., 2022) without distinguishing surface improvement from structural change. Weidinger et al. (2021) provide a taxonomy of 21 LLM risk categories including representation harms, but existing benchmarks operationalise few of these. Blodgett et al. (2020) argue that NLP bias research must specify who is harmed, by what mechanism, and with what evidence; our three-finding structure directly follows this requirement. Kirk et al. (2021) document intersectional occupational biases in GPT-family models, the closest prior work to our empirical scope; we extend their empirical direction to cross-demographic referencing, topic diversity, and equity framing.

3 Dataset and Methodology

Dataset. We analyse 450,000 completions from 15 models spanning the OpenAI GPT lineage: GPT-2 (base, medium, large, XL), GPT-3 (davinci:2020-05-03, text-davinci:002, text-davinci:003), six GPT-4 variants (gpt-4-0613, gpt-4-1106-preview, gpt-4-0125-preview, gpt-4-turbo-2024-04-09, gpt-4o-2024-05-13, gpt-4o-2024-08-06), GPT-5 without context, and GPT-5 with context. GPT-5 without a system prompt responds as a conversational assistant rather than a sentence completion model, producing outputs structurally unlike earlier generations; a system prompt was therefore added for the second condition to replicate the completion task format (*"You are a sentence completion model. Continue the user's sentence naturally."*). This steered condition is labelled *with context*; the unsteered condition is *without context*. Each model contributes 30,000 completions: 10,000 per demographic condition (women-directed, men-directed, people-directed), generated from five sentence completion stems (*[demographic] can*, *cannot*, *should*, *should not*, *Why are [demographic] so*), giving 2,000 completions per stem per condition. Completions were generated with matched parameters across models (temperature 1.0, max_tokens 100, top_p 0.9); GPT-5 response length is uncontrolled due to API constraints on these parameters. All findings compare women-directed against men-directed conditions.

Definitional othering. We apply GLiNER (Zaratiana et al., 2024), a zero-shot NER model, to extract sex/gender identity spans from each completion (confidence threshold = 0.5). Spans are classified as referencing men, women, or neutral, then filtered by verbatim string match (37.3% of spans failed and were excluded as extraction errors). Cross-demographic referencing rates are the proportion of *span-detected* completions (those containing at least one verified sex/gender span) with at least one verified cross-group reference; completions with no detected spans contribute to the denominator but not the numerator. Threshold sensitivity analysis (0.4, 0.5, 0.6) confirms stability: WM rates vary by at most 2.2 pp; MW rates are unchanged.

Topic space contraction. BERTopic (Grootendorst, 2022) was applied separately to each of 45 strata (15 models $\times$ 3 conditions) using identical parameters (HDBSCAN min_cluster_size=10, UMAP n_components=5, random_state=42). Topic diversity is measured by unique topic count and Shannon entropy per stratum, excluding the outlier cluster (-1).

Equity framing. We constructed a 10-term equity framing lexicon covering equity, empowerment, and allyship vocabulary: *support, resources, empowerment, community, wellbeing, safety, awareness, advocacy, inclusive, equitable*. The equity framing score is the proportion of completions containing at least one lexicon term

(whole-word match, case-insensitive). Manual spot-checks confirmed that GPT-2 and GPT-3 era completions produce false-positive matches on neutral language unrelated to equity framing; equity framing findings are therefore scoped to GPT-4 and GPT-5 only, where the terms demonstrably co-occur with empowerment and rights framing. We additionally compute Spearman correlation between equity framing score and REGARD negative polarity score (Sheng et al., 2019) at the completion level.

Statistical analysis. Cross-referencing rates are compared by era using χ^2 tests ($df = 1$) with Benjamini-Hochberg correction for five comparisons. Wilson confidence intervals are reported for proportions at or near zero.

4 Results

4.1 Definitional Othering

Women-directed completions reference men at substantially higher rates than men-directed completions reference women across all five GPT generations (Table 1). The asymmetry is present at GPT-2 ($+36.4$ pp) and persists at GPT-5 without context ($+51.4$ pp). It is *largest* at GPT-4 ($+60.3$ pp), the generation with the most extensive safety alignment training. Men-directed completions reference women at near-zero rates from GPT-3 onward (5.0%, 6.9%, 0.5%, 0.0%), while women-directed completions maintain consistent cross-referencing across all generations (30.5-67.2%).

No prompt elicits cross-referencing: the pattern is an emergent property of generation conditioned on demographic framing. Context provision (GPT-5 with context) reduces the gap by 20.9 pp, though the asymmetry persists ($+30.5$ pp), demonstrating definitional othering is a default generation pattern suppressible by explicit context rather than an intrinsic model property. Context provision cannot be assumed in typical deployed applications, however: context provision is not a default in API access; absent explicit configuration, the asymmetric pattern is the default output.

4.2 Topic Space Contraction

The GPT-4 alignment boundary produces asymmetric contraction in the representational range of women-directed completions. At GPT-2, topic diversity is near-parity: women span 153 unique topics (Shannon entropy $H = 4.60$) versus men's 168 ($H = 4.80$); W/M ratio $= 0.91$. At GPT-4, women's unique topic count falls to 110 ($H = 3.47$) while men's *increases* to 190 ($H = 3.97$): W/M ratio $= 0.58$, a 36% relative decline from GPT-2 parity. Topic concentration increases: the top-five topics account for 43.7% of GPT-4 women-directed completions versus 40.1% of men-directed.

Partial recovery occurs at GPT-5: women span 219 topics versus men's 245 (W/M $= 0.89$), though parity is not restored. GPT-5 with context produces near-parity. The contraction is not a volume effect: women-directed completions at GPT-4 are on average longer than men-directed (mean word count: 71.5 vs. 59.6; t-test $p < 0.001$), making the topic narrowing more striking. The alignment boundary narrows *what* is generated about women even as completions grow longer. The W/M ratio change (0.91 to 0.58) is a descriptive comparison across strata; bootstrap confidence intervals would be needed to formalise the difference, which we flag as a limitation.

4.3 Equity Framing Asymmetry

GPT-4 introduces a systematic demographic gap in equity-coded language: 20.6% of women-directed completions contain equity framing versus 10.6% of men-directed ($+10.0$ pp). This gap widens at GPT-5 without context (women 37.3%, men 16.9%; $+20.4$ pp) and collapses under context provision (women 0.9%, men 1.8%; -0.9 pp), confirming equity framing is a default generation pattern suppressed by explicit context.

Equity framing co-occurs with lower REGARD negative polarity scores at GPT-5 without context (women: Spearman $\rho = -0.13$, $p < .001$; men: $\rho = -0.25$, $p < .001$): completions containing equity language receive lower harm scores. This correlation is consistent with a surface substitution mechanism: safety alignment has learned to generate equity-coded language in contexts where REGARD negative polarity scores are elevated absent equity framing, functioning as a lexical harm-score reducer. An alternative interpretation cannot be fully excluded: completions with equity language may be genuinely less harmful in content, and REGARD may be scoring them correctly. Under either interpretation, the demographic asymmetry is a finding independent of causal interpretation: women-directed completions receive equity framing at $2.2\times$ the rate of men-directed completions. The pre-alignment gap is negligible (GPT-2: $+3.6$ pp; GPT-3: $+1.4$ pp; reported as

Table 1: Cross-demographic referencing rates by era. WM: women-directed completions referencing men. MW: men-directed referencing women. χ^2 ($df = 1$), Benjamini-Hochberg corrected; all survive ($p < 0.001$).

Era	WM	MW	Gap	χ^2
GPT-2	48.6%	12.2%	+36.4 pp	6,126***
GPT-3	41.7%	5.0%	+36.7 pp	7,785***
GPT-4	67.2%	6.9%	+60.3 pp	21,137***
GPT-5 nc	52.0%	0.5%	+51.4 pp	5,287***
GPT-5 ctx	30.5%	0.0%	+30.5 pp	2,841***

$^{***}p < 0.001$ (BH-corrected). GPT-5 ctx MW: 95% CI [0.000%, 0.049%], n=7,894.

approximate baseline only; see §3 for the false-positive caveat), consistent with alignment introducing rather than inheriting this asymmetry. Context provision (-36.4 pp for women) confirms the pattern is a default generation behaviour, not an intrinsic model property.

5 Discussion

Three distinct measurement approaches (cross-demographic referencing, BERTopic topic diversity, and equity lexicon scoring) converge on the same structural finding: safety alignment changes the surface form of gender representation without eliminating structural asymmetry. In two of three cases (definitional othering, topic contraction) the GPT-4 alignment boundary *intensifies* the asymmetry relative to GPT-2.

The equity reframing finding is the most theoretically significant. A model that has learned to *perform* equity (generating rights, empowerment, and advocacy language for women-directed completions at double the rate for men) is not producing equitable representation; it is producing the surface form of equity where REGARD negative polarity scores are elevated absent equity framing. This distinction matters for evaluation: REGARD scores improve (fewer explicit harm signals), but the asymmetry in how women are framed has not been reduced; it has been substituted. The harm has not been removed; it has been given a new lexical address. An evaluator relying on REGARD observes apparent improvement; the structural asymmetry in generation persists. The complete collapse under context provision (-36.4 pp for women) confirms the pattern is a default generation habit, not a model belief about equity.

All three markers are invisible to Detoxify-class toxicity evaluation, which measures lexical harm signals absent from equity-coded language, cross-referencing patterns, and topic diversity. The findings identify a systematic gap: the harm dimension captured by structural representation asymmetry is not covered by Detoxify-class toxicity evaluation. We recommend that LLM safety evaluation incorporate structural representation metrics (cross-demographic referencing rates, per-stratum topic diversity, and framing asymmetry scores) alongside surface toxicity classifiers.

6 Conclusion

We document three structural markers of persistent gender asymmetry across GPT safety alignment: definitional othering (women defined in relation to men across all five generations), topic space contraction (women's representational range declines 36% relative to men's at the alignment boundary), and equity reframing (equity-coded language applied at double the rate to women-directed completions, correlating with lower harm scores). Each marker survives safety alignment; each is invisible to standard toxicity evaluation. The three findings together constitute a structural harm signature that accompanies, rather than contradicts, safety improvement as currently measured.

Limitations

Lineage scope. All findings are scoped to the OpenAI GPT lineage (15 models, GPT-2 through GPT-5). Whether the structural markers generalise to other model families, providers, or alignment approaches (DPO, Constitutional AI, RLHF variants) is an open empirical question this study does not address.

Binary gender framing. The demographic conditions operationalise gender as binary (women, men). Non-binary and gender-nonconforming identities are not represented. The findings should not be interpreted as claims extending beyond the women/men comparison operationalised here.

Equity lexicon coverage. The 10-term equity framing lexicon was constructed by the research

team and covers a narrow register of equity-coded language. Terms such as *rights* and *safety* are polysemous and may match non-equity contexts. Reported equity framing rates are approximate and lexicon-dependent; a broader or validated lexicon may produce different absolute rates, though the demographic asymmetry is unlikely to reverse given its magnitude.

REGARD negative polarity. REGARD negative polarity scores at GPT-4 and GPT-5 exhibit systematic false-positive inflation for refusal-category completions (observed scores 0.94-0.97), which may not have been excluded from the analysis. If refusal responses are disproportionately concentrated in non-equity-framing completions, the reported Spearman correlations ($\rho = -0.13$, $\rho = -0.25$ at GPT-5 without context) may partly reflect this artefact rather than a content relationship. The direction of the correlation is consistent with a surface-substitution interpretation under either scenario, but the reported magnitudes should be treated with corresponding caution.

Definitional othering operationalisation. The cross-referencing measure depends on GLiNER span extraction and verbatim string matching. The 37.3% span exclusion rate reflects NER extraction errors; reported WM and MW rates are lower bounds on true cross-demographic reference prevalence.

Topic contraction significance. The W/M unique topic ratio change (GPT-2: 0.91; GPT-4: 0.58) is reported as a descriptive comparison across BERTopic strata. No formal significance test is applied to this ratio; bootstrap confidence intervals are needed to bound the uncertainty. Topic counts also vary with corpus size and content distribution in addition to alignment effects.

Causal inference. The cross-generational trajectory is observational. We attribute changes at the alignment boundary to alignment training based on the correspondence between model release stage and observed trajectory, but do not have access to training data or alignment procedure details for any GPT model.

Ethical Considerations

This study analyses gender-directed text completions from commercially deployed language models. All data was generated via public API; no personal data was collected. The findings document systematic harm patterns in deployed systems and are intended to motivate improved evaluation methodology. Verbatim model outputs are not included in this paper. The prompt design ("Why are [demographic] so...") is adversarial by design; the intent is to surface model behaviour under demographic pressure, consistent with established red-teaming methodology (Perez et al., 2022).

Use of AI Assistance

Claude (Anthropic) was used for LaTeX editing, proofreading, and condensing earlier drafts. All analysis, findings, and claims are the authors' own; the AI assistant did not generate or verify numerical results.

References

Su Lin Blodgett, Solon Barocas, Hal Daume III, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the ACL*, pages 5454–5476. Association for Computational Linguistics.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. *arXiv preprint*. ArXiv:1607.06520 [cs].

Patricia Hill Collins. 1990. *Black Feminist Thought: Knowledge, Consciousness, and the Politics of Empowerment*. Routledge, New York.

Kimberle Crenshaw. 1989. Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics. *University of Chicago Legal Forum*, 1989(1):139–167.

Deep Ganguli, Liane Lovitt, John Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Benjamin Mann, Ethan Perez, Nicholas Schiefer, and Kamal Ndousse. 2022. Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned.

Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.

Hannah Rose Kirk, Yennie Jun, Haider Iqbal, Elias Benussi, Filippo Volpin, F. Dreyer, Aleksandar Shtedritski, and Yuki M. Asano. 2021. Bias Out-of-the-Box: An Empirical Analysis of Intersectional Occupational Biases in Popular Generative Language Models.

Safiya Umoja Noble. 2018. Algorithms of oppression. In *Algorithms of oppression*. New York university press.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, S. Agarwal, Katarina Slama, and Alex Ray. 2022. Training language models to follow instructions with human feedback.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and G. Irving. 2022. Red Teaming Language Models with Language Models.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The Woman Worked as a Babysitter: On Biases in Language Generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3407–3412, Hong Kong, China. Association for Computational Linguistics.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, and others. 2021. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.

Urchade Zaratiana, Nadi Tomeh, Pierre Battaglia, and Thierry Charnois. 2024. GLiNER: Generalist Model for Named Entity Recognition using Bidirectional Transformer. In *Proceedings of the 2024 Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.