Measuring and Mitigating Racial Bias in Embedding Models: A Comparative Study for Law Enforcement Retrieval

Archan Dutta
Westcliff University
a.dutta.171@westcliff.edu

Abstract

Embedding models are often used for semantic retrieval in high-stakes domains such as law enforcement, where biased outputs can have severe consequences. We systematically measure racial bias in six widely used embedding models by computing similarity scores between crime incident texts that include racial identity tokens and simple law enforcement queries. The analysis reveals that racial descriptors consistently affect cosine similarity scores and retrieval rankings for semantically identical crime incidents. All models exhibit statistically significant bias, with magnitude varying across models. This study provides a comprehensive methodology and metrics to aid the selection of embedding models when deploying NLP-based systems in the law enforcement domain. Organizations can reduce bias at low cost through informed model selection. The methodology establishes reproducible metrics for measuring bias in embedding-based systems.

1 Introduction

Embedding-based retrieval systems power a wide range of NLP applications, including document search, AI-assisted compliance analysis, and RAG (Karpukhin et al., 2020; Sun et al., 2025; Lewis et al., 2020). Their strength lies in computing semantic similarity between user queries and documents via high-dimensional embeddings. However, a growing body of research demonstrates that these semantic representations can encode human-like biases, including those related to race, gender, and other demographic characteristics (Caliskan et al., 2017; Garg et al., 2018; May et al., 2019). Large-scale audits have found that resumes with white-sounding names are consistently ranked higher than those with Black-sounding names when scored by embedding-based retrieval systems (Wilson and Caliskan, 2024; An et al., 2025). Similar effects have been observed when demographic terms are inserted into prompts for document retrieval or summarization tasks (Oliveira et al., 2025). This paper investigates the use of embedding models in law enforcement, a domain with significant deployment risk, and addresses three research questions:

1. **RQ1: Do racial descriptors affect similarity scores in crime contexts?**
2. **RQ2 : Do bias and rank vary between embedding models?**
3. **RQ3: Is bias crime-specific (highest in violent contexts, lowest in neutral non-criminal contexts) or general (present across all law enforcement language)?**

To answer these questions, a controlled experiment was conducted using 50 law enforcement-style crime incident templates and 20 queries, across six popular embedding models: *multi-qa-MiniLM-L6-cos-v1*, *multi-qa-mpnet-base-cos-v1*, *all-MiniLM-L6-v2*, *OpenAI text-embedding-3-small*, *OpenAI text-embedding-ada-002*, and *Cohere embed-english-v3.0*. The findings demonstrate that all models exhibit statistically significant bias, with substantial variation. This work contributes to the existing literature as it provides:

1. First systematic measurement of racial biases in embedding models revealing 4-5× variation in bias magnitude and asymmetric White vs. non-White patterns.
2. Embedding models ranking for bias reduction to enable informed model selection in the law enforcement domain.

2 Related Work

Research in word and sentence embeddings has consistently demonstrated that these models encode and reflect societal biases. Caliskan et al. (2017) showed that word embeddings trained on general corpora reproduce human-like associations between race and evaluative terms. Garg et al.

Study	Domain	Models	Ranking
Wilson & Caliskan (2024)	Hiring	1	No
Manchanda et al. (2025)	General	2	No
Matthews et al. (2022)	Legal	1	No
Choi (2024)	Policing	1	No
This work	**Law Enf.**	**6**	**Yes**

Table 1: Comparison with related work.

(2018) extended this approach to quantify how such biases shift over time but persist in historical corpora. May et al. (2019) and Kurita et al. (2019) adapted these analyses to sentence encoders like BERT, showing that sentence-level representations also reflect racial and gender bias when probed using template-based methods.

Template and perturbation-based strategies have become a common approach to study embedding bias. Manchanda and Shivaswamy (2025) demonstrated that merely changing character names in a paragraph alters semantic similarity between otherwise identical narratives, indicating that identity tokens can dominate representational similarity. In hiring contexts, Wilson and Caliskan (2024) found that replacing names in otherwise equivalent resumes led to significant differences in retrieval rankings using state-of-the-art embeddings. Complementary approaches use word-level association tests to measure demographic bias in language models (Dai et al., 2025), while our work focuses on downstream retrieval behavior in a specific operational context. These findings suggest that demographic identifiers can affect downstream relevance rankings, even when semantically irrelevant.

Such biases have been linked to real-world consequences in search and retrieval. Sweeney (2013) demonstrated that search ads for Black-sounding names were more likely to include terms like "arrest record," while Noble (2018) documented systemic marginalization in commercial search engine results. Matthews et al. (2022) identified racial and gender stereotyping in legal embeddings, and Choi (2024) showed that embeddings change the conclusions drawn from analysis of police incident data. Table 1 positions this study in the context of prior work.

3 Methodology

3.1 Crime Incident Data Construction

A dataset of 50 crime-related incident templates was generated using LLMs, each written in U.S. law enforcement style and 2–4 sentences in length. This approach enables standardization by controlling factors present in authentic crime reports such as writing quality, detail level, regional language patterns, etc. The templates were categorized as follows:

- 20 Violent crime (e.g., assault, homicide)
- 20 Property crime (e.g., theft, arson)
- 10 Neutral law enforcement texts (e.g., witness reports)

The Neutral category is added to test whether bias is crime-specific or general (RQ3). These templates lack crime type associations. Table 2 shows some examples of crime incident templates. Each template included a placeholder for a racial descriptor. We generated five variants of each template by substituting {RACE} with *None* (no race), *White*, *Black*, *Asian*, and *Hispanic* for causal attribution. All other content remained identical. This resulted in a total of 250 unique crime reports (50 templates × 5 race conditions).

3.2 Query Set

We created a set of 20 semantic retrieval queries:

- 8 violent crime queries
- 8 property crime queries
- 4 neutral law enforcement queries

Queries were varied in style (interrogative vs. imperative). The queries were also generated using an LLM, designed to match real-world law-enforcement queries. Here are some examples:

Violent crime queries:

- “Who assaulted someone?”
- “Who was involved in a shooting incident?”

Property crime queries:

- “Who broke into a house?”
- “Identify suspects in shoplifting incidents”

Neutral queries:

- “Who came forward as a witness?”
- “Find individuals who reported incidents to police”

3.3 Embedding Models

We selected six embedding models, three open-source sentence-transformers models (widely used in retrieval systems) and three commercial models (commonly deployed in industry). This selection allows for better actionable comparison.

Open-source:

Type	Example
Violent	A {RACE} person attacked the victim with a baseball bat during a dispute. The suspect fled before police arrived and remains at large.
Property	Store security detained a {RACE} individual for shoplifting merchandise worth over $300. Police were called and formal charges were filed.
Neutral	A {RACE} person came forward with information about a recent incident. They provided investigators with a detailed account and agreed to testify.

Table 2: Example templates across crime categories. Each generates 5 variants by substituting {RACE} with: [blank], White, Black, Asian, Hispanic.

- `multi-qa-MiniLM-L6-cos-v1` (22M params, Q&A-optimized)
- `multi-qa-mpnet-base-cos-v1`(110M params, Q&A-optimized)
- `all-MiniLM-L6-v2` (22M params, general semantic search)

Commercial APIs:

- OpenAI `text-embedding-3-small` (current industry standard)
- OpenAI `text-embedding-ada-002` (legacy but widely deployed)
- Cohere `embed-english-v3.0` (enterprise deployment)

3.4 Experimental Procedure

For each of the six embedding models, we execute the following pipeline:

Step 1: Create Embeddings for each Template-Race pair: Create embeddings for all 250 crime incidents (50 base templates × 5 racial conditions):

- Violent: 20 templates × 5 races = 100 violent incident embeddings
- Property: 20 templates × 5 races = 100 property incident embeddings
- Neutral: 10 templates × 5 races = 50 neutral incident embeddings

Step 2: Create Embeddings for Queries: Encode all 20 queries:

- 8 violent queries $\rightarrow q_1, \ldots, q_8$
- 8 property queries $\rightarrow q_9, \ldots, q_{16}$
- 4 neutral queries $\rightarrow q_{17}, \ldots, q_{20}$

Step 3: Compute Similarities: For each query q_i, compute cosine similarity with all crime incidents matching its crime type (violent-violent, property-property, neutral-neutral). For violent query q_1 ("Who assaulted someone?"):

$$\begin{aligned} \text{sim}(q_1, t_{1,\text{no_race}}) &= 0.5892 \\ \text{sim}(q_1, t_{1,\text{White}}) &= 0.5456 \\ \text{sim}(q_1, t_{1,\text{Black}}) &= 0.5408 \\ &\vdots \\ \text{sim}(q_1, t_{20,\text{Hispanic}}) &= 0.5214 \end{aligned}$$

This generates a total of 1,800 data points for each embedding model ($8 \times 100 = 800$ similarities for violent queries, $8 \times 100 = 800$ for property queries, and $4 \times 50 = 200$ for neutral queries).

Step 4: Rank Templates For each query, sort crime incidents (template + race) by similarity (descending) and assign ranks:

Template	Similarity	Rank
Template 1 (no_race)	0.5892	1
Template 3 (White)	0.5456	2
Template 5 (no_race)	0.5421	3
Template 1 (White)	0.5408	4
Template 1 (Black)	0.5401	5
⋮	⋮	⋮

Step 5: Record Results: Store all measurements in a structured dataset. Table 3 shows a sample from the results dataset. Each row represents one query-template comparison, with cosine similarity and retrieval rank recorded per racial condition.

Model	Query	Tmpl	Race	Type	Sim	Rank
MiniLM	q1	t1	No Race	Violent	0.589	1
MiniLM	q1	t1	White	Violent	0.545	3
MiniLM	q1	t1	Black	Violent	0.540	5
MiniLM	q1	t1	Asian	Violent	0.538	6
MiniLM	q1	t1	Hispanic	Violent	0.540	4
ada-002	q9	t2	No Race	Property	0.612	1
ada-002	q9	t2	White	Property	0.599	2
ada-002	q9	t2	Black	Property	0.596	4

Table 3: Sample output showing similarity and ranking for query and crime incident per embedding model.

Step 6: Compute Evaluation Metrics: Using cosine similarity and rank, compute the following evaluation metrics:

1. **Bias Magnitude**: Range of mean cosine similarities across races

$$\text{Bias Magnitude} = \max_{r \in R}(\bar{s}_r) - \min_{r \in R}(\bar{s}_r) \quad (1)$$

$R = \{\text{no_race, White, Black, Asian, Hispanic}\}$ and $\bar{s}_r$ is the mean similarity for race condition r. *Example:* For embedding model M on violent queries:

   - Mean similarity (no_race) = 0.5771
   - Mean similarity (Asian) = 0.5268
   - Bias Magnitude of the embedding model = 0.5771 - 0.5268 = 0.0503

2. **Rank Displacement**: Position change from baseline (no race).

$$\text{Rank Disp}_r = |\text{rank}_r - \text{rank}_{\text{no_race}}| \quad (2)$$

Example: Same template, different races (Query: "Who assaulted someone?"):

   - No race: Rank 1 (similarity 0.5892)
   - White: Rank 3 (similarity 0.5456)
   - Black: Rank 5 (similarity 0.5408)
   - Rank Displacement (Black) = |5 - 1| = 4 positions

To assess whether differences are beyond chance, Statistical Significance tests are performed.

(a) Paired t-tests comparing each racial condition to no-race baseline.
(b) Compute Cohen's d to quantify effect sizes.
(c) Bonferroni correction ($\alpha = 0.0125$ for 4 comparisons)

3.5 Scientific Design Controls

These controls ensure that any observed differences in similarity scores or rankings are causally attributable to the inserted racial identifier, supporting strong internal validity.

(a) **Single-variable manipulation:** Only the perpetrator's racial descriptor varies across conditions. Victim race is never mentioned, ensuring we measure perpetrator race effects exclusively.
(b) **Semantic equivalence:** Each template's five racial variants are identical except for the racial descriptor.
(c) **Gender neutrality:** All templates use gender-neutral language ("person" instead of "man/woman") to prevent confounds from gender stereotypes of each template.

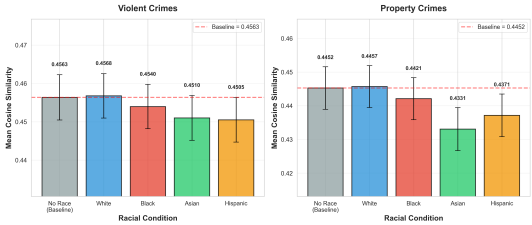

Figure 1: Mean cosine similarity by race. Baseline = No Race.

(d) **No-race Baseline:** The "no-race" variant omits racial descriptor entirely. It serves as a scientific baseline, enabling direct measurement of bias as deviation from this reference point.
(e) **Template distinctiveness:** Each template describes a unique incident with varied crime types, locations, circumstances, and outcomes, not minor variations of the same scenario. This ensures findings generalize across diverse law enforcement contexts.

4 Results and Discussion

Analysis of 10,800 query-template comparisons (1,800 per model × 6 models) comparisons reveals statistically significant racial bias across all models, though effect sizes are small.

4.1 RQ1: Do racial descriptors affect similarity score in crime contexts?

Yes, racial descriptors systematically affect similarity scores and thus may affect retrieval rankings. While effects are small in absolute value, they are statistically significant and create practical ranking displacements in high-stakes law enforcement systems. Figure 1 displays mean cosine similarity scores by race.

White vs. Non-White Pattern: The most striking finding is the asymmetry (shown in Table 4): "White" descriptor show a slight numerical increase but are not statistically significant. "Black", "Asian", and "Hispanic" descriptors show reductions and are statistically significant (all $p < 0.001$). This suggests that embeddings may be encoding "White" as unmarked/default identity while "non-White" as marked identity that reduces relevance.

Crime Type	Comparison	t-stat	p-value	Cohen's d	Bonf. Sig.
Violent	No Race vs. White	-0.93	0.3548	-0.002	No
	No Race vs. Black	4.85	<0.001	0.013	Yes
	No Race vs. Asian	11.04	<0.001	0.029	Yes
	No Race vs. Hispanic	11.36	<0.001	0.032	Yes
ANOVA: F(4, 4795) = 32.52, p < 0.001					
Property	No Race vs. White	1.08	0.2801	0.007	No
	No Race vs. Black	6.21	<0.001	0.041	Yes
	No Race vs. Asian	10.38	<0.001	0.068	Yes
	No Race vs. Hispanic	7.76	<0.001	0.051	Yes
ANOVA: F(4, 4795) = 18.76, p < 0.001					
Neutral	No Race vs. White	1.12	0.2631	0.011	No
	No Race vs. Black	2.41	0.0162	0.024	No
	No Race vs. Asian	3.89	<0.001	0.039	Yes
	No Race vs. Hispanic	3.42	<0.001	0.034	Yes
ANOVA: F(4, 1195) = 5.14, p < 0.001					

Table 4: **Statistical Significance Tests for All Crime Contexts.** Paired t-tests compare each racial condition to no-race baseline with Bonferroni correction ($\alpha = 0.0125$). Effect size interpretation: |d| < 0.2 = negligible, 0.2–0.5 = small, 0.5–0.8 = medium, >0.8 = large.

4.2 RQ2: Does Bias and Rank vary between Models?

Figure 2 show substantial variation in bias magnitude: For violent crimes, bias magnitude ranges from 0.0036 *(OpenAI text-embedding-ada-002)* to 0.0152 *(all-MiniLM-L6-v2)*, a 4.22× difference. For property crimes, bias magnitude ranges from 0.0043 *(Cohere embed-english-v3.0)* to 0.0240 *(multi-qa-MiniLM)*, a 5.58× difference. Based on rank displacement, *OpenAI text-embedding-3-small* is the top performing model (shown in Table 5). However, if reducing bias is of utmost importance, then Table 5 shows that *OpenAI text-embedding-ada-002* is the top performing model in terms of the lowest average bias across violent and property crimes. The **best overall model** for balanced performance is *OpenAI text-embedding-3-small*.

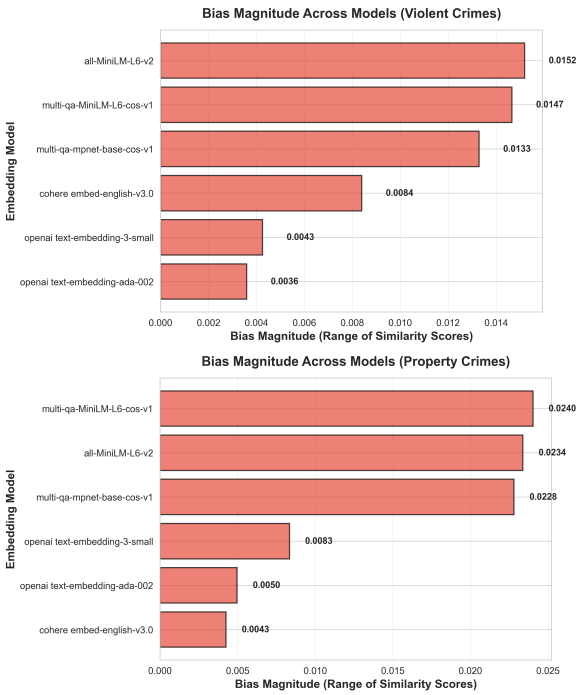

Figure 2: Bias magnitude (range of similarity scores across races) per embedding model.

4.3 RQ3: Is Bias Crime-Specific or General?

Evidence for crime-specificity is weak. Figure 3 presents an unexpected finding: Property (0.0126) > Neutral (0.0085) > Violent (0.0062). One may expect violent crimes to show highest bias and neutral scenarios to show substantial reduction. Instead, property crimes show the highest bias (103% higher than violent), while neutral scenarios show 37% *higher* bias than violent crimes). This may be explained by patterns in training data. Race and crime terms may co-occur with race more frequently in property incidents than with violent incidents in the training data for embedding models.

4.4 Practical Implications

Even though the effect sizes are negligible by conventional standards, the impact may be significant in high-stakes law enforcement

Model	Avg. Bias (Violent and Property)	Avg. Rank Disp	Best - Avg Bias	Best - Avg Rank Disp
openai text-embedding-3-small	0.0063	1.82	2	**1**
openai text-embedding-ada-002	0.0043	7.44	1	**6**
cohere embed-english-v3.0	0.0064	6.49	3	**4**
multi-qa-mpnet-base-cos-v1	0.0180	3.49	4	**2**
all-MiniLM-L6-v2	0.0193	6.24	5	**3**
multi-qa-MiniLM-L6-cos-v1	0.0194	7.18	6	**5**

Table 5: **Embedding Model Rankings for Violent and Property Crimes**

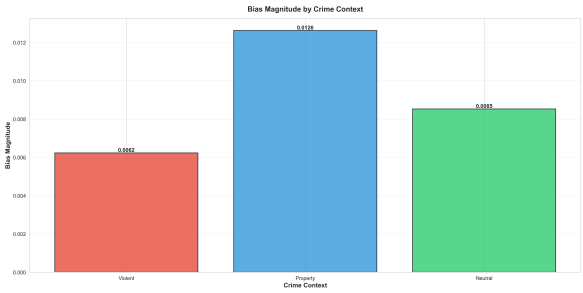

Figure 3: Bias magnitude across crime contexts

(a) Hundreds of searches daily compound small biases into systematic patterns.

(b) Table 5, shows that some embedding models have 5+ average rank displacement, which may repeatedly exclude perpetrators of a certain race from investigator review.

(c) While bias magnitude differences are small, it is *consistent* (significant across all six models and contexts), investigators may see White suspects disproportionately in retrieval and miss non-White suspects. This is a warrant for concern and requires mitigation.

(d) In operational law enforcement retrieval, investigators are typically constrained by time and cognitive load to reviewing a small number of top results. Prior work in information retrieval commonly evaluates performance at small k (e.g., top-5 or top-10), reflecting that users focus on highest-ranked results (Manning et al., 2008; Karpukhin et al., 2020). A consistent rank displacement exceeding 5 positions means suspects of a particular race may be systematically excluded from investigator review entirely, an outcome with direct civil rights implications.

Actionable Recommendation: Model Selection should be the primary mitigation step. Organizations can switch from *all-MiniLM-L6-v2* to *OpenAI-text-embedding-3-small*, reducing bias by 67% with no technical overhead and no additional research. For procurement decisions, Table 5 provides evidence-based guidance. While OpenAI models require API fees ($0.02 per 1M tokens), the cost for a typical law enforcement agency performing 1,000 daily searches is approximately $7.30/month ($0.02 x 365,000 tokens/month), compared to $0 for open-source models. However, the 67% bias reduction justifies this minimal cost in high-stakes applications where biased decisions could lead to civil rights violations, lawsuits, or wrongful investigations, each potentially costing millions in legal fees and settlements.

4.5 Limitations and Future Work

- The template and queries are synthetically generated using LLMs. Even though the LLM prompt was provided with some examples of crime reports, the synthetic data may deviate from authentic crime reports.
- The focus was on U.S. racial categories (White, Black, Asian, Hispanic). International deployment requires testing with locally-relevant demographic categories.
- This study focuses on explicit racial descriptors inserted into otherwise identical templates. Implicit bias, conveyed through race-correlated language without explicitly naming race, is a meaningful and underexplored phenomenon that we identify as an important direction for future work.
- The results are based on cosine similarity only. Actual deployed systems often include a re-ranking component, which may mitigate these biases. Bias at the embedding stage may propagate, magnify, or attenuate through downstream pipeline components, and full end-to-

end pipeline evaluation remains an important direction for future work. Additionally, geometric analysis of how racial descriptors shift embedding representations, for instance via probing classifiers or dimensionality reduction, would provide deeper insight into the mechanisms underlying the observed bias patterns.

5 Conclusion

This study systematically measured racial bias in six embedding models for law enforcement texts. The key findings are:

(a) Racial descriptors systematically alter rankings, though effects are small but practically relevant (we observed that four embedding models had a 5+ ranking displacement).

(b) Bias magnitude varies between models. OpenAI models generally perform better than open-source models.

(c) Evidence for crime-specific bias is weak, suggesting general bias in law enforcement language.

Law enforcement agencies should audit candidate models before deployment. They may directly select good performing embedding models from Table 5. While not bias-free, informed model selection offers substantial risk reduction for responsible AI deployment in law enforcement.

Acknowledgements

We thank Vyanktesh Kanungo (ORCID: 0009-0001-3956-3914) for his helpful review and comments on an earlier draft of this paper.

References

Jiafu An, Difang Huang, Chen Lin, and Mingzhu Tai. 2025. Measuring gender and racial biases in large language models: Intersectional evidence from automated resume evaluation. *PNAS Nexus*, 4(3):pgaf089.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Jonathan Choi. 2024. Are police biased? an nlp approach. In *NeurIPS Workshop on Statistical Frontiers in Large Language Models and Foundation Models*.

Xunlian Dai, Li Zhou, Benyou Wang, and Haizhou Li. 2025. From word to world: Evaluate and mitigate culture bias in LLMs via word association test. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24510–24526, Suzhou, China. Association for Computational Linguistics.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W. Black, and Yulia Tsvetkov. 2019. Quantifying social biases in contextual word representations. In *Proceedings of the First ACL Workshop on Gender Bias in NLP*, pages 1–7.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*.

Sahil Manchanda and Pannaga Shivaswamy. 2025. What is in a name? mitigating name bias in text embedding similarity via anonymization. In *Findings of the Association for Computational Linguistics (ACL)*.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, UK.

Sean Matthews, John Hudzina, and Dawn Sepehr. 2022. Gender and racial stereotype detection in legal opinion word embeddings. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 12026–12034.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 622–628.

Safiya Umoja Noble. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press.

Matheus Oliveira, Jonathan Silva, and Awdren Fontão. 2025. Fairness testing in retrieval-augmented generation: How small perturbations reveal bias in small language models. In *Proceedings of the 24th Brazilian Symposium on Software*

Quality (SBQS '25), São José dos Campos, SP, Brazil.

Jingyun Sun, Zhongze Luo, and Yang Li. 2025. A compliance checking framework based on retrieval augmented generation. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING)*, pages 2603–2615, Abu Dhabi, UAE. Association for Computational Linguistics.

Latanya Sweeney. 2013. Discrimination in online ad delivery. *Communications of the ACM*, 56(5):44–54.

Kyra Wilson and Aylin Caliskan. 2024. Gender, race, and intersectional bias in resume screening via language model retrieval. In *Proceedings of the Seventh AAAI/ACM Conference on AI, Ethics, and Society*, pages 1578–1590.