Gender Bias in LLaMA-3 Embeddings: Implications for LinkedIn-Style Retrieval Systems

Christopher S. Penn
Chief Data Scientist
TrustInsights.ai
cspenn@trustinsights.ai

Katie Robbert
Chief Executive Officer
TrustInsights.ai
ceo@trustinsights.ai

Abstract

Large language models increasingly power professional network retrieval systems, with LinkedIn recently deploying LLaMA-3 for member and content embeddings. We investigate whether these embeddings exhibit gender bias by measuring semantic drift when we attribute identical professional content to names of different perceived genders. We constructed 406 paired LinkedIn-style posts with identical text content, headlines, and professional context, differing only in author name (male versus female variants), and extracted embeddings using LLaMA-3.2-3B's hidden states with mean pooling—replicating LinkedIn's published methodology. We find systematic bias: mean cosine similarity between paired embeddings is 0.994 (not the expected 1.0), with Cohen's $d = -0.93$ (large effect) and $p < 0.0001$ across both parametric and non-parametric statistical tests. This approximately 0.6% embedding deviation, while small per-pair, represents systematic differential treatment that compounds across retrieval, ranking, and recommendation systems—potentially affecting search visibility for millions of professionals. We release our dataset, code, and statistical framework to enable reproducible bias auditing of LLM-based retrieval systems.

1 Introduction

LinkedIn serves over one billion professionals worldwide, functioning as a critical infrastructure for job seeking, professional networking, and career advancement (LinkedIn Corporation, 2024). Recent research reveals that LinkedIn has deployed LLaMA-3, Meta's large language model, to power their retrieval system for member and content search (Gupta et al., 2024). Critically, author name serves as an explicit feature in embedding construction, raising questions about whether the model treats identical content differently based on the perceived gender of the author's name.

If embeddings differ for identical professional content based solely on name, the retrieval system exhibits bias at its foundational layer. This bias propagates through search results, connection recommendations, and job matching algorithms. Unlike downstream ranking systems that can potentially correct for bias using behavioral signals, embedding-stage bias affects all users at the first filter stage. Cold-start users—new professionals without established engagement histories—face particular vulnerability, as LinkedIn's own research acknowledges significant performance gaps for this population (Team, 2025).

Our research addresses a fundamental question:

> Does LLaMA-3 produce systematically different embeddings for identical professional content when only the author's perceived gender differs?

To answer this question, we constructed 406 paired posts containing identical professional content, differing only in author name (male versus female phonetic variants). This sample size provides exceptional statistical power, exceeding the theoretical minimum required to detect this effect size by over 30 times. We extracted embeddings using LinkedIn's published methodology—hidden state extraction with mean pooling (Gupta et al., 2024)—and applied rigorous statistical analysis with multiple robustness checks. We found large, systematic bias: Cohen's $d = -0.93$, indicating that the model consistently encodes gendered names differently even when all other content remains identical.

Our contributions are fourfold:

1. We present the first public audit of gender bias in LinkedIn's LLaMA-3-based retrieval embeddings, replicating their published methodology.

2. We apply the established paired-content methodology from fairness research (Bertrand and Mullainathan, 2004; Wilson and Caliskan, 2024) to professional network profile embeddings.

3. We release an open-source auditing tool and dataset enabling community verification and extension of our findings.

4. We provide a statistical framework combining parametric and non-parametric validation suitable for embedding bias research.

The remainder of this paper proceeds as follows: Section 2 reviews related work on embedding bias and retrieval fairness. Section 3 describes our experimental design, data collection, and statistical methods. Section 4 presents our findings. Section 5 discusses implications and limitations. Section 6 concludes with directions for future work.

2 Related Work

2.1 Bias in Word and Sentence Embeddings

Research on bias in language model embeddings has established that these representations encode societal stereotypes. The Word Embedding Association Test (WEAT) demonstrated that word embeddings trained on large corpora contain human-like biases, associating male names with career terms and female names with family terms (Caliskan et al., 2017). Subsequent work extended these findings to sentence-level encoders, showing that contextualized representations from BERT and similar models exhibit gender bias in downstream tasks (May et al., 2019; Kurita et al., 2019).

The foundational work of Bertrand and Mullainathan (2004) on name-based discrimination in hiring—finding that resumes with stereotypically white names received 50% more callbacks than identical resumes with stereotypically Black names—established the real-world impact of name-based differential treatment. Our work extends this methodology to the embedding space of modern LLMs.

2.2 Fairness in Information Retrieval

Information retrieval fairness has emerged as a critical research area as search systems increasingly mediate access to opportunities (Gao and Shah, 2020). Studies have documented bias in search engine results, particularly for queries related to people and professions (Kay et al., 2015). Algorithmic hiring systems have faced scrutiny for perpetuating or amplifying historical biases present in training data (Dastin, 2018).

Most relevant to our work, Wilson and Caliskan (2024) audited gender and racial bias in resume screening using language model retrieval. Their study used paired resumes with 120 demographically-associated names to measure bias in Massive Text Embedding models, finding significant disparities in retrieval ranking. Our work extends this methodology to professional network search, specifically targeting LinkedIn's LLaMA-3-based architecture.

Professional networks present unique fairness challenges because retrieval quality directly affects career outcomes. Search visibility influences job opportunities, networking connections, and professional reputation. Bias at the retrieval stage affects who appears in search results before any downstream ranking or personalization.

2.3 LinkedIn's Retrieval Architecture

LinkedIn recently published details of their LLM-based retrieval system (Gupta et al., 2024). Their architecture fine-tunes LLaMA-3 as a dual encoder to generate dense embeddings for both members (queries) and content (items), enabling semantic search across their billion-user platform. In this dual encoder setup, a single shared LLM processes member and item prompts separately, producing embeddings that are compared via cosine similarity. The system constructs embeddings using member features including name, headline, and content text, applying mean pooling over hidden states from the model's final transformer layer.

Their published results show that the base LLaMA-3 model achieves Recall@10 of 0.24,

while fine-tuning on LinkedIn-specific data improves this to 0.42—a 74% relative improvement. This gap highlights that while fine-tuning substantially improves retrieval quality, the base model's representations form the foundation upon which these improvements build.

LinkedIn's 360Brew paper describes their downstream ranking system, a 150-billion parameter model that re-ranks retrieval results (Team, 2025). While this system can potentially correct some biases through personalization, it operates after the initial retrieval filter and cannot recover candidates excluded at the embedding stage.

Additional LinkedIn engineering publications describe their feed infrastructure, including FishDB for generic retrieval (Engineering, 2024b), speculative decoding optimizations for their hiring assistant (Engineering, 2024a), and recent work on accelerating recommendation systems using SGLang (Shimizu et al., 2025).

2.4 Gap in the Literature

While Wilson and Caliskan (2024) demonstrated bias in retrieval embeddings for resume screening, their work used general-purpose embedding models rather than platform-specific systems. No public audit has examined gender bias specifically in LinkedIn's LLaMA-3-based retrieval architecture, despite LinkedIn's recent publication of their methodology (Gupta et al., 2024). Our work addresses this gap by replicating LinkedIn's published embedding approach and testing for gender bias using the paired-content methodology established in prior fairness research.

3 Methodology

3.1 Experimental Design

We employ a within-subject paired comparison design to isolate the effect of gendered names on embedding representations. Our independent variable is author name (male versus female phonetic variant). Our dependent variable is the cosine similarity between embedding vectors for paired posts. Under the null hypothesis of no bias, embeddings for identical content should be identical regardless of author name, yielding cosine similarity of 1.0.

The key strength of this design is its control for content confounders. Because each pair contains identical text content, headline, and professional context, any difference in embeddings can only arise from the author name. This eliminates alternative explanations based on content quality, topic, writing style, or other textual factors.

3.2 Data Collection

3.2.1 Source Material

We collected public LinkedIn posts using screen recording on an iPhone 15 running the LinkedIn mobile application. The researcher scrolled through their public feed twice, capturing visible posts. Each session lasted 5 minutes and captured approximately 75 posts. To mitigate potential bias from the researcher's AI-focused feed, we supplemented with hashtag-based searches across diverse professional domains, gathering an additional 250 posts:

- Healthcare: #healthcare, #nursing
- Social issues: #black, #poverty
- Creative: #music
- Education: #teachers
- Professional: #leadership

We processed screen recordings using Google Gemini 3 Pro in Google AI Studio for video-to-JSON transcription. The model extracted three fields from each visible post: author name, professional headline (as displayed, which may be truncated), and post text content. The model excluded sponsored content and advertisements during transcription. See Appendix A for the complete transcription prompt.

This collection strategy ensures topic diversity beyond the researcher's algorithmically-curated feed and enables claims about bias across professional domains.

3.2.2 Pair Construction

We used Google Gemini 3 Pro to generate gender-coded name variants from the transcribed data. For each post, we produced both a male-coded and female-coded version

Attribute	Value
Total pairs	406
Unique content pieces	406
Collection period	December 2025
Sources	Feed + hashtag search
Professional domains	5+
Name transformation	Male → Female

Table 1: Dataset characteristics.

by prompting the model to transform names while preserving ethnic and cultural consistency. For example, "Christina Applegate" becomes "Christian Applegate," and "Robert Miller" becomes "Roberta Miller." Critically, Gemini automatically maintained ethnic consistency during transformation—names from non-Western naming conventions received culturally appropriate gender transformations, not Western substitutions.

The transformation process used separate prompts for male and female dataset generation (see Appendix A for complete prompts). We stripped emoji characters from name fields during processing. All other fields—professional headline and post text content—remained identical between pairs. The final dataset contains 406 unique content pairs.

3.2.3 Data Quality

We performed duplicate detection to ensure each content piece appears exactly once, removing one exact duplicate we discovered during quality analysis. Field validation confirmed consistent structure across all records. We determined sample size through statistical power analysis: 406 pairs provides greater than 99% power to detect effect sizes of $d = 0.2$ or larger at $\alpha = 0.05$.

Table 1 summarizes dataset characteristics.

3.3 Embedding Generation

3.3.1 Model Selection

We use LLaMA-3.2-3B (Meta AI, 2024), Meta's open-weights large language model, in bfloat16 precision. We intentionally test the base model (not fine-tuned) to assess inherent bias in the foundation model that underlies LinkedIn's production system. LinkedIn's published work uses a fine-tuned variant; our results thus represent a lower bound on what bias exists before any domain-specific training that could either reduce or amplify these effects.

3.3.2 Extraction Method

We replicate LinkedIn's published dual encoder methodology for embedding generation (Gupta et al., 2024). Following their Figure 2, we apply the item-side (content) encoding process:

1. Tokenize the input prompt
2. Pass tokens through the model with hidden state output enabled
3. Extract hidden states from the final transformer layer
4. Apply mean pooling across all token positions
5. Obtain a 3072-dimensional embedding vector

The following code illustrates our implementation:

```
# Embedding extraction
outputs = model(
    **inputs,
    output_hidden_states=True
)
hidden_states = outputs.hidden_states
    [-1]
embedding = hidden_states.mean(dim=1)
```

3.3.3 Prompt Construction

We format input prompts to match LinkedIn's retrieval schema:

Author Name: {name}
Author Headline: {headline}
Post Text: {text_content}

This structure mirrors how LinkedIn constructs member and content representations for retrieval, ensuring our methodology aligns with production usage.

3.4 Statistical Analysis

3.4.1 Primary Analysis

For each pair, we compute cosine similarity between the male-name and female-name embeddings:

$$\text{similarity} = \frac{\mathbf{v}_m \cdot \mathbf{v}_f}{\|\mathbf{v}_m\|\|\mathbf{v}_f\|} \tag{1}$$

where $\mathbf{v}_m$ and $\mathbf{v}_f$ are the embedding vectors for male and female name variants respectively.

We test against the null hypothesis $H_0 : \mu = 1.0$ using a one-sample t-test, with significance threshold $\alpha = 0.05$.

3.4.2 Assumption Testing

We assess normality of the similarity distribution using the Shapiro-Wilk test. When the data violate normality assumptions, we rely primarily on non-parametric alternatives.

3.4.3 Robustness Checks

We employ multiple complementary statistical approaches:

- Wilcoxon signed-rank test: Non-parametric alternative to the t-test, robust to non-normal distributions.
- Permutation test: We compute empirical p-values from 10,000 random resamplings, making no distributional assumptions.
- Bootstrap confidence intervals: Bias-corrected and accelerated (BCa) method with 10,000 resamples for robust uncertainty quantification.

3.4.4 Effect Size

We compute Cohen's d as our primary effect size measure:

$$d = \frac{\bar{x} - \mu_0}{s} \tag{2}$$

where $\bar{x}$ is the observed mean similarity, $\mu_0 = 1.0$ is the expected value under no bias, and s is the sample standard deviation.

We interpret effect sizes following standard conventions: $|d| < 0.2$ negligible, $0.2 \leq |d| < 0.5$ small, $0.5 \leq |d| < 0.8$ medium, and $|d| \geq 0.8$ large (Cohen, 1988).

We also compute a WEAT-style effect size for comparability with the embedding bias literature.

3.5 Reproducibility

To ensure reproducibility, we fix random seeds for NumPy and PyTorch. All experiments run on Apple Silicon hardware using the MPS backend. Complete code and data are publicly available at https://github.com/trustinsights/linkedinbias.

4 Results

4.1 Primary Findings

Table 2 presents summary statistics for cosine similarity across all 406 pairs. The observed mean similarity of 0.9940 deviates significantly from the expected value of 1.0 under the null hypothesis of no bias.

Metric	Value
N (pairs)	406
Mean similarity	0.9940
Standard deviation	0.0064
95% CI	[0.9934, 0.9946]
Expected (no bias)	1.0000
Deviation	-0.0060

Table 2: Summary statistics for cosine similarity between paired embeddings.

Table 3 presents results from our statistical tests. All tests reject the null hypothesis with $p < 0.0001$, indicating that the observed deviation from perfect similarity is highly unlikely to occur by chance.

Test	Statistic	p-value
One-sample t-test	$t = -18.9$	< 0.0001
Wilcoxon signed-rank	$W = 0$	< 0.0001
Permutation test	—	< 0.0001

Table 3: Statistical test results. All tests reject $H_0 : \mu = 1.0$.

The Shapiro-Wilk test indicated non-normality in the similarity distribution ($p < 0.05$), validating our use of non-parametric alternatives. The agreement between parametric and non-parametric tests strengthens confidence in our findings.

4.2 Effect Size Analysis

Cohen's $d = -0.927$ indicates a large effect size by conventional standards. The negative sign reflects that observed similarity falls below the expected value of 1.0. Bootstrap 95% confidence interval for the effect size is $[-1.02, -0.83]$, excluding zero and confirming the robustness of this estimate.

The WEAT-style effect size of 0.93 places this bias in the "large" category within the embedding bias literature, comparable to gender-

career associations found in word embeddings (Caliskan et al., 2017).

4.3 Robustness Analysis

To assess whether our findings reflect a stable phenomenon rather than a sampling artifact, we analyzed effect size stability as the sample grew from initial collection through final dataset. Table 4 presents this progression.

N	Mean Sim.	Std Dev	Cohen's d
76	0.9933	0.0083	-0.808
158	0.9934	0.0074	-0.894
210	0.9939	0.0067	-0.905
285	0.9937	0.0065	-0.960
334	0.9939	0.0068	-0.902
406	0.9940	0.0064	-0.927

Table 4: Effect size stability as sample size increased. The effect remained in the "large" range throughout, with mean similarity stable at approximately 0.994.

Several observations support the robustness of our findings:

1. Mean similarity remained remarkably stable at approximately 0.994 across all sample sizes, from 76 to 406 pairs.

2. Standard deviation decreased as expected with larger samples (0.0083 to 0.0064), indicating the effect is consistent across posts rather than driven by outliers.

3. Effect size stabilized in the "large" range ($d \approx -0.9$) rather than regressing toward zero as would occur if the initial finding were a statistical fluke.

4. The more than fivefold increase in sample size—incorporating diverse professional domains via hashtag search—did not diminish the effect.

4.4 Distribution Characteristics

The distribution of cosine similarities across all 406 pairs centers at 0.994, clearly separated from the expected value of 1.0 under the null hypothesis. No pair achieved perfect similarity of 1.0, and all pairs showed some degree of embedding divergence based on gendered name. The minimum observed similarity was 0.971 and the maximum was 0.999, indicating consistent bias across all content types rather than outlier-driven effects.

5 Discussion

5.1 Interpretation of Findings

Our results demonstrate that LLaMA-3 produces systematically different embeddings for identical professional content when author names differ by perceived gender. The approximately 0.6% deviation from perfect similarity, while numerically small, represents a large effect by statistical standards ($d = -0.93$) because this deviation occurs consistently across all 406 tested pairs.

This finding indicates that the model encodes gender-associated information from names into the embedding representation, even when the actual content—the professional headline and post text that should determine relevance—remains identical. The embedding space itself treats gendered names differently, not as a function of content quality or relevance, but as an inherent property of how the model processes names.

The consistency of this effect across our sample is particularly notable. We did not find that some content types showed bias while others did not; rather, the bias manifested uniformly across diverse professional domains, from technology to healthcare to education. This suggests the phenomenon reflects a fundamental property of how LLaMA-3 represents gendered names rather than an artifact of particular content categories.

5.2 Implications for Professional Networks

The implications of embedding-stage bias extend throughout the retrieval pipeline. Embeddings serve as the first filter in search systems—they determine which candidates enter consideration before any downstream ranking or personalization. A systematic difference in how male-named and female-named profiles embed means that identical professional qualifications may receive different initial relevance scores.

At LinkedIn's scale of over one billion users, even small systematic biases aggregate to substantial effects. If embedding bias causes female-named profiles to rank slightly lower

in initial retrieval, this compounds across millions of daily searches. The affected individuals may never appear in search results, connection suggestions, or job recommendations—not because of their qualifications, but because of how the model encodes their name.

Cold-start users face particular vulnerability. LinkedIn's own research on their 360Brew ranking system notes that behavioral signals improve ranking quality, but new users lack engagement history to generate these signals (Team, 2025). For new professionals entering the job market, embedding-stage bias cannot be corrected by downstream systems that rely on behavioral data that does not yet exist.

5.3 Limitations

We acknowledge several important limitations of this work:

Base model versus fine-tuned model. We tested LLaMA-3.2-3B in its base configuration, while LinkedIn uses a fine-tuned variant. Fine-tuning on LinkedIn-specific data could either reduce bias (if training data is balanced and objectives penalize differential treatment) or amplify it (if historical engagement data reflects existing societal biases). Our results represent the foundation model's inherent behavior; production bias may differ.

Name selection. Our name transformations used Western gendered name pairs (e.g., Robert/Roberta, Michael/Michelle). Results may not generalize to other naming conventions, including names from non-Western cultures, names that do not follow binary gender patterns, or names where gender association varies by cultural context.

Content domain. We tested LinkedIn-style professional posts exclusively. The model may behave differently for other content types, longer documents, or different prompt formats. Our findings speak specifically to the retrieval use case we examined.

No downstream measurement. We measured embedding similarity, not actual search rankings, job recommendations, or user outcomes. While embedding bias is a necessary condition for retrieval bias, we did not directly measure effects on end-user experience. Demonstrating real-world harm would require access to LinkedIn's production systems or user studies.

Association not causation. We demonstrate that bias exists in embeddings; we cannot prove that this bias causes harm in production systems. LinkedIn's downstream ranking and recommendation systems may partially or fully correct for embedding-stage bias, though this would require system-level transparency we do not have.

Single model family. We tested only LLaMA-3. Other model families (GPT, Claude, Mistral) may exhibit different bias patterns. Our methodology enables testing other models, but we report only LLaMA-3 results here.

5.4 Relation to LinkedIn's Published Research

LinkedIn's 360Brew paper describes their downstream ranking system as capable of learning personalized preferences that could potentially correct some biases (Team, 2025). However, several factors limit this correction:

First, the ranking model trains on historical engagement data, which itself may reflect societal biases in who users choose to connect with or whose content they engage with. Correcting embedding bias with a model trained on biased engagement data may simply relocate rather than eliminate the bias.

Second, cold-start users cannot benefit from personalization signals. For new professionals—precisely the population most dependent on algorithmic discovery for career opportunities—embedding-stage bias propagates uncorrected.

Third, our finding demonstrates that bias enters the system at the embedding stage, before any correction is possible. Downstream ranking cannot recover candidates excluded from initial retrieval, regardless of how fair the ranking system is.

5.5 Broader Implications

Our findings contribute to growing evidence that LLM-based systems require bias auditing at each stage of their pipeline, not just in final outputs. Researchers have scrutinized the embedding stage less than generation or classification tasks, yet embeddings increasingly

power high-stakes applications including hiring, lending, and content recommendation.

Platforms deploying LLM-based retrieval face a transparency challenge: users cannot easily detect or verify whether their profile embeddings differ systematically from comparable profiles with different demographic characteristics. This information asymmetry limits individual recourse and places responsibility on platforms to proactively audit and address bias.

The methodology we introduce—paired-content comparison with comprehensive statistical analysis—offers a template for embedding bias auditing that others can apply to their systems. We encourage platforms to conduct and publish similar audits.

6 Conclusion

We present the first public audit of gender bias in LLM embeddings for professional network retrieval. Using LinkedIn's published methodology, we generated embeddings for 406 paired posts—identical content with only author name varying by perceived gender—and found systematic bias: mean cosine similarity of 0.994 (versus expected 1.0), Cohen's $d = -0.93$ (large effect), $p < 0.0001$.

This finding held robustly across a more than fivefold increase in sample size and agreement between parametric and non-parametric statistical tests. The bias manifested consistently across diverse professional domains, suggesting it reflects fundamental properties of how LLaMA-3 represents gendered names rather than artifacts of particular content categories.

Our contributions include: (1) the first public audit of LinkedIn's LLaMA-3-based retrieval embeddings for gender bias, (2) application of established paired-content methodology to professional network search, (3) an open-source auditing tool and dataset, and (4) a statistical framework for embedding bias research.

Future work should examine whether fine-tuning reduces or amplifies base model bias, extend analysis to other demographic attributes and naming conventions, measure downstream effects on actual search rankings and user outcomes, and develop bias mitigation techniques applicable to embedding models.

We release all code and data at https://github.com/trustinsights/linkedinbias to enable verification, replication, and extension of this work.

Ethics Statement

This research used publicly available LinkedIn posts collected via screen recording of the LinkedIn mobile application. Our released dataset contains author first and last names, professional headlines, and post text content. No other identifying information is present—we did not collect profile URLs, usernames, employment history, or contact information.

Our dataset contains a mixture of original and transformed names. When an original author's name already matched the target gender coding (e.g., a male-coded name for the male dataset), we retained that name as-is. The LLM modified only names requiring gender transformation. Consequently, the male-coded dataset contains real names for originally male-coded authors, while the female-coded dataset contains real names for originally female-coded authors. The gender-flipped counterparts in each dataset are synthetic transformations.

We collected all posts from public LinkedIn feeds. We release our methodology and tools to enable responsible bias auditing, with the goal of improving fairness in AI-powered professional platforms. We recognize that bias auditing research can itself be misused; we encourage use of these methods for improving system fairness rather than exploiting identified biases.

Acknowledgments

We thank the open-source community for the tools that made this research possible, including the Hugging Face Transformers library and the broader PyTorch ecosystem.

References

Marianne Bertrand and Sendhil Mullainathan. 2004. Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination. American Economic Review, 94(4):991–1013.

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

Jacob Cohen. 1988. Statistical Power Analysis for the Behavioral Sciences, 2nd edition. Lawrence Erlbaum Associates, Hillsdale, NJ.

Jeffrey Dastin. 2018. Amazon scraps secret ai recruiting tool that showed bias against women. Reuters.

LinkedIn Engineering. 2024a. Accelerating llm inference with speculative decoding: Lessons from linkedin's hiring assistant. LinkedIn Engineering Blog.

LinkedIn Engineering. 2024b. Fishdb: A generic retrieval engine for scaling linkedin's feed. LinkedIn Engineering Blog.

Ruoyuan Gao and Chirag Shah. 2020. Toward a fair marketplace: Counterfactual evaluation of the trade-off between relevance, fairness and satisfaction in recommendation systems. Proceedings of the 29th ACM International Conference on Information and Knowledge Management, pages 2145–2148.

Ravi Gupta and 1 others. 2024. Large scale retrieval for the linkedin feed using causal language models. arXiv preprint arXiv:2510.14223.

Matthew Kay, Cynthia Matuszek, and Sean A Munson. 2015. Unequal representation and gender stereotypes in image search results for occupations. In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, pages 3819–3828. ACM.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. In Proceedings of the First Workshop on Gender Bias in Natural Language Processing, pages 166–172. Association for Computational Linguistics.

LinkedIn Corporation. 2024. About linkedin. LinkedIn. Over 1 billion members worldwide.

Chandler May, Alex Wang, Shikha Borber, 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. Association for Computational Linguistics.

Meta AI. 2024. Llama 3.2: Lightweight, open models for on-device ai. Meta AI Blog.

Steven Shimizu, Qing Lan, Tejas Dharamsi, Sundara Raman Ramachandran, Arup De, Yubo Wang, Akhilesh Gupta, Yanning Chen, Ata Fatahi, Zhipeng Wang, and Biao H. 2025. Turbocharging linkedin's recommendation systems with sglang. LinkedIn Engineering Blog.

LinkedIn AI Team. 2025. 360brew: A decoder-only foundation model for personalized ranking and recommendation. arXiv preprint arXiv:2501.16450.

Kyra Wilson and Aylin Caliskan. 2024. Gender, race, and intersectional bias in resume screening via language model retrieval. In Proceedings of the AAAI Conference on Artificial Intelligence. AAAI. Also published as Brookings Institution report.

A Data Collection Prompts

We used Google Gemini 3 Pro in Google AI Studio for both transcription and name transformation. We used the following prompts in our data collection pipeline.

A.1 Video Transcription Prompt

> You are a transcription expert skilled in transcribing social media content. Your remit today is to transcribe the LinkedIn posts shown in the associated video attached. You will transcribe three fields. First is the name of the poster. Second is the headline of the poster. Note that the headline may be truncated due to display constraints. Transcribe exactly what you see. Third is the text content of their post. Exclude sponsored content and advertisements. For posts which contain multimedia such as images or video, transcribe the text to the post exactly and then add a placeholder for audio or video indicating where the multimedia content is. You will return your content in JSON format with three fields name, headline, and text content.

A.2 Male Dataset Generation Prompt

> Our next task is to produce 2 sets of data based on this JSON data. Here's what we're going to do. For each name field in the JSON object, we are going to produce a male name dataset first, then a female name dataset. You'll substitute the nearest male-coded name in each field if the existing name is ambiguous or female coded.
>
> Example: "name": "Christina Applegate" becomes "name": "Christian Applegate"
>
> Produce the male dataset first. If a name is already male-coded, leave it as is. The new dataset should be in the same JSON format and have identical headline and text_content. The only change we are making is the name field.

Critically important: strip away emoji from the name field in the results.

Produce ONLY the male dataset now.

A.3 Female Dataset Generation Prompt

Now perform the same task, but making each name female-coded. Keep all other data the same, as you did in the male data set.