Calibrating WEAT Against Anisotropy: ZCA Whitening as a Geometric Pre-Processing Step for Embedding Association Tests

Seitaro Ono[1], **Senna Ross**[2], **Jun Saiki**[1]

[1]Graduate School of Human and Environmental Studies, Kyoto University, Kyoto, Japan
[2]Brock University, St. Catharines, Ontario, Canada
ono@cv.jinkan.kyoto-u.ac.jp, sz23hr@brocku.ca, saiki.jun.8e@kyoto-u.ac.jp

arXiv:2608.06908v1 [cs.CL] 7 Aug 2026

Abstract

We propose Zero-phase Component Analysis (ZCA) whitening as a geometric pre-processing step for the Word Embedding Association Test (WEAT). WEAT is a bias measurement method widely used in both computational social science and AI fairness research. It relies on cosine similarity as a measure of semantic association, which assumes that the embedding space is approximately isotropic. However, prior work has reported that many widely used language models do not satisfy this assumption, raising concerns about the reliability of bias measurements. ZCA whitening transforms the covariance of the embedding space into the identity matrix while minimizing perturbation to the original vectors. This transformation restores the isotropy condition on which WEAT relies. We evaluate our approach on ten standard WEAT test suites and seven models spanning three architectural families, yielding 70 model–task combinations. The results show that ZCA whitening substantially reduces the anisotropy of the embedding spaces across all models. Particularly for highly anisotropic models, we further observe improvements on standard semantic similarity benchmarks, indicating that the calibrated space better captures semantic associations. After calibration, over 30% of WEAT results change significance status, and effect sizes shift in both directions depending on bias category. These shifts suggest that uncalibrated measurements may both overestimate and underestimate the associations encoded in the embedding space. These findings indicate that previously reported bias measurements in anisotropic embedding spaces should be interpreted with caution and may benefit from re-evaluation with calibrated methods. Our approach contributes to restoring the measurement foundation of WEAT across both computational social science and AI fairness research.

Code — https://github.com/seigit/zca-weat

Accepted at the 9th AAAI/ACM Conference on AI, Ethics, and Society (AIES 2026). This is an extended version including appendices; the final published version will appear in the AAAI Digital Library.

1 Introduction

Word embeddings (Mikolov et al. 2013; Pennington, Socher, and Manning 2014) encode statistical regularities of language, including social biases that can propagate to downstream applications (Bolukbasi et al. 2016; Caliskan, Bryson, and Narayanan 2017; Garg et al. 2018). Measuring these biases accurately is necessary for developing equitable NLP systems. The Word Embedding Association Test (WEAT; Caliskan, Bryson, and Narayanan 2017) transposes the logic of the Implicit Association Test (IAT; Greenwald, McGhee, and Schwartz 1998) from human reaction-time studies to the geometry of word embedding spaces, measuring differential cosine-similarity associations between target concepts (e.g., European-American vs. African-American names) and attribute concepts (e.g., pleasant vs. unpleasant words). The WEAT has attracted significant attention in both computational social science and AI fairness research, and has been applied in a large number of studies to measure racial, gender, age, and other biases across diverse models and modalities.

However, the validity of any measurement metric depends on the assumptions under which it operates. WEAT assumes that cosine similarity faithfully captures semantic associations between words. This assumption holds when the embedding space is approximately isotropic (Ethayarajh 2019; Timkey and Van Schijndel 2021; Rudman et al. 2022). In an isotropic space, vectors are distributed with roughly uniform variance in all directions. Under this condition, the angle between two vectors reflects their semantic relationship rather than a geometric artifact.

Recent work has shown that this isotropy assumption fails in many language models: static embeddings exhibit dominant mean directions (Mu and Viswanath 2018), contextualized representations occupy narrow cones with near-unity pairwise similarities (Ethayarajh 2019), and a small number of "rogue dimensions" dominate cosine similarity in Transformers (Timkey and Van Schijndel 2021).

Despite this accumulating evidence, the consequences of anisotropy for bias measurement have received surprisingly little direct attention. Wolfe, Hiniker, and Howe (2024) recently introduced ML-EAT, which provides a comprehensive multilevel evaluation framework and correctly identifies anisotropy as a threat to the validity of cosine-based bias tests. However, the purpose of ML-EAT was to improve the interpretability of EAT measurements, and proposing a correction for anisotropy was outside its scope. A calibration method that restores the geometric conditions under which cosine-based bias measurement becomes reliable remains a significant challenge.

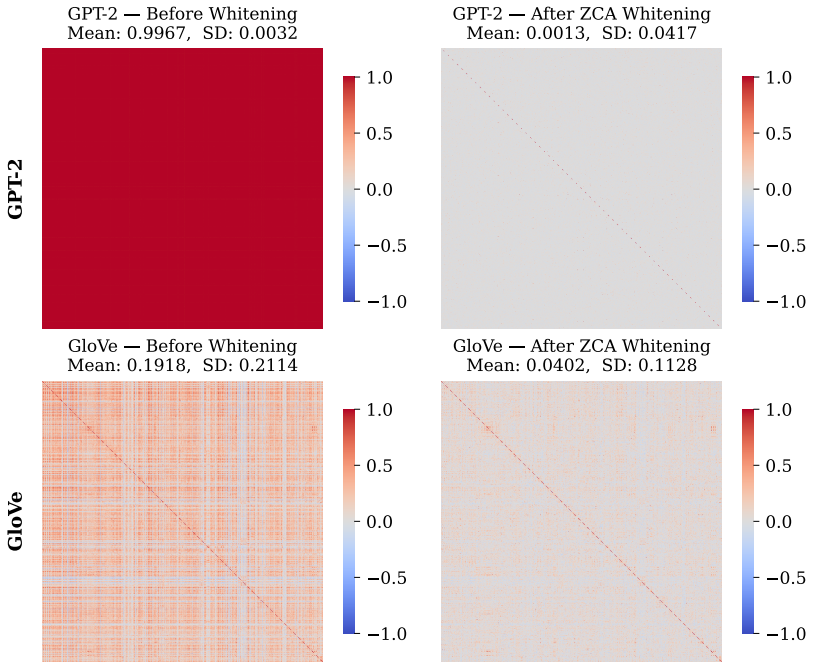

Figure 1: Pairwise cosine similarity matrices for 2,000 randomly sampled embedding vectors from WikiText-103 before (left column) and after (right column) ZCA whitening. **Top:** GPT-2, whose extreme anisotropy (mean cosine similarity $= 0.997$) renders the matrix a uniform red field. **Bottom:** GloVe (mean $= 0.192$), which exhibits moderate anisotropy before whitening. After calibration, both models converge to near-zero mean similarity with meaningful pairwise variation restored.

In this paper, we introduce a calibration step that removes anisotropy from language model embedding spaces, restoring the isotropy assumption under which cosine-based bias measurement is valid. Specifically, we propose ZCA (Zero-phase Component Analysis) whitening as a pre-processing calibration for embedding association tests. ZCA whitening transforms the embedding space to have identity covariance while minimizing the perturbation to the original vectors, making it well suited for effectively reducing anisotropy. Moreover, whitening has been shown to improve the quality of embeddings on standard semantic similarity benchmarks (Huang et al. 2021; Su et al. 2021), suggesting that calibrated embedding spaces may better capture the semantic associations on which bias measurement depends.

We evaluate our approach across seven embedding models spanning three architectural families (static, contextualized, and contrastive unsupervised models) using the ten standard WEAT test suites from Caliskan, Bryson, and Narayanan (2017). We make the following contributions:

1. We propose ZCA whitening as a novel pre-processing calibration method for embedding association tests and show that it substantially reduces anisotropy across all seven models in the embedding space.
2. We show that cosine similarity in the calibrated space more faithfully captures semantic associations on standard similarity benchmarks, supporting the validity of bias measurements in the calibrated space.
3. We systematically show that calibration changes WEAT outcomes in both directions across 70 model–task combinations, reducing measured associations in some configurations and increasing them in others.

2 Related Work

2.1 The Embedding Association Test

Caliskan, Bryson, and Narayanan (2017) introduced the Word Embedding Association Test (WEAT), a measurement of intrinsic bias in word embeddings drawing on the design of the IAT (Greenwald, McGhee, and Schwartz 1998). The WEAT quantifies the relative association of two target groups (such as European-American and African-American names) with two attribute groups (such as pleasant and unpleasant words). Given target sets X and Y and attribute sets A and B, the individual association for each target word w is:

$$s(w, A, B) = \frac{1}{|A|} \sum_{a \in A} \cos(\vec{w}, \vec{a}) - \frac{1}{|B|} \sum_{b \in B} \cos(\vec{w}, \vec{b}) \quad (1)$$

The test statistic aggregates individual associations across the two target sets:

$$S(X, Y, A, B) = \sum_{x \in X} s(x, A, B) - \sum_{y \in Y} s(y, A, B) \quad (2)$$

The effect size d normalizes this difference:

$$d = \frac{\text{mean}_{x \in X}\ s(x, A, B) - \text{mean}_{y \in Y}\ s(y, A, B)}{\text{std}_{w \in X \cup Y}\ s(w, A, B)} \quad (3)$$

This effect size is analogous to Cohen's d (Cohen 1992), where values of 0.2, 0.5, and 0.8 are conventionally interpreted as small, medium, and large effects. Statistical significance is assessed via a one-sided permutation test that computes the p-value as:

$$p = \Pr_i\big[S(X_i, Y_i, A, B) > S(X, Y, A, B)\big] \tag{4}$$

where $\{(X_i, Y_i)\}$ denotes the set of all equally sized partitions of $X \cup Y$.

Since its introduction, the WEAT has been extended across modalities and architectures, in order to measure social biases in AI systems. Text-based variants include SC-WEAT for single-target associations (Caliskan, Bryson, and Narayanan 2017), SEAT for sentence-level embeddings (May et al. 2019), a pretraining-objective-based adaptation (Kurita et al. 2019), and CEAT, which treats contextualization as a random effect at the word level (Guo and Caliskan 2021). The framework has also been extended beyond text to image encoders (iEAT; Steed and Caliskan 2021), grounded vision-and-language models (Grounded-WEAT; Ross, Katz, and Barbu 2021), CLIP (Wolfe and Caliskan 2022; Wolfe, Banaji, and Caliskan 2022a; Radford et al. 2021), speech models (SpEAT; Slaughter et al. 2023), and text-to-video generators (VEAT; Sun et al. 2026). Recently, Wolfe, Hiniker, and Howe (2024) proposed ML-EAT, which provides multilevel interpretable bias measurements and includes an anisotropy diagnostic. All of these variants rely on cosine similarity and are therefore subject to the same geometric artifacts that this paper addresses. Their widespread adoption across NLP, computer vision, and speech processing underscores the importance of ensuring that their measurements are geometrically sound.

2.2 Applications of EATs in Social Science

Embedding association tests have been widely adopted in the social sciences as a tool for studying societal phenomena at scale. A growing body of work uses EATs to probe contemporary cultural patterns in large corpora, including gender defaults and stereotypes in internet English (Caliskan et al. 2022; Bailey, Williams, and Cimpian 2022), the relationship between gender bias and societal structure (Napp 2023), and temporal shifts in Wikipedia (Schmahl et al. 2020). Cross-cultural and cross-linguistic studies have extended this reach to dozens of languages (Lewis and Lupyan 2020; Charlesworth et al. 2021; Mukherjee et al. 2023). Historical corpora have enabled the study of how these associations evolve over time, including a century of gender and ethnic stereotypes (Garg et al. 2018), 200 years of social-group stereotypes (Charlesworth, Caliskan, and Banaji 2022), intersectional stereotypes (Charlesworth et al. 2024a; Borenstein et al. 2023), xenophobia in 19th–20th century travel literature (Sunsay 2023), and the expansion of the societal "moral circle" (Leach, Kitchin, and Sutton 2023). Collectively, these applications demonstrate that EATs can serve not only as a measure of contemporary associations but also as a tool for tracing long-term societal change.

In applied domains, EATs have been used to measure biases in biomedical research (Rios, Joshi, and Shin 2020), clinical notes (Cobert et al. 2024), legal opinions (Matthews, Hudzina, and Sepehr 2022), and court proceedings (Dutta et al. 2023). Gray and Wu (2025) further proposed SD-WEAT, a variant handling multi-level attribute groups for healthcare bias benchmarking. Recent work has also linked embedding-based measurements to human-level implicit associations, demonstrating meaningful correspondence between embedding biases and societal stereotypes (Charlesworth et al. 2024b; Morehouse et al. 2023; Manzini et al. 2019).

These diverse applications underscore that the validity of EAT measurements is not a narrow technical concern but a matter of scientific integrity across multiple disciplines. However, several studies have identified potential problems with methods that rely on cosine similarity between embedding vectors, suggesting that the geometric properties of the embedding space may affect the reliability of such measurements.

2.3 Embedding Anisotropy and Its Implications for WEAT

Anisotropy in word embeddings describes a geometric property in which vectors cluster in a narrow region of the embedding space (Cai et al. 2021), causing cosine similarities between arbitrary word pairs to be systematically high. Ethayarajh (2019) demonstrated that contextualized representations from BERT, ELMo, and GPT-2 occupy a narrow cone with average cosine similarities approaching 1.0 in upper layers, while Mu and Viswanath (2018) identified dominant principal components in word2vec and GloVe, and Timkey and Van Schijndel (2021) showed that a few "rogue dimensions" dominate cosine computation in Transformers. Further work has documented additional effects (Zhou et al. 2022; Godey, de la Clergerie, and Sagot 2024; Machina and Mercer 2024; Rajaee and Pilehvar 2022).

For WEAT, such geometric distortion can compress or inflate the differential cosine associations that define its effect size, producing measurements that do not accurately reflect the underlying bias structure. Wolfe, Hiniker, and Howe (2024) incorporated an anisotropy diagnostic into ML-EAT that flags unreliable models, but ML-EAT aims at interpretability rather than correction. Mu and Viswanath (2018) proposed removing top principal components, but this requires explicitly discarding information and selecting a model-dependent hyperparameter, limiting its practical applicability. Thus, a calibration method that restores the reliability of cosine-based measurement without requiring such choices is desirable.

3 Approach

3.1 How Anisotropy Distorts Cosine-Based Association Tests

Cosine similarity between two vectors $\vec{u}, \vec{v} \in \mathbb{R}^d$ is defined as:

$$\cos(\vec{u}, \vec{v}) = \frac{\vec{u} \cdot \vec{v}}{\|\vec{u}\| \, \|\vec{v}\|} \tag{5}$$

For cosine similarity to serve as a reliable measure of semantic association, the embedding space should be approx-

imately isotropic (Ethayarajh 2019; Wolfe, Hiniker, and Howe 2024). Specifically, the covariance matrix Σ of the embedding vectors should be approximately proportional to the identity matrix (Rudman et al. 2022):

$$\Sigma \approx \sigma^2 I \tag{6}$$

When Σ departs from proportionality to I, cosine similarity may become systematically distorted. Because cosine similarity is dominated by high-variance directions, vectors can appear more similar than they semantically are, while meaningful differences along low-variance directions are masked (Mu and Viswanath 2018; Timkey and Van Schijndel 2021).

Since the WEAT computes differential cosine-similarity associations between target and attribute word sets (Equations 1–3), the effect size d may be affected by anisotropy if the distortion is not uniform across the word sets involved in the test. This suggests that anisotropy could both inflate and compress genuine bias signals, depending on the relationship between the word sets and the covariance structure of the embedding space.

To address this potential distortion, a calibration step that transforms the covariance matrix of the embedding space toward the identity matrix is needed, thereby restoring the geometric conditions under which cosine similarity can function as a reliable measure of semantic association.

3.2 ZCA Whitening

We adopt Zero-phase Component Analysis (ZCA) whitening as the pre-processing calibration step.

Given a set of embedding vectors $\{x_1, x_2, \ldots, x_n\} \subset \mathbb{R}^d$, we first compute the sample mean μ and the sample covariance matrix Σ:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \Sigma = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)(x_i - \mu)^\top \tag{7}$$

We then center each vector by subtracting the mean and apply the ZCA whitening matrix W_{ZCA} to obtain the transformed vectors:

$$z_i = W_{\text{ZCA}}(x_i - \mu), \quad W_{\text{ZCA}} = \Sigma^{-1/2} \tag{8}$$

where $\Sigma^{-1/2}$ denotes the symmetric matrix square root of Σ^{-1}. To compute this matrix, we perform the eigendecomposition of the covariance matrix $\Sigma = U\Lambda U^\top$, where U is an orthogonal matrix whose columns are the eigenvectors and Λ is a diagonal matrix of the corresponding eigenvalues. The whitening matrix then takes the explicit form:

$$W_{\text{ZCA}} = U\Lambda^{-1/2}U^\top \tag{9}$$

This transformation rescales each eigenvector direction by the inverse square root of its eigenvalue: dimensions with disproportionately large variance are compressed, while dimensions with small variance are expanded, resulting in uniform variance across all directions.

The transformed vectors $\{z_i\}$ have identity covariance by construction:

$$\text{Cov}(z) = W_{\text{ZCA}}\,\Sigma\,W_{\text{ZCA}}^\top = \Sigma^{-1/2}\,\Sigma\,\Sigma^{-1/2} = I, \tag{10}$$

where the final equality follows from the symmetry of $W_{\text{ZCA}} = \Sigma^{-1/2}$ (see Appendix A for a full derivation). Since $\text{Cov}(z) = I$, the calibrated embedding space satisfies the isotropy condition (Equation 6 with $\sigma^2 = 1$) under which cosine similarity can function as a reliable measure of semantic association.

Among the family of whitening transformations that produce identity covariance, ZCA uniquely minimizes the expected squared distance between original and transformed vectors (Kessy, Lewin, and Strimmer 2018), making it particularly suitable for bias measurement calibration where preserving the original semantic structure is essential. Moreover, whitening has been shown to substantially improve scores on standard semantic textual similarity benchmarks (Huang et al. 2021; Su et al. 2021), suggesting that the calibrated embedding space more faithfully captures semantic associations between words and may therefore provide a more reliable basis for extracting social associations that were obscured by anisotropy in the original space.

3.3 Calibration Pipeline

Our full measurement pipeline operates in three stages.

In the first stage, we estimate the whitening matrix from a large reference sample of embedding vectors. We adopt WikiText-103 (Merity et al. 2017) as the reference corpus for collecting these samples. From this corpus, we draw 100,000 samples and feed them into the model under evaluation to obtain the corresponding embedding vectors. Using the resulting embeddings, we compute the sample mean μ and the covariance matrix Σ (Equation 7), from which we estimate the ZCA whitening matrix W_{ZCA} (Equation 9). This estimation is performed once per model and the resulting whitening statistics are reused across all test suites.

In the second stage, we apply the estimated whitening transformation to the specific word vectors involved in the WEAT test suites. Each word vector is centered by subtracting the reference mean μ and then transformed by W_{ZCA} to obtain the calibrated vector (Equation 8).

In the third stage, we compute WEAT scores (effect size d and permutation p-value) using the calibrated vectors.

This pipeline is modular: it can be applied to any embedding model and any cosine-similarity-based bias metric without modifying the metric itself.

4 Experimental Setup

4.1 Models

We evaluate seven embedding models spanning three architectural families.

Static embeddings. GloVe (Pennington, Socher, and Manning 2014) trained on Wikipedia and Gigaword (6B tokens, 100 dimensions), and word2vec (Mikolov et al. 2013) trained on Google News (approximately 100B tokens, 300 dimensions). We use the 6B-token, 100-dimensional GloVe model rather than the 840B-token, 300-dimensional variant used in Caliskan, Bryson, and Narayanan (2017). Our purpose is not to replicate their specific results but to evaluate the impact of anisotropy calibration, and the 6B model exhibits sufficient anisotropy (Table 2) to serve this purpose.

Table 1: The ten standard WEAT test suites (Caliskan, Bryson, and Narayanan 2017). Each suite specifies two target sets (X, Y) and two attribute sets (A, B). $|\cdot|$ denotes the number of stimulus words in each set, following the lists provided in Caliskan, Bryson, and Narayanan (2017). EA and AA denote European-American and African-American, respectively.

ID	**Bias Type**	**Target** X	**Target** Y	**Attribute** A	**Attribute** B	$\lvert X\rvert$	$\lvert Y\rvert$	$\lvert A\rvert$	$\lvert B\rvert$
W1	Valence	Flowers	Insects	Pleasant	Unpleasant	25	25	25	25
W2	Valence	Instruments	Weapons	Pleasant	Unpleasant	25	25	25	25
W3	Race	EA names	AA names	Pleasant	Unpleasant	32	32	25	25
W4	Race	EA names	AA names	Pleasant	Unpleasant	16	16	25	25
W5	Race	EA names	AA names	Pleasant	Unpleasant	16	16	8	8
W6	Gender	Male names	Female names	Career	Family	8	8	8	8
W7	Gender	Math	Arts	Male terms	Female terms	8	8	8	8
W8	Gender	Science	Arts	Male terms	Female terms	8	8	8	8
W9	Health	Mental disease	Physical disease	Temporary	Permanent	6	6	7	7
W10	Age	Young names	Old names	Pleasant	Unpleasant	8	8	8	8

Contextualized embeddings. BERT-base-cased (Devlin et al. 2019), RoBERTa-base (Liu et al. 2019), and GPT-2 (Radford et al. 2019), each with 12 Transformer layers and 768-dimensional hidden representations. For all Transformer models, we use representations from the final layer, consistent with prior work applying EATs to contextualized models (May et al. 2019; Guo and Caliskan 2021; Wolfe, Hiniker, and Howe 2024). Word-level embeddings are obtained following the Aggregated procedure of Bommasani, Davis, and Cardie (2020): for each target or attribute word, we sample up to 20 sentences containing that word from a preprocessed WikiText-103 corpus. Preprocessing retains sentences of 7–75 tokens and excludes section headers, yielding up to 200,000 indexed sentences with a fixed random seed for reproducibility. Within each sentence, subword tokens of the target word are mean-pooled into a single contextualized vector, and the resulting vectors across the 20 sentences are then averaged to produce one word-level representation. Bommasani, Davis, and Cardie (2020) reported that bias evaluation results were fairly stable across $n_i \in \{20, 50, 100\}$ contexts per word. We therefore adopted $n_i = 20$ to maintain consistency with their protocol and to minimize computational overhead. We applied the canonical WEAT to these word-level representations rather than sentence-level variants such as SEAT (May et al. 2019) or distributional variants such as CEAT (Guo and Caliskan 2021). This choice allows us to use a unified measurement procedure across all seven models, including static embeddings for which SEAT and CEAT are not applicable.

Unsupervised contrastive embeddings. We adopt two unsupervised contrastive models from the SimCSE framework (Gao, Yao, and Chen 2021): unsup-SimCSE-BERT-base-uncased (unsup-BERT) and unsup-SimCSE-RoBERTa-base (unsup-RoBERTa). Both models share the same Transformer architecture as their supervised counterparts (12 layers, 768-dimensional hidden representations), and word-level embeddings are obtained using the same aggregation procedure of Bommasani, Davis, and Cardie (2020) described above. These models differ from the standard contextualized models in two key aspects: training signal (unsupervised vs. supervised) and objective function (contrastive vs. language modeling). We include them in our analysis to examine how these factors influence the degree of anisotropy and the patterns of bias detected by WEAT.

Whitening statistics. The whitening matrix W_{ZCA} is estimated from WikiText-103 (Merity et al. 2017). For static models, we sample up to 100,000 vocabulary items. For contextualized and unsupervised contrastive models, we sample up to 100,000 sentences. These contextualized models are pretrained on sentence-level inputs. Feeding isolated words may therefore produce out-of-distribution representations (Bommasani, Davis, and Cardie 2020). To avoid this issue, we use sentence-level inputs to estimate the covariance matrix. To ensure numerical stability when computing the whitening matrix (Equation 9), we regularize the eigenvalues by replacing Λ with $\Lambda+\varepsilon I$ before inversion, yielding:

$$W_{\text{ZCA}} = U(\Lambda + \varepsilon I)^{-1/2}U^{\top} \tag{11}$$

where $\varepsilon = 10^{-3}$. This prevents near-zero eigenvalues from producing excessively large entries in the whitening matrix.

4.2 WEAT Test Suites

We use the ten standard WEAT test suites from Caliskan, Bryson, and Narayanan (2017) (Table 1). These test suites correspond to well-established IAT findings in social psychology and have been used in numerous studies to benchmark bias in NLP models (Kurita et al. 2019; May et al. 2019; Guo and Caliskan 2021; Wolfe, Hiniker, and Howe 2024).

4.3 Evaluation Protocol

Experiment 1: Anisotropy measurement. We quantify anisotropy by computing the mean and standard deviation of pairwise cosine similarities among 2,000 randomly sampled embedding vectors from WikiText-103, both before and after ZCA whitening. An isotropic space yields a mean pairwise cosine similarity near zero. In addition, we qualitatively compare the embedding spaces before and after whitening by examining the explained variance ratio of the covariance matrices, which illustrates how variance is distributed across dimensions. We also visualize pairwise cosine similarity matrices over the sampled vectors.

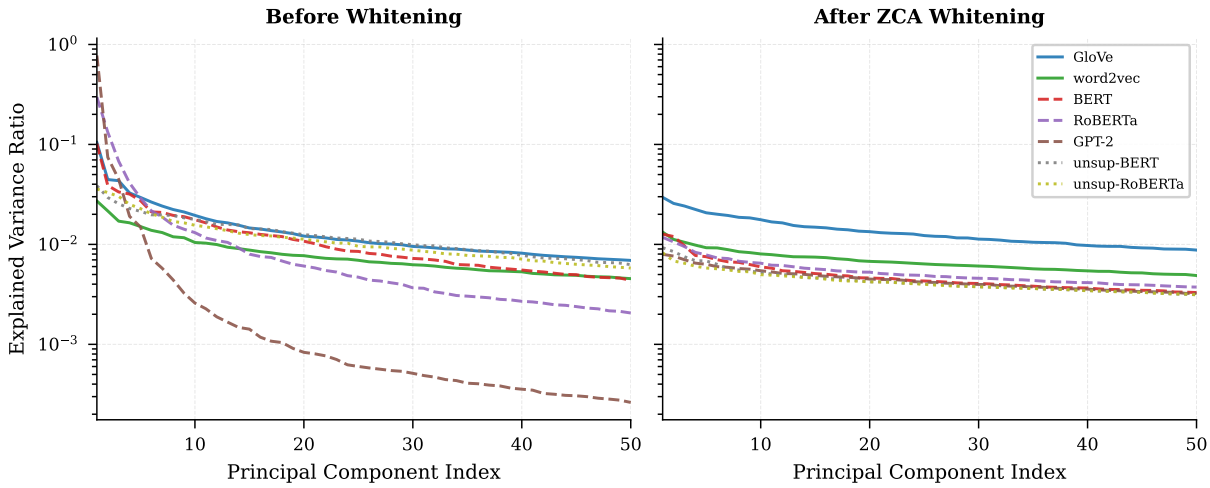

Figure 2: Explained variance ratio of the top 50 principal components (log scale) before (left) and after (right) ZCA whitening for all seven models. **Left:** Contextualized models (GPT-2, RoBERTa, BERT) show extreme variance concentration in the first few components, while static models (GloVe, word2vec) and contrastive models (unsup-BERT, unsup-RoBERTa) exhibit flatter distributions. **Right:** After whitening, all models converge to a near-uniform eigenvalue distribution.

Experiment 2: Impact of anisotropy on semantic structure. We evaluate how anisotropy affects the quality of cosine-similarity-based semantic measurements by computing Spearman rank correlations (ρ) on WordSim-353 (Finkelstein et al. 2002), SimLex-999 (Hill, Reichart, and Korhonen 2015), and STS-B (Cer et al. 2017) before and after whitening. Each benchmark provides word or sentence pairs with human-assigned similarity scores. A higher Spearman correlation indicates that cosine similarity better captures the semantic associations in these benchmarks. Statistical significance is assessed via paired bootstrap resampling (10,000 iterations). If calibration improves these correlations, it suggests that cosine similarity in the calibrated space more faithfully captures semantic associations, supporting the validity of bias measurements conducted in that space.

Experiment 3: Bias measurement with WEAT. For each model–task combination, we compute the WEAT effect size d and permutation p-value (100,000 permutations) on both raw and whitened embeddings. Results are classified based on the change in statistical significance at $p < 0.05$: **Stable**, significance unchanged; **Disappearing**, significant before whitening but non-significant after; **Emerging**, non-significant before but significant after whitening. We then analyze, for each bias category, how these classifications distribute across models and identify the systematic patterns of distortion induced by anisotropy.

5 Results

5.1 ZCA Whitening Reduces Anisotropy

Table 2 presents the mean pairwise cosine similarity before and after ZCA whitening for all seven models. Static embeddings exhibit mild anisotropy (GloVe: 0.192; word2vec: 0.059), while contextualized models display severe anisotropy, with GPT-2 reaching an extreme value of 0.997 (SD = 0.003) in which virtually all vectors point in nearly the same direction. This pattern is consistent with prior reports of cosine similarities concentrating near 1.0 between arbitrary word pairs in GPT-2 Base (Timkey and Van Schijndel 2021; Wolfe, Hiniker, and Howe 2024). The unsupervised contrastive models show intermediate anisotropy (unsup-BERT: 0.386; unsup-RoBERTa: 0.475), consistent with the uniformity-promoting nature of contrastive learning (Gao, Yao, and Chen 2021) but still substantial enough to distort cosine-based measurements. After ZCA whitening, all models converge to near-zero mean pairwise cosine similarity (0.001–0.040), confirming that calibration reduces anisotropy regardless of the initial degree of distortion. This reduction is also visible in the pairwise cosine similarity matrices (Figure 1). GPT-2's matrix appears as a nearly uniform red field before whitening, while GloVe exhibits moderate anisotropy. After calibration, both center near zero with restored pairwise variation. The same qualitative pattern holds for the other five models (see Appendix B).

Table 2: Anisotropy measured by mean pairwise cosine similarity (± standard deviation, SD) among 2,000 randomly sampled vectors before (Raw) and after (White) ZCA whitening.

Model	Raw Mean	Raw SD	Wh. Mean	Wh. SD
GloVe	0.192	0.211	0.040	0.113
word2vec	0.059	0.074	0.013	0.067
BERT	0.821	0.068	0.002	0.043
RoBERTa	0.954	0.024	0.001	0.048
GPT-2	0.997	0.003	0.001	0.042
unsup-BERT	0.386	0.094	0.001	0.042
unsup-RoBERTa	0.475	0.090	0.001	0.041

Figure 2 provides further confirmation through the explained variance ratio of each model's embedding covariance matrix. Before whitening, the contextualized models show extreme variance concentration in the first few principal components, whereas the static and contrastive models exhibit flatter but still non-uniform distributions. After whitening, the explained variance ratio flattens substantially across all models, indicating that dominant directional components have been removed and variance is distributed uniformly across dimensions.

5.2 Calibrated Embeddings Better Capture Semantic Associations

Table 3 reports Spearman rank correlations on three semantic similarity benchmarks before and after ZCA whitening. For models with high anisotropy (GPT-2, RoBERTa, and BERT), whitening produces substantial improvements on word-level benchmarks (WordSim-353 and SimLex-999). The most dramatic gains appear for GPT-2, where Spearman ρ on WordSim-353 increases from 0.263 to 0.620 ($\Delta = +0.358$, $p < 0.001$) and on SimLex-999 from 0.097 to 0.411 ($\Delta = +0.314$, $p < 0.001$), suggesting that the original embedding space was so anisotropic that cosine similarity was largely non-functional and whitening restored much of its discriminative capacity. This pattern is consistent with Timkey and Van Schijndel (2021), who showed that correcting for rogue dimensions substantially improves cosine similarity as a measure of semantic association. RoBERTa and BERT show similar word-level gains ($\Delta = +0.054$ to $+0.191$), while their STS-B scores are either unchanged (RoBERTa) or modestly decreased (BERT: $\Delta = -0.055$).

For word2vec, which has the lowest initial anisotropy (mean cosine 0.059), the results are mixed and small in magnitude ($|\Delta| \leq 0.027$), suggesting that whitening neither substantially improves nor degrades an already near-isotropic space. The unsupervised contrastive models show essentially unchanged word-level scores but small, significant decreases on STS-B ($\Delta = -0.055$ for unsup-BERT; $\Delta = -0.034$ for unsup-RoBERTa), possibly reflecting a domain mismatch between the reference corpus (WikiText-103) and the data on which these models were originally optimized.

Taken together, these results indicate that ZCA whitening preserves the semantic structure of the embedding space. In many cases, it also enables cosine similarity to more faithfully capture semantic associations. This improvement is particularly pronounced for highly anisotropic models, where calibration is most needed. The modest STS-B decreases warrant acknowledgment but are small relative to the word-level gains. Because WEAT measures bias through cosine similarity, these findings support the interpretation that WEAT measurements in the calibrated space are more reliable indicators of genuine semantic associations, particularly for models with high initial anisotropy.

5.3 Anisotropy Distorts WEAT in Both Directions

Figure 3 visualizes WEAT effect sizes before and after ZCA whitening across all 70 model–task combinations, and Figure 4 plots whitened versus raw effect sizes grouped by bias type. Full numerical results, including effect sizes, p-values, and significance change classifications for all combinations, are provided in Appendix C. Across the 70 combinations, we observe 48 Stable cases (68.6%), 12 Disappearing cases (17.1%), and 10 Emerging cases (14.3%). Effect size shifts also occur within Stable cases. Among the Stable cases that remain significant in both spaces, 7 exhibit $|\Delta d| > 0.5$, including GPT-2 on W3 ($d = 1.24 \rightarrow 0.57$) and unsup-RoBERTa on W6 ($d = 1.61 \rightarrow 0.99$). In total, 29 of 70 combinations (41.4%) show either a significance change or a substantial effect-size shift ($|\Delta d| > 0.5$) after calibration, suggesting that anisotropy is a practical source of measurement error.

Table 3: Spearman ρ on semantic similarity benchmarks before (Raw) and after (White) ZCA whitening. Δ denotes the change. Significance: $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, n.s. = not significant (paired bootstrap, 10k iterations).

Model	Bench.	Raw	White	Δ	Sig.
GloVe	WS-353	.477	.586	+.109	***
	SL-999	.297	.380	+.083	***
	STS-B	.560	.599	+.039	***
word2vec	WS-353	.686	.661	−.025	*
	SL-999	.441	.469	+.027	***
	STS-B	.707	.684	−.023	***
BERT	WS-353	.560	.658	+.098	***
	SL-999	.412	.465	+.054	***
	STS-B	.633	.578	−.055	***
RoBERTa	WS-353	.446	.559	+.113	***
	SL-999	.266	.457	+.191	***
	STS-B	.650	.648	−.002	n.s.
GPT-2	WS-353	.263	.620	+.358	***
	SL-999	.097	.411	+.314	***
	STS-B	.428	.596	+.168	***
unsup-BERT	WS-353	.718	.734	+.016	n.s.
	SL-999	.536	.545	+.009	n.s.
	STS-B	.840	.786	−.055	***
unsup-RoBERTa	WS-353	.536	.532	−.004	n.s.
	SL-999	.436	.407	−.029	n.s.
	STS-B	.845	.811	−.034	***

Valence (W1–W2). Most valence measurements cluster near or above the diagonal in Figure 4(a), indicating that effect sizes are preserved or increase after whitening. Static models retain significance throughout, though word2vec shows compression of inflated effects (W2: $d = 1.63 \rightarrow 1.03$). The contextualized models show a markedly different pattern: four of six contextualized model–task combinations exhibit Emerging bias (BERT W2, RoBERTa W2, GPT-2 W1, GPT-2 W2). The most striking case is GPT-2 on W2 ($d = -0.87 \rightarrow d = 1.05$). BERT and RoBERTa on W2, and GPT-2 on W1, follow the same pattern (see Appendix C). Because valence associations such as flowers–pleasant and weapons–unpleasant are among the most robust findings in the IAT literature (Greenwald, McGhee, and Schwartz 1998), their absence in uncalibrated contextual-

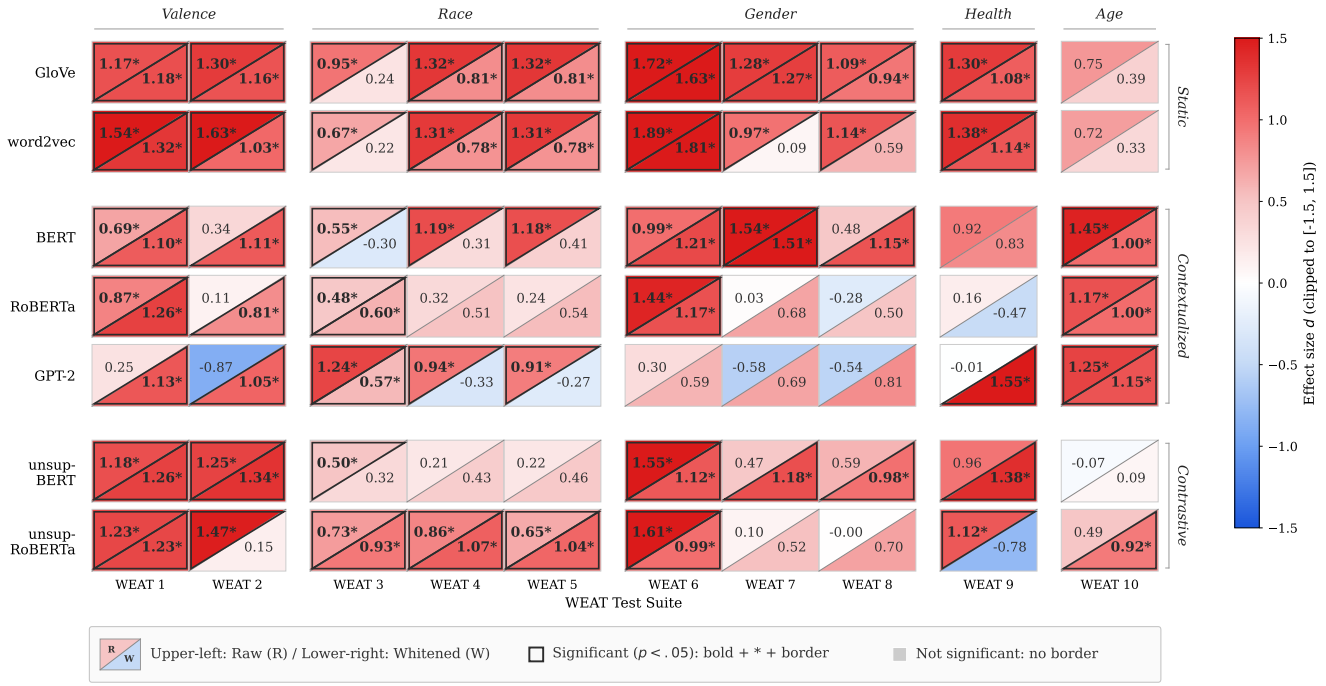

Figure 3: WEAT effect sizes (d) before (upper-left triangle) and after (lower-right triangle) ZCA whitening across seven models and ten test suites. Color intensity encodes effect size magnitude (clipped to $[-1.5, 1.5]$). Bold values with asterisks and bordered cells indicate statistical significance ($p < 0.05$); regular-weight values without borders indicate non-significant results. Models are grouped by architectural family: static (top), contextualized (middle), and unsupervised contrastive (bottom). Test suites are grouped by bias type: Valence (W1–W2), Race (W3–W5), Gender (W6–W8), Health (W9), and Age (W10).

ized embeddings likely reflects geometric distortion rather than a genuine lack of association. The contrastive models are mostly Stable, with one Disappearing case (unsup-RoBERTa W2: $d = 1.47 \rightarrow 0.15$).

Race (W3–W5). Race measurements are predominantly located below the diagonal in Figure 4(b), indicating systematic decreases in effect size after whitening. This category concentrates the bulk of overestimation: 8 of 12 Disappearing cases (66.7%) occur in W3–W5. Both static models show Disappearing bias on W3 (e.g., GloVe: $d = 0.95 \rightarrow 0.24$), with reduced but still significant effects on W4 and W5. Among the contextualized models, BERT exhibits Disappearing bias across all three race tests, and GPT-2 on W4 and W5 (full transitions in Appendix C); RoBERTa is the exception, maintaining stable significance throughout. The contrastive models show only one Disappearing case within the race tests (unsup-BERT W3), and unsup-RoBERTa even shows slightly increased race effects after whitening (W4: $d = 0.86 \rightarrow 1.07$).

Gender (W6–W8). Gender measurements show considerable scatter around the diagonal in Figure 4(c), reflecting coexistence of overestimated and underestimated bias. Static models retain strong significance on W6 (word2vec: $d = 1.89 \rightarrow 1.81$; GloVe: $d = 1.72 \rightarrow 1.63$), but word2vec shows Disappearing cases on W7 and W8. The contextualized models reveal a more nuanced pattern: only one of nine combinations shows Emerging bias (BERT on W8: $d = 0.48 \rightarrow 1.15$). GPT-2 shows directional reversals on W7 ($d = -0.58 \rightarrow 0.69$) and W8 ($d = -0.54 \rightarrow 0.81$), where raw associations flip from negative to positive after whitening. This suggests that severe anisotropy can invert the apparent sign of association. The contrastive models provide further evidence: unsup-BERT shows Emerging bias on both W7 ($d = 0.47 \rightarrow 1.18$) and W8 ($d = 0.59 \rightarrow 0.98$). These emerging effect sizes are comparable to or larger than those in the standard contextualized models, suggesting that while contrastive training reduces geometric distortion, it does not eliminate the bias encoded in the pre-training data.

Health and Age (W9–W10). For health (W9), the most notable result is GPT-2's Emerging case ($d = -0.01 \rightarrow d = 1.55$), where whitening uncovers a strong association that was geometrically masked in the raw space. Unsup-BERT also shows an Emerging case and unsup-RoBERTa a Disappearing one on W9, while static models show mixed patterns (see Appendix C). For age (W10), six of seven models show Stable results, with one Emerging case (unsup-RoBERTa: $d = 0.49 \rightarrow d = 0.92$); the contextualized models maintain large significant effects ($d > 1.0$ after whitening).

5.4 The Role of Anisotropy Severity

The frequency of significance changes appears broadly consistent with the degree of initial anisotropy: GPT-2 (mean cosine 0.997) shows changes in 5 of 10 tests, BERT (0.821)

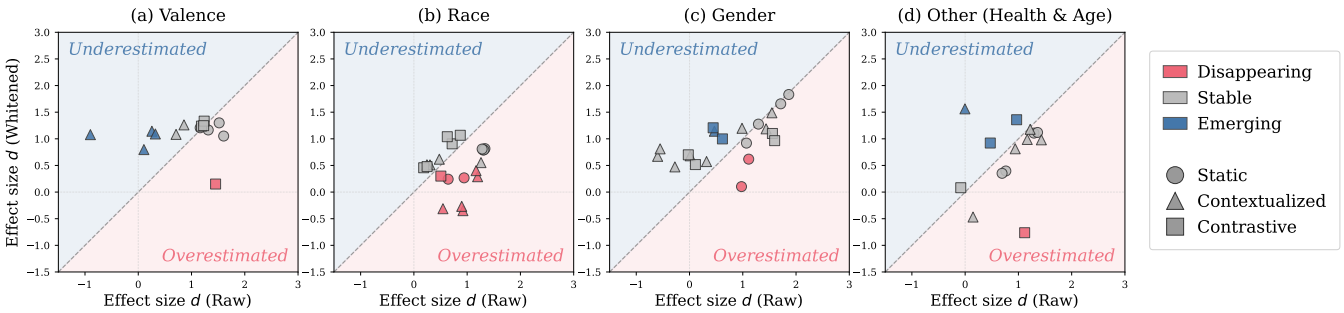

Figure 4: Scatter plots of WEAT effect sizes before (x-axis) and after (y-axis) ZCA whitening, separated by bias type: (a) Valence, (b) Race, (c) Gender, and (d) Other (Health & Age). The diagonal line represents no change; points above the line indicate increased effect sizes after whitening (underestimation in the raw space), and points below indicate decreased effect sizes (overestimation). Marker color denotes significance change: Disappearing (red), Stable (pink), and Emerging (blue). Marker shape denotes architectural family: circle (static), triangle (contextualized), and square (contrastive).

in 5, unsup-BERT (0.386) in 4, while word2vec (0.059) shows changes in 3 of 10 tests. However, RoBERTa (0.954) shows only 1 change despite its high anisotropy, and GloVe (0.192) likewise shows only 1 change, indicating that the relationship between anisotropy severity and the frequency of significance changes is not strictly monotonic and may depend on additional factors such as the specific geometry of the embedding space relative to the WEAT stimulus words.

5.5 Patterns of Distortion Across Bias Categories and Embedding Families

The category-level analysis reveals two systematic asymmetries, summarized visually in Figure 4. First, distortion direction depends on bias category: Disappearing cases concentrate in the race tests, which account for 8 of 12 (66.7%), while Emerging cases are most frequent in the valence tests (4 of 10), with the remainder split evenly between gender (3) and health/age (3). This asymmetry suggests that anisotropy may interact differently with different stimulus configurations. However, the specific mechanism behind this pattern is not fully clear from our experimental results alone. One possibility is that the geometric relationship between target sets and the dominant variance directions of the embedding space differs across bias categories. As a result, anisotropy may amplify cosine differences in some cases and compress them in others. Identifying the precise factors that determine the direction of distortion is an important direction for future work.

Second, distortion direction depends on architectural family. The contextualized models, which exhibit the highest anisotropy (mean cosine 0.821–0.997), account for 6 of 10 Emerging cases (60.0%), consistent with the interpretation that severe anisotropy compresses the cosine similarity range until genuine differential associations fall below the significance threshold. The static models, despite their relatively low anisotropy, contribute 4 of 12 Disappearing cases (33.3%), indicating that even moderate distortion can inflate specific measurements. The contrastive models exhibit both directions of distortion across categories, indicating that contrastive training reduces geometric distortion but neither prevents it uniformly nor eliminates the bias encoded in the pre-training data.

6 Discussion

6.1 Implications for Computational Social Science

Our findings have direct implications for intrinsic bias measurement. The observation that approximately 30% of WEAT measurements change significance status after geometric calibration suggests that a substantial fraction of assessments based on uncalibrated embeddings may be unreliable. Practitioners conducting bias audits should therefore perform an isotropy check before applying cosine-based measurements and calibrate when anisotropy is detected.

The WEAT has been widely used in computational social science to study societal biases that are difficult or costly to measure experimentally, including historical shifts in stereotypes over 200 years (Charlesworth, Caliskan, and Banaji 2022), 100 years of gender and ethnic stereotypes (Garg et al. 2018), gender stereotypes across 25 languages (Lewis and Lupyan 2020), and intersectional stereotypes (Charlesworth et al. 2024a). Our results suggest that conclusions drawn from highly anisotropic models may need to be revisited with calibrated measurements to determine whether reported biases were overestimated, underestimated, or accurately captured. This does not invalidate prior findings; rather, geometric calibration offers a tool to assess their robustness.

The bidirectional nature of the distortion makes this concern particularly pressing. If anisotropy only inflated bias measurements, uncalibrated results could be interpreted as upper bounds. However, because anisotropy can also mask genuine biases, uncalibrated results cannot be treated as conservative estimates either.

Our work parallels methodological debates in psychology about how implicit bias should be measured. Just as Greenwald et al. (2022) provided recommended best practices for research using the IAT, we argue that computational bias measures require analogous methodological discipline, including correction for known geometric distortions. More-

house, Swaroop, and Pan (2025) likewise called for importing social science best practices into LLM bias probing. Our ZCA calibration addresses one specific aspect of measurement invariance: ensuring that WEAT scores are comparable across models with differing degrees of anisotropy.

The importance of calibrating intrinsic bias measurements is reinforced by growing evidence that such biases propagate to downstream behavior. EAT-measured biases in vision-language models correlate with downstream task performance and propagate to zero-shot retrieval (Ghate et al. 2025), image classification (Wolfe, Banaji, and Caliskan 2022b), visual question answering, captioning, and generation (Wolfe and Caliskan 2022; Friedrich et al. 2025), and sentiment classification (Mei, Fereidooni, and Caliskan 2023). Biased AI outputs have also been shown to shape humans' own implicit associations (Sim et al. 2025), establishing a pathway from embedding-level bias to changes in human implicit associations. If intrinsic measurements are distorted by anisotropy, the resulting inaccuracies may lead to erroneous assessments of downstream risk, making geometric calibration a prerequisite for reliable bias auditing.

As AI systems come under regulatory oversight, bias auditing tools must meet high standards of reliability. The EU AI Act (European Parliament and Council of the European Union 2024) mandates bias examination as part of its data governance requirements for datasets used in high-risk AI systems, and the NIST AI Risk Management Framework (National Institute of Standards and Technology 2023) emphasizes valid and reliable measurement in AI risk assessment. ZCA calibration could be integrated into such frameworks as a standard preprocessing step for embedding-based bias measurement, ensuring comparable audit results across models.

6.2 Compatibility with EAT Variants

ZCA whitening operates as a pre-processing calibration step that is agnostic to the specific form of the association test applied afterward. Because it transforms only the embedding space geometry without modifying the test procedure itself, it is compatible with any cosine-similarity-based bias metric. It can thus be applied to sentence embeddings in SEAT (May et al. 2019), contextualized vectors in CEAT (Guo and Caliskan 2021), image embeddings in iEAT (Steed and Caliskan 2021), and speech representations in SpEAT (Slaughter et al. 2023). For ML-EAT (Wolfe, Hiniker, and Howe 2024), our approach provides a corrective step that complements its anisotropy diagnostic: when ML-EAT identifies a model as highly anisotropic, ZCA whitening can restore the conditions under which cosine-based evaluation becomes meaningful. More broadly, any future variant of embedding association tests relying on cosine similarity would benefit from the same calibration procedure.

6.3 Limitations

Several limitations should be acknowledged. First, the whitening matrix is estimated from WikiText-103, and different reference corpora may yield different transformations; the sensitivity of WEAT outcomes to the choice and size of the reference corpus warrants systematic investigation. Second, our evaluation is limited to seven models, and whether the calibration behavior we observe generalizes to a broader range of models requires further investigation. Third, our implementation uses a fixed regularization parameter ε for numerical stability, and we do not provide a sensitivity analysis of how this choice affects calibrated WEAT scores. Fourth, the relationship between calibrated WEAT measurements and downstream task bias remains an open question. Prior work has reported mixed findings: Goldfarb-Tarrant et al. (2021) found frequent divergence, while Ghate et al. (2025) demonstrated correlation in vision-language models. Whether calibration improves this relationship requires future investigation. Addressing these open questions constitutes an important direction for future research.

Beyond these limitations, we emphasize the scope of our contribution. This work evaluates the *geometric reliability* of WEAT measurements. We examine whether cosine similarity is a valid metric within the embedding space, not how the associations it measures shape the outputs of a deployed system. We do not evaluate the alignment between WEAT measurements and human implicit association measures. The latter is a separate question that applies equally to raw and calibrated measurements and has been examined in prior work (Caliskan, Bryson, and Narayanan 2017; Morehouse et al. 2023; Charlesworth et al. 2024b). Extending human alignment analysis to calibrated measurements is a natural next step.

7 Conclusion

This research proposed ZCA whitening as a pre-processing calibration step for the Word Embedding Association Test (WEAT). The method transforms the covariance of the embedding space into the identity matrix while minimizing perturbation to the original vectors, restoring the isotropy condition on which WEAT relies. We evaluated our approach on ten WEAT test suites and seven models (70 model–task combinations). ZCA whitening reduced anisotropy across all models, and for highly anisotropic models it also improved standard semantic similarity benchmark scores. After calibration, over 30% of WEAT results changed significance status, and effect sizes shifted in both directions depending on bias category.

These results suggest that bias measurements obtained in strongly anisotropic spaces may partly reflect geometric properties of the space rather than social associations alone. We do not claim that previously reported findings are invalid. Rather, studies conducted under such conditions may benefit from re-examination with a calibrated metric.

ZCA whitening operates as a model-agnostic preprocessing step. It can be applied not only to WEAT but also to its variants, such as CEAT, iEAT, and SpEAT, and in principle to any bias metric that relies on cosine similarity. The reliability of such measurements is inseparable from the geometry of the space in which they are computed, and our approach offers a practical step toward geometrically valid bias measurement in computational social science and AI fairness research.

Ethical Statement

This work aims to improve the reliability of bias measurement in word embeddings. However, several ethical considerations deserve attention.

Risk of misinterpretation. Our finding that some previously significant bias measurements may not survive geometric calibration could be misinterpreted as evidence that AI systems are less biased than previously reported. We caution against this interpretation. Calibration reveals that while some specific measurements were inflated, others were deflated, and the overall picture is one of measurement unreliability rather than bias absence. The appropriate response is to re-measure with calibrated tools, not to conclude that bias is less prevalent. This concern is particularly acute given the bidirectional nature of the distortion we document: practitioners, journalists, or model developers who selectively cite Disappearing cases as evidence of model improvement would be misrepresenting our findings. As Blodgett et al. (2020) argued, NLP bias research must be grounded in clear normative reasoning about who is harmed and how; our calibration method addresses one specific source of measurement error but does not resolve deeper questions about what constitutes bias or how bias measurements should inform practice.

Scope of intrinsic bias metrics. Even with proper geometric calibration, intrinsic bias measurements capture only one dimension of the complex ways in which AI systems can perpetuate social inequities. Embedding-level bias metrics should be used alongside downstream task evaluations (Goldfarb-Tarrant et al. 2021; Cabello, Jørgensen, and Søgaard 2023), qualitative audits, and participatory assessments involving affected communities (Selbst et al. 2019). The communities most likely to be harmed by biased AI systems include racial, gender, and other minoritized groups represented in WEAT target sets. Their perspectives should inform what counts as a meaningful bias measurement in the first place.

Responsibility in auditing contexts. As discussed in Section 6.1, embedding-based bias measurements are increasingly used in regulatory and auditing contexts. This dual-use character introduces specific responsibilities: a calibration method that changes which biases are flagged as significant could, if applied uncritically, either expose previously hidden harms or obscure documented ones. We therefore recommend that calibrated measurements be reported alongside uncalibrated ones rather than replacing them, so that audit trails preserve both perspectives and allow stakeholders to assess the geometric reliability of each measurement.

Acknowledgments

We thank the reviewers of AIES 2026 for their thoughtful and constructive feedback. We also thank our colleagues for helpful discussions throughout this work.

References

Bailey, A. H.; Williams, A.; and Cimpian, A. 2022. Based on Billions of Words on the Internet, People = Men. *Science Advances*, 8(13): eabm2463.

Blodgett, S. L.; Barocas, S.; Daumé III, H.; and Wallach, H. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5454–5476. Online: Association for Computational Linguistics.

Bolukbasi, T.; Chang, K.-W.; Zou, J. Y.; Saligrama, V.; and Kalai, A. T. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in Neural Information Processing Systems*, 29.

Bommasani, R.; Davis, K.; and Cardie, C. 2020. Interpreting Pretrained Contextualized Representations via Reductions to Static Embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4758–4781. Online: Association for Computational Linguistics.

Borenstein, N.; Stańczak, K.; Rolskov, T.; Käfer, N. K.; da Silva Perez, N.; and Augenstein, I. 2023. Measuring intersectional biases in historical documents. In *Findings of the Association for Computational Linguistics: ACL 2023*, 2711–2730.

Cabello, L.; Jørgensen, A. K.; and Søgaard, A. 2023. On the independence of association bias and empirical fairness in language models. In *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency*, 370–378.

Cai, X.; Huang, J.; Bian, Y.; and Church, K. 2021. Isotropy in the Contextual Embedding Space: Clusters and Manifolds. In *International Conference on Learning Representations*.

Caliskan, A.; Ajay, P. P.; Charlesworth, T.; Wolfe, R.; and Banaji, M. R. 2022. Gender Bias in Word Embeddings: A Comprehensive Analysis of Frequency, Syntax, and Semantics. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 156–170.

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

Cer, D.; Diab, M.; Agirre, E.; Lopez-Gazpio, I.; and Specia, L. 2017. SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, 1–14. Vancouver, Canada: Association for Computational Linguistics.

Charlesworth, T. E.; Caliskan, A.; and Banaji, M. R. 2022. Historical representations of social groups across 200 years of word embeddings from Google Books. *Proceedings of the National Academy of Sciences*, 119(28): e2121798119.

Charlesworth, T. E.; Ghate, K.; Caliskan, A.; and Banaji, M. R. 2024a. Extracting intersectional stereotypes from embeddings: Developing and validating the Flexible Intersectional Stereotype Extraction procedure. *PNAS Nexus*, 3(3): pgae089.

Charlesworth, T. E.; Yang, V.; Mann, T. C.; Kurdi, B.; and Banaji, M. R. 2021. Gender stereotypes in natural language: Word embeddings show robust consistency across child and adult language corpora of more than 65 million words. *Psychological Science*, 32(2): 218–240.

Charlesworth, T. E. S.; Morehouse, K.; Rouduri, V.; and Cunningham, W. A. 2024b. Echoes of Culture: Relationships of Implicit and Explicit Attitudes with Contemporary English, Historical English, and 53 Non-English Languages. *Social Psychological and Personality Science*, 15(7): 812–823.

Cobert, J.; Mills, H.; Lee, A.; Gologorskaya, O.; Espejo, E.; Jeon, S. Y.; Boscardin, W. J.; Heintz, T. A.; Kennedy, C. J.; Ashana, D. C.; Chapman, A. C.; and Lee, S. J. 2024. Measuring Implicit Bias in ICU Notes Using Word-Embedding Neural Network Models. *Chest*, 165(6): 1481–1490.

Cohen, J. 1992. A Power Primer. *Psychological Bulletin*, 112(1): 155–159.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Burstein, J.; Doran, C.; and Solorio, T., eds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Dutta, S.; Srivastava, P.; Solunke, V.; Nath, S.; and KhudaBukhsh, A. R. 2023. Disentangling Societal Inequality from Model Biases: Gender Inequality in Divorce Court Proceedings. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI-23)*, 5959–5967.

Ethayarajh, K. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. 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)*, 55–65.

European Parliament and Council of the European Union. 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence (AI Act). Official Journal of the European Union, L 1689. https://eur-lex.europa.eu/eli/reg/2024/1689/oj. Accessed: 2026-08-04.

Finkelstein, L.; Gabrilovich, E.; Matias, Y.; Rivlin, E.; Solan, Z.; Wolfman, G.; and Ruppin, E. 2002. Placing Search in Context: The Concept Revisited. *ACM Transactions on Information Systems*, 20(1): 116–131.

Friedrich, F.; Brack, M.; Struppek, L.; Hintersdorf, D.; Schramowski, P.; Luccioni, S.; and Kersting, K. 2025. Auditing and Instructing Text-to-Image Generation Models on Fairness. *AI and Ethics*, 5(3): 2103–2123.

Gao, T.; Yao, X.; and Chen, D. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In Moens, M.-F.; Huang, X.; Specia, L.; and Yih, S. W.-t., eds., *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 6894–6910. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Garg, N.; Schiebinger, L.; Jurafsky, D.; and Zou, J. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16): E3635–E3644.

Ghate, K.; Charlesworth, T.; Diab, M. T.; and Caliskan, A. 2025. Biases Propagate in Encoder-based Vision-Language Models: A Systematic Analysis From Intrinsic Measures to Zero-shot Retrieval Outcomes. In Che, W.; Nabende, J.; Shutova, E.; and Pilehvar, M. T., eds., *Findings of the Association for Computational Linguistics: ACL 2025*, 18562–18580. Vienna, Austria: Association for Computational Linguistics. ISBN 979-8-89176-256-5.

Godey, N.; de la Clergerie, É.; and Sagot, B. 2024. Anisotropy Is Inherent to Self-Attention in Transformers. In Graham, Y.; and Purver, M., eds., *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, 35–48. St. Julian's, Malta: Association for Computational Linguistics.

Goldfarb-Tarrant, S.; Marchant, R.; Muñoz Sánchez, R.; Pandya, M.; and Lopez, A. 2021. Intrinsic Bias Metrics Do Not Correlate with Application Bias. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1926–1940. Association for Computational Linguistics.

Gray, M.; and Wu, L. 2025. Benchmarking Bias in Embeddings of Healthcare AI Models: Using SD-WEAT for Detection and Measurement Across Sensitive Populations. *BMC Medical Informatics and Decision Making*, 25(1): 258.

Greenwald, A. G.; Brendl, M.; Cai, H.; Cvencek, D.; Dovidio, J. F.; Friese, M.; Hahn, A.; Hehman, E.; Hofmann, W.; Hughes, S.; Hussey, I.; Jordan, C.; Kirby, T. A.; Lai, C. K.; Lang, J. W. B.; Lindgren, K. P.; Maison, D.; Ostafin, B. D.; Rae, J. R.; Ratliff, K. A.; Spruyt, A.; and Wiers, R. W. 2022. Best Research Practices for Using the Implicit Association Test. *Behavior Research Methods*, 54(3): 1161–1180.

Greenwald, A. G.; McGhee, D. E.; and Schwartz, J. L. 1998. Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, 74(6): 1464.

Guo, W.; and Caliskan, A. 2021. Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 122–133.

Hill, F.; Reichart, R.; and Korhonen, A. 2015. SimLex-999: Evaluating Semantic Models With (Genuine) Similarity Estimation. *Computational Linguistics*, 41(4): 665–695.

Huang, J.; Tang, D.; Zhong, W.; Lu, S.; Shou, L.; Gong, M.; Jiang, D.; and Duan, N. 2021. WhiteningBERT: An easy unsupervised sentence embedding approach. In *Findings of the association for computational linguistics: EMNLP 2021*, 238–244.

Kessy, A.; Lewin, A.; and Strimmer, K. 2018. Optimal whitening and decorrelation. *The American Statistician*, 72(4): 309–314.

Kurita, K.; Vyas, N.; Pareek, A.; Black, A. W.; and Tsvetkov, Y. 2019. Measuring bias in contextualized word representations. In *Proceedings of the first workshop on gender bias in natural language processing*, 166–172.

Leach, S.; Kitchin, A. P.; and Sutton, R. M. 2023. Word embeddings reveal growing moral concern for people, animals and the environment. *British Journal of Social Psychology*, 62(4): 1925–1938.

Lewis, M.; and Lupyan, G. 2020. Gender stereotypes are reflected in the distributional structure of 25 languages. *Nature Human Behaviour*, 4(10): 1021–1028.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692.

Machina, A.; and Mercer, R. 2024. Anisotropy is Not Inherent to Transformers. In Duh, K.; Gomez, H.; and Bethard, S., eds., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 4892–4907. Mexico City, Mexico: Association for Computational Linguistics.

Manzini, T.; Yao Chong, L.; Black, A. W.; and Tsvetkov, Y. 2019. Black is to Criminal as Caucasian is to Police: Detecting and Removing Multiclass Bias in Word Embeddings. In Burstein, J.; Doran, C.; and Solorio, T., eds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 615–621. Minneapolis, Minnesota: Association for Computational Linguistics.

Matthews, S.; Hudzina, J.; and Sepehr, D. 2022. Gender and racial stereotype detection in legal opinion word embeddings. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 12026–12033.

May, C.; Wang, A.; Bordia, S.; Bowman, S.; and Rudinger, R. 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, Volume 1 (Long and Short Papers)*, 622–628.

Mei, K.; Fereidooni, S.; and Caliskan, A. 2023. Bias Against 93 Stigmatized Groups in Masked Language Models and Downstream Sentiment Classification Tasks. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*, 1699–1710. Chicago, IL, USA: ACM.

Merity, S.; Xiong, C.; Bradbury, J.; and Socher, R. 2017. Pointer Sentinel Mixture Models. In *Proceedings of the 5th International Conference on Learning Representations (ICLR 2017)*.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.; and Dean, J. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems*, volume 26.

Morehouse, K.; Swaroop, S.; and Pan, W. 2025. Position: Rethinking LLM Bias Probing Using Lessons from the Social Sciences. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, 81841–81860. PMLR.

Morehouse, K. N.; Charlesworth, T. E. S.; Rouduri, V.; and Cunningham, W. A. 2023. Traces of Human Attitudes in Contemporary and Historical Word Embeddings (1800–2000). *Research Square*. Preprint.

Mu, J.; and Viswanath, P. 2018. All-but-the-Top: Simple and Effective Post-processing for Word Representations. In *Proceedings of the 6th International Conference on Learning Representations (ICLR 2018)*.

Mukherjee, A.; Raj, C.; Zhu, Z.; and Anastasopoulos, A. 2023. Global voices, local biases: Socio-cultural prejudices across languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 15828–15845.

Napp, C. 2023. Gender stereotypes embedded in natural language are stronger in more economically developed and individualistic countries. *PNAS Nexus*, 2(11): pgad355.

National Institute of Standards and Technology. 2023. Artificial Intelligence Risk Management Framework (AI RMF 1.0). Technical Report NIST AI 100-1, National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.100-1.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 1532–1543.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 8748–8763. PMLR.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language Models are Unsupervised Multitask Learners. *OpenAI blog*, 1(8): 9.

Rajaee, S.; and Pilehvar, M. T. 2022. An Isotropy Analysis in the Multilingual BERT Embedding Space. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Findings of the Association for Computational Linguistics: ACL 2022*, 1309–1316. Dublin, Ireland: Association for Computational Linguistics.

Rios, A.; Joshi, R.; and Shin, H. 2020. Quantifying 60 Years of Gender Bias in Biomedical Research with Word Embeddings. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing (BioNLP 2020)*, 1–13. Association for Computational Linguistics.

Ross, C.; Katz, B.; and Barbu, A. 2021. Measuring social biases in grounded vision and language embeddings. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 998–1008.

Rudman, W.; Gillman, N.; Rayne, T.; and Eickhoff, C. 2022. IsoScore: Measuring the Uniformity of Embedding Space Utilization. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Findings of the Association for Computational Linguistics: ACL 2022*, 3325–3339. Dublin, Ireland: Association for Computational Linguistics.

Schmahl, K. G.; Viering, T. J.; Makrodimitris, S.; Jahfari, A. N.; Tax, D.; and Loog, M. 2020. Is Wikipedia succeeding in reducing gender bias? Assessing changes in gender bias in Wikipedia using word embeddings. In *Proceedings of the fourth workshop on natural language processing and computational social science*, 94–103.

Selbst, A. D.; Boyd, D.; Friedler, S. A.; Venkatasubramanian, S.; and Vertesi, J. 2019. Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 59–68.

Sim, M.; Brigham, N. G.; Kohno, T.; Charlesworth, T. E. S.; and Caliskan, A. 2025. Biased AI Outputs Can Impact Humans' Implicit Bias: A Case Study of the Impact of Gender-Biased Text-to-Image Generators. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, volume 8, 2375–2386.

Slaughter, I.; Greenberg, C.; Schwartz, R.; and Caliskan, A. 2023. Pre-trained speech processing models contain human-like biases that propagate to speech emotion recognition. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 8967–8989.

Steed, R.; and Caliskan, A. 2021. Image representations learned with unsupervised pre-training contain human-like biases. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 701–713.

Su, J.; Cao, J.; Liu, W.; and Ou, Y. 2021. Whitening Sentence Representations for Better Semantics and Faster Retrieval. arXiv:2103.15316.

Sun, Y.; Saxon, M.; Yang, I.; Gueorguieva, A.-M.; and Caliskan, A. 2026. VEAT Quantifies Implicit Associations in Text-to-Video Generator Sora and Reveals Challenges in Bias Mitigation. arXiv:2601.00996.

Sunsay, C. 2023. A historical evaluation of the disease avoidance theory of xenophobia. *PLOS ONE*, 18(12): e0294816.

Timkey, W.; and Van Schijndel, M. 2021. All bark and no bite: Rogue dimensions in transformer language models obscure representational quality. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 4527–4546.

Wolfe, R.; Banaji, M. R.; and Caliskan, A. 2022a. Evidence for hypodescent in visual semantic AI. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 1293–1304.

Wolfe, R.; Banaji, M. R.; and Caliskan, A. 2022b. Markedness in Visual Semantic AI. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 1269–1279.

Wolfe, R.; and Caliskan, A. 2022. American = White in Multimodal Language-and-Image AI. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 800–812.

Wolfe, R.; Hiniker, A.; and Howe, B. 2024. ML-EAT: A Multilevel Embedding Association Test for Interpretable and Transparent Social Science. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, 1608–1620.

Zhou, K.; Ethayarajh, K.; Card, D.; and Jurafsky, D. 2022. Problems with Cosine as a Measure of Embedding Similarity for High Frequency Words. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 401–423. Dublin, Ireland: Association for Computational Linguistics.

Supplementary Material for:

Calibrating WEAT Against Anisotropy: ZCA Whitening as a Geometric Pre-Processing Step for Embedding Association Tests

A Derivation of the Identity Covariance Property of ZCA-Whitened Vectors

This appendix provides the complete derivation showing that the ZCA-whitened vectors $z_i = W_{\text{ZCA}}(x_i - \mu)$ have identity covariance, as summarized in Equation 10 of the main paper.

Starting from the definition of the sample covariance of $\{z_i\}$:

$$\begin{aligned}
\text{Cov}(z) &= \frac{1}{n-1}\sum_{i=1}^{n} z_i z_i^\top \\
&= \frac{1}{n-1}\sum_{i=1}^{n} W_{\text{ZCA}}(x_i-\mu)\big(W_{\text{ZCA}}(x_i-\mu)\big)^\top \\
&= \frac{1}{n-1}\sum_{i=1}^{n} W_{\text{ZCA}}(x_i-\mu)(x_i-\mu)^\top W_{\text{ZCA}}^\top \\
&= W_{\text{ZCA}}\left(\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)^\top\right) W_{\text{ZCA}}^\top \\
&= W_{\text{ZCA}}\,\Sigma\,W_{\text{ZCA}}^\top \\
&= \Sigma^{-1/2}\,\Sigma\,\Sigma^{-1/2} \\
&= I. \qquad \text{(A1)}
\end{aligned}$$

The third equality uses the transpose identity $(AB)^\top = B^\top A^\top$. The sixth equality substitutes $W_{\text{ZCA}} = \Sigma^{-1/2}$, using the fact that this matrix is symmetric (so $W_{\text{ZCA}}^\top = W_{\text{ZCA}}$). The final equality follows from

$$\Sigma^{-1/2}\,\Sigma\,\Sigma^{-1/2} = \Sigma^{-1/2}\,\Sigma^{1/2}\,\Sigma^{1/2}\,\Sigma^{-1/2} = I.$$

B Pairwise Cosine Similarity Heatmaps for All Models

Figures A1–A3 show pairwise cosine similarity matrices for 2,000 randomly sampled embedding vectors, grouped by architectural family. Across all seven models, ZCA whitening substantially reduces pairwise similarity regardless of the initial anisotropy severity.

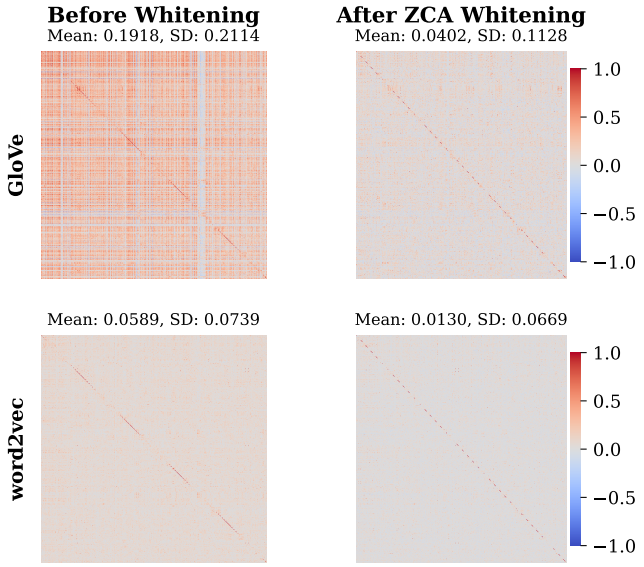

Figure A1: Pairwise cosine similarity matrices for static embedding models (GloVe, word2vec) before (left) and after (right) ZCA whitening.

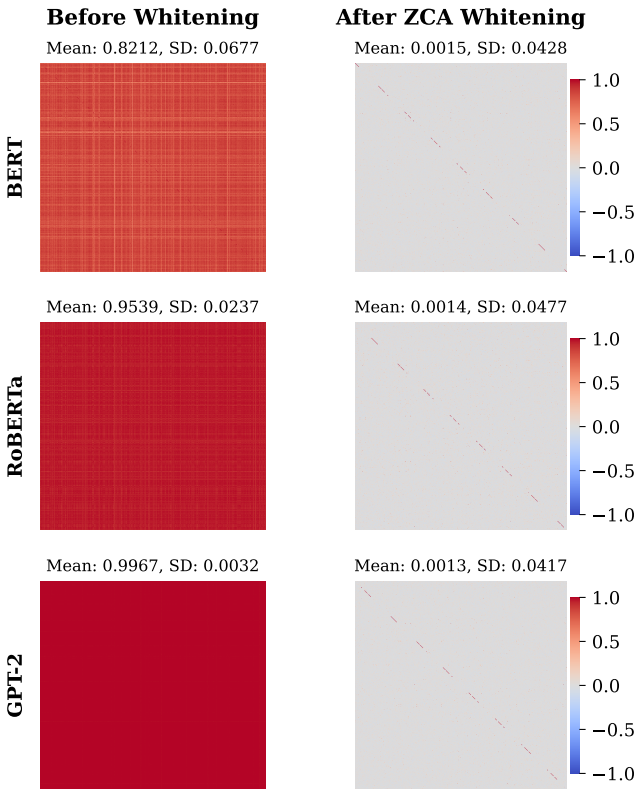

Figure A2: Pairwise cosine similarity matrices for contextualized models (BERT, RoBERTa, GPT-2) before (left) and after (right) ZCA whitening.

Figure A3: Pairwise cosine similarity matrices for contrastive models (unsup-BERT, unsup-RoBERTa) before (left) and after (right) ZCA whitening.

C Full WEAT Results

Table A1: WEAT effect sizes (d) and permutation p-values (100,000 permutations, one-sided) before (Raw) and after (White) ZCA whitening across seven models and ten test suites (see Table 1 for full definitions). Significance at $p < 0.05$. **S** = Stable, **D** = Disappearing (significant $\rightarrow$ non-significant), **E** = Emerging (non-significant $\rightarrow$ significant). Bold d values indicate $p < 0.05$. p-values smaller than 10^{-4} are reported as their order of magnitude.

Static Embeddings

		GloVe					word2vec				
ID	**Bias**	d_R	p_R	d_W	p_W	Chg	d_R	p_R	d_W	p_W	Chg
W1	Valence	**1.17**	10^{-5}	**1.18**	10^{-5}	S	**1.54**	10^{-5}	**1.32**	10^{-5}	S
W2	Valence	**1.30**	10^{-5}	**1.16**	10^{-5}	S	**1.63**	10^{-5}	**1.03**	10^{-5}	S
W3	Race	**0.95**	10^{-5}	0.24	.1660	D	**0.67**	.0031	0.22	.1989	D
W4	Race	**1.32**	10^{-5}	**0.81**	.0098	S	**1.31**	10^{-5}	**0.78**	.0125	S
W5	Race	**1.32**	10^{-5}	**0.81**	.0095	S	**1.31**	10^{-5}	**0.78**	.0131	S
W6	Gender	**1.72**	10^{-5}	**1.63**	10^{-5}	S	**1.89**	10^{-5}	**1.81**	10^{-5}	S
W7	Gender	**1.28**	.0020	**1.27**	.0023	S	**0.97**	.0233	0.09	.4353	D
W8	Gender	**1.09**	.0091	**0.94**	.0280	S	**1.14**	.0100	0.59	.1263	D
W9	Health	**1.30**	.0069	**1.08**	.0254	S	**1.38**	.0025	**1.14**	.0148	S
W10	Age	0.75	.0691	0.39	.2294	S	0.72	.0769	0.33	.2636	S

Contextualized Embeddings

		GPT-2					BERT					RoBERTa				
ID	**Bias**	d_R	p_R	d_W	p_W	Chg	d_R	p_R	d_W	p_W	Chg	d_R	p_R	d_W	p_W	Chg
W1	Valence	0.25	.1869	**1.13**	10^{-5}	E	**0.69**	.0063	**1.10**	10^{-5}	S	**0.87**	.0008	**1.26**	10^{-5}	S
W2	Valence	−0.87	.9993	**1.05**	10^{-5}	E	0.34	.1211	**1.11**	10^{-5}	E	0.11	.3486	**0.81**	.0018	E
W3	Race	**1.24**	10^{-5}	**0.57**	.0109	S	**0.55**	.0142	−0.30	.8829	D	**0.48**	.0272	**0.60**	.0082	S
W4	Race	**0.94**	.0022	−0.33	.8209	D	**1.19**	10^{-5}	0.31	.1886	D	0.32	.1867	0.51	.0772	S
W5	Race	**0.91**	.0032	−0.27	.7706	D	**1.18**	.0001	0.41	.1256	D	0.24	.2530	0.54	.0644	S
W6	Gender	0.30	.2901	0.59	.1259	S	**0.99**	.0217	**1.21**	.0051	S	**1.44**	.0008	**1.17**	.0077	S
W7	Gender	−0.58	.8641	0.69	.0867	S	**1.54**	.0003	**1.51**	.0007	S	0.03	.4803	0.68	.0880	S
W8	Gender	−0.54	.8531	0.81	.0525	S	0.48	.1742	**1.15**	.0084	E	−0.28	.7113	0.50	.1658	S
W9	Health	−0.01	.5044	**1.55**	.0011	E	0.92	.0573	0.83	.0800	S	0.16	.3974	−0.47	.7795	S
W10	Age	**1.25**	.0057	**1.15**	.0058	S	**1.45**	.0004	**1.00**	.0208	S	**1.17**	.0060	**1.00**	.0225	S

Unsupervised Contrastive Embeddings

		unsup-BERT					unsup-RoBERTa				
ID	**Bias**	d_R	p_R	d_W	p_W	Chg	d_R	p_R	d_W	p_W	Chg
W1	Valence	**1.18**	.0008	**1.26**	10^{-5}	S	**1.23**	10^{-5}	**1.23**	10^{-5}	S
W2	Valence	**1.25**	10^{-5}	**1.34**	10^{-5}	S	**1.47**	10^{-5}	0.15	.2990	D
W3	Race	**0.50**	.0220	0.32	.1002	D	**0.73**	.0014	**0.93**	10^{-5}	S
W4	Race	0.21	.2792	0.43	.1132	S	**0.86**	.0059	**1.07**	.0007	S
W5	Race	0.22	.2725	0.46	.1006	S	**0.65**	.0315	**1.04**	.0009	S
W6	Gender	**1.55**	.0005	**1.12**	.0099	S	**1.61**	10^{-4}	**0.99**	.0218	S
W7	Gender	0.47	.1887	**1.18**	.0055	E	0.10	.4236	0.52	.1553	S
W8	Gender	0.59	.1277	**0.98**	.0185	E	−0.00	.5038	0.70	.0864	S
W9	Health	0.96	.0517	**1.38**	.0020	E	**1.12**	.0228	−0.78	.9073	D
W10	Age	−0.07	.5591	0.09	.4343	S	0.49	.1823	0.92	.0314	E