Accurate and Efficient Statistical Testing for Word Semantic Breadth

Yo Ehara
Tokyo Gakugei University
ehara@u-gakugei.ac.jp

Abstract

Measuring the breadth of a word's meaning, or how widely it spreads across contexts, has become feasible with contextualized token embeddings. A word type can be represented as a cloud of token vectors, with dispersion-based statistics serving as proxies for contextual diversity (Nagata and Tanaka-Ishii, 2025). These dispersion-based measurements are useful for deciding appropriate sense distinctions when constructing thesauri and domain-specific dictionaries. However, when comparing the breadth of two word types, naive hypothesis testing on dispersion can be misleading, such that differences in semantic direction (mean embedding direction) can masquerade as dispersion differences, inflating the Type-I error and yielding "statistically significant" outcomes even when there is no true breadth difference. This failure is especially problematic because significance testing is meant to distinguish genuine effects from incidental fluctuations in such small-difference regimes. We propose a Householder-aligned permutation test to isolate dispersion differences from directional differences. Our method applies a single Householder reflection to align the mean direction of each word type with a shared reference direction and then performs a permutation test on a dispersion statistic computed from the aligned token clouds, yielding calibrated, model-agnostic p-values. To make permutation testing practical, we further introduce a graphics processing unit (GPU)-oriented implementation that batches permutations and linear algebra operations. Empirically, our alignment reduced the Type-I error by 32.5% compared with the naive permutation test while preserving sensitivity to genuine breadth differences. Our GPU implementation achieved a 23× speedup over the central processing unit (CPU) baseline.

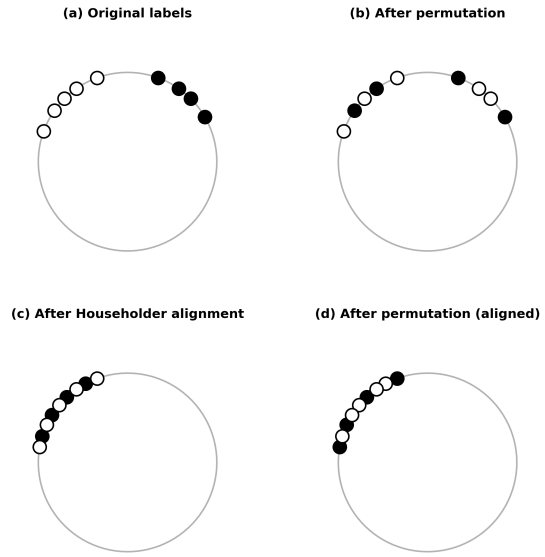

Figure 1: Illustration of Type-I error inflation in naive permutation tests and resolution by Householder alignment. Each circle represents a unit sphere (depicted in 2D); black and white points are ℓ_2-normalized embeddings from two word types. **Top row (naive test):** (a) The two groups occupy different regions of the sphere (different mean directions) but have similar dispersion. (b) After randomly permuting labels, black points now span both regions, making them appear more dispersed than they actually are. This spurious inflation of dispersion leads to false positives. **Bottom row (Householder-aligned test):** (c) After applying Householder reflection, both groups are aligned with the same mean direction and occupy the same region. (d) Permuting labels within the aligned data preserves the true dispersion, yielding calibrated p-values.

1 Introduction

Contextualized embeddings have become a standard representation for modeling lexical meaning in natural language processing (NLP), supporting fine-grained analyses of how meanings vary across contexts and corpora (Periti and Tahmasebi, 2024; Giulianelli et al., 2020; Kutuzov and Giulianelli,

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 48110–48125

2020). Previous work has largely focused on *semantic differences*, for example, whether two usages instantiate the same sense, or how a word's meaning shifts across domains or time. This can often be annotated by directly asking native speakers to judge semantic relatedness.

In contrast, *semantic breadth* (how widely a word extends across diverse contextual realizations) is harder to elicit by immediate inspection: even when speakers can tell that two usages differ in meaning, deciding whether a word is "broad" typically requires examining many naturally occurring usages. Therefore, recent studies have advocated corpus-driven, embedding-based proxies for meaning breadth, such as contextual diversity, which are derived from the geometry of contextualized token vectors (Nagata and Tanaka-Ishii, 2025). To compare such breadth signals across words, corpora, or models, reliable statistical tests are required in the small-difference regime.

A natural choice is a non-parametric permutation test, which is already used in NLP significance testing because it avoids strong distributional assumptions (Dror et al., 2018). However, such tests are also widely recognized as computationally intensive (Dror et al., 2018). More importantly, in our setting, naively applying a permutation test to compare the dispersion of two embedding sets can produce an inflated Type-I error, such that the test may declare "significant" differences even between two sets that have the same breadth.

Figure 1 illustrates the reasons for this inflation. Consider two sets of unit-normalized embedding vectors on the unit sphere, depicted as black and white points in Panel a. Although both groups have similar dispersions, they occupy different regions of the sphere; that is, they have different mean directions. When the labels (Panel b) are naively permuted, some black points are reassigned to positions originally occupied by white points, and vice versa. Consequently, the black group now spans both regions and thus appears much more dispersed than the original. When building a null distribution from such permutations, this mixing of directions systematically inflates the dispersion estimates, causing the test to reject the null hypothesis even when both groups have identical true dispersions.

We address this problem by aligning the two embedding sets *only in the mean direction* before conducting a standard permutation test on dispersion statistics. Specifically, given the unit mean directions $\hat{\mu}_X = \mu_X/\|\mu_X\|_2$ and $\hat{\mu}_Y = \mu_Y/\|\mu_Y\|_2$, a Householder matrix is constructed as follows:

$$H = I - 2\frac{vv^\top}{v^\top v}, \qquad v = \hat{\mu}_X - \hat{\mu}_Y, \quad (1)$$

which satisfies $H\hat{\mu}_X = \hat{\mu}_Y$. Applying H to all the vectors in X removes the mean-direction mismatch while preserving norms and relative geometry.

As shown in Figure 1(c), after Householder reflection, both groups occupy the same region of the sphere. Thus, when the labels (Panel d) are permuted, points are exchanged within this common region; therefore, the dispersion of each permuted group remains comparable to that of the original. Hence, the subsequent permutation test targets the breadth (dispersion) rather than the semantic difference (mean direction), yielding calibrated p values that correctly maintain the nominal Type-I error rate.

Empirically, the proposed Householder-aligned test reduces Type-I error by 32.5% compared to a naive set-level permutation test while retaining sensitivity to genuine breadth differences. To make such resampling practical, we also provide a GPU implementation for the permutation test, yielding a 23× speed-up over the naive CPU-oriented baseline.

Applications The statistical testing task that we focus on in this paper also has practical applications in NLP. One direct application of our work is resource allocation for sense inventory construction in lexicography. As noted in a previous study (Nagata and Tanaka-Ishii, 2025), even for the same word, the granularity of sense distinctions can vary substantially across dictionaries, and producing a high-quality dictionary with carefully standardized sense granularity over a wide vocabulary is difficult because of the associated annotation costs. Another practical use case is comparing contextual breadth across corpora or domains (e.g., general vs. technical text) to identify words whose usage becomes noticeably narrower or broader beyond sampling variation, which can support vocabulary and material selection and corpus-based analyses.

In our setting, we treat a word whose sense inventory has already been curated with high accuracy as a reference and a word whose senses are uncurated or only partially curated as a target. We then statistically test whether the difference in semantic breadth (contextual breadth) derived from contextual embeddings is likely to be more than a

sampling artifact. This enables dictionary builders to distinguish between (i) words that are plausibly well covered by the current number or granularity of senses (small difference) and (ii) words that are more likely to require additional sense splitting or example collection (large difference). In other words, the test provides evidence regarding which words should be prioritized for further lexicographic effort (i.e., which words may warrant finer-grained sense annotation).

Therefore, the contribution of this work is not to estimate the number of senses per se but to provide a well-calibrated testing procedure that can assess the need for additional sense annotation, even in the small-difference regime, while avoiding spurious detections.

Contribution Summary The main contributions of this study are as follows:

- A critical failure mode of naive permutation testing is identified for semantic breadth: mean-direction (semantic) differences inflate Type-I error, undermining the purpose of significance testing.
- A Householder-aligned permutation test is proposed to provably remove the mean-direction confounder and empirically reduce Type-I error by 32.5%.
- A GPU implementation of permutation testing is introduced for embedding-set statistics, achieving a 23 $\times$ speedup over a naive baseline.

Code and other resources are available at `https://rebrand.ly/WordSemanticBreadth`.

2 Related Work

2.1 Meaning–Frequency Law and Sense Inventories

The meaning–frequency law has been repeatedly tested with lexicographic sense inventories such as WordNet (Miller, 1995). A key recurring observation is that the law is more robust when analyzed at aggregated scales (e.g., by averaging senses within frequency/rank bins) than at the level of individual lemmas. Bond et al. explicitly report that bin-level averaging can support the law across multiple languages, while predicting sense counts for a single lemma fails dramatically (Bond et al., 2019). This motivates evaluation protocols that distinguish global trends from word-level uncertainty.

2.2 Contextual Diversity from Contextualized Embeddings

Contextualized encoders such as BERT (Devlin et al., 2019) produce token embeddings that vary across contexts, enabling type-level statistics computed from token clouds. Nagata and Tanaka-Ishii propose a new formulation of Zipf's law using contextual diversity computed from such clouds, positioning it as a corpus-driven proxy for meaning that can be compared across model sizes and architectures (Nagata and Tanaka-Ishii, 2025). At the same time, contextual embeddings exhibit non-trivial geometry; for example, anisotropy and layer-wise differences complicate the interpretation of distances and norms, potentially confounding naive dispersion comparisons (Ethayarajh, 2019). Related work also explores polysemy quantification from contextual embedding geometry (Xypolopoulos et al., 2021), though these approaches often emphasize clustering or multi-sense structure rather than calibrated hypothesis testing. Recent analyses further study relationships between the norm of mean contextualized embeddings and variance (Yamagiwa and Shimodaira, 2025), which is closely related to dispersion statistics based on mean vectors.

2.3 Statistical Significance Testing in NLP

Randomized significance tests, including permutation tests and approximate randomization, have a long history in NLP evaluation, particularly in machine translation (Koehn, 2004; Graham et al., 2014) and have been critically discussed for pitfalls and misinterpretations (Riezler and Maxwell, 2005). More generally, Dror et al. provide practical guidance for choosing and reporting significance tests in NLP (Dror et al., 2018), and Berg-Kirkpatrick et al. analyze when significance claims do and do not transfer across datasets (Berg-Kirkpatrick et al., 2012).

Recent work aims to make permutation testing more computationally tractable or exact for certain statistics (Zmigrod et al., 2022). However, their method is limited to discrete-valued data and does not accommodate continuous quantities such as the dispersion of embedding vectors (e.g., Mean Resultant Length) that we consider. Therefore, their approach is not applicable to the objectives of the present study.

In parallel, uncertainty-aware embedding methods have been proposed to enable hypothesis testing directly in embedding space (Vallebueno et al.,

2024). These threads motivate a testing-centric approach to dispersion-based meaning proxies, where the goal is not only to compute a scalar score, but also to quantify whether observed differences are likely to be genuine rather than sampling artifacts.

Regarding the choice of permutation testing, a non-parametric randomization or permutation approach is a natural choice in NLP evaluation because our statistic is computed from token-embedding clouds without a reliable parametric distributional form. Dror et al. (2018) discuss that while there are sampling-free tests for certain scalar settings, such approaches do not directly extend to our setting where each word is represented as a set of high-dimensional vectors. In such cases, they note that resampling-based procedures such as permutation tests (and related bootstrap-style methods) are often used despite their computational cost. Our contribution is to make this resampling-based testing practical at scale by introducing an efficient alignment-and-testing pipeline for calibrated dispersion comparison.

2.4 Relation to Lexical Semantic Relations

If one wishes to connect contextual dispersion to specific lexical semantic relations (e.g., graded lexical entailment), datasets such as HyperLex provide a benchmark for entailment strength rather than sense counts (Vulić et al., 2017). Such relations may partially overlap with dispersion-based "breadth," but they are not equivalent; testing frameworks can help disentangle what dispersion-based proxies capture in practice.

3 Proposed Methods

3.1 Householder-Aligned Permutation Test

Let w be a word type and $\{\mathbf{h}_{w,i}\}_{i=1}^{n_w} \subset \mathbb{R}^d$ be the contextual embeddings extracted from a pretrained LM at the token positions corresponding to the occurrences of w in a corpus. We normalize each vector onto a unit sphere:

$$\mathbf{x}_{w,i} = \frac{\mathbf{h}_{w,i}}{\|\mathbf{h}_{w,i}\|_2} \in \mathbb{S}^{d-1}, \tag{2}$$

and interpret $\{\mathbf{x}_{w,i}\}$ as samples from a word-specific directional distribution on $\mathbb{S}^{d-1}$. Following prior work on directional statistics, within-word dispersion of these unit vectors is used as a proxy for semantic breadth. Intuitively, polysemous words appear in a wider variety of contexts and thus yield more dispersed contextual vectors.

Problem setting. Given two words, u and k (e.g., an "unknown" word u and a "known" reference k), a *word-level* statistical test is required to determine whether u is semantically broader than k. The two samples are expressed as

$$X = \{\mathbf{x}_i\}_{i=1}^{n} \quad \text{and} \quad Y = \{\mathbf{y}_j\}_{j=1}^{m}, \tag{3}$$

where $\mathbf{x}_i = \mathbf{x}_{u,i}$ and $\mathbf{y}_j = \mathbf{x}_{k,j}$, with n, m fixed (typically, we subsample to a common size to control the variance across words). A natural null hypothesis is that the two words have the same dispersion but different mean directions.

$$H_0 : \text{disp}(X) = \text{disp}(Y) \text{ with } \mathbb{E}[X] \neq \mathbb{E}[Y] \text{ allowed.} \tag{4}$$

This is because the mean direction is a strong nuisance factor in contextual embedding spaces. Even if two words have identical dispersion, their average contextual directions can differ substantially.

Why alignment is required. A standard two-sample permutation test relies on the exchangeability of the pooled samples under H_0. Here, if the two distributions differ in the mean direction, naive label permutation can break exchangeability because the permuted groups become mixtures of different directions and exhibit artificially increased dispersion, such that the resulting permutation distribution no longer corresponds to the intended null. We can show that the transformation that maximizes the mean resultant length of the merged set is given by the Householder transform: its proof sketch is in Appendix D. The Procrustes alignment (Schönemann, 1966) is not directly applicable in our setting because it requires a pointwise correspondence-based objective, typically assuming paired observations. In contrast, our task compares two *unpaired* token-embedding sets for two different words, potentially with different sample sizes, and no natural token-to-token correspondence is available.

Householder mean-direction alignment. The nuisance mean-direction difference is removed by mapping the sample mean direction of X onto that of Y via Householder reflection. Let $\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i$, $\bar{\mathbf{y}} = \frac{1}{m}\sum_{j=1}^{m} \mathbf{y}_j$, $\hat{\boldsymbol{\mu}}_x = \frac{\bar{\mathbf{x}}}{\|\bar{\mathbf{x}}\|_2}$, $\hat{\boldsymbol{\mu}}_y = \frac{\bar{\mathbf{y}}}{\|\bar{\mathbf{y}}\|_2}$. If $\hat{\boldsymbol{\mu}}_x \neq \hat{\boldsymbol{\mu}}_y$, the Householder axis is defined as

$$\mathbf{u} = \frac{\hat{\boldsymbol{\mu}}_x - \hat{\boldsymbol{\mu}}_y}{\|\hat{\boldsymbol{\mu}}_x - \hat{\boldsymbol{\mu}}_y\|_2}, \tag{5}$$

and the reflection matrix is

$$\mathbf{H} = \mathbf{I} - 2\mathbf{u}\mathbf{u}^\top, \tag{6}$$

which satisfies $\mathbf{H}\hat{\boldsymbol{\mu}}_x = \hat{\boldsymbol{\mu}}_y$ and $\mathbf{H}^\top\mathbf{H} = \mathbf{I}$. We then align X by applying $\mathbf{H}$ to every vector in X:

$$\mathbf{x}'_i = \mathbf{H}\mathbf{x}_i \quad (i = 1, \ldots, n), \tag{7}$$

and Y is left unchanged. Because $\mathbf{H}$ is orthogonal, it preserves all within-set distances and rotation-invariant dispersion statistics, changing only the mean direction of X.

Dispersion proxy and test statistics. Let the mean resultant length (MRL) be

$$r(X) = \left\| \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \right\|_2 \in [0, 1], \tag{8}$$

where a larger r indicates a *lower* dispersion (vectors concentrated around a direction), and a smaller r indicates a *higher* dispersion. Many directional models (e.g., von Mises-Fisher) relate r monotonically to a concentration parameter κ by a mapping $\kappa = g_d(r)$ (we use the same estimator as in prior work, such that any monotonic g_d yields an equivalent ordering). We define semantic breadth proxy v as an inverse concentration measure.

$$v(X) = \frac{1}{g_d(r(X))}, \tag{9}$$

and compare words via the log-volume difference

$$T_{\text{obs}} \;=\; \log v(X') - \log v(Y), \tag{10}$$

where $X' = \{\mathbf{x}'_i\}$ is the Householder-aligned version of X. A one-sided alternative "u is broader than k" corresponds to $T_{\text{obs}} > 0$.

Permutation procedure (fixed-space test). The aligned samples $Z = X' \cup Y$ of total size $N = n + m$ are pooled, and through B random permutations, the N vectors are reassigned into two groups of size n and m. For permutation $b \in \{1, \ldots, B\}$, let $(X^{(b)}, Y^{(b)})$ be the permuted split in computing

$$T^{(b)} = \log v(X^{(b)}) - \log v(Y^{(b)}). \tag{11}$$

Using the standard Monte Carlo permutation p-value with a $+1$ correction,

$$p \;=\; \frac{1 + \sum_{b=1}^{B} \mathbb{I}\big[T^{(b)} \geq T_{\text{obs}}\big]}{B + 1}. \tag{12}$$

This test directly returns a word-level significance statement without relying on binning, regression assumptions, or corpus-level trend visualizations.

Mitigating label-dependent preprocessing. A subtlety is that $\mathbf{H}$ is estimated from the observed split, which can, in principle, distort the permutation p-values. To improve calibration with minimal complexity, we adopt a *cross-fitted alignment*: Each word occurrence is randomly split into two halves, estimating $\mathbf{H}$ from the first half only, applying it to the second half, and running a permutation test on the held-out half. This decouples the alignment estimation from the data used for hypothesis testing and can be complemented by a sanity check to determine whether split-half "same-word vs same-word" comparisons yield approximately uniform p-values. Appendix E shows an example of this sanity check experiment.

3.2 GPU-Accelerated Permutation Inference

The Householder-aligned test is computationally dominated by the permutation stage: a naïve implementation loops over B permutations and repeatedly recomputes group means and norms, incurring $O(BNd)$ operations per word pair (with $N = n + m$). Crucially, the required operations almost entirely comprise dense linear algebra (matrix multiplication, reduction, and normalization) in GPU execution.

Vectorized formulation with sign matrices. Let $\mathbf{X} \in \mathbb{R}^{N \times d}$ be a pooled matrix with rows of aligned vectors in $Z = X' \cup Y$. Each permutation b can be represented by a sign vector $\mathbf{s}^{(b)} \in \{+1, -1\}^N$ with exactly n entries $+1$ (assigned to group 1) and m entries -1 (group 2). These are stacked into a sign matrix

$$\mathbf{S} = \begin{bmatrix} (\mathbf{s}^{(1)})^\top \\ \vdots \\ (\mathbf{s}^{(B)})^\top \end{bmatrix} \in \{+1, -1\}^{B \times N}. \tag{13}$$

We define the total sum vector $\mathbf{t} \in \mathbb{R}^d$ and the signed group-difference sums $\mathbf{U} \in \mathbb{R}^{B \times d}$ as

$$\mathbf{t} = \mathbf{1}^\top\mathbf{X}, \qquad \mathbf{U} = \mathbf{S}\mathbf{X}. \tag{14}$$

For each permutation b, $\mathbf{U}_{b,:} = \sum_{i=1}^{N} s_i^{(b)} \mathbf{X}_{i,:}$ equals *(group1 sum) − (group2 sum)*. Because (group1 sum) + (group2 sum) = $\mathbf{t}$ for every permutation, both group sums are recovered without a second matrix multiplication, as follows:

$$\boldsymbol{\sigma}_1^{(b)} = \frac{\mathbf{t} + \mathbf{U}_{b,:}}{2}, \qquad \boldsymbol{\sigma}_2^{(b)} = \frac{\mathbf{t} - \mathbf{U}_{b,:}}{2}. \tag{15}$$

Thus, the permuted mean vectors are:

$$\bar{\boldsymbol{\mu}}_1^{(b)} = \frac{1}{n}\boldsymbol{\sigma}_1^{(b)}, \qquad \bar{\boldsymbol{\mu}}_2^{(b)} = \frac{1}{m}\boldsymbol{\sigma}_2^{(b)}, \tag{16}$$

and the mean resultant lengths are:

$$r_1^{(b)} = \|\bar{\boldsymbol{\mu}}_1^{(b)}\|_2, \qquad r_2^{(b)} = \|\bar{\boldsymbol{\mu}}_2^{(b)}\|_2. \tag{17}$$

We then apply the same monotone mapping $g_d(\cdot)$ and compute $T^{(b)}$ for all b using batched element-wise operations. In practice, Eq. (14) is a single general matrix multiplication (GEMM) call that is typically the fastest kernel available for modern GPUs.

Chunked execution and streaming p-values. Storing $\mathbf{S}$ explicitly can be memory-intensive for large B or N. Therefore, the permutations are processed in blocks of size B_0, as follows: For each block, we generate $\mathbf{S}_{\text{blk}} \in \{+1, -1\}^{B_0 \times N}$ on the fly, compute $\mathbf{U}_{\text{blk}} = \mathbf{S}_{\text{blk}}\mathbf{X}$, obtain $\{T^{(b)}\}$ for the block, and update the exceedance count in (12) without storing all $T^{(b)}$. This yields an $O(BNd)$-time algorithm with $O(Nd + B_0 d)$ working memory (plus the transient block of signs, which can be stored in int8 and cast to float on the GPU).

Efficient Householder application on GPU. Even without using the $d \times d$ matrix $\mathbf{H}$, each aligned vector can be computed from $\mathbf{H} = \mathbf{I} - 2\mathbf{u}\mathbf{u}^\top$ as

$$\mathbf{x}' = \mathbf{x} - 2\mathbf{u}(\mathbf{u}^\top \mathbf{x}), \tag{18}$$

which requires only one dot product and scaled vector subtraction. For a batch matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$, Eq. (18) becomes

$$\mathbf{X}' = \mathbf{X} - 2(\mathbf{X}\mathbf{u})\mathbf{u}^\top, \tag{19}$$

Notably, a matrix-vector product followed by an outer product is GPU-friendly.

Practical notes. (1) **Precision.** Performing GEMM in half precision when available, means, norms, and the final test statistic are accumulated in float32 to avoid numerical drift near $r \approx 1$. (2) **Reusability** When testing many word pairs with the same (n, m), the same randomly generated sign blocks can be reused across pairs, thereby amortizing the sign generation cost. (3) **End-to-end pipeline.** In our workloads, the LM forward passes dominate when extracting contextual vectors, whereas permutation inference becomes dominant when comparing many word pairs or when using a large B. The above vectorization allows permutation testing to remain feasible without sacrificing statistical resolution.

In summary, we present our proposed methods in Algorithm 1 and Algorithm 2. Both algorithms are mathematically equivalent; however, Algorithm 2 achieves faster execution through GPU acceleration.

4 Experiments

4.1 Main Experimental Setup

In the study by Nagata and Tanaka-Ishii (2025), it has been shown that examining the correlation between contextual diversity and sense counts at the individual word level tends to be noisy and does not reveal clear differences. For this reason, it has become established practice to rank words by frequency, construct bins of 100 words each, and perform analyses using the average values within each bin (Bond et al., 2019). In our experiments, we similarly rank words based on dispersion (a measure of semantic breadth) and classify word pairs by the ranking gap.

Word pairs with a dispersion ranking gap of 100 or less are treated as belonging to the same bin in existing research (Nagata and Tanaka-Ishii, 2025). That is, there is an implicit assumption that there is "no difference in semantic breadth" between these words. Under this setting, when pairs with small gaps (e.g., gap=1–10) are tested and rejected, this constitutes a Type-I Error (false positive), since we are declaring a "significant difference" where none should exist.

Since our test examines "whether there is a difference in semantic breadth," we can use WordNet synset counts as a gold standard. Rejected pairs are those judged to have "significant differences in semantic breadth," while pairs with WordNet sense count differences are those that actually differ in semantic breadth. Thus, this task can be evaluated using precision: the ability to identify pairs with actual WordNet sense count differences from embedding vectors. Precision is defined as the proportion of rejected pairs that actually have different WordNet sense counts.

While (Nagata and Tanaka-Ishii, 2025) conducted experiments using the older bert-large-uncased (Devlin et al., 2019) to capture linguistic relationships, ModernBERT (Warner et al., 2025) has since been proposed, which shares the same BERT architecture as bert-large-uncased but captures more semantic information. Therefore, in

this study, we first employed ModernBERT for our experiments. We use BNC (British National Corpus) (BNC Consortium, 2007) as our corpus for English and ModernBERT (Warner et al., 2025) as the language model. For each gap setting, we sample 300 pairs and conduct tests at significance level $\alpha = 0.01$. We compare two methods: the baseline (naive permutation test without alignment) and the proposed method (permutation test after Householder reflection alignment).

Importantly, small gaps correspond to situations where the difference in contextual diversity is small. Statistical tests are fundamentally designed to determine whether small differences are significant. Since our focus is on evaluating the test's behavior in such borderline cases where one would want to conduct a statistical test, we restrict our evaluation to gap values from 1 to 10. Larger gaps represent increasingly obvious differences that would not typically require statistical testing to confirm.

4.2 Experimental Results

Figure 2 shows the Type-I Error rate and Precision across dispersion ranking gaps from 1 to 10. The proposed method (Householder alignment) demonstrates clear advantages over the baseline in controlling Type-I Error while maintaining the ability to detect genuine sense differences.

At gap=10, where relative differences in contextual diversity become more pronounced within the small-gap regime, the proposed method reduces Type-I Error from 4.0% to 2.67% while simultaneously improving Precision from 41.7% to 62.5%. Similarly, at gap=5, the proposed method achieves both lower Type-I Error (1.67% vs. 1.33%) and substantially higher Precision (60.0% vs. 75.0%). These improvements demonstrate that the Householder alignment effectively removes the confounding effect of mean-direction differences, allowing the test to focus on genuine dispersion differences.

In the smallest gap ranges (gap=1–3), where contextual diversity differences are minimal, both methods appropriately control Type-I Error near or below the nominal level. This indicates that both methods behave correctly when there is truly no difference to detect. The key advantage of the proposed method emerges as the gap increases within the small-gap regime, where the baseline begins to show elevated false positive rates.

Table 1 presents detailed comparisons for each gap value. In the Type-I Error metric, the proposed method wins in 3 cases (gap=5, 9, 10), ties in 6 cases, and loses only in 1 case (gap=4). In the Precision metric, the proposed method again wins in 3 cases (gap=4, 5, 10), ties in 6 cases, and loses in 1 case (gap=9). These results validate the effectiveness of the proposed Householder-aligned permutation test in the challenging regime where differences are subtle and statistical testing is most needed. As shown in Table 1 for Gap=10, the Type-I error decreased from 0.040 in the Baseline to 0.027 in the Proposed method, representing a 32.5% reduction in Type-I error.

4.3 Additional Experiments

We verified whether similar trends to those observed with ModernBERT and BNC could be found using models for other languages and corpora. For the Japanese corpus, we used BCCWJ (Maekawa et al., 2014). The proposed alignment method was evaluated using permutation tests to compare word-embedding spaces across different corpora. We conducted experiments using three corpus model combinations: BNC with BERT-tiny, BNC with ModernBERT, and BCCWJ with BERT-large. For each setting, we sampled 500 word pairs and performed permutation tests with $B = 5{,}000$ permutations at a significance level of $\alpha = 0.05$.

Similarly to the previous settings, to assess the statistical properties of the permutation test, we considered two gap conditions between embedding pairs: When **Gap=50**, word pairs are sampled from positions that differ by 50 ranks in their frequency distributions. Since both samples are drawn from the same underlying distribution (with only minor rank differences), rejecting the null hypothesis constitutes a Type-I error. Thus, the rejection rate under this condition estimates the **Type-I error rate**.

When **Gap=100**, word pairs are sampled from positions that differ by 100 ranks, representing a meaningful distributional difference. Since the null hypothesis of identical distributions is genuinely false, the rejection rate under this condition estimates the **statistical power** of the test.

4.4 Results

Table 2 presents the rejection rates for both the baseline permutation test (without alignment) and the proposed method (with alignment).

Type-I Error Control (Gap=50) Under the null hypothesis condition, a well-calibrated test should reject at a rate close to the nominal significance

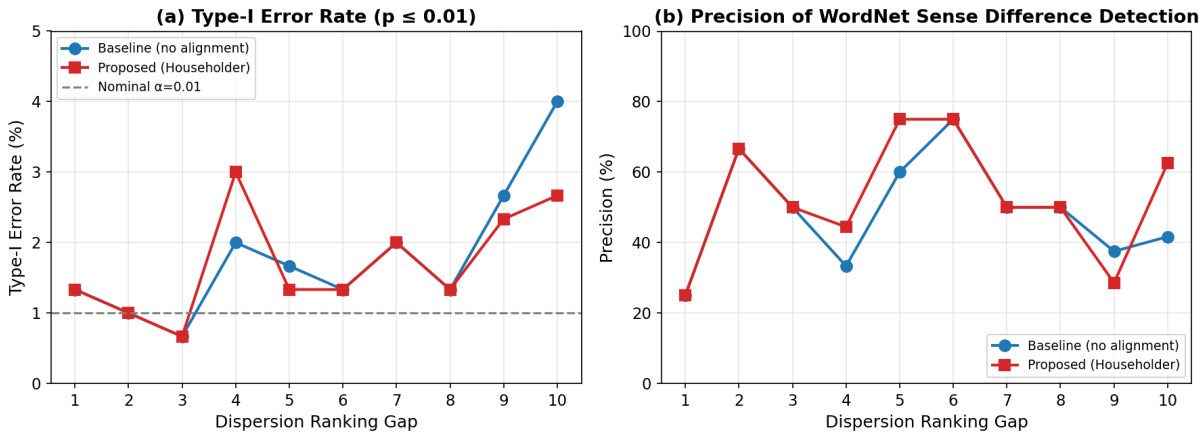

Figure 2: Comparison of (a) Type-I Error rate and (b) Precision across dispersion ranking gaps from 1 to 10. The dashed line in (a) indicates the nominal significance level $\alpha = 0.01$. The proposed Householder-aligned method (red) achieves lower Type-I Error at gap=5, 9, 10 while maintaining higher or comparable precision in detecting actual WordNet sense differences.

	Type-I Error (↓)		**Precision (↑)**	
Gap	Baseline	Proposed	Baseline	Proposed
1	0.013	0.013	0.250	0.250
2	0.010	0.010	0.667	0.667
3	0.007	0.007	0.500	0.500
4	**0.020**	0.030	0.333	**0.444**
5	0.017	**0.013**	0.600	**0.750**
6	0.013	0.013	0.750	0.750
7	0.020	0.020	0.500	0.500
8	0.013	0.013	0.500	0.500
9	0.027	**0.023**	**0.375**	0.286
10	0.040	**0.027**	0.417	**0.625**

Table 1: Type-I Error and Precision comparison at dispersion ranking gaps 1–10 ($\alpha \leq 0.01$). Bold indicates better performance (lower Type-I Error or higher Precision).

Corpus / Model	**Gap**	**Baseline**	**Proposed**
BNC / BERT-tiny	50	0.034	0.024
BNC / BERT-tiny	100	0.064	0.044
BNC / ModernBERT	50	0.040	0.022
BNC / ModernBERT	100	0.090	0.062
BCCWJ / BERT-large	50	0.032	0.022
BCCWJ / BERT-large	100	0.090	0.046

Table 2: Rejection rates of the permutation test under null (gap=50) and alternative (gap=100) hypotheses. Baseline refers to the standard permutation test without alignment, and Proposed applies our alignment strategy.

level ($\alpha = 0.05$). The baseline method showed rejection rates ranging from 3.2% to 4.0%, whereas the proposed alignment method achieved lower rates between 2.2% and 2.4%. Both methods maintained rejection rates below the nominal level; however, the proposed method demonstrated more conservative behavior, reducing the risk of false positives. This conservative calibration is particularly desirable in scientific applications where controlling Type-I errors is crucial.

Statistical Power (Gap=100) Under the alternative hypothesis condition, the baseline method achieved rejection rates between 6.4% and 9.0%, whereas the proposed method yielded rates between 4.4% and 6.2%. Although our alignment strategy resulted in a slightly lower power compared to the baseline, this trade-off is acceptable given the improved Type-I error control. The reduced power reflects the more stringent null distribution induced by proper alignment, which corrects for potential confounds that may artificially inflate rejection rates.

Implementation	Time per permutation	Total time ($B = 20{,}000$)
CPU	1.588 ms	31,750 ms
GPU	0.069 ms	1,377 ms

Table 3: Runtime comparison between CPU and GPU implementations.

Word	v	**WordNet senses**
articulate	≈ 0.000740	7
triple	≈ 0.000813	7
mark	≈ 0.000718	30
colitis	≈ 0.000103	1
debtor	≈ 0.000123	1

Table 4: Representative examples relating dispersion and WordNet sense counts. Larger dispersion often coincides with more senses (e.g., *articulate*, *triple*, and *mark*), whereas specialized nouns tend to have smaller dispersion and fewer senses (e.g., *colitis* and *debtor*).

Qualitative Examples Table 4 shows representative examples relating dispersion and WordNet sense counts. Larger dispersion often coincides with more WordNet senses, whereas specialized nouns tend to have smaller dispersion and fewer senses. For additional intuition, a qualitative t-SNE visualization before and after Householder alignment for one representative pair is provided in Appendix F.

Discussion The results demonstrate that the proposed alignment method provides better calibration of the permutation test. The baseline method, while showing rejection rates nominally below $\alpha = 0.05$ under the null hypothesis, may be susceptible to inflated Type-I errors in more challenging scenarios that violate distributional assumptions. Our alignment strategy addresses this issue by ensuring that the permutation distribution accurately reflects the null hypothesis, leading to more reliable statistical inferences.

4.5 Computational Efficiency

We evaluated the runtime performance of our GPU-accelerated permutation test implementation against the CPU baseline. Benchmarks were conducted on BCCWJ with BERT-large embeddings using the gap50 metric (Type-I error evaluation), with $B = 20{,}000$ permutations.

The GPU implementation achieves a speedup factor of: $\frac{1.588 \text{ ms}}{0.069 \text{ ms}} = \mathbf{23.0}$ times. This **23-fold speedup** is critical for large-scale permutation testing, where thousands of hypothesis tests must be performed across multiple corpora and model configurations.

5 Conclusion

This study introduces a testing framework for comparing the semantic breadth of two word types from their contextualized token-embedding clouds. We show that without controlling for semantic-direction differences, dispersion-based testing can suffer from inflated Type I errors, undermining the reliability of significance claims in the small-effect regime where statistical testing is required the most. By aligning the mean directions with a single Householder transformation, our approach targets dispersion differences more directly and yields better-calibrated permutation test p values. In addition, because permutation testing is well established in NLP but often considered computationally expensive in practice (Dror et al., 2018), we present a GPU-oriented implementation strategy that makes large-scale experimentation feasible.

Future Work Despite focusing on word-type token clouds, the proposed alignment and permutation framework applies to any two sets of representations, including sentence-embedding sets. We did not evaluate sentence vectors in this study because of concerns that anisotropy and other geometric artifacts would complicate the interpretation of dispersion at the sentence level (Ethayarajh, 2019). If future work mitigates these anisotropy concerns and enables reliable dispersion-based inference for sentence embeddings, the proposed test can be transferred to sets of prompts or questions used in LLM evaluation, providing a principled foundation for quantifying the *breadth of knowledge* targeted by an evaluation suite independent of subjective human judgments.

Another possible direction is to apply the proposed framework to second-language vocabulary learning support (Ehara, 2025, 2022; Ehara et al., 2012). Among words with similar frequency, words with larger dispersion may be more difficult to learn, as they tend to appear across a wider range of contexts and senses.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 22K12287 and by JST, PRESTO Grant Number JPMJPR2363. We are deeply grateful to the anonymous reviewers for their constructive feedback.

Limitations

Our empirical evaluation is necessarily limited in coverage. We tested a restricted set of corpora and model families; however, broader validation across genres, languages, and encoder architectures remains essential for assessing generality. In particular, expanding the range of the corpora would clarify the robustness of the observed calibration improvements under different frequency profiles and contextual distributions.

A second limitation concerns the dispersion statistics themselves. Similar to prior contextual diversity formulations based on von Mises-Fisher modeling, the underlying assumptions effectively treat the distribution of (normalized) token vectors for a word type as unimodal and isotropic, whereas real token clouds can be multimodal and/or anisotropic (Nagata and Tanaka-Ishii, 2025). Such geometric deviations may affect both effect size and calibration of tests based on simplified distributional assumptions. Prior large-scale evidence suggests that these assumptions can still yield observable and useful regularities at the word-type level, even when the true distributions are not perfectly isotropic (Nagata and Tanaka-Ishii, 2025). However, our work remains within this scope and does not claim to resolve anisotropy in general.

Finally, the practical constraints of pretrained tokenizers and vocabularies can limit the word types that can be analyzed cleanly (e.g., subword splitting), and this can introduce additional variability when comparing specific lexical items (Nagata and Tanaka-Ishii, 2025).

Ethical Considerations

This study relies solely on publicly available pretrained language models and existing linguistic resources such as WordNet. We do not collect, process, or generate any personal data, nor do we conduct experiments involving human subjects. The corpora used in our experiments (BNC, BCCWJ) are well-established linguistic resources obtained through appropriate academic licenses. Therefore, we do not foresee any significant ethical concerns arising from this work.

References

Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein. 2012. An empirical investigation of statistical significance in NLP. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 995–1005, Jeju Island, Korea. Association for Computational Linguistics.

BNC Consortium. 2007. The British National Corpus, XML edition. License: http://www.natcorp.ox.ac.uk/docs/licence.html.

Francis Bond, Arkadiusz Janz, Marek Maziarz, and Ewa Rudnicka. 2019. Testing Zipf's meaning-frequency law with wordnets as sense inventories. In *Proceedings of the 10th Global Wordnet Conference*, pages 342–352, Wroclaw, Poland. Global Wordnet Association.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, Melbourne, Australia. Association for Computational Linguistics.

Yo Ehara. 2022. An intelligent interactive support system for word usage learning in second languages. In *Artificial Intelligence in Education - 23rd International Conference, AIED 2022, Durham, UK, July 27-31, 2022, Proceedings, Part I*, Lecture Notes in Computer Science, pages 453–464. Springer.

Yo Ehara. 2025. Educational cone model in embedding vector spaces. In *Proceedings of ICCE 2025: The 33rd International Conference on Computers in Education (short paper)*.

Yo Ehara, Issei Sato, Hidekazu Oiwa, and Hiroshi Nakagawa. 2012. Mining words in the minds of second language learners: Learner-specific word difficulty. In *Proceedings of COLING 2012*, pages 799–814, Mumbai, India. The COLING 2012 Organizing Committee.

Kawin Ethayarajh. 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), pages 55–65, Hong Kong, China. Association for Computational Linguistics.

Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. 2020. Analysing lexical semantic change with contextualised word representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3960–3973, Online. Association for Computational Linguistics.

Yvette Graham, Nitika Mathur, and Timothy Baldwin. 2014. Randomized significance tests in machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 266–274, Baltimore, Maryland, USA. Association for Computational Linguistics.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Andrey Kutuzov and Mario Giulianelli. 2020. UiO-UvA at SemEval-2020 task 1: Contextualised embeddings for lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 126–134, Barcelona (online). International Committee for Computational Linguistics.

Kikuo Maekawa, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den. 2014. Balanced corpus of contemporary written Japanese. *Language Resources and Evaluation*, 48:345–371.

George A. Miller. 1995. WordNet: A lexical database for English. *Communications of the ACM*, 38(11):39–41.

Ryo Nagata and Kumiko Tanaka-Ishii. 2025. A new formulation of Zipf's meaning-frequency law through contextual diversity. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15323–15335, Vienna, Austria. Association for Computational Linguistics.

Francesco Periti and Nina Tahmasebi. 2024. A systematic comparison of contextualized word embeddings for lexical semantic change. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4262–4282, Mexico City, Mexico. Association for Computational Linguistics.

Stefan Riezler and John T. Maxwell. 2005. On some pitfalls in automatic evaluation and significance testing for MT. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 57–64, Ann Arbor, Michigan. Association for Computational Linguistics.

Peter H. Schönemann. 1966. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10.

Andrea Vallebueno, Cassandra Handan-Nader, Christopher D Manning, and Daniel E. Ho. 2024. Statistical uncertainty in word embeddings: GloVe-V. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9032–9047, Miami, Florida, USA. Association for Computational Linguistics.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605.

Ivan Vulić, Daniela Gerz, Douwe Kiela, Felix Hill, and Anna Korhonen. 2017. HyperLex: A large-scale evaluation of graded lexical entailment. *Computational Linguistics*, 43(4):781–835.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. 2025. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547, Vienna, Austria. Association for Computational Linguistics.

Christos Xypolopoulos, Antoine Tixier, and Michalis Vazirgiannis. 2021. Unsupervised word polysemy quantification with multiresolution grids of contextual embeddings. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3391–3401, Online. Association for Computational Linguistics.

Hiroaki Yamagiwa and Hidetoshi Shimodaira. 2025. Norm of mean contextualized embeddings determines their variance. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7778–7808, Abu Dhabi, UAE. Association for Computational Linguistics.

Ran Zmigrod, Tim Vieira, and Ryan Cotterell. 2022. Exact paired-permutation testing for structured test statistics. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4894–4902, Seattle, United States. Association for Computational Linguistics.

A Checklist Answers

This appendix provides answers to the ARR Responsible NLP Research checklist. All section titles below correspond to the checklist item IDs.

A.1 ID: A2 Potential Risks

Elaboration: Potential risks are discussed in the Limitations section. Our method is a general-purpose statistical testing framework for comparing the semantic breadth of word embeddings. Therefore, we do not foresee dual-use concerns or direct societal harms from this statistical methodology.

A.2 ID: B1 Cite Creators Of Artifacts

Elaboration: As for corpora, we used the BNC (BNC Consortium, 2007), BCCWJ (Maekawa et al., 2014), and WordNet (Miller, 1995).

As for models we used: BERT-tiny: https://huggingface.co/prajjwal1/bert-tiny. BERT-large-uncased: https://huggingface.co/google-bert/bert-large-uncased. BERT for Japanese, namely bert-large-japanese-v2: https://huggingface.co/tohoku-nlp/bert-large-japanese-v2.

A.3 ID: B2 Discuss The License For Artifacts

Elaboration: We have confirmed that all corpora and models are distributed under licenses that permit their use for research purposes.

A.4 ID: B3 Artifact Use Consistent With Intended Use

Elaboration: We have confirmed that all corpora and models are distributed under licenses that permit their use for research purposes. Therefore, we consider their use in this study to be consistent with the intended use specified by their respective licenses.

A.5 ID: B4 Data Contains Personally Identifying Information Or Offensive Content

Elaboration: We use only publicly available, well-established corpora (such as BNC and BCCWJ) and models that have been previously released for research purposes under their respective licenses. We do not create new datasets or collect any personal data. These corpora have undergone standard curation processes by their original creators. Therefore, we do not consider that our paper contains personally identifying information or offensive content.

A.6 ID: B5 Documentation of Artifacts and ID: B6 Statistics For Data

Elaboration: We documented basic information of the artifacts within the body texts of this paper. The more detailed documentation is available from the urls and papers in the artifact citations.

Table 5 provides statistics for the corpora used in our experiments.

Corpus	Size (tokens)	License
BNC	109,369,848	BNC User Licence[1]
BCCWJ	124,102,859	Academic License[2]

Table 5: Statistics of corpora used in experiments.

[1] http://www.natcorp.ox.ac.uk/docs/licence.html
[2] https://clrd.ninjal.ac.jp/bccwj/en/index.html

Pretrained Models:

- `BERT-tiny`: 4.4M parameters (Apache 2.0)
- `BERT-large`: 340M parameters (Apache 2.0)
- `ModernBERT`: Apache 2.0

Lexical Resource:

- WordNet: Used for sense count evaluation (WordNet License)

A.7 ID: C1 Model Size And Budget

Elaboration: We conducted computational experiments to evaluate the proposed Householder-aligned permutation test:

- **Main experiments (Section 4):** Type-I error and precision evaluation using WordNet sense counts
- **Additional experiments:** Type-I error control and statistical power across multiple corpus-model combinations
- **GPU acceleration:** Benchmark comparisons between CPU and GPU implementations

All models that we used in this paper are below 1 billion parameters. The electricity fee for running two H100 GPUs are the budget for our computation. For both BNC and BCCWJ, extracting all contextual embeddings for the target set of words took approximately 4 hours of GPU computation. The proposed GPU-based permutation tests take 20 minutes for 20,000 permutations of 300 words.

A.8 ID: C2 Experimental Setup And Hyperparameters

Elaboration: All experimental setups and hyperparameters are described in Appendix A and Section 4.

Our experimental setup uses the following configurations:

Main Experiments (Section 4):

Parameter	Value
Corpus	BNC
Model	ModernBERT
Number of word pairs	300 per gap setting
Dispersion ranking gap	1–10
Significance level (α)	0.01
Gold standard	WordNet synset counts

Additional Experiments:

Parameter	Value
Corpus–Model combinations	BNC + BERT-tiny BNC + ModernBERT BCCWJ + BERT-large
Number of word pairs	500
Number of permutations (B)	5,000
Significance level (α)	0.05
Gap conditions	50 (Type-I error) 100 (Statistical power)

Algorithm Parameters:

- Embedding normalization: ℓ_2-normalization to unit sphere
- Dispersion statistic: Mean resultant length (MRL)
- Test statistic: Log-volume difference (Eq. 11)
- Alignment method: Householder reflection (single transformation)

A.9 ID: C3 Descriptive Statistics

Elaboration: We report comprehensive descriptive statistics for our experimental results:

- **Table 1:** Type-I Error and Precision comparison at dispersion ranking gaps 1–10
- **Table 2:** Rejection rates under null (gap=50) and alternative (gap=100) hypotheses for three corpus–model combinations
- **Figure 2:** Visualization of Type-I Error rate and Precision across gap values

Key Results Summary:

- Type-I error reduction: 32.5% compared to baseline (naive permutation test)
- GPU speedup: 23× over CPU baseline
- At gap=10: Type-I Error reduced from 4.0% to 2.67%; Precision improved from 41.7% to 62.5%

A.10 ID: C4 Parameters For Packages

Elaboration: We used PyTorch https://pytorch.org/, HuggingFace https://huggingface.co/ models. All parameters are described in the experiment setup descriptions of Section 4 and this Appendix.

A.11 ID: E1 Information About Use Of AI Assistants

Elaboration: We used codex CLI and Claude code to implement our proposed methods. All code is manually checked. We used ChatGPT and Claude to improve the English of the paper.

B Algorithm Description

This appendix presents the algorithms used in our method: Algorithm 1 describes the basic Householder-aligned permutation test, and Algorithm 2 gives its GPU-oriented implementation.

C Qualitative Examples of Statistical Testing

To make the calibration-power trade-off in Table 2 more concrete (Japanese; BCCWJ/BERT-large), we inspected the pairs whose significance changed after alignment and presented representative examples.

Baseline-only rejections (rejected by the baseline but not by the proposed method) include several content-word pairs, e.g., 普段–毎年 (usually / in everyday situations – every year), 書物–気管 (books / written works – trachea / airway), and 家電–騒音 (home appliances – noise / noise pollution). In contrast, many pairs that remain statistically significant under both methods involve verb-/function-like tokens versus content nouns, e.g., 避

Word	Total	Baseline / greater	Proposed / greater	Baseline / two-sided	Proposed / two-sided
ablaze	160	0.2441	0.2474	0.4895	0.4956
accolade	160	0.6356	0.6345	0.7330	0.7356
actuality	156	0.7650	0.7644	0.4702	0.4714
acheson	154	0.4623	0.4613	0.9205	0.9191
adaptability	153	0.5581	0.5578	0.8827	0.8841
adenauer	152	0.0369	0.0342	0.0756	0.0709
additive	151	0.5910	0.5918	0.8148	0.8139
absentee	150	0.6509	0.6534	0.6997	0.6950
abnormally	149	0.3397	0.3412	0.6888	0.6901
abstinence	147	0.5419	0.5395	0.9168	0.9186
aberration	145	0.9040	0.9056	0.1845	0.1815
acetate	145	0.6496	0.6489	0.7017	0.7019
abstracted	139	0.2506	0.2562	0.5095	0.5173
abyss	137	0.1744	0.1724	0.3501	0.3466
abatement	136	0.4426	0.4416	0.8877	0.8860
according	135	0.3434	0.3443	0.6827	0.6846
acrimonious	135	0.9729	0.9628	0.0519	0.0725
abstain	132	0.0420	0.0363	0.0841	0.0734
accomplishment	130	0.4508	0.4498	0.8928	0.8919
acne	130	0.7160	0.7159	0.5633	0.5644

Table 6: Same-word split-half sanity-check p-values for 20 English words. "Baseline" corresponds to `align=none`, and "Proposed" corresponds to `align=once`.

け–最低 (avoid(ing) – the worst / lowest) and 叫ん–サラリーマン (shout(ing) – office worker / salary-man) (we also observe a few noun–noun cases such as 仏教–官僚 (Buddhism – bureaucrat(s)) and 下着–村人 (underwear – villager(s))).

D Proof Sketch

In this section, we provide a proof sketch that the transformation $\mathbf{R}$ that maximizes the MRL of the merged set is given by the Householder transform. To avoid display/rendering issues, we write the sums in a simplified form.

The MRL for the merged set can be written as

$$\text{MRL} = \frac{1}{n+m} \left\| \sum_i \mathbf{R}\mathbf{x}_i + \sum_j \mathbf{y}_j \right\| \tag{20}$$

$$= \frac{1}{n+m} \left\| \mathbf{R} \sum_i \mathbf{x}_i + \sum_j \mathbf{y}_j \right\| \tag{21}$$

$$= \frac{1}{n+m} \left\| n\,\mathbf{R}\bar{\mathbf{x}} + m\,\bar{\mathbf{y}} \right\|. \tag{22}$$

Thus, the MRL reduces to the norm of a sum of two vectors. By the triangle inequality, this quantity is maximized if and only if the two vectors are parallel and not in opposite directions:

$$\begin{aligned} \left\| n\,\mathbf{R}\bar{\mathbf{x}} + m\,\bar{\mathbf{y}} \right\| &\le \left\| n\,\mathbf{R}\bar{\mathbf{x}} \right\| + \left\| m\,\bar{\mathbf{y}} \right\| \\ &= n\left\|\mathbf{R}\bar{\mathbf{x}}\right\| + m\left\|\bar{\mathbf{y}}\right\|, \end{aligned} \tag{23}$$

if and only if $\mathbf{R}\bar{\mathbf{x}}$ and $\bar{\mathbf{y}}$ are parallel, i.e., $\mathbf{R}\bar{\mathbf{x}} = C\bar{\mathbf{y}}$ where $C > 0$.

In particular, when $\mathbf{R} = \mathbf{H}$, Line 267 of the paper implies that

$$\mathbf{H}\bar{\mathbf{x}}/\|\bar{\mathbf{x}}\| = \bar{\mathbf{y}}/\|\bar{\mathbf{y}}\|. \tag{24}$$

Therefore, the maximum MRL of the merged set is achieved by rotating $\mathbf{X}$ using the Householder transform. Note that the MRL is upper-bounded by

$$\begin{aligned} &\frac{1}{n+m}\left(n\|\mathbf{R}\bar{\mathbf{x}}\| + m\|\bar{\mathbf{y}}\|\right) \\ &= \frac{1}{n+m}\left(n\|\bar{\mathbf{x}}\| + m\|\bar{\mathbf{y}}\|\right), \end{aligned} \tag{25}$$

since $\mathbf{R}$ is orthogonal.

E Sanity Check

In this section, we provide an example of sanity check experiments that we explained in Section 3.1. As shown in Table 6, when occurrences of the same word are randomly divided into two groups, the two groups already have very similar mean vectors even without alignment. As a result, the baseline and the proposed method yield almost identical p-values in this same-word null setting, indicating that little difference in performance is expected between the two methods in such cases.

F Qualitative visualization

In this section, we revisit how the Householder transform aligns the mean vectors of the contextualized word-embedding sets for the two words. Figure 3 and Figure 4 show a qualitative t-SNE

Algorithm 1 Householder-Aligned Permutation Test

Require: Unit-normalized embeddings $X = \{\mathbf{x}_i\}_{i=1}^n$, $Y = \{\mathbf{y}_j\}_{j=1}^m$; number of permutations B; significance level α

Ensure: p-value for testing H_0: $\mathrm{disp}(X) = \mathrm{disp}(Y)$

// Step 1: Compute sample mean directions

$\bar{\mathbf{x}} \leftarrow \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i$; $\bar{\mathbf{y}} \leftarrow \frac{1}{m}\sum_{j=1}^m \mathbf{y}_j$

$\hat{\boldsymbol{\mu}}_x \leftarrow \bar{\mathbf{x}}/\|\bar{\mathbf{x}}\|_2$; $\hat{\boldsymbol{\mu}}_y \leftarrow \bar{\mathbf{y}}/\|\bar{\mathbf{y}}\|_2$

// Step 2: Construct Householder reflection

$\mathbf{u} \leftarrow (\hat{\boldsymbol{\mu}}_x - \hat{\boldsymbol{\mu}}_y)/\|\hat{\boldsymbol{\mu}}_x - \hat{\boldsymbol{\mu}}_y\|_2$

// Step 3: Align X to Y's mean direction

for $i = 1$ to n **do**

  $\mathbf{x}'_i \leftarrow \mathbf{x}_i - 2\mathbf{u}(\mathbf{u}^\top \mathbf{x}_i)$ ▷ Eq. (18)

end for

$X' \leftarrow \{\mathbf{x}'_i\}_{i=1}^n$

// Step 4: Compute observed test statistic

$r_{X'} \leftarrow \left\|\frac{1}{n}\sum_{i=1}^n \mathbf{x}'_i\right\|_2$; $r_Y \leftarrow \left\|\frac{1}{m}\sum_{j=1}^m \mathbf{y}_j\right\|_2$

$T_{\mathrm{obs}} \leftarrow \log(1/g_d(r_{X'})) - \log(1/g_d(r_Y))$ ▷ Eq. (10)

// Step 5: Permutation test on aligned data

$Z \leftarrow X' \cup Y$ ▷ Pool aligned samples

count $\leftarrow 0$

for $b = 1$ to B **do**

  Randomly partition Z into $(X^{(b)}, Y^{(b)})$ with sizes (n, m)

  $r_1^{(b)} \leftarrow \left\|\frac{1}{n}\sum_{\mathbf{z}\in X^{(b)}} \mathbf{z}\right\|_2$; $r_2^{(b)} \leftarrow \left\|\frac{1}{m}\sum_{\mathbf{z}\in Y^{(b)}} \mathbf{z}\right\|_2$

  $T^{(b)} \leftarrow \log(1/g_d(r_1^{(b)})) - \log(1/g_d(r_2^{(b)}))$

  if $T^{(b)} \geq T_{\mathrm{obs}}$ **then**

    count $\leftarrow$ count $+ 1$

  end if

end for

// Step 6: Compute p-value

$p \leftarrow (1 + \mathrm{count})/(B + 1)$ ▷ Eq. (12)

return p

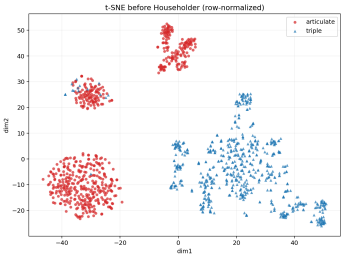

Figure 3: Baseline, i.e., t-SNE visualization before applying Householder transform.

Algorithm 2 GPU-Accelerated Permutation Statistics

Require: Aligned pooled matrix $\mathbf{X} \in \mathbb{R}^{N\times d}$ (rows of $Z = X' \cup Y$); group sizes (n, m) with $N = n + m$; number of permutations B; block size B_0

Ensure: Exceedance count for p-value computation

// Precompute total sum (once)

$\mathbf{t} \leftarrow \mathbf{1}^\top \mathbf{X} \in \mathbb{R}^d$ ▷ Sum of all rows

count $\leftarrow 0$

// Process permutations in GPU-friendly blocks

for each block $k = 1, 2, \ldots, \lceil B/B_0 \rceil$ **do**

  // Generate sign matrix for this block

  $\mathbf{S}_{\mathrm{blk}} \leftarrow$ random $\{+1, -1\}^{B_0\times N}$ with exactly n entries $+1$ per row

  // Single GEMM call for all permutations in block

  $\mathbf{U}_{\mathrm{blk}} \leftarrow \mathbf{S}_{\mathrm{blk}}\mathbf{X} \in \mathbb{R}^{B_0\times d}$ ▷ Eq. (14)

  // Recover group sums via broadcasting

  for $b = 1$ to B_0 **in parallel do**

    $\boldsymbol{\sigma}_1^{(b)} \leftarrow (\mathbf{t} + \mathbf{U}_{\mathrm{blk},b,:})/2$

    $\boldsymbol{\sigma}_2^{(b)} \leftarrow (\mathbf{t} - \mathbf{U}_{\mathrm{blk},b,:})/2$

  // Compute mean resultant lengths

    $r_1^{(b)} \leftarrow \|\boldsymbol{\sigma}_1^{(b)}/n\|_2$; $r_2^{(b)} \leftarrow \|\boldsymbol{\sigma}_2^{(b)}/m\|_2$

  // Compute test statistic

    $T^{(b)} \leftarrow \log(1/g_d(r_1^{(b)})) - \log(1/g_d(r_2^{(b)}))$

  end for

  // Update exceedance count (streaming)

  count $\leftarrow$ count $+ \sum_{b=1}^{B_0} \mathbb{I}[T^{(b)} \geq T_{\mathrm{obs}}]$

end for

return count

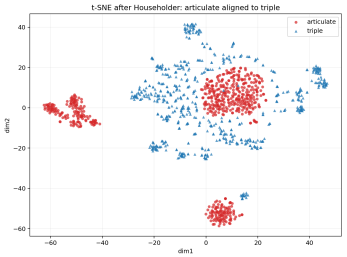

Figure 4: Proposed, i.e., t-SNE visualization after applying Householder transform.

visualization (van der Maaten and Hinton, 2008) for the word pair *articulate-triple* before and after Householder alignment. Before alignment, the two

token clouds occupy more clearly separated regions in this 2D visualization. After Householder alignment, the two sets appear more mixed, which is consistent with the role of the proposed method in reducing nuisance mean-direction mismatch before permutation. We emphasize, however, that this figure is provided only for intuition: low-dimensional projections such as t-SNE do not faithfully preserve the original high-dimensional geometry or dispersion statistics, and should not be treated as primary evidence.