Hierarchical Concept Geometry in Language Representations Emerges from Word Co-occurrence

Andres Nava
Johns Hopkins University
anava1@jh.edu

Matthieu Wyart
EPFL & Johns Hopkins University
mwyart1@jh.edu

Abstract

We propose a distributional theory of how hypernymy—the "is-a" relation between general and specific concepts—is encoded geometrically in language representations. Starting from the empirically verified assumption that words closer on the WordNet hypernym graph co-occur more often, we characterize theoretically the spectrum of the resulting embedding Gram matrix of word2vec embeddings. For regular tree hierarchies, under mild positivity and decay conditions on the co-occurrence kernel, we prove that the leading eigenvectors first separate broad taxonomic branches and then progressively finer sub-branches, producing a *hierarchical splitting geometry* with a coarse-to-fine spectral organization that mirrors the tree. We confirm these predictions in word2vec embeddings across many sampled WordNet subtrees, and show empirically that the same signature extends strikingly well to Gemma 2B unembeddings. Our results establish that pairwise word statistics can generate hierarchical concept geometry without a hierarchy-specific functional mechanism, and provide empirical support for a similar distributional account of the geometry observed in LLMs.

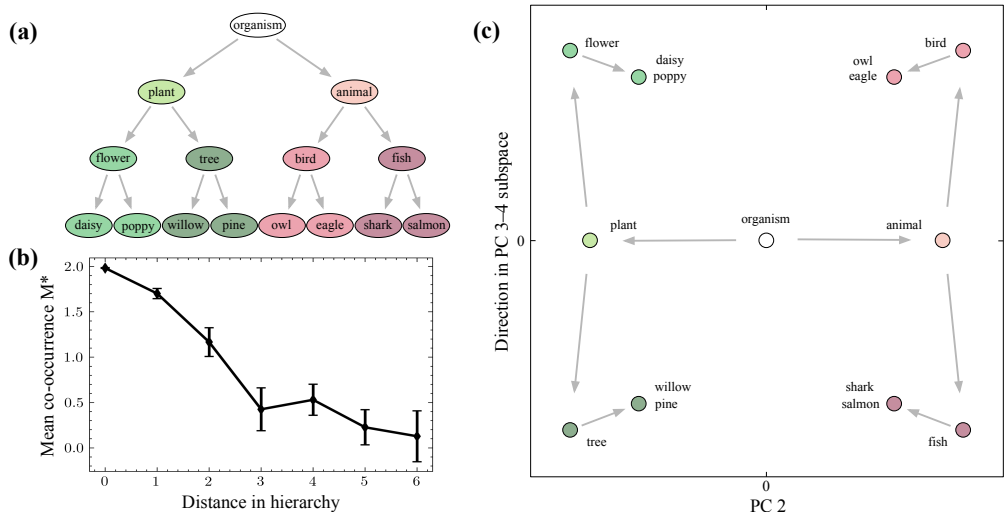

Figure 1: (a) WordNet hierarchy for a taxonomy of organisms. (b) Mean co-occurrence statistic $M^\star$ as a function of graph distance (number of edges) in the organism taxonomy. Error bars indicate one standard error. Nearby concepts co-occur more often. (c) Projections of taxonomy nodes onto principal components (PCs) predicted by the theory. PC 2 splits the plant and animal subtrees, while the degenerate PC 3-4 subspace contains directions that split flower vs. tree and bird vs. fish. We show one selected direction in this subspace.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

1 Introduction

While the success of Large Language Models (LLMs) in learning languages and reasoning from examples is mesmerizing, understanding the reasons for this success remains a challenge. A common belief is that LLMs build a representation of language that allows for simple manipulation. Recent work has shown that semantic variables often appear geometrically in representation spaces, forming directions, subspaces, loops, prisms, and other low-dimensional structures [1, 2, 3, 4, 5], some of which were recently found in the brain [6]. Such organization indeed allows one to manipulate concepts with simple arithmetic operations, such as vector addition or rotation. In this paper, we ask what geometric structure is induced by hypernymy, the "is-a" relation between more specific and more general concepts. It is one of the most basic organizing principles of meaning: an *owl* is a *bird*, a *bird* is an *animal*, and an *animal* is an *organism*. We take the WordNet hypernym graph [7] as our operational definition of semantic hierarchy, with hypernyms corresponding to more general concepts and hyponyms to more specific concepts. This relation is exemplified on a taxonomy of organisms in Figure 1(a).

Specific algorithms have been introduced to bias representations to be hierarchical [8, 9], however, LLMs already spontaneously form hierarchical representations in some form. Indeed, linear transformations can predict ancestor or hypernym representations from descendant or hyponym representations in both skip-gram embeddings [10] and LLM hidden activations [11], while linear transformations of LLM activations can recover WordNet hypernym-graph distances [12]. Park et al. [13] give a functional interpretation of hierarchical geometry in LLMs. Their starting point is the linear representation hypothesis: the idea that high-level semantic variables are encoded by linear directions or concept vectors in the model's representation space. In their framework, a concept vector is meaningful because of the operations it supports: moving in that direction should increase the model's probability of expressing the target concept, while leaving unrelated semantic variables unchanged. Thus, steering between *bird* and *fish* should change which kind of animal is represented, but should not change the probability of being an animal at all. For semantic hierarchies, Park et al.'s postulate implies a particularly elegant geometry. A descendant concept such as *bird* should preserve the broader *animal* component while adding an orthogonal refinement that distinguishes birds from other animals. In such a geometry, inner products with concept vectors act as clean membership readouts: an animal direction detects animalhood, a bird direction detects birdhood, and progressively finer directions detect progressively more specific classes.

Our contribution is to give a simple statistical mechanism that suffices to produce this approximate geometry. Rather than postulating hierarchical orthogonality from functional desiderata, we show that hierarchical geometry arises naturally from co-occurrence statistics, i.e., how often words appear near one another in text. This account is also more predictive: it implies not only that the same geometry should appear outside LLMs, in simple word embeddings such as word2vec, but also that the geometry should have a specific coarse-to-fine spectral organization.

To see this, we assume and confirm that words closer on the WordNet graph tend to co-occur more often, as illustrated in Figure 1(b): for example, *tree* and *plant* co-occur more often than *tree* and *organism*. Because word2vec embeddings are determined by co-occurrence statistics, this assumption leads to detailed predictions for their geometry. Extracting symmetric subtrees of WordNet, we predict that the corresponding word2vec embeddings exhibit what we call *hierarchical splitting geometry*: successive principal components separate subtrees from coarse to fine levels of the taxonomy. These PCs, which correspond to successive eigendirections of the Gram matrix, first separate broad taxonomic branches and then progressively finer subbranches. In the organism tree of Figure 1, the second PC separates animals from plants by taking opposite signs on the two subtrees; the third and fourth PCs, which are degenerate, split birds from fish and flowers from trees. These hierarchical splits are visualized in Figure 1(c): projecting each node's vector onto the relevant PCs separates nodes according to their hierarchical relationships. The remaining modes encode progressively finer splits, such as separating *daisy* from *poppy*.

We first confirm these predictions in word2vec embeddings. We then show empirically that the same hierarchical splitting geometry extends strikingly well to LLM representations. Overall, our results refine the functional picture of Park et al.: hierarchical geometry is real and robust, but it need not arise from a hierarchy-specific functional mechanism. Instead, the approximate orthogonality observed in practice is reproduced in word2vec as a consequence of the spectral structure induced by pairwise word statistics. This yields a more mechanistic and quantitatively predictive account of hierarchical

geometry in representation spaces. It also supplies a null hypothesis for geometric probing. These signatures alone do not identify a hierarchy-specific mechanism or establish functional use, so such claims require additional evidence.

2 Related work

Hierarchical representation. Hierarchical organization of embeddings was investigated theoretically in a supervised setting [14, 15]. In this view, branching points in the taxonomy can lead to different features; interesting representations can emerge in machines seeking to predict those. These works do not consider the structure of the data distribution, however, and as such do not apply to self-supervised methods central to word embeddings.

Co-occurrence and embedding geometry. It is well known that word2vec approximately performs a spectral decomposition of the Pointwise Mutual Information (PMI) matrix characterizing word co-occurrence (related to $M^\star$ in eq. (1)) [16]. Two recent works make assumptions on the structure of co-occurrence to predict the embedding geometry of certain relations. Korchinski et al. [17] argue that if words are characterized by discrete attributes (such as gender) that control co-occurrence, a parallelogram structure of the kind $king - man + woman = queen$ must occur. Instead, if words are characterized by continuous attributes (such as the seasonality of months of the year or the latitude and longitude of cities), embedding geometry is characterized by smooth manifolds such as circles or maps [18]. This work extends these approaches to hierarchical semantic notions, in particular hypernymy.

Hyperbolic embeddings. Hyperbolic embeddings [8] exploit negative curvature to represent hierarchical structures faithfully in low dimensions. Our analysis considers the full-dimensional Euclidean geometry induced by co-occurrence statistics, where inner products decrease with graph distance and the hierarchy is naturally represented. The distinction becomes important under dimensionality reduction. In Euclidean space, truncating the spectrum preferentially removes finer hierarchical splitting modes, leading to distortions of the tree geometry. Hyperbolic methods can preserve hierarchical distances in substantially fewer dimensions [19].

3 Theory

3.1 From co-occurrence statistics to hierarchical embedding geometry

Word embedding models such as word2vec and GloVe learn from co-occurrence statistics [20, 21]. Recent work has shown that their learned representations can be expressed in terms of the top eigenvectors of a normalized co-occurrence matrix $M^\star$ [22]. This matrix is defined by

$$M^\star_{ij} := \frac{P_{ij} - P_i P_j}{\frac{1}{2}(P_{ij} + P_i P_j)} \approx \log\left(\frac{P_{ij}}{P_i P_j}\right), \tag{1}$$

where the approximation connects $M^\star$ to PMI when $P_{ij}/(P_i P_j)$ is close to one.

Writing the eigendecomposition $M^\star = \Phi \Lambda^\star \Phi^\top$, let $\lambda^\star_1, \ldots, \lambda^\star_n$ denote the eigenvalues in decreasing order, and let $\phi_1, \ldots, \phi_n$ be the corresponding eigenvectors. For an embedding of dimension d, the word2vec representation is constructed from the top d eigenvectors with positive eigenvalues of $M^\star$. Concretely, if $\Phi^{(d)} \in \mathbb{R}^{|\mathcal{S}| \times d}$ contains these eigenvectors and $\Lambda^{\star(d)}$ is the corresponding diagonal matrix of eigenvalues, then the embedding coordinates are

$$W^{(d)}_{i\mu} = \Phi^{(d)}_{i\mu} \sqrt{\Lambda^{\star(d)}_{\mu\mu}}, \qquad \mu = 1, \ldots, d. \tag{2}$$

We denote the resulting word2vec Gram matrix by $G_{\text{word2vec}} = W^{(d)}(W^{(d)})^\top$. Thus, the embedding dimension d determines how many positive spectral modes of $M^\star$ are retained: increasing d incorporates progressively finer eigenmodes of the co-occurrence structure. The retained eigenvectors define the principal directions of the embedding point cloud, while the corresponding eigenvalues determine the variance along those directions.

Our starting assumption is that words that are closer in the WordNet hypernym graph tend to co-occur in text more often. We take the WordNet hypernym graph as our operational notion of semantic hierarchy and ask what geometric structure this induces in word2vec embeddings.

Assumption 3.1 (Hierarchical co-occurrence model)**.** Let $\mathcal{S}$ denote the selected vocabulary. Each word is identified with a node in the WordNet hypernym graph, and for $i, j \in \mathcal{S}$, let $\mathrm{dist}(i, j)$ denote their distance. We assume that

$$P_{ij} = P_i P_j \left[\widetilde{C}(\mathrm{dist}(i, j)) + \varepsilon_{ij}\right], \tag{3}$$

where P_{ij} is the co-occurrence probability within a fixed context window, P_i is the unigram probability, $\widetilde{C}$ is a function of distance, and ε_{ij} is a fluctuation term.

This assumption formalizes the idea that, after normalizing for unigram frequencies, the mean co-occurrence signal is controlled primarily by distance in the semantic hierarchy. In the analysis below, we neglect the fluctuation term and work with the mean co-occurrence structure, which isolates the effect of distance and yields a tractable model. The effect of noise is discussed later.

Under Assumption 3.1, there exists a function f such that

$$M^\star_{ij} = f(\mathrm{dist}(i, j)). \tag{4}$$

After neglecting ε_{ij}, the ratio $P_{ij}/P_i P_j$ depends only on distance, and therefore $M^\star_{ij}$, being a deterministic transform of that ratio, also depends only on distance. The kernel f thus describes how distance on the tree affects co-occurrence. We denote this theoretical kernel matrix by G_{th}. The empirical comparisons use G_{word2vec} and G_{LLM}, the Gram matrix of processed LLM unembedding vectors.

Empirically, we observe that the mean co-occurrence value is positive at small distances and decays rapidly towards zero within a standard error in Figure 2. This motivates the following assumption on the function f. The decay also persists across context-window choices (Appendix E; see Figure 9).
Assumption 3.2 (Positive decaying kernel)**.** The function f is positive and strictly decreasing as a function of distance.

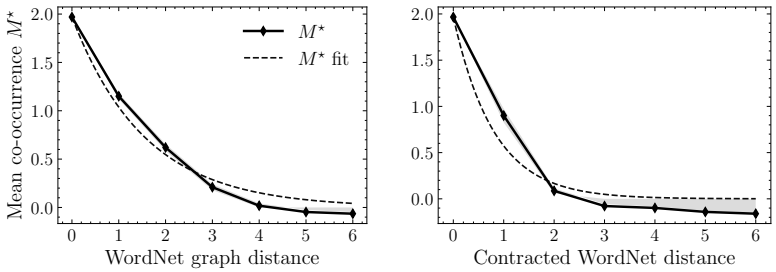

Figure 2: **Mean co-occurrence statistic $M^\star_{ij}$ decays with semantic distance under two taxonomy constructions.** Both the original WordNet-distance measure (left) and the contracted arborescence used in our experiments (right) show monotone, approximately exponential decay. Dashed curves show fitted exponential kernels $f(d) = \alpha e^{-\beta d}$; shaded bands denote one standard error.

We first consider the idealized setting in which $d \geq \mathrm{rank}(M^\star)$ and $M^\star$ is positive semidefinite (PSD), so that all nonzero eigenmodes of $M^\star$ are retained, and the Gram matrix of the embeddings equals $M^\star$. In practice, however, one typically works with $d \leq \mathrm{rank}(M^\star)$, so the embedding is a rank-d truncation built from the top d eigenvectors. Appendix D shows that the same qualitative predictions extend to non-idealized settings where $M^\star$ is indefinite (see Figure 7).

Our theory treats the word embeddings associated with any regular subtree of WordNet. We focus on perfect binary subtrees, as many of them can be constructed, allowing us to quantitatively test our theory. Theoretical extensions to more general regular trees are deferred to Appendix B.

3.2 Hierarchy-adapted spectral decomposition

Observation: In the idealized binary-tree setting under the assumptions of section 3.1, the leading eigenvectors of $M^\star$ organize hierarchically: the first mode varies only with depth, while the dominant split modes are ordered from coarse to fine, as illustrated in the top row of Figure 3. We now explain why this structure arises.

Consider a binary tree of depth L, and identify the vocabulary $\mathcal{S}$ with its nodes.

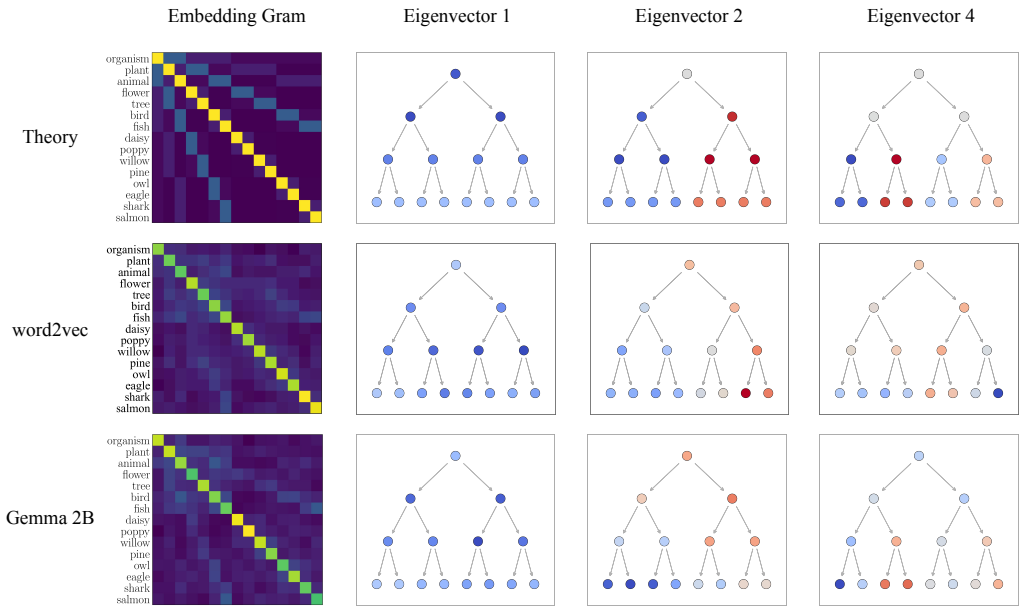

Figure 3: **Hierarchical splitting geometry in word2vec and LLM representation vectors; example: taxonomy of organisms** (Left) Gram matrices from theory G_{th}: a fitted exponential kernel $f(d) = 1.967 \cdot e^{-1.235 \cdot d}$ (top), word2vec G_{word2vec} (middle), and Gemma G_{LLM} (bottom). Inspection of the eigenstructure reveals qualitative agreement with theory. (Right) Top eigenvectors of each Gram matrix, visualized by projecting each node's representation and coloring by sign (blue: positive; red: negative; white: neutral). The eigenvectors show hierarchical splitting geometry: the first mode is approximately constant within each level; the second splits animals from plants; the third and fourth, which are degenerate in the theory, split finer branches. One direction within each degenerate eigenspace is selected for visualization.

Multi-resolution structure on the tree. Functions on the nodes of a tree admit a natural multi-resolution decomposition, analogous to Haar wavelets on the line. This basis consists of:

Scaling basis modes. For each level $\ell = 0, 1, \ldots, L$, let ϕ_ℓ denote the normalized indicator of the set of nodes at depth ℓ: ϕ_ℓ is constant on level ℓ, zero elsewhere, and has unit norm. We refer to ϕ_ℓ as the *scaling basis mode at level* ℓ.

Wavelet basis modes. Fix an internal node u, and let $T(u)$ denote the subtree rooted at u. For each relative depth $r = 1, \ldots, h(u)$, where $h(u)$ is the height of $T(u)$, let $\psi_{u,r}$ denote the normalized Haar wavelet on $T(u)$ at relative depth r: it is supported on the descendants of u at that level, has constant magnitude on that support, takes opposite signs on the two sides of the split at u, and has unit norm. We refer to $\psi_{u,r}$ as the *wavelet basis mode at node* u *and relative level* r. These modes capture hierarchical contrasts inside the subtree rooted at u.

Examples. On the organism tree in Figure 1(a), a scaling basis mode assigns the same value to all nodes on a fixed level. For example, the level-2 scaling basis mode is

$$\phi_{\ell=2} = (e_{\text{flower}} + e_{\text{tree}} + e_{\text{bird}} + e_{\text{fish}}) \ / \ 2.$$

A wavelet basis mode instead contrasts two sibling subtrees. For example,

$$\psi_{u=\text{organism},r=2} = (e_{\text{flower}} + e_{\text{tree}} - e_{\text{bird}} - e_{\text{fish}}) \ / \ 2.$$

Block structure of $M^\star$. Under Assumption 3.1, $M^\star$ is invariant under swapping the two child subtrees of any internal node. To state the resulting invariant subspaces explicitly, define the scaling space

$$S_{\text{sc}} := \text{span}\{\phi_{\ell=0}, \ldots, \phi_{\ell=L}\}, \tag{5}$$

and, for each internal node u, define the split space

$$S^u := \text{span}\{\psi_{u,r=1}, \ldots, \psi_{u,r=h(u)}\}. \tag{6}$$

$M^\star$ preserves these subspaces. Multiplication by $M^\star$ sends a vector x to $(M^\star x)_i = \sum_j f(\mathrm{dist}(i,j))x_j$. For $x \in S_{\mathrm{sc}}$, two nodes at the same depth have equal distance profiles with respect to the other nodes and thus transform equally. If $x = \psi_{u,r}$, then for any node outside $T(u)$, the two child subtrees of u contribute equal weights with opposite signs and cancel. Inside $T(u)$, the result remains antisymmetric across the same split and depends only on relative depth. Hence

$$M^\star S_{\mathrm{sc}} \subseteq S_{\mathrm{sc}}, \qquad M^\star S^u \subseteq S^u.$$

Thus, in the hierarchy-adapted Haar basis, $M^\star$ decomposes into a scaling block $A^{\mathrm{sc}} \in \mathbb{R}^{(L+1)\times(L+1)}$ on S_{sc} and wavelet blocks $A^u \in \mathbb{R}^{h(u)\times h(u)}$ on each S^u. By symmetry, subtrees of the same height h have identical wavelet blocks, denoted $A^{(h)}$. Appendix A.3 illustrates the scaling and wavelet blocks and their dimensions (see Figure 6). This gives the following structural result.

Theorem 1 (Hierarchy-aligned eigenvectors)**.** *The eigenvectors of $M^\star$ can be chosen to lie entirely within either the scaling subspace or a wavelet subspace associated with the subtree of height h. Specifically, the k-th scaling eigenvector has the form*

$$v_k^{\mathrm{sc}} = \sum_{\ell=0}^{L} c_{\ell,k}\phi_\ell, \tag{7}$$

while the k-th eigenvector in the wavelet subspace associated with an internal node u of height h has the form

$$v_k^{(u)} = \sum_{r=1}^{h} a_{r,k}^{(h)}\psi_{u,r}. \tag{8}$$

Ordering of eigenvalues. We now use Assumption 3.2 to constrain the ordering of these modes and further characterize the dominant ones. We write λ_i^{sc} for the i-th largest eigenvalue of A^{sc}, and $\lambda_i^{(h)}$ for the i-th largest eigenvalue of $A^{(h)}$.

Theorem 2 (Coarse-to-fine spectral ordering)**.** *Under Assumption 3.2:*

1. *The largest eigenvalue of $M^\star$ is λ_1^{sc}. The corresponding eigenvector can be written as in eq. (7) with $c_{\ell,1} > 0$ for all ℓ.*
2. *The leading eigenvector of each wavelet block $A^{(h)}$ can be written as in eq. (8) with $a_{r,1}^{(h)} > 0$ for all r.*
3. *The wavelet blocks are nested across subtree height: for each $h \geq 1$, $A^{(h)}$ is exactly the upper-left $h \times h$ corner of $A^{(h+1)}$. Therefore, $A^{(h)}$ is a principal submatrix of $A^{(h+1)}$. Consequently, their eigenvalues satisfy Cauchy interlacing:*

$$\lambda_1^{(h+1)} \geq \lambda_1^{(h)} \geq \cdots \geq \lambda_h^{(h+1)} \geq \lambda_h^{(h)} \geq \lambda_{h+1}^{(h+1)}.$$

4. *In particular, the leading wavelet eigenvalues are monotone in subtree height:*

$$\lambda_1^{(L)} \geq \lambda_1^{(L-1)} \geq \cdots \geq \lambda_1^{(1)}.$$

Thus, dominant non-scaling modes appear coarse-to-fine in the spectrum.

The positivity in theorem 2, item 2, means the leading split direction is sign-coherent. Positive on all descendants of one child, negative on all descendants of the other. It is a clean, linearly readable split direction separating the two branches at every depth.

Interpretation. The top row of Figure 3 shows the hierarchical splitting geometry defined by theorems 1 and 2. The first mode is a scaling eigenvector and only varies along depth, so it is less informative about the branching structure. The dominant wavelet modes resolve the hierarchy from coarse to fine. Among these modes, the root split comes first, followed by splits associated with its children and then deeper nodes. Other spectral modes may interleave with these dominant splits. In this sense, the top of the spectrum reveals the major taxonomic contrasts before finer within-subtree distinctions. Linear readability of individual concepts does not imply a low-dimensional representation of the entire hierarchy, since distinct branches require independent split directions.

Degenerate eigenspaces. For subtrees of the same height, dominant wavelet modes are degenerate by symmetry, so individual eigenvectors are identifiable only up to rotation within the degenerate eigenspace. This explains why empirical eigenvectors may appear as mixtures of same-level splits, such as the plant- and animal-rooted wavelets in Figure 3. Proofs of each statement are given in Appendix A.

3.3 Parametric decay model

The results above characterize the qualitative spectral structure under general decay assumptions. To obtain quantitative predictions, we instantiate the kernel f with a simple parametric form. The parametric model resolves the relative placement of the non-leading scaling modes, which is not fixed by the qualitative theory alone.

As shown in Figure 2, mean normalized co-occurrence decays approximately exponentially in distance over the range relevant to our experiments. We therefore model the decay with the following parametric form:

$$f(\mathrm{dist}(i,j)) = \alpha \cdot e^{-\beta \cdot \mathrm{dist}(i,j)} \tag{9}$$

with fitted parameters $\alpha,\ \beta > 0$. This form satisfies the positivity and monotonic-decay assumptions of the preceding analysis while providing a flexible two-parameter family that captures the observed behavior.

3.4 Effect of noise

In the idealized setting, the hierarchical splitting geometry is exact, but empirical co-occurrence statistics contain fluctuations. Such noise can be treated by regular perturbation theory: it lifts exact degeneracies and mixes directions within nearly degenerate eigenspaces. If small enough, it does not destroy the underlying coarse-to-fine structure. In our empirical tests, we thus use observables designed to be resilient to these effects. These observables are defined in section 4.3.

4 Empirical testing in word2vec embeddings

4.1 Empirical setup

4.1.1 Vocabulary and hierarchy construction

We restrict our analysis to words satisfying three conditions: (i) they appear as noun lemmas in WordNet, (ii) they can be matched to a single token in Gemma's vocabulary, allowing us to test our predictions in this model, and (iii) they have an unambiguous word-sense assignment within the candidate set. The last condition ensures that the hypernym relations of each retained word are well defined.

To obtain a tree-compatible hierarchy, we convert the WordNet noun hypernym DAG into a rooted arborescence by selecting parent assignments that maximize depth from the global root. We then restrict to eligible words and contract paths through ineligible intermediate nodes. The resulting contracted arborescence defines the distances used in all empirical tests. Construction details are given in Appendix C.3. Under this distance, mean normalized co-occurrence decays approximately exponentially as shown in the right panel of Figure 2. We fit the exponential kernel in eq. (9) to this empirical decay curve to use as the theoretical object compared to the empirical embedding Gram matrices below.

4.1.2 word2vec embeddings

We construct word2vec embeddings from the top $d = 2048$ positive eigenmodes of the restricted $M^\star$, following eq. (2). This matches Gemma's representation dimension. Appendix C.6 gives construction details.

4.1.3 Binary subtree sampling

For quantitative tests, we sample perfect binary subtrees of depth $L = 3$. We evaluate all eligible roots at $L = 2, 3, 4, 5$ in the Gemma-2B depth sweep (Appendix E.3; see Figures 12 and 13). We focus on $L = 3$ because its 15-node trees resolve more hierarchical levels than $L = 2$ while retaining 21 eligible roots, compared with 5 at $L = 4$ and 2 at $L = 5$. A valid tree is defined recursively by choosing two distinct valid child subtrees at each internal node. We sample uniformly over valid tree structures using dynamic programming counts, avoiding bias toward individual word sets.

4.2 Example: the taxonomy of organisms

We first compare the word2vec Gram matrix G_{word2vec} for the organism taxonomy of Figure 1 with our theory G_{th}. The top two rows of Figure 3 compare the Gram matrices and leading eigenvectors from each. Our central prediction that the leading eigenvectors organize into scaling and split modes is confirmed in the organism example. The first mode is approximately constant at each level and has a consistent sign, while subsequent modes separate subtrees through sign at progressively finer resolutions, one separating plants from animals, others resolving finer contrasts such as bird versus fish or flower versus tree.

4.3 Quantitative eigenspace alignment across word2vec

We now test whether our predictions hold systematically across many binary subtrees of depth $L = 3$. We sample eligible subtrees and compare the empirical Gram matrix from the word2vec embeddings with the fitted exponential Gram matrix.

To quantify alignment, we compare subspaces rather than individual eigenvectors. The theoretical spectrum contains exponentially many degeneracies as tree depth increases, so individual same-level split directions are only defined up to rotation within their eigenspace. Empirical fluctuations further perturb these directions, making eigenvector-level comparisons unstable. Let $U_k \in \mathbb{R}^{15\times k}$ contain the top k orthonormal eigenvectors of G_{emp}, with G_{emp} equal to G_{word2vec} or G_{LLM}, and let $V_k \in \mathbb{R}^{15\times k}$ contain those of G_{th}, ordered by decreasing eigenvalue. We therefore use top-k eigenspace alignment:

$$g(k) := \frac{1}{k}\|U_k^\top V_k\|_F^2, \tag{10}$$

This quantity approaches unity when the leading eigenspaces of the two matrices are perfectly aligned.

For organism- and cognition-rooted trees, empirical top-k eigenspaces align substantially above shuffled-label baselines, obtained by globally permuting the word and vector pairs, as shown in the left and middle panels of Figure 4.

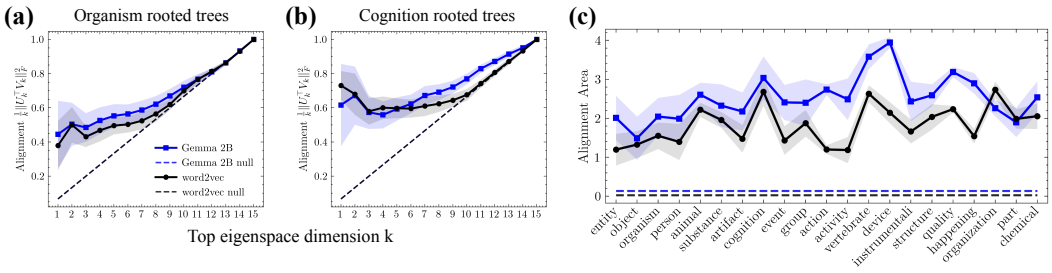

Figure 4: **Hierarchical splitting geometry in binary trees sampled from WordNet.** Top-k eigenspace alignment between theoretical G_{th} and empirical G_{word2vec} or G_{LLM} Gram matrices for (a) organism-rooted trees, and (b) cognition-rooted trees. Panel (c) summarizes all eligible roots using the alignment area from eq. (11). Both word2vec and Gemma unembeddings align substantially above shuffled-label baselines. Shaded bands indicate one standard deviation. Each root uses at most 5,000 sampled trees.

To assess the generality of this finding, we repeat this procedure over all eligible root words with constructible binary trees. For each root, we summarize alignment by the area between the top-k alignment curve and a null diagonal baseline:

$$\text{Alignment Area} = \sum_{k=1}^{15}\left(\frac{1}{k}\|U_k^\top V_k\|_F^2 - \frac{k}{15}\right). \tag{11}$$

We average this statistic over sampled trees from each root. The right panel of Figure 4 shows the distribution across eligible roots together with the shuffled baseline; the empirical alignment area (solid black) is consistently above the null (dashed).

Substantial spectral alignment appears across many semantic hierarchies, not only organism-rooted trees, confirming that a hierarchical splitting geometry is a recurring feature of co-occurrence-based embeddings.

5 Empirical testing for LLMs

We test whether hierarchical splitting geometry extends to Gemma 2B unembeddings [23].

5.1 LLM representation processing

Following Park et al. [24], we center and whiten unembeddings v_i as $\tilde{v}_i = \Sigma^{-1/2}(v_i - \mu)$, where μ and Σ are the full-vocabulary mean and covariance. Alignment also persists without whitening (Appendix E.4; see Figure 14).

5.2 Testing our theory across LLM unembeddings

We compare the processed unembedding Gram matrix G_{LLM} with the fitted co-occurrence kernel G_{th}. The organism example exhibits the predicted scaling and split modes as seen in Figure 3. Across sampled binary trees, top-k eigenspace alignment (eqs. (10) and (11)) substantially exceeds the shuffled baseline (Figure 4). Remarkably, the same signature appears in Llama, Qwen, and Pythia unembeddings and Pythia's untied input embeddings (Appendix E.4; see Figure 15).

Connection to corpus statistics. These observations motivate a possible link between corpus statistics and LLM geometry. An affine softmax head expresses next-token log-probabilities as inner products between context and unembedding vectors, up to normalization and bias terms [25]. For single-word contexts, the corresponding corpus target is determined by directed bigram statistics, an adjacent-word analogue of the windowed co-occurrences underlying $M^\star$ (eq. (1)). In Appendix G, we show that an empirical PMI matrix derived from these bigrams already exhibits hierarchical spectral structure. If the model predicts these statistics accurately and its two factors share the target's structure evenly, each inherits the corresponding singular geometry. This provides a natural route from corpus statistics to the hierarchy observed in unembeddings and Gemma's final-layer context representations. Appendix G develops this account in detail. Verifying prediction accuracy and factor balance remains future work.

5.3 What geometric probes do and do not establish

We ask whether the geometric diagnostics of Park et al. [13] require a hierarchy-specific functional mechanism. Their concept vector $\bar{\ell}_w$ represents membership in w or its descendants, and their functional account predicts that a child's innovation $\bar{\ell}_w - \bar{\ell}_p$ is orthogonal to its parent vector $\bar{\ell}_p$. Intuitively, steering between *dog* and *cat* should change which animal is represented without changing how strongly *animal* applies. Orthogonality expresses this selective control, since moving along the child-specific innovation leaves the linear readout defined by the parent vector unchanged. The diagnostic

$$\cos\big(\bar{\ell}_w - \bar{\ell}_p, \bar{\ell}_p\big) \tag{12}$$

should therefore concentrate near zero for true parent-child pairs.

Following Park et al., we independently select 70% of descendant tokens for each synset to form $Y_{\mathrm{tr}}(w)$. Let μ_w be their mean representation and Σ_w their Ledoit–Wolf shrinkage covariance. The estimated direction and concept vector are

$$\tilde{g}_w = \frac{\Sigma_w^\dagger \mu_w}{\|\Sigma_w^\dagger \mu_w\|_2}, \qquad \bar{\ell}_w = (\tilde{g}_w^\top \mu_w)\tilde{g}_w, \tag{13}$$

where $\dagger$ denotes the pseudoinverse. We apply the same estimator to Gemma unembeddings and theoretical word2vec embeddings constructed from the fitted exponential kernel using eq. (2).

Both representations yield nearly orthogonal parent-child innovations, while random-parent baselines are displaced (Figure 5(d)). Note that orthogonality is approximate even in our noiseless theoretical model. Implementation details are in Appendix F. The theoretical word2vec embeddings also reproduce held-out concept separation and hypernym-chain diagnostics (Figure 5).

Shuffled controls test whether these signatures are nonrandom. Our theoretical word2vec embeddings provide a co-occurrence-based null hypothesis, reproducing them without a hierarchy-specific functional objective. Even held-out success therefore establishes accessibility, not origin or use. These signatures alone provide no discriminating evidence for such a mechanism, but do not rule it out. Establishing functional use requires additional causal evidence, such as interventions.

The theory also predicts that top-PC probes favor broad categories, individual PCs can be unstable within degenerate eigenspaces, and linear readability need not imply a compact representation of the whole hierarchy.

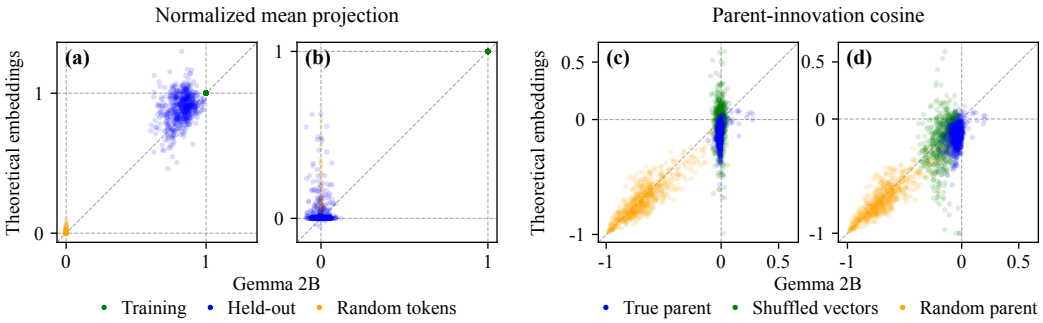

Figure 5: **Concept-vector diagnostics show qualitative agreement between Gemma and theoretical word2vec embeddings.** Horizontal axes show Gemma values and vertical axes show theoretical values. **(a,b)** Mean projections along $\tilde{g}_w$ (eq. (13)), normalized by their training means, for original (a) and shuffled (b) embeddings. Held-out descendants separate from random tokens in (a). This separation largely disappears after shuffling in (b). **(c,d)** Innovation cosines (eq. (12)), with concept vectors estimated from whole descendant sets (c) or training subsets (d). True-parent cosines remain near zero, whereas random-parent cosines are negative. Shuffled-vector controls also lie near zero in (c), but shift negative in (d).

6 Limitations

Our theory applies directly to co-occurrence-driven word2vec embeddings. For LLMs, testing corpus agreement and explaining the transfer without assuming factor balance remain open (Appendix G). Further directions include contextual representations of ambiguous words, mean token representations over corpus contexts, and multilingual and cross-resource tests. For domain-restricted corpora, the theory predicts geometric changes corresponding to changes in the measured co-occurrence kernel. Deviations from corpus-predicted geometry need not be errors. Whether data interventions guided by them alter geometry and downstream behavior remains to be tested.

7 Conclusion

We presented a distributional account of hierarchical semantic geometry. Starting from the assumption that word co-occurrence decays with distance on the WordNet hypernym graph, we derived predictions of hierarchical splitting geometry for word2vec embeddings on regular tree hierarchies. The resulting PCs encode the taxonomy from coarse to fine levels, as confirmed not only in word2vec but also in Gemma unembeddings.

Together with [17, 18], this work supports the following broader view: each word is characterized by discrete attributes (such as gender or singular vs. plural), continuous attributes (such as season of the year or geographical location), and hierarchical attributes (position in WordNet). Words with similar attributes co-occur more often, and this alone gives rise to the elegant geometrical organization of word embeddings. Our results establish this mechanism for co-occurrence-driven embeddings and provide empirical support for a similar distributional account of LLM geometry.

Acknowledgments

This work was supported by the Simons Foundation through the Simons Collaboration on the Physics of Learning and Neural Computation (Award ID: SFI-MPS-POL00012574-05), PI Wyart.

Competing interests

The authors declare no competing interests.

References

[1] Joshua Engels, Eric J Michaud, Isaac Liao, Wes Gurnee, and Max Tegmark. Not all language model features are one-dimensionally linear. In *The Thirteenth International Conference on Learning Representations*, 2025.

[2] Jack Merullo, Carsten Eickhoff, and Ellie Pavlick. Language models implement simple Word2Vec-style vector arithmetic. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5030–5047, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[3] Alexander Modell, Patrick Rubin-Delanchy, and Nick Whiteley. The origins of representation manifolds in large language models. *arXiv preprint arXiv:2505.18235*, 2025.

[4] Wes Gurnee and Max Tegmark. Language models represent space and time. In *The Twelfth International Conference on Learning Representations*, 2024.

[5] Wes Gurnee, Emmanuel Ameisen, Isaac Kauvar, Julius Tarng, Adam Pearce, Chris Olah, and Joshua Batson. When models manipulate manifolds: The geometry of a counting task. *Transformer Circuits Thread*, 2025.

[6] Hanlin Zhu, Melissa Franch, Elizabeth A. Mickiewicz, James L. Belanger, Rhiannon L. Cowan, Kalman A. Katlowitz, Ana G. Chavez, Assia Chericoni, Danika Paulo, Xinyuan Yan, Shervin Rahimpour, Ben Shofty, Eleonora Bartoli, Jay A. Hennig, Nicole R. Provenza, Elliot H. Smith, Steven T. Piantadosi, Benjamin Y. Hayden, and Sameer A. Sheth. A geometric foundation for word meaning in the brain. *bioRxiv preprint 2026.01.28.702241*, 2026.

[7] George A. Miller. Wordnet: a lexical database for english. *Commun. ACM*, 38(11):39–41, November 1995.

[8] Maximillian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[9] Nghia Nguyen, Tianjiao Ding, and René Vidal. Hierarchical concept embedding & pursuit for interpretable image classification. *arXiv preprint arXiv:2602.11448*, 2026.

[10] Ruiji Fu, Jiang Guo, Bing Qin, Wanxiang Che, Haifeng Wang, and Ting Liu. Learning semantic hierarchies via word embeddings. In Kristina Toutanova and Hua Wu, editors, *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1199–1209, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

[11] Masaki Sakata, Benjamin Heinzerling, Takumi Ito, Sho Yokoi, and Kentaro Inui. Linear representations of hierarchical concepts in language models. *arXiv preprint arXiv:2604.07886*, 2026.

[12] Pierre Orhan, Pablo Diego-Simón, Emmanuel Chemla, Yair Lakretz, Yves Boubenec, and Jean-Rémi King. Emergence of phonemic, syntactic, and semantic representations in artificial neural networks. *arXiv preprint arXiv:2601.18617*, 2026.

[13] Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. The geometry of categorical and hierarchical concepts in large language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

[14] Andrew M. Saxe, James L. McClelland, and Surya Ganguli. A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences*, 116(23):11537–11546, 2019.

[15] Andrew M. Saxe, James L. McClelland, and Surya Ganguli. Learning hierarchical categories in deep neural networks. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 35, 2013.

[16] Omer Levy and Yoav Goldberg. Neural word embedding as implicit matrix factorization. *Advances in Neural Information Processing Systems*, 27, 2014.

[17] Daniel J Korchinski, Dhruva Karkada, Yasaman Bahri, and Matthieu Wyart. On the emergence of linear analogies in word embeddings. In *Advances in Neural Information Processing Systems*, 2025.

[18] Dhruva Karkada, Daniel J Korchinski, Andres Nava, Matthieu Wyart, and Yasaman Bahri. Symmetry in language statistics shapes the geometry of model representations. *arXiv preprint arXiv:2602.15029*, 2026.

[19] Rik Sarkar. Low distortion delaunay embedding of trees in hyperbolic plane. In *Graph Drawing*, volume 7034, pages 355–366, United Kingdom, 2011. Springer.

[20] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 26, 2013.

[21] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[22] Dhruva Karkada, James B Simon, Yasaman Bahri, and Michael R DeWeese. Closed-form training dynamics reveal learned features and linear structure in word2vec-like models. In *Advances in Neural Information Processing Systems*, 2025.

[23] Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. Gemma: Open models based on gemini research and technology, 2024.

[24] Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

[25] Zhilin Yang, Zihang Dai, Ruslan Salakhutdinov, and William W. Cohen. Breaking the softmax bottleneck: A high-rank RNN language model. In *International Conference on Learning Representations*, 2018.

[26] Leslie Kish. *Survey sampling.* Wiley, 1965.

[27] Neel Nanda and Joseph Bloom. Transformerlens. https://github.com/TransformerLensOrg/TransformerLens, 2022.

[28] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, et al. The llama 3 herd of models, 07 2024.

[29] Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, February 2004.

[30] Simon Du, Wei Hu, and Jason D Lee. Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[31] Sanjeev Arora, Nadav Cohen, and Elad Hazan. On the optimization of deep networks: Implicit acceleration by overparameterization. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 244–253. PMLR, 10–15 Jul 2018.

A Proofs for the hierarchy-aligned spectral theory

This appendix gives the formal proofs for the theoretical claims in sections 3.1 to 3.3. The purpose of the organization below is to make explicit which assumptions are used in each main-text result.

A.1 From the co-occurrence model to a distance kernel

Proposition 3 (Distance-kernel reduction)**.** *Under Assumption 3.1, if the fluctuation term is neglected, i.e.* $\varepsilon_{ij} = 0$*, then there exists a scalar function* f *such that*

$$M^\star_{ij} = f(\mathrm{dist}(i,j)). \tag{14}$$

Moreover, if the mean normalized co-occurrence is positive and decreasing as a function of distance, then f *satisfies Assumption 3.2.*

Proof. With $\varepsilon_{ij} = 0$, Assumption 3.1 gives

$$\frac{P_{ij}}{P_i P_j} = \widetilde{C}(\mathrm{dist}(i,j)). \tag{15}$$

The statistic in eq. (1) is a deterministic scalar transform of this ratio:

$$M^\star_{ij} = \frac{P_{ij} - P_i P_j}{\frac{1}{2}(P_{ij} + P_i P_j)} = \frac{\widetilde{C}(\mathrm{dist}(i,j)) - 1}{\frac{1}{2}(\widetilde{C}(\mathrm{dist}(i,j)) + 1)}. \tag{16}$$

Thus $M^\star_{ij} = f(\mathrm{dist}(i,j))$ with

$$f(d) := \frac{\widetilde{C}(d) - 1}{\frac{1}{2}(\widetilde{C}(d) + 1)}. \tag{17}$$

The final statement is exactly the empirical positivity and monotonicity condition stated as Assumption 3.2. □

A.2 Binary-tree notation and Haar basis

Let T_L be the full binary tree of depth L, containing all nodes at depths $0, 1, \ldots, L$. Let V be the real vector space of functions on the nodes of T_L, with standard basis vectors e_i. For a node u, let $|u|$ denote its depth, let $h(u) := L - |u|$ denote the height of the subtree rooted at u, and let $T_r(u)$ denote the descendants of u at relative depth r. Thus $|T_r(u)| = 2^r$.

The scaling modes are

$$\phi_\ell := 2^{-\ell/2} \sum_{i \in T_\ell(0)} e_i, \qquad \ell = 0, 1, \ldots, L. \tag{18}$$

If u is internal, write its two children as u_+ and u_-. The wavelet modes rooted at u are

$$\psi_{u,r} := 2^{-r/2} \left(\sum_{i \in T_{r-1}(u_+)} e_i - \sum_{i \in T_{r-1}(u_-)} e_i \right), \qquad r = 1, \ldots, h(u). \tag{19}$$

Equivalently, $\psi_{u,r}$ is supported on $T_r(u)$, is constant on each side of the split at u, takes opposite signs on the two sides, and has unit norm.

Define

$$S_{\mathrm{sc}} := \mathrm{span}\{\phi_0, \ldots, \phi_L\}, \qquad S^u := \mathrm{span}\{\psi_{u,1}, \ldots, \psi_{u,h(u)}\}. \tag{20}$$

These are the spaces defined in eqs. (5) and (6).

Lemma 4 (Orthogonal Haar decomposition)**.** *The collection*

$$\mathcal{H} := \{\phi_0, \ldots, \phi_L\} \cup \{\psi_{u,r} : u \text{ internal},\ r = 1, \ldots, h(u)\} \tag{21}$$

is an orthonormal basis of V*. Equivalently,*

$$V = S_{\mathrm{sc}} \oplus \bigoplus_{u \text{ internal}} S^u. \tag{22}$$

Proof. The normalization in eqs. (18) and (19) gives unit norm. Scaling modes at distinct depths have disjoint support, so they are orthogonal. A scaling mode and a wavelet mode either have disjoint support or lie on the same depth layer; in the latter case the inner product is proportional to the sum of the $+$ and $-$ coefficients of the wavelet, which cancel.

Now consider two wavelet modes. If their supports are on different depth layers, their inner product is zero. If their rooted subtrees are disjoint, their supports are disjoint. If one root is an ancestor of the other, then the ancestor wavelet is constant on the descendant subtree at the relevant depth, while the descendant wavelet has zero sum across the two children of its root; hence their inner product is zero. Finally, if they have the same root, then distinct relative depths have disjoint support and the same relative depth gives unit norm.

It remains only to count dimensions. The number of modes is

$$(L+1) + \sum_{d=0}^{L-1} 2^d (L-d) = 2^{L+1} - 1, \tag{23}$$

which equals the number of nodes of T_L. Hence the orthonormal collection is a basis. □

Throughout the rest of this appendix, let

$$M_{ij} = f(\mathrm{dist}(i,j)) \tag{24}$$

be the idealized distance-kernel matrix on T_L. In the main text, this matrix is $M^\star$ after applying proposition 3.

A.3 Proof of theorem 1: hierarchy-aligned eigenvectors

Lemma 5 (Invariance of the scaling and split spaces)**.** *For M defined by eq. (24),*

$$MS_{\mathrm{sc}} \subseteq S_{\mathrm{sc}}, \qquad MS^u \subseteq S^u \quad \textit{for every internal node } u. \tag{25}$$

Consequently, in the basis $\mathcal{H}$, M is block diagonal with one block on S_{sc} and one block on each S^u.

Proof. First let $x \in S_{\mathrm{sc}}$. Then x is constant on every depth layer. If i and i' have the same depth, there is a tree automorphism sending i to i' and preserving all depth layers. Since $\mathrm{dist}(i,j)$ is invariant under tree automorphisms, the sums

$$(Mx)_i = \sum_j f(\mathrm{dist}(i,j))x_j, \qquad (Mx)_{i'} = \sum_j f(\mathrm{dist}(i',j))x_j \tag{26}$$

are equal. Hence Mx is also constant on depth layers, so $Mx \in S_{\mathrm{sc}}$.

Now fix an internal node u and let $x \in S^u$. Such an x is supported on $T(u)$, is antisymmetric under the swap of the two children of u, and is constant on each relative depth layer within each child subtree. If $i \notin T(u)$, then every node j on the $+$ side of u can be paired with the corresponding node j' on the $-$ side. These paired nodes satisfy $\mathrm{dist}(i,j) = \mathrm{dist}(i,j')$ and $x_j = -x_{j'}$, so their contributions cancel. Thus $(Mx)_i = 0$ for all $i \notin T(u)$.

For $i \in T(u)$, the same child-swap symmetry gives antisymmetry of Mx across the two sides of the split at u. Moreover, automorphisms acting inside either child subtree show that $(Mx)_i$ depends only on the side of u containing i and on the relative depth of i below u. Therefore Mx is supported on $T(u)$, is antisymmetric across the split at u, and is constant on relative depth layers within each side. This is exactly the space S^u. □

Proof of theorem 1. By lemma 4, V decomposes orthogonally as

$$V = S_{\mathrm{sc}} \oplus \bigoplus_{u \text{ internal}} S^u. \tag{27}$$

By lemma 5, each summand is invariant under the symmetric matrix M. Therefore M restricts to a symmetric linear operator on each summand, and each restriction admits an orthonormal eigenbasis. Taking the union of these eigenbases gives an eigenbasis of M whose vectors lie entirely within the scaling space or within a single split space.

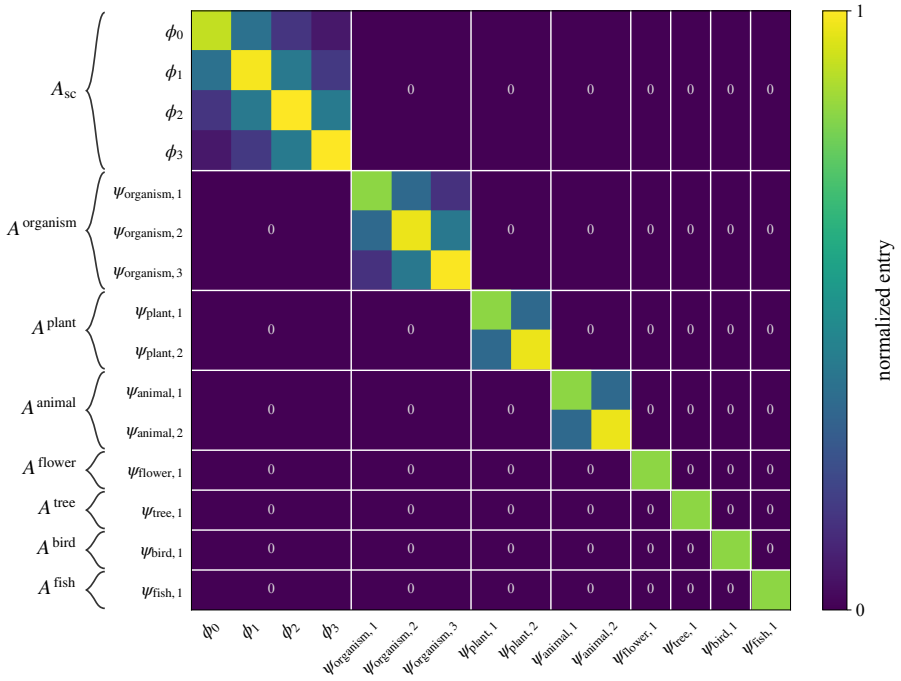

Figure 6: **Invariant blocks in the hierarchy-adapted Haar basis.** For the 15-node organism taxonomy, the transformed distance-kernel matrix has one 4×4 scaling block, one 3×3 root split block, two 2×2 split blocks, and four 1×1 split blocks. Off-block entries vanish exactly; split blocks at the same height coincide. Colors show entries normalized by the largest matrix entry.

The eigenvectors in the scaling block therefore have the form

$$v_k^{\text{sc}} = \sum_{\ell=0}^{L} c_{\ell,k} \phi_\ell, \tag{28}$$

and the eigenvectors in the split block rooted at u have the form

$$v_k^{(u)} = \sum_{r=1}^{h(u)} a_{r,k}^{(h(u))} \psi_{u,r}. \tag{29}$$

These are exactly eqs. (7) and (8). □

A.4 Closed-form block entries

This subsection records the block formulas used in the proof of theorem 2. Let A^{sc} be the matrix of $M|_{S_{\text{sc}}}$ in the basis $(\phi_0, \dots, \phi_L)$, and let $A^{(h)}$ be the common matrix of $M|_{S^u}$ for any node u of height h, in the basis $(\psi_{u,1}, \dots, \psi_{u,h})$.

Lemma 6 (Scaling block)**.** *For $\ell, m = 0, \dots, L$,*

$$A_{\ell,m}^{\text{sc}} = 2^{-(\ell+m)/2} \left[\sum_{a=0}^{\min(\ell,m)-1} 2^{\ell+m-a-1} f(\ell + m - 2a) + 2^{\max(\ell,m)} f(|\ell - m|) \right], \tag{30}$$

where the empty sum is interpreted as zero.

Proof. By definition,

$$A_{\ell,m}^{\text{sc}} = 2^{-(\ell+m)/2} \sum_{i \in T_\ell(0)} \sum_{j \in T_m(0)} f(\text{dist}(i,j)). \tag{31}$$

Group ordered pairs (i, j) by the depth a of their lowest common ancestor. If the lowest common ancestor is strictly above both nodes, then $a = 0, \ldots, \min(\ell, m) - 1$, the distance is $\ell + m - 2a$, and the number of ordered pairs with that value of a is $2^{\ell+m-a-1}$. The remaining pairs are those in which one node is ancestral to the other, including $i = j$ when $\ell = m$. Their distance is $|\ell - m|$, and their number is $2^{\max(\ell,m)}$. Substitution gives eq. (30). □

Lemma 7 (Split block)**.** *For $r, s = 1, \ldots, h$,*

$$A^{(h)}_{r,s} = 2^{-(r+s)/2} \left[-2^{r+s-1} f(r+s) + \sum_{a=1}^{\min(r,s)-1} 2^{r+s-a-1} f(r+s-2a) + 2^{\max(r,s)} f(|r-s|) \right]. \tag{32}$$

Proof. Fix a node u of height h. In the bilinear form

$$\psi_{u,r}^{\top} M \psi_{u,s}, \tag{33}$$

pairs on opposite sides of the split at u enter with negative sign, while pairs on the same side enter with positive sign. There are 2^{r+s-1} ordered pairs on opposite sides, and every such pair has distance $r + s$, giving the first term.

For pairs on the same side of the split, group by the depth a of the lowest common ancestor relative to u. If the lowest common ancestor lies strictly below u and strictly above both nodes, then $a = 1, \ldots, \min(r, s) - 1$, the distance is $r + s - 2a$, and the number of ordered pairs is $2^{r+s-a-1}$. The remaining same-side pairs are those in which one node is ancestral to the other, including the diagonal case, and there are $2^{\max(r,s)}$ such ordered pairs at distance $|r - s|$. Including the normalization of the two wavelets gives eq. (32). □

A.5 Proof of theorem 2: coarse-to-fine spectral ordering

We prove the four items of theorem 2 in order.

Item 1: the largest eigenvalue lies in the scaling block. Assume $f(d) > 0$ for all relevant distances. Then M is a strictly positive symmetric matrix. By the Perron-Frobenius theorem, its largest eigenvalue is simple and has an eigenvector v_{PF} with strictly positive entries.

By theorem 1, every eigenvector of M can be chosen to lie in one of the blocks in eq. (22). Since the Perron-Frobenius eigenvalue is simple, v_{PF} itself must lie entirely in one of these invariant summands. It cannot lie in any split space S^u: if u is not the root, every vector in S^u vanishes outside $T(u)$, and if u is the root, every nonzero vector in S^u has opposite signs on the two sides of the root split. Therefore $v_{\mathrm{PF}} \in S_{\mathrm{sc}}$.

Writing

$$v_{\mathrm{PF}} = \sum_{\ell=0}^{L} c_{\ell,1} \phi_\ell, \tag{34}$$

we have $c_{\ell,1} > 0$ for all ℓ, because v_{PF} is strictly positive and the supports of the ϕ_ℓ are disjoint depth layers. Hence the largest eigenvalue of M is the largest eigenvalue λ_1^{sc} of the scaling block A^{sc}, and its eigenvector has the form stated in item 1.

Item 2: the leading eigenvector of each split block is sign-coherent. Assume now that f is strictly decreasing. We show first that every entry of every split block $A^{(h)}$ is strictly positive. In eq. (32), the negative contribution comes from opposite-side pairs, all at distance $r + s$. The positive contributions come from same-side pairs. The total coefficient of the same-side terms equals the total coefficient of the opposite-side term:

$$\sum_{a=1}^{\min(r,s)-1} 2^{r+s-a-1} + 2^{\max(r,s)} = 2^{r+s-1}. \tag{35}$$

Every same-side distance appearing in the positive terms is strictly smaller than $r + s$. Since f is strictly decreasing, each same-side contribution is larger than the corresponding contribution at distance $r + s$. Therefore

$$A^{(h)}_{r,s} > 0 \qquad \text{for all } r, s = 1, \ldots, h. \tag{36}$$

Thus $A^{(h)}$ is a strictly positive symmetric matrix. Perron-Frobenius implies that its leading eigenvalue is simple and that the corresponding eigenvector can be chosen with strictly positive coordinates in the basis $(\psi_{u,1}, \dots, \psi_{u,h})$. This proves item 2.

Item 3: split blocks are nested and satisfy interlacing. The formula in eq. (32) contains no dependence on h except through the allowed index range $1 \le r, s \le h$. Therefore

$$A^{(h)} = A^{(h+1)}_{1:h,\,1:h}. \tag{37}$$

Thus $A^{(h)}$ is a leading principal submatrix of $A^{(h+1)}$. Since these matrices are symmetric, Cauchy's interlacing theorem gives

$$\lambda_1^{(h+1)} \ge \lambda_1^{(h)} \ge \lambda_2^{(h+1)} \ge \cdots \ge \lambda_h^{(h+1)} \ge \lambda_h^{(h)} \ge \lambda_{h+1}^{(h+1)}. \tag{38}$$

This proves item 3.

Item 4: leading split eigenvalues are ordered coarse-to-fine. Taking the first inequality from item 3 for $h = 1, 2, \dots, L-1$ gives

$$\lambda_1^{(L)} \ge \lambda_1^{(L-1)} \ge \cdots \ge \lambda_1^{(1)}. \tag{39}$$

A split block of height h corresponds to a subtree of height h; larger h means a coarser split higher in the hierarchy. Hence the dominant split modes are ordered from coarser to finer subtrees, proving item 4.

A.6 Exponential kernel used in the parametric model

The qualitative theory above only requires f to be positive and decreasing. The parametric model in section 3.3 uses

$$f(d) = \alpha e^{-\beta d}, \qquad \alpha, \beta > 0. \tag{40}$$

This subsection records the additional algebraic structure of that special case.

Proposition 8 (Rank-one relation for the exponential kernel)**.** *Let f be given by eq. (40). Define*

$$q_r := 2^{r/2} e^{-\beta r}, \qquad \theta := \frac{e^{2\beta}}{2}, \qquad q_0 := 1. \tag{41}$$

For a height-h subtree, the local scaling block and split block satisfy

$$A^{\mathrm{sc},(h)} = \begin{pmatrix} 0 & 0 \\ 0 & A^{(h)} \end{pmatrix} + \alpha \begin{pmatrix} 1 \\ q \end{pmatrix} \begin{pmatrix} 1 \\ q \end{pmatrix}^{\top}, \qquad q = (q_1, \dots, q_h)^{\top}. \tag{42}$$

Consequently, for the full tree,

$$\lambda_1^{\mathrm{sc}} \ge \lambda_1^{(L)} \ge \lambda_2^{\mathrm{sc}} \ge \lambda_2^{(L)} \ge \lambda_3^{\mathrm{sc}} \ge \cdots \ge \lambda_L^{(L)} \ge \lambda_{L+1}^{\mathrm{sc}}. \tag{43}$$

Proof. Substituting $f(d) = \alpha e^{-\beta d}$ into eqs. (30) and (32) gives, with $m = \min(r, s)$,

$$A^{\mathrm{sc},(h)}_{r,s} = \alpha q_r q_s \left[\theta^m + \frac{1}{2} \sum_{a=0}^{m-1} \theta^a \right], \qquad r, s = 0, \dots, h, \tag{44}$$

and

$$A^{(h)}_{r,s} = \alpha q_r q_s \left[\theta^m + \frac{1}{2} \sum_{a=1}^{m-1} \theta^a - \frac{1}{2} \right], \qquad r, s = 1, \dots, h. \tag{45}$$

Comparing the two displays yields

$$A^{\mathrm{sc},(h)}_{0,0} = \alpha, \qquad A^{\mathrm{sc},(h)}_{0,r} = \alpha q_r, \qquad A^{\mathrm{sc},(h)}_{r,s} = A^{(h)}_{r,s} + \alpha q_r q_s \quad r, s \ge 1. \tag{46}$$

This is exactly eq. (42).

Taking $h = L$, the full scaling block is a positive rank-one perturbation of $0 \oplus A^{(L)}$. For the exponential kernel, $A^{(L)}$ is positive semidefinite because $e^{-\beta \mathrm{dist}(i,j)}$ is a positive-semidefinite tree kernel; hence the eigenvalues of $0 \oplus A^{(L)}$ are ordered as $\lambda_1^{(L)} \ge \cdots \ge \lambda_L^{(L)} \ge 0$. Cauchy's interlacing theorem for a rank-one symmetric perturbation then gives eq. (43). □

Corollary 9 (Placement of non-leading scaling modes)**.** *Under the exponential kernel, for* $k = 1, \dots, L-1$*,*

$$\lambda_1^{(L-k)} \geq \lambda_{k+1}^{(L)} \geq \lambda_{k+2}^{\mathrm{sc}}. \tag{47}$$

In particular,

$$\lambda_1^{(L-1)} \geq \lambda_3^{\mathrm{sc}}, \qquad \lambda_1^{(L-2)} \geq \lambda_4^{\mathrm{sc}}, \qquad \textit{and so on.} \tag{48}$$

Proof. The inequality $\lambda_{k+1}^{(L)} \geq \lambda_{k+2}^{\mathrm{sc}}$ follows from eq. (43). The inequality $\lambda_1^{(L-k)} \geq \lambda_{k+1}^{(L)}$ follows by iterating the split-block interlacing inequalities from theorem 2. Combining the two inequalities gives the result. □

A.7 Perturbation statement used in the noise discussion

The main text uses the standard perturbative intuition that small empirical fluctuations lift degeneracies and rotate directions inside nearly degenerate eigenspaces, while preserving well-separated spectral subspaces. The precise statement needed for our use of top-k eigenspace alignment is the following standard consequence of the Davis-Kahan sin-Θ theorem.

Proposition 10 (Stability of clustered eigenspaces)**.** *Let* M_0 *be the ideal distance-kernel matrix and* $M = M_0 + E$ *a symmetric perturbation. Let* U *have orthonormal columns spanning a cluster of consecutive eigenvalues of* M_0*, separated from the rest of its spectrum by a gap* $\gamma > 0$*. Let* $\widehat{U}$ *span the eigenvectors with the corresponding indices in the ordered spectrum of* M*. If* $\|E\|_2 < \gamma/2$*, then*

$$\|\sin\Theta(U, \widehat{U})\|_2 \leq \frac{2\|E\|_2}{\gamma}. \tag{49}$$

Proof. By Weyl's inequality, the selected unperturbed cluster and the complementary perturbed spectrum are separated by at least $\gamma - \|E\|_2 > \gamma/2$. Applying the Davis–Kahan theorem gives the stated bound. Individual eigenvectors may rotate within the cluster, while the cluster's invariant subspace remains stable. □

B Extension of theory to more general trees

B.1 Spherically symmetric trees

Let the tree have depth L, and suppose that every node at depth d has b_d children, for $d = 0, \dots, L-1$. For a node u at depth d, define

$$N_{d,r} := \prod_{q=d}^{d+r-1} b_q, \qquad N_{d,0} := 1, \tag{50}$$

so that $N_{d,r} = |T_r(u)|$ is the number of descendants of u at relative depth r.

As in the binary case, define the scaling modes by

$$\phi_\ell = \frac{1}{\sqrt{N_{0,\ell}}} \sum_{i \in T_\ell(0)} e_i, \qquad \ell = 0, \dots, L, \tag{51}$$

and set

$$S_{\mathrm{sc}} := \mathrm{span}\{\phi_0, \dots, \phi_L\}. \tag{52}$$

For an internal node u at depth d, let $u_1, \dots, u_{b_d}$ denote its children, and choose an orthonormal basis

$$c^{(1)}, \dots, c^{(b_d - 1)} \in \left\{ c \in \mathbb{R}^{b_d} : \sum_{q=1}^{b_d} c_q = 0 \right\}. \tag{53}$$

For $a = 1, \dots, b_d - 1$ and $r = 1, \dots, L - d$, define

$$\psi_{u,r,a} = \frac{1}{\sqrt{N_{d+1,r-1}}} \sum_{q=1}^{b_d} c_q^{(a)} \sum_{i \in T_{r-1}(u_q)} e_i. \tag{54}$$

These modes contrast the child subtrees of u while remaining constant on relative-depth layers inside each child subtree. Define

$$S^u := \mathrm{span}\{\psi_{u,r,a} : r = 1, \dots, L-d,\ a = 1, \dots, b_d - 1\}. \tag{55}$$

Theorem 11 (Hierarchy-adapted decomposition for spherically symmetric trees)**.** *The modes* $\{\phi_\ell\}$ *and* $\{\psi_{u,r,a}\}$ *form an orthonormal basis of the vertex space, giving*

$$\mathbb{R}^{|T|} = S_{\mathrm{sc}} \oplus \bigoplus_{u \in \mathrm{Int}(T)} S^u. \tag{56}$$

Moreover, if

$$M_{ij} = f(\mathrm{dist}(i,j)), \tag{57}$$

then M *preserves* S_{sc} *and every* S^u*. In the corresponding hierarchy-adapted basis,*

$$M \cong A^{\mathrm{sc}} \oplus \bigoplus_{d=0}^{L-1} \bigoplus_{u:\,|u|=d} \left(A^{[d]} \otimes I_{b_d - 1} \right), \tag{58}$$

where $A^{[d]} \in \mathbb{R}^{(L-d)\times(L-d)}$ *acts on the relative-depth coordinate and* I_{b_d-1} *acts on the child-contrast coordinate.*

Proof. The proof follows the same orthogonality and cancellation argument as in the binary case. Scaling modes at different depths have disjoint support. Each split mode has zero sum across the children of its root, so it is orthogonal to every scaling mode. Two split modes are orthogonal if their supports are disjoint or lie on different depth layers. If one split root is an ancestor of the other, then the ancestor split mode is constant on the descendant subtree while the descendant split mode has zero child-sum, so their inner product vanishes. Counting the modes gives

$$(L+1) + \sum_{d=0}^{L-1} \left(\prod_{q=0}^{d-1} b_q \right) (L-d)(b_d - 1) = |T|, \tag{59}$$

so the orthonormal collection is a basis.

Now let $M_{ij} = f(\mathrm{dist}(i,j))$. Since distance is invariant under every rooted-tree automorphism, and in particular under permutations of sibling subtrees at any internal node, M sends depth-layer-constant vectors to depth-layer-constant vectors. Hence S_{sc} is invariant.

For a split space S^u, contributions from the child subtrees of u cancel outside $T(u)$, because the child-contrast coefficients sum to zero and paired nodes have equal distance to any point outside $T(u)$. Inside $T(u)$, symmetry implies that the result is still supported on $T(u)$, lies in the child-contrast subspace at u, and can only mix the relative-depth coordinate and the contrast coordinate. Thus S^u is invariant.

Finally, for fixed depth $d = |u|$, all nodes u have isomorphic descendant trees. The action of permuting the b_d children of u is the standard representation on the contrast index a. Since M is invariant under these permutations, Schur's lemma implies that the restriction of M to S^u acts as the identity on the child-contrast factor and only mixes the relative-depth coordinate. Therefore the block has the form $A^{[d]} \otimes I_{b_d-1}$. □

Irregular trees The exact invariant decomposition above uses spherical symmetry. For a fully irregular rooted tree, local Haar-like contrasts can still be defined, but a distance kernel $M_{ij} = f(\mathrm{dist}(i,j))$ need not preserve each local split space exactly, because sibling subtrees may have different sizes or shapes. Thus irregular WordNet-induced trees should be understood as empirical perturbations of the symmetric idealization.

B.2 s-ary trees

The full s-ary tree is the special case $b_0 = \cdots = b_{L-1} = s$. In this setting the depth-dependent blocks $A^{[d]}$ depend only on the remaining subtree height $h = L - d$, so we write them as $A^{(h)}$.

Corollary 12 (Fixed s-ary trees)**.** *Suppose* $b_0 = \cdots = b_{L-1} = s$*. Then*

$$M \cong A^{\mathrm{sc}} \oplus \bigoplus_{u \in \mathrm{Int}(T)} \left(A^{(h(u))} \otimes I_{s-1} \right). \tag{60}$$

Moreover,

$$A^{(h)} = A^{(h+1)}_{1:h,\,1:h}. \tag{61}$$

Therefore Cauchy interlacing gives

$$\lambda_1^{(h+1)} \geq \lambda_1^{(h)} \geq \lambda_2^{(h+1)} \geq \cdots \geq \lambda_h^{(h+1)} \geq \lambda_h^{(h)} \geq \lambda_{h+1}^{(h+1)}. \tag{62}$$

In particular,

$$\lambda_1^{(L)} \geq \lambda_1^{(L-1)} \geq \cdots \geq \lambda_1^{(1)}. \tag{63}$$

Thus the dominant split modes are ordered from coarser to finer subtrees for every fixed branching factor $s \geq 2$. The binary case in the main text is the special case $s = 2$.

Proof. When $b_0 = \cdots = b_{L-1} = s$, every height-$h$ subtree is isomorphic to every other height-h subtree. Hence the split block depends only on h. The height-h block is obtained from the height-$(h+1)$ block by restricting the relative-depth indices to $1, \ldots, h$, so it is a leading principal submatrix. The interlacing and monotonicity statements then follow from Cauchy's interlacing theorem. □

Explicit block entries. We now record the explicit block entries for the fixed-branching case. To avoid overloading notation, let b denote the fixed branching factor of the full b-ary tree. For a node u, recall that

$$T_r(u) := \{i \in T(u) : i \text{ is at relative depth } r \text{ below } u\}, \tag{64}$$

so that $|T_r(u)| = b^r$.

The local scaling block for a height-h subtree has entries

$$A_{r,t}^{\mathrm{sc},(h)} = \frac{1}{\sqrt{b^r b^t}} \sum_{i \in T_r(u)} \sum_{j \in T_t(u)} f(\mathrm{dist}(i,j)), \qquad r, t = 0, \ldots, h. \tag{65}$$

Grouping pairs by the relative depth ℓ of their lowest common ancestor gives

$$A_{r,t}^{\mathrm{sc},(h)} = b^{-(r+t)/2} \left[\sum_{\ell=0}^{\min(r,t)-1} (b-1) b^{r+t-\ell-1} f(r+t-2\ell) + b^{\max(r,t)} f(|r-t|) \right], \tag{66}$$

where the empty sum is interpreted as zero. The first term counts pairs whose lowest common ancestor is strictly above both nodes; the final term counts pairs for which one node is ancestral to the other, including $i = j$.

For split modes, fix an internal node u. Let $u_1, \ldots, u_b$ be its children, and choose an orthonormal basis

$$c^{(1)}, \ldots, c^{(b-1)} \in \left\{ c \in \mathbb{R}^b : \sum_{q=1}^{b} c_q = 0 \right\}. \tag{67}$$

For contrast direction $a = 1, \ldots, b-1$, define

$$\psi_{u,r,a} = \frac{1}{\sqrt{b^{r-1}}} \sum_{q=1}^{b} c_q^{(a)} \sum_{i \in T_{r-1}(u_q)} e_i, \qquad r = 1, \ldots, h(u). \tag{68}$$

The split space at u is

$$S^u = \mathrm{span}\{\psi_{u,r,a} : r = 1, \ldots, h(u),\ a = 1, \ldots, b-1\}, \tag{69}$$

so $\dim S^u = h(u)(b-1)$.

Because $M_{ij} = f(\mathrm{dist}(i,j))$ is invariant under permutations of the b child subtrees, Schur's lemma implies that M acts as the identity on the contrast index a. Thus the full split block on S^u has the form

$$A^{(h)} \otimes I_{b-1}, \tag{70}$$

up to the ordering of the basis. Equivalently,

$$\psi_{u,r,a}^\top M \psi_{u,t,a'} = A_{r,t}^{(h)} \delta_{a,a'}. \tag{71}$$

The depth block $A^{(h)} \in \mathbb{R}^{h \times h}$ has entries

$$A_{r,t}^{(h)} = \psi_{u,r,a}^\top M \psi_{u,t,a}, \qquad r, t = 1, \ldots, h, \tag{72}$$

for any fixed contrast direction a. This quantity is independent of a.

Grouping pairs by the relative depth ℓ of their lowest common ancestor gives

$$A^{(h)}_{r,t} = b^{-(r+t-2)/2}\Bigg[- b^{r+t-2} f(r+t) + \sum_{\ell=1}^{\min(r,t)-1} (b-1) b^{r+t-\ell-2} f(r+t-2\ell) + b^{\max(r,t)-1} f(|r-t|)\Bigg]. \tag{73}$$

The negative term comes from pairs lying in different child subtrees of u; the positive terms come from pairs lying in the same child subtree. For $b = 2$, $\dim S^u = h(u)$ and this reduces to the binary split block used in the main text.

Exponential kernel. For the exponential kernel

$$f(d) = \alpha e^{-\beta d}, \qquad \alpha, \beta > 0, \tag{74}$$

the s-ary block formulas take a particularly compact form. Let

$$x := e^{-\beta}, \qquad q_r := s^{r/2} x^r, \qquad \theta := \frac{1}{s x^2}. \tag{75}$$

Then, for $m = \min(r, t)$,

$$A^{\mathrm{sc},(h)}_{r,t} = \alpha q_r q_t \left[\theta^m + \frac{s-1}{s} \sum_{a=0}^{m-1} \theta^a\right], \qquad r, t = 0, \dots, h, \tag{76}$$

and

$$A^{(h)}_{r,t} = \alpha q_r q_t \left[\theta^m + \frac{s-1}{s} \sum_{a=1}^{m-1} \theta^a - \frac{1}{s}\right], \qquad r, t = 1, \dots, h. \tag{77}$$

For $s = 2$, these reduce to the binary formulas above.

Most importantly, the same rank-one relation survives for every s. Comparing the two displays, we obtain

$$A^{\mathrm{sc},(h)}_{0,0} = \alpha, \qquad A^{\mathrm{sc},(h)}_{0,r} = \alpha q_r, \qquad A^{\mathrm{sc},(h)}_{r,t} = A^{(h)}_{r,t} + \alpha q_r q_t \quad r, t \geq 1. \tag{78}$$

Equivalently,

$$A^{\mathrm{sc},(h)} = \begin{pmatrix} 0 & 0 \\ 0 & A^{(h)} \end{pmatrix} + \alpha \begin{pmatrix} 1 \\ q \end{pmatrix} \begin{pmatrix} 1 \\ q \end{pmatrix}^{\top}, \qquad q = (q_1, \dots, q_h)^{\top}. \tag{79}$$

Thus the same spectral consequences hold. The local scaling block is a positive rank-one perturbation of $0 \oplus A^{(h)}$, so its eigenvalues interlace with those of $A^{(h)}$. Since the exponential tree kernel $e^{-\beta \mathrm{dist}(i,j)}$ is positive semidefinite, the split block $A^{(h)}$ is also positive semidefinite, and the eigenvalues of $0 \oplus A^{(h)}$ are ordered with the additional zero eigenvalue at the end. Combined with the principal-submatrix relation above, this again implies that dominant split modes are ordered from coarser to finer subtrees, while the non-leading scaling modes are interleaved with the split spectrum. In particular, the qualitative hierarchy-aligned spectral picture derived for binary trees is not tied to binary branching; it is a consequence of distance-dependent co-occurrence on a symmetric rooted tree.

C Experimental details

Code availability. Code is available at https://github.com/andresnava1000/hierarchical-concept-geometry.

This appendix gives implementation details for the main-text experiments. We first describe the construction of each main-text figure, then describe the shared preprocessing, WordNet hierarchy, co-occurrence, word2vec, and Gemma-unembedding pipelines. The treatment of indefinite $M^\star$ and its positive spectral component is deferred to Appendix D.

C.1 Main-text figure construction

C.1.1 Figure 1: illustrative organism taxonomy

Figure 1 is illustrative and is not used as a separate quantitative result. The displayed taxonomy is chosen to make the predicted splitting geometry visually interpretable and comparable to the organism example of Park et al. [13]. We fix the top-level structure to *organism* branching into *plant* and *animal*; among valid descendant trees, we select a tree using simple frequency and WordNet-structure heuristics, described in Appendix C.10. The quantitative evidence comes from the sampled binary-subtree experiments in Figure 4.

Panel (b) plots the mean normalized co-occurrence statistic $M^\star_{ij}$ as a function of edge distance within this taxonomy. Panel (c) diagonalizes the theoretical Gram matrix obtained from a distance kernel evaluated on the full binary-tree distance matrix.

C.1.2 Figure 2: global co-occurrence decay

Figure 2 estimates the relationship between semantic distance and normalized co-occurrence. We compute $M^\star$ from the corpus co-occurrence probabilities described in Appendix C.5. For each distance $d = 0, \dots, 6$, we estimate the mean of $M^\star_{ij}$ by sampling 5000 eligible synset pairs at that distance.

Pairs are sampled by first drawing an eligible synset u, then using breadth-first search to count and sample an eligible synset v from the distance-d frontier of u. For $d = 0$, we take $v = u$. For $d > 0$, samples with empty eligible frontiers are skipped. To correct for the fact that the procedure samples a start node and then a frontier element, each sampled pair is weighted by the size of the eligible distance-d frontier from which v was drawn. Means and standard errors are computed from these weighted samples, using the Kish effective sample size [26].

The left panel uses minimum undirected distance in the original WordNet noun hypernym DAG, restricted to eligible synsets. The right panel uses undirected edge distance in the contracted WordNet hierarchy described in Appendix C.3. Error bars indicate one standard error of the weighted binned mean.

We fit the exponential kernel in eq. (9) by weighted nonlinear least squares on the binned means. For bins with positive finite mean $\bar{M}_d$ and positive finite standard error s_d, we minimize

$$\sum_d \left(\frac{\bar{M}_d - \alpha e^{-\beta d}}{s_d} \right)^2, \qquad \alpha > 0, \quad \beta \geq 0.$$

The contracted-hierarchy fit is used as the default theoretical kernel for the main empirical comparisons.

C.1.3 Figure 3: organism example in theory, word2vec, and Gemma

Figure 3 compares the theoretical, word2vec, and Gemma Gram matrices on the same illustrative organism taxonomy used in Figure 1. The theoretical Gram matrix is obtained by evaluating the fitted distance kernel on tree distances. The word2vec Gram matrix is $G_{\text{word2vec}} = WW^\top$, where W is constructed from the top positive eigenmodes of the restricted $M^\star$ matrix according to eq. (2). The Gemma Gram matrix is computed from centered and whitened Gemma unembedding vectors restricted to the same tokens.

For the eigenvector visualizations, node colors represent the signed value of the corresponding eigenvector coordinate. The color scale runs from blue for positive values through white for values near zero to red for negative values. Since degenerate eigenspaces have no canonical basis, we display one representative direction from each degenerate eigenspace; such directions may appear as mixtures of same-level splits.

C.1.4 Figure 4: top-k eigenspace alignment

Figure 4 is the main quantitative experiment. We sample perfect binary subtrees of depth $L = 3$ from the contracted WordNet hierarchy. For each sampled tree, we compute a theoretical Gram matrix from the fitted distance kernel and empirical Gram matrices from word2vec and Gemma unembeddings.

We use $g(k)$ and the alignment area defined in eqs. (10) and (11).

The ambient dimension is 15, since a depth-3 full binary tree has $2^4 - 1 = 15$ nodes.

The shuffled-label baseline globally permutes the mapping from synsets to empirical vectors while keeping the theoretical tree fixed. This preserves the global distribution and spectrum of empirical vectors but destroys the semantic correspondence between positions in the hierarchy and representation vectors.

In addition to the global shuffled-label baseline, we also compute a within-tree shuffle baseline. For each sampled tree, let $x_1, \dots, x_n$ denote the empirical vectors assigned to the $n = 15$ tree positions in breadth-first order. The within-tree shuffle samples a random permutation π of $\{1, \dots, n\}$ and forms the permuted empirical Gram matrix

$$G_{ij}^{\text{within}} = x_{\pi(i)}^\top x_{\pi(j)}.$$

This baseline preserves the sampled tree, the multiset of empirical vectors appearing in that tree, the empirical Gram spectrum, and all pairwise similarities among those vectors, but destroys the assignment between particular semantic nodes and particular tree positions. It is therefore a more local control than the global shuffled-label baseline: the global shuffle tests whether the vocabulary-level semantic correspondence matters, while the within-tree shuffle tests whether the ordering of the vectors within each sampled hierarchy matters beyond the unordered vector content of that tree.

For every root and every representation setting, we sample up to 5000 valid binary subtrees, using all valid subtrees when fewer than 5000 are available. For each sampled tree, we compute one shuffled-label baseline by permuting the mapping from synsets to empirical vectors while keeping the theoretical tree fixed. Thus each root contributes up to 5000 original tree-level alignment curves and up to 5000 shuffled tree-level alignment curves.

C.1.5 Figure 5: concept-vector diagnostics

We apply the concept-vector estimator of Park et al. [13] to Gemma unembeddings and theoretical embeddings constructed from the fitted co-occurrence kernel, using eq. (13).

For panels (a,b), we independently select 70% of descendant tokens for each synset to estimate its concept direction. We plot mean projections of training, held-out, and random tokens along this direction, normalized by the training-set mean. Panel (b) repeats the procedure after shuffling token-vector assignments.

Panels (c,d) show the parent-child innovation cosine from eq. (12), with concept vectors estimated from whole descendant sets in (c) and training subsets in (d). We compare true-parent, random-parent, and shuffled-vector controls. The random-parent control replaces each true parent with a randomly selected eligible concept; the shuffled-vector control permutes token-vector assignments before concept estimation. Covariance regularization and implementation details are given in Appendix F.

C.2 Vocabulary preprocessing

C.2.1 Token matching

We begin with WordNet noun lemmas that have co-occurrence statistics. We enforce token hygiene by lowercasing lemma strings, excluding digits, and excluding multiword or hyphenated lemmas unless otherwise stated. We then require that each lemma match a single Gemma tokenizer entry after removing tokenizer boundary markers, stripping whitespace, and lowercasing.

When multiple tokenizer entries match the same cleaned lemma, we prefer the exact leading-whitespace token, then other leading-boundary tokens, then shorter deterministic matches; lowercased surfaces are preferred to all-caps variants. After intersecting WordNet noun lemmas with the co-occurrence vocabulary and Gemma single-token vocabulary, $17{,}566$ lemmas remain.

C.2.2 Polysemy filtering

To obtain an unambiguous synset-lemma assignment, we apply a polysemy filter. For a lemma ℓ and synset s, define

$$\text{monosemy}(\ell, s) = \frac{\text{count}(\ell, s)}{\sum_{s'} \text{count}(\ell, s')}, \tag{80}$$

where counts are WordNet lemma counts summed across all parts of speech. If the total count is zero or unavailable, the score is set to zero. For each synset, we choose the eligible lemma with highest monosemy score, breaking ties by English Zipf frequency from `wordfreq`. Synsets for which no eligible lemma is uniquely preferred are removed. This produces $11{,}735$ unique synset-lemma pairs.

C.3 WordNet hierarchy construction

WordNet's noun hypernym graph is a DAG because some synsets have multiple hypernyms. We convert it into a rooted arborescence before sampling binary subtrees. Starting from the global WordNet noun root, we assign each synset a single parent by choosing a hypernym parent that maximizes depth from the root. Ties are broken deterministically by sorting candidate parent synset names lexicographically and taking the first candidate.

We then restrict this arborescence to the eligible synsets from Appendix C.2.2. Ineligible intermediate nodes are contracted out: the parent of an eligible synset becomes its nearest eligible ancestor in the rooted arborescence. The resulting contracted arborescence is the hierarchy used for the right panel of Figure 2, the fitted theoretical kernel, and the binary-subtree experiments. Unless otherwise stated, semantic distance means undirected edge distance in this contracted hierarchy.

C.4 Binary subtree sampling

A valid depth-L binary subtree rooted at u is defined recursively. For $L = 0$, the only valid tree is u. For $L > 0$, a valid tree consists of u and two distinct valid child subtrees of depth $L - 1$. We use $L = 3$ in the main quantitative experiments.

To sample uniformly over valid tree structures, we compute dynamic-programming counts $N(u, L)$, the number of valid depth-L binary subtrees rooted at u. At each internal node, child pairs are selected with probability proportional to the number of valid completions below them. This samples uniformly over admissible binary tree structures rather than over individual word sets.

C.5 Co-occurrence statistics

We construct co-occurrence statistics from the November 2023 English Wikipedia dump accessed through `https://huggingface.co/datasets/wikimedia/wikipedia`. The corpus is lowercased and tokenized by extracting contiguous alphabetic spans using the regular expression `[a-z]+`. We discard articles with fewer than 500 retained tokens.

The co-occurrence vocabulary consists of observed clean WordNet noun lemmas. Concretely, we enumerate English WordNet noun lemmas, retain only lemmas matching `[a-z]+`, and intersect this set with tokens observed in the preprocessed Wikipedia corpus. We then write the filtered corpus as a memory-mapped token stream, removing tokens outside this vocabulary and storing explicit article-boundary offsets so that co-occurrence pairs are never counted across articles.

Given corpus tokens $C[\nu]$, we estimate a symmetric weighted skip-gram co-occurrence distribution with context window $L = 16$. For a token pair separated by distance d, we use weight $r(d) = L/d$. The unnormalized co-occurrence count is

$$\#(i,j) = \sum_{\nu} \sum_{d=1}^{L} \frac{L}{d} \left[\mathbf{1}\{C[\nu] = i, C[\nu + d] = j\} + \mathbf{1}\{C[\nu] = j, C[\nu + d] = i\} \right], \tag{81}$$

where the sums are restricted to positions within the same article. We normalize by the total co-occurrence mass,

$$P_{ij} = \frac{\#(i,j)}{\sum_{i',j'} \#(i',j')}. \tag{82}$$

The unigram probabilities used in $M^\star$ are the marginals of this co-occurrence distribution,

$$P_i = \sum_j P_{ij}.$$

Finally, we form the normalized co-occurrence statistic as in eq. (1). All matrices are stored sparsely on the observed co-occurrence support. The relationship between this possibly indefinite matrix and the positive semidefinite Gram matrix realized by Euclidean embeddings is discussed in Appendix D.

C.6 word2vec embeddings

We construct the word2vec embeddings directly from the normalized co-occurrence matrix using eq. (2). We first restrict $M^\star$ to the eligible synset-lemma set after the token-matching and polysemy filters of Appendices C.2.1 and C.2.2, obtaining $M^\star_S$. Since a single Euclidean embedding space realizes a positive semidefinite Gram matrix, we retain the positive spectral modes of $M^\star_S$. In the

main experiments, we keep the top $d = 2048$ positive eigenmodes, matching the dimension of the Gemma unembedding vectors. The treatment of negative eigenvalues of $M^\star$, and the corresponding PSD component of the co-occurrence matrix, is given in Appendix D.

C.7 Gemma unembeddings and LLM internal activations

We use the pretrained `google/gemma-2b` checkpoint of Gemma Team [23], accessed through the Hugging Face `transformers` library. This checkpoint is a two-billion-parameter decoder-only language model trained on 3T tokens, with a vocabulary of 256,128 tokens and representation dimension 2048. We analyze its unembedding vectors and, in robustness checks, contextual residual-stream activations through `TransformerLens` [27].

For each eligible lemma, we identify its matched single token and extract the corresponding vector from the model's output embedding matrix. In implementation, this is the matrix returned by `model.get_output_embeddings().weight`. We then center and whiten the full vocabulary of unembedding vectors before restricting to WordNet tokens, as described in section 5.1. The covariance inverse square root is computed by eigendecomposition of the empirical full-vocabulary covariance. The whitening transform is fit once on the full Gemma vocabulary and then applied to the eligible WordNet tokens. Robustness checks using globally centered but unwhitened Gemma vectors are reported in Appendix E.4.

For the LLM internal-activation controls, we also extract residual-stream activations from Gemma and Llama using `TransformerLens`. For each WordNet synset, we convert the selected best lemma to a raw title-cased string with no prefix or suffix; for example, `animal.n.01`, `cat.n.01`, and a multiword lemma such as `sea_turtle` are represented by the prompts "Animal", "Cat", and "Sea Turtle", respectively. We then extract the final-token activation from the post-block residual stream,

$$\text{blocks}.\ell.\text{hook_resid_post},$$

where $\ell = \lfloor n_{\text{layers}}/2 \rfloor$ is the exact middle layer of the model. The extracted activation vectors are globally centered across the eligible synset vocabulary before computing Gram matrices and alignment scores. We do not tune over layers or prompts; the middle layer and raw lexical prompt are used as a simple, model-agnostic choice.

For the Gemma internal-activation control, the middle layer is $\ell = 9$. Its residual-stream dimension is 2048, and its vocabulary size is 256,128.

For the Llama internal-activation control, we use `meta-llama/Llama-3.2-1B`, a member of the Llama 3 family of autoregressive transformer language models [28]. In the `TransformerLens` configuration used here, this model has 16 transformer layers, residual-stream dimension 2048, and vocabulary size 128,256; hence the exact middle layer is $\ell = 8$. We apply the same extraction protocol as for Gemma: raw title-cased best-lemma prompts, final-token residual-stream activations, and global centering across the eligible synset vocabulary.

Additional model checkpoints. We use the base checkpoints `Qwen/Qwen2.5-1.5B` (28 layers, dimension 1536, tied input/output weights) and `EleutherAI/pythia-1.4b` (24 layers, dimension 2048, untied weights; final step 143,000). Their cached Hugging Face revisions are `8faed761d45a263340a0528343f099c05c9a4323` and `fedc38a16eea3bd36a96b906d78d11d2ce18ed79`, respectively. We analyze Qwen's output matrix and both Pythia matrices, using the same full-vocabulary centering and whitening procedure.

The purpose of these internal-activation controls is not to optimize a model-specific semantic layer, but to test whether the hierarchical alignment signal appears in a simple contextual representation of modern decoder-only language models.

C.8 Computational resources

All experiments were run on an Apple M4 CPU machine with approximately 20GB of available RAM and no discrete GPU. The most expensive preprocessing step was the construction of Wikipedia co-occurrence statistics, which took roughly 48 hours. The eigendecomposition step took approximately 30 minutes. All other figure-generation and alignment computations took at most 30 minutes each. Extracting the Gemma and Llama internal activations used in the robustness controls took approximately 30 minutes in total.

C.9 Existing assets and licenses.

We use WordNet, the November 2023 English Wikipedia dump, pretrained Gemma/Llama model resources, TransformerLens, and standard scientific Python libraries. These assets are cited in the paper, and the accompanying code/supplement lists the versions, sources, and available license or terms-of-use information. We use these assets for non-commercial research analysis and do not redistribute pretrained model weights or the Wikipedia dump.

C.10 Illustrative organism taxonomy

The organism taxonomy in Figures 1 and 3 is used only for visualization. We fixed the top-level structure to match Park et al. [13]: *organism* branches into *plant* and *animal*. Among valid descendant trees satisfying this upper structure, we selected a tree using a lexicographic heuristic that first avoids low-frequency lemmas. The quantitative results do not depend on this illustrative choice and instead use the uniformly sampled binary subtrees described in Appendix C.4.

D The normalized co-occurrence matrix $M^\star$

D.1 Definition, relation to PMI, and connection to word embeddings

This appendix gives additional background on the normalized co-occurrence matrix $M^\star$ used throughout the paper. The starting point is the classical connection between word embedding algorithms and co-occurrence matrix factorization. Levy and Goldberg [16] showed that the skip-gram with negative sampling objective implicitly targets a shifted pointwise mutual information matrix: at optimum, word-context inner products approximate

$$\mathrm{PMI}_{ij} - \log k, \tag{83}$$

where k is the number of negative samples and

$$\mathrm{PMI}_{ij} := \log \frac{P_{ij}}{P_i P_j}. \tag{84}$$

Thus, in an idealized low-rank symmetric matrix-factorization view, the leading spectral modes of the PMI matrix determine the geometry of the learned embeddings. This is the sense in which word embedding geometry can be studied through the eigenspaces of a co-occurrence-derived target matrix.

In the main text we use a closely related normalized co-occurrence matrix, $M^\star$, which arises in the symmetric-embedding, quadratic-loss approximation to skip-gram analyzed by Karkada et al. [22]. Let P_{ij} denote the probability that words i and j co-occur within the chosen context window, and let P_i and P_j denote the corresponding unigram probabilities. We use the definition of $M^\star$ in eq. (1). Karkada et al. show that, under this quadratic symmetric formulation, training can be viewed as approximately solving the matrix factorization problem

$$\widehat{W} = \arg\min_W \left\| WW^\top - M^\star \right\|_F^2. \tag{85}$$

Consequently, when $M^\star$ is positive semidefinite and the embedding dimension is large enough, the learned Gram matrix $WW^\top$ recovers $M^\star$; at finite dimension it recovers its leading positive spectral modes. Thus $M^\star$, like PMI, is a co-occurrence-derived matrix target for word embedding algorithms, but it corresponds to a slightly different objective and has useful normalization properties.

To see its relation to PMI explicitly, write the co-occurrence ratio as

$$R_{ij} := \frac{P_{ij}}{P_i P_j}. \tag{86}$$

Then

$$M^\star_{ij} = \frac{2(R_{ij} - 1)}{R_{ij} + 1}. \tag{87}$$

Thus $M^\star_{ij}$ is a monotone transform of the same pairwise co-occurrence signal measured by PMI. If $x_{ij} := \log R_{ij} = \mathrm{PMI}_{ij}$, then $R_{ij} = e^{x_{ij}}$, so

$$M^\star_{ij} = 2\frac{e^{x_{ij}} - 1}{e^{x_{ij}} + 1} = 2\tanh\left(\frac{x_{ij}}{2}\right) = 2\tanh\left(\frac{\mathrm{PMI}_{ij}}{2}\right). \tag{88}$$

Consequently, when R_{ij} is close to 1, or equivalently when PMI_{ij} is close to zero,

$$M^{\star}_{ij} \approx \mathrm{PMI}_{ij} \,. \tag{89}$$

The advantage of $M^{\star}$ is that it preserves the ordering of pairwise co-occurrence strength while remaining bounded:

$$-2 \leq M^{\star}_{ij} \leq 2. \tag{90}$$

It is therefore a bounded, symmetric normalization of the PMI signal. In the experiments, we work with $M^{\star}$ because it is the matrix target associated with the symmetric quadratic embedding objective in eq. (85), while retaining the same local interpretation as PMI as a normalized measure of excess co-occurrence.

D.2 Positive spectral component and Euclidean embedding geometry

The main text presents the cleanest version of the theory in the idealized case where $M^{\star}$ is positive semidefinite. In that setting, a Euclidean embedding Gram matrix can be identified directly with $M^{\star}$, and the hierarchy-aligned spectral theory applies to $M^{\star}$ itself. Empirically, however, $M^{\star}$ need not be positive semidefinite. This matters because a single Euclidean embedding space can only realize positive semidefinite Gram matrices.

Let

$$M^{\star} = U \Lambda U^{\top} \tag{91}$$

be the eigendecomposition of $M^{\star}$. Define the positive and negative spectral components

$$M^{+} := U \Lambda_{+} U^{\top}, \qquad M^{-} := U \Lambda_{-} U^{\top}, \tag{92}$$

where

$$(\Lambda_{+})_{aa} := \max(\Lambda_{aa}, 0), \qquad (\Lambda_{-})_{aa} := \max(-\Lambda_{aa}, 0). \tag{93}$$

Then

$$M^{\star} = M^{+} - M^{-}, \tag{94}$$

with both M^{+} and M^{-} positive semidefinite.

The positive component M^{+} is the part of $M^{\star}$ that can be represented as a Euclidean Gram matrix. If all positive eigenmodes are retained, the corresponding Gram matrix is M^{+}. If only the top r positive eigenmodes are retained, the realized rank-r Gram matrix is

$$M^{+}_{r} = U_r \Lambda_r U_r^{\top} = W_r W_r^{\top}, \qquad W_r := U_r \Lambda_r^{1/2}, \tag{95}$$

where Λ_r contains the top r positive eigenvalues and U_r contains the corresponding eigenvectors. This is the finite-dimensional Euclidean geometry recovered by a single-space spectral embedding.

This motivates the following softened version of the distance-kernel assumption used in the main text.
Assumption D.1 (PSD hierarchical distance structure)**.** Let S be a selected set of WordNet nodes, and let $\mathrm{dist}(i, j)$ denote their tree distance. The positive spectral component of the normalized co-occurrence matrix satisfies

$$M^{+}_{ij} = C^{+}(\mathrm{dist}(i, j)), \qquad i, j \in S, \tag{96}$$

for some positive decreasing function C^{+}.

This assumption is weaker than requiring $M^{\star}$ itself to be both positive semidefinite and distance-dependent. The idealized assumption in the main text corresponds to the special case $M^{-} = 0$, so that $M^{\star} = M^{+}$. More generally, $M^{\star}$ may contain negative spectral components, but the single-space Euclidean embedding geometry is governed by M^{+}. Therefore, the block-decomposition and hierarchy-aligned eigenspace arguments apply directly to M^{+} whenever Assumption D.1 holds.

Empirically, this softened assumption is supported by Figure 7. We restrict $M^{\star}$ to the eligible lemma set, symmetrize the restricted matrix, compute its leading positive eigenmodes, and construct low-rank PSD truncations M^{+}_{r} for several ranks r. For each rank, we estimate the mean entry of M^{+}_{r} as a function of WordNet distance by the same distance-bin sampling procedure used in Figure 2. Across ranks, M^{+}_{r} decays monotonically with semantic distance, showing that the hierarchical distance signal is present in the Gram-realizable part of the co-occurrence matrix.

In the main text, we fit the exponential distance kernel using the raw pairwise statistic $M^{\star}$, because $M^{\star}$ is computed directly from pairwise co-occurrence and unigram probabilities. By contrast, M^{+} is defined through a global spectral transformation of the full matrix. Figure 7 shows that this simplification does not drive the phenomenon: the same distance-decay structure appears in the PSD component that determines Euclidean embedding geometry.

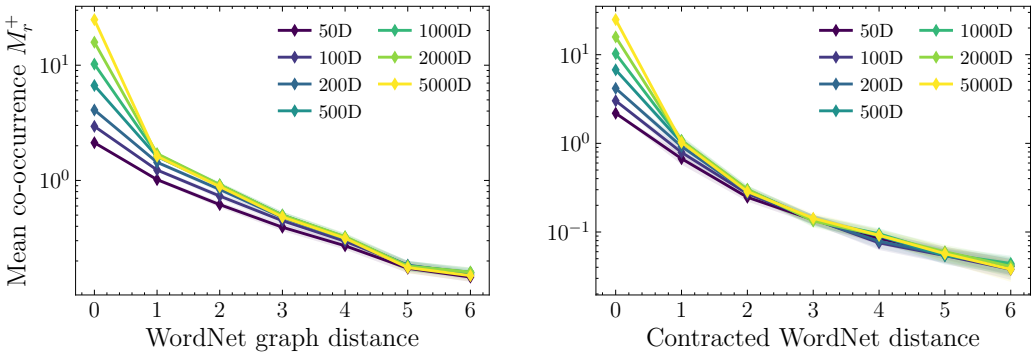

Figure 7: **Distance decay in low-rank PSD truncations of $M^\star$.** We restrict $M^\star$ to the eligible lemma set, retain the leading r positive eigenmodes, and form rank-r PSD Gram matrices $M_r^+ = U_r \Lambda_r U_r^\top$. Curves show the mean entry of M_r^+ as a function of semantic distance in the original WordNet graph (left) and in the contracted arborescence used in our experiments (right). The decay persists across ranks and distance constructions, supporting the PSD hierarchical distance structure in Assumption D.1.

E Additional robustness analyses

This appendix collects additional checks on the empirical assumptions and alignment results used in the main text. We begin with a diagnostic for the basic distance-kernel approximation $M^\star_{ij} \approx f(D_{ij})$. A possible concern is that the apparent decay with induced tree distance D_{ij} could instead be explained by a coarser topological variable, such as the depth of the lowest common ancestor of i and j. In that case, D_{ij} would not be the right one-dimensional summary of co-occurrence structure: pairs with small distance would appear to co-occur more often only because they tend to share deeper, more specific ancestors.

Figure 8 addresses this concern by conditioning on lowest-common-ancestor depth. Importantly, this diagnostic is computed over the same distribution of induced tree geometries used in the alignment experiments, rather than over uniformly random pairs from WordNet. For each sampled complete binary tree under a candidate root, we evaluate $M^\star_{ij}$ for every unordered pair of nodes in the sampled tree, including diagonal pairs, and assign each pair its induced tree distance D_{ij} and induced $\mathrm{depthLCA}_{ij}$. We then bin pairs jointly by $(D_{ij}, \mathrm{depthLCA}_{ij})$, compute root-level bin means after aggregating over that root's sampled trees, and average these root-level means across roots. Within each fixed-$\mathrm{depthLCA}$ stratum, the mean co-occurrence statistic still decreases with D_{ij}. Thus the decay captured by $f(D_{ij})$ is not merely a consequence of comparing pairs whose common ancestor lies at different depths.

Having checked this structural assumption, we next ask whether the reported eigenspace-alignment signal depends on other modeling or representation choices. The remaining robustness checks vary the null baseline, the parametric form of the distance kernel, the sampled tree depth, and the LLM representation used to form the empirical Gram matrix.

We also recompute the co-occurrence kernel using context widths $W \in \{1, 2, 4, 8, 16, 32\}$ (Figure 9). The decay with WordNet distance persists under both original and contracted tree distances. These windows are counted in the filtered noun stream, so $W = 1$ here does not mean adjacent tokens in the original corpus.

E.1 Additional baseline

Before varying kernel form, tree depth, or representation construction, we first show that the main eigenspace-alignment result remains above a stricter within-tree shuffle control. This control is stricter than the global shuffled-label baseline because it preserves the exact empirical vectors appearing in each sampled tree and only destroys their assignment to tree positions. This comparison is shown in Figure 10.

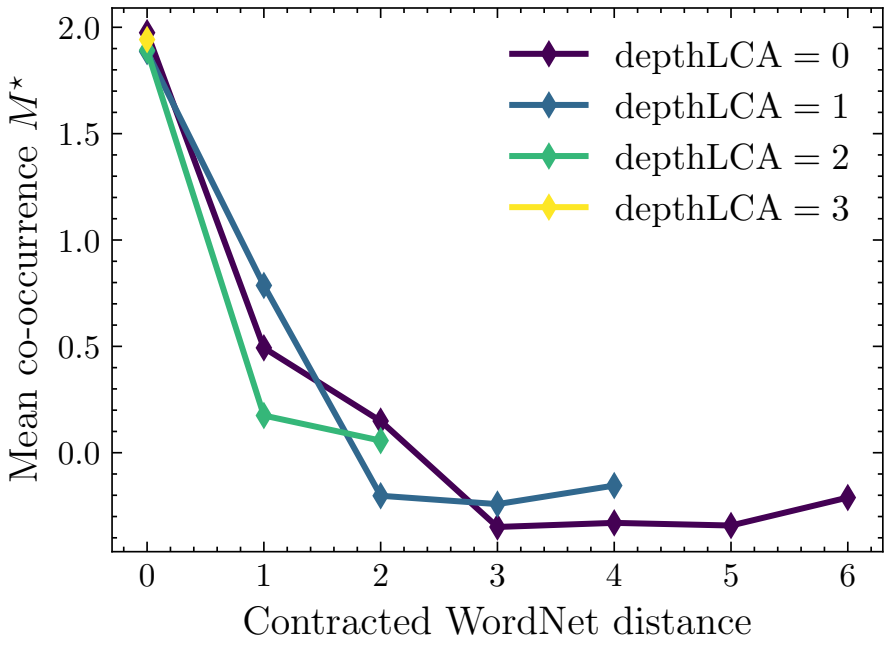

Figure 8: **Distance decay persists after conditioning on lowest-common-ancestor depth.** For each sampled complete binary tree, we evaluate $M^\star_{ij}$ for every unordered node pair (i, j), including diagonal pairs, and assign each pair its induced tree distance D_{ij} and lowest-common-ancestor depth $\mathrm{depthLCA}_{ij}$. Pairs are then grouped by $(D_{ij}, \mathrm{depthLCA}_{ij})$, and bin means are computed after aggregating over sampled trees for each root and then averaged across roots. Colors indicate fixed $\mathrm{depthLCA}$ strata. Within each stratum, the mean co-occurrence statistic decreases with D_{ij}, showing that the observed decay is not explained solely by common-ancestor depth and supporting the use of a distance-based model $f(D_{ij})$.

E.2 Kernel form

We repeated the analysis using an alternative kernel form for the theoretical tree covariance. In addition to the exponential kernel used in the main analysis, we fit a shifted power-law kernel of the form

$$f(d) = \alpha(1 + d)^{-\beta}.$$

The resulting alignment curves remain above both the global and within-tree baselines, indicating that the qualitative effect is not specific to the exponential decay assumption. This control is shown in Figure 11.

E.3 Sampled tree depth

Second, we repeated the root-sweep analysis with tree depth $L = 2$, corresponding to binary trees with $2^{L+1} - 1 = 7$ nodes rather than the 15-node trees used when $L = 3$. This smaller tree size substantially increases the number of eligible roots, because fewer descendants are required to construct a complete sampled binary subtree. Despite this larger and less restrictive root set, the alignment area remains elevated above both baselines for many roots. Thus, the effect is not driven by a small subset of roots that happen to support larger complete trees. This control is shown in Figure 12.

For deeper trees, Figure 13 compares Gemma-2B at $L = 3, 4, 5$. The eligible-root counts fall to $21, 5, 2$, respectively, and the deeper roots are partly nested rather than independent replications. The reported mean alignment remains above both shuffle controls for every eligible deeper root.

E.4 Alternative representations

Third, we evaluated alternative representations beyond the whitened Gemma embedding space used in the main analysis. Specifically, we repeated the alignment analysis using Gemma unembeddings,

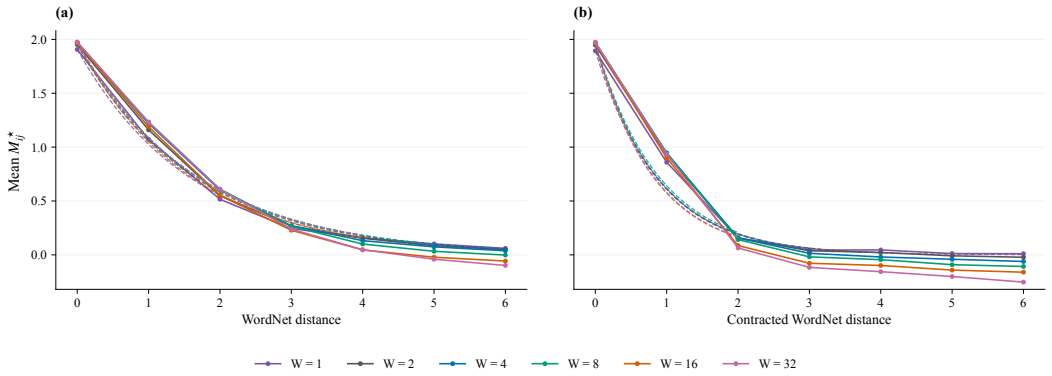

Figure 9: **Context-window robustness of co-occurrence decay.** Mean $M^\star_{ij}$ entries are plotted against **(a)** original WordNet distance and **(b)** contracted WordNet distance for $W \in \{1, 2, 4, 8, 16, 32\}$, measured in the filtered noun stream. Solid curves connect distance-bin means for $d = 0, \ldots, 6$, including self-pairs at $d = 0$. Dashed curves show $\alpha e^{-\beta d}$ fits to positive bins. The sweep shows how the measured decay changes with context width under each distance definition.

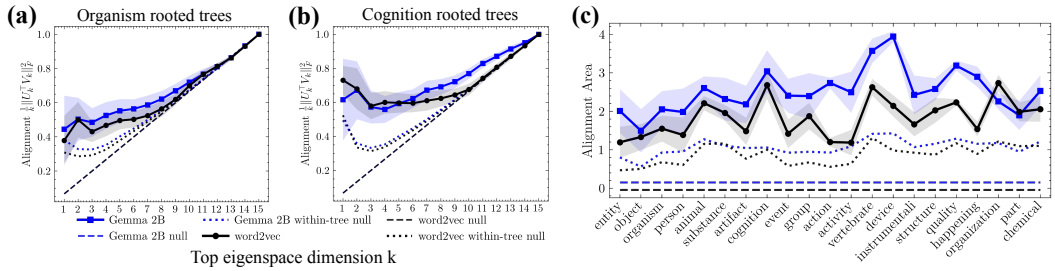

Figure 10: **Recreation of the main eigenspace-alignment experiment with an additional within-tree shuffle baseline.** We repeat the top-k eigenspace-alignment analysis from Figure 4 while displaying both null baselines. Solid curves show the empirical alignment between the theoretical tree-kernel Gram matrix and the representation Gram matrix. The global shuffled-label baseline randomly permutes the mapping from synsets to empirical vectors across the full eligible vocabulary, preserving the global vector distribution while destroying the semantic correspondence between WordNet nodes and representation vectors. The within-tree shuffle baseline is computed separately for each sampled tree: it preserves the tree, the multiset of empirical vectors in that tree, and the empirical Gram spectrum, but randomly permutes those vectors across the tree positions before recomputing alignment to the same theoretical Gram matrix. Thus, the within-tree baseline controls for unordered vector content within each sampled tree, whereas the global shuffle controls for vocabulary-level label correspondence. The empirical alignment remains above both baselines, showing that the effect is not explained solely by the global distribution of vectors or by the unordered collection of vectors appearing in a sampled tree. Shaded regions indicate one standard deviation across sampled trees.

Gemma internal residual-stream activations, and Llama internal residual-stream activations. In each case, we subtracted the global mean vector from the representation before computing alignment.

For the internal activation controls, we used the exact middle transformer layer in each model, chosen as $\lfloor n_{\text{layers}}/2 \rfloor$, and used the raw title-cased best lemma as the input string, e.g. "Animal", "Cat", or "Sea Turtle". This choice is intentionally simple and model-agnostic: it avoids tuning the layer index and avoids adding sentence-level context that could introduce additional uncontrolled structure.

All representation variants show alignment above the null baselines, although the alignment is much weaker for the internal representations. These representation controls are shown in Figure 14.

We further repeat the output-space alignment analysis for Llama-3.2-1B, Qwen2.5-1.5B, and Pythia-1.4B, and evaluate Pythia's untied input embeddings (Figure 15). We apply the same token-matching and whitening procedure separately to each model, matching its eligible lemmas against the Gemma reference rather than requiring a token to occur in every model. The different root counts therefore

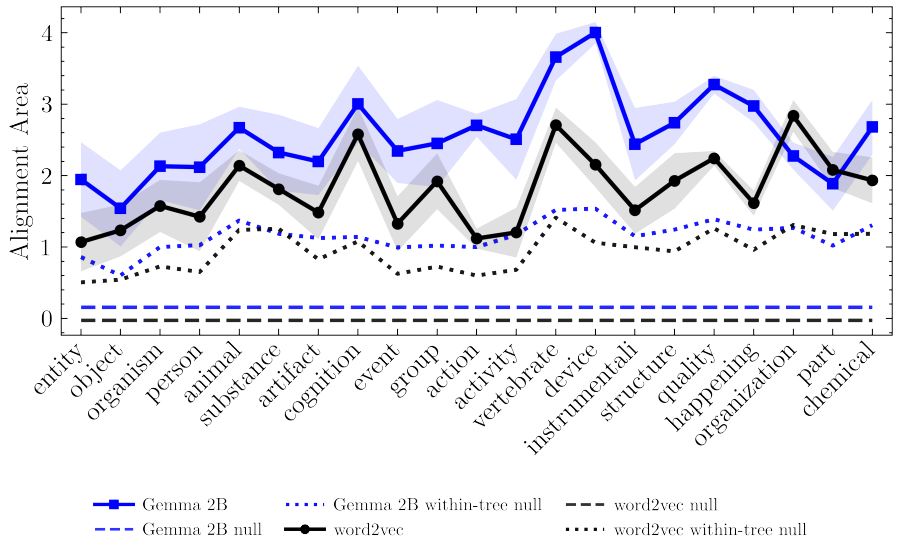

Figure 11: **Robustness to kernel parameterization.** We compare the alignment obtained using the exponential distance kernel from the main text with the alignment obtained using a shifted power-law kernel, $f(d) = 1.967 \cdot (1 + d)^{-2.153}$. The empirical alignment remains above both the global shuffle and within-tree shuffle baselines, indicating that the result is driven by distance-dependent hierarchical decay rather than by the specific exponential parameterization.

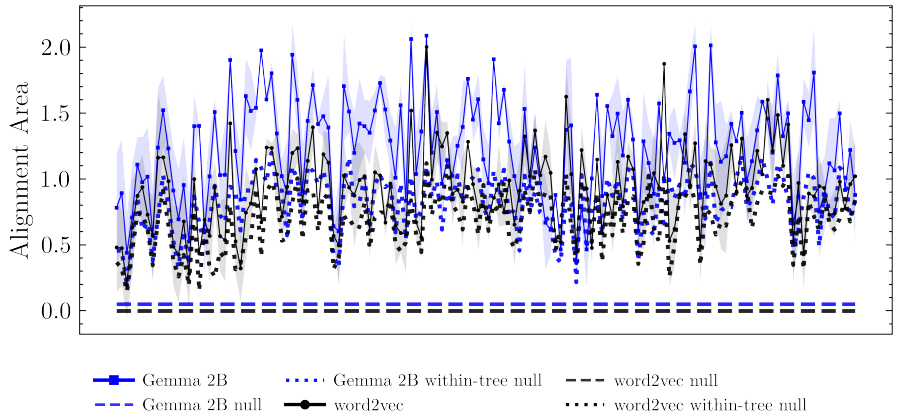

Figure 12: **Root-sweep robustness for $L = 2$.** We repeat the root-sweep analysis using complete binary trees of depth $L = 2$, corresponding to $2^{L+1} - 1 = 7$ nodes per sampled tree. This relaxation substantially increases the number of eligible roots relative to $L = 3$, from 21 to 144. The alignment area remains above the within-tree shuffle baseline for most roots.

reflect single-token eligibility, not omitted runs. All evaluated spaces have mean alignment above both shuffle controls for their eligible roots.

Together, these checks show that the alignment is robust to kernel form, sampled-tree depth, and representation choice. It persists across four LLM families, in both tied and untied embedding settings, and under the stricter within-tree shuffle control.

F Concept-vector diagnostics of hierarchical geometry

The main text shows that the fitted co-occurrence model reproduces the parent-child orthogonal-innovation diagnostic of Park et al. [13].

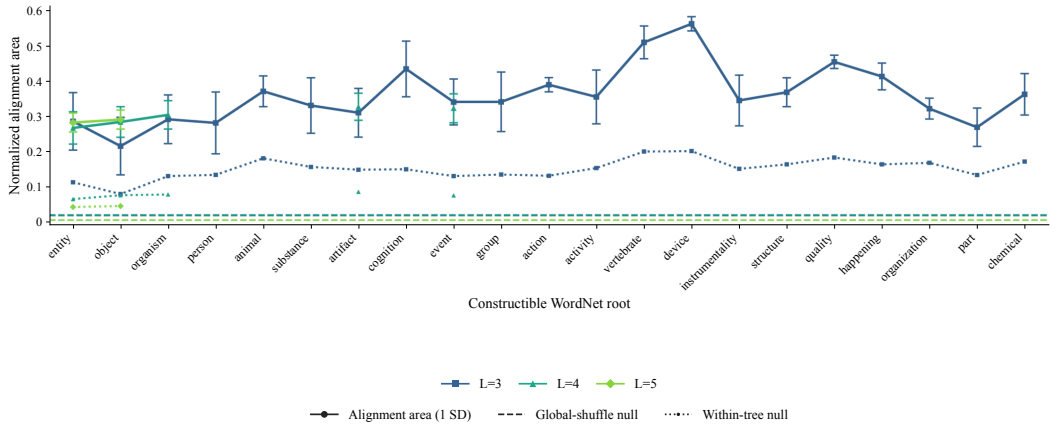

Figure 13: **Depth robustness of Gemma-2B unembedding alignment.** Complete binary trees of depths $L = 3, 4, 5$ contain 15, 31, and 63 nodes and are constructible under 21, 5, and 2 eligible roots, respectively. Alignment areas above the random-subspace null are divided by their maximum $2^L - 1$ to compare depths. Solid curves show means and bars indicate one SD across sampled trees. Dashed lines show mean global-shuffle nulls across roots; dotted curves show within-tree shuffle nulls. Deeper roots are partly nested and are not independent replications.

Here we give the implementation details of the concept-vector estimator.

For each WordNet synset w, let $Y(w)$ denote the set of tokens corresponding to w or one of its descendants.

Concept-vector estimator. Let $g(y) \in \mathbb{R}^d$ denote the representation vector associated with token y. In the LLM experiments, $g(y)$ is the centered and whitened Gemma unembedding vector. In the synthetic experiments, $g(y)$ is the corresponding theoretical co-occurrence embedding vector. For each concept w, we estimate the concept vector from a training subset $Y_{\mathrm{tr}}(w) \subset Y(w)$, formed by independently selecting 70% of descendant tokens for each synset, following Park et al.

Define

$$\mu_w := \mathbb{E}[g(y) \mid y \in Y_{\mathrm{tr}}(w)], \qquad \Sigma_w := \mathrm{Cov}(g(y) \mid y \in Y_{\mathrm{tr}}(w)).$$

Operationally, the estimator chooses a direction whose projected target-set mean is large relative to its projected target-set variance:

$$\max_{u \in \mathbb{R}^d} \frac{(u^\top \mu_w)^2}{u^\top \Sigma_w u}.$$

The direction $\tilde{g}_w$ and concept vector $\bar{\ell}_w$ are given in eq. (13), where $\Sigma_w^\dagger$ denotes the Moore-Penrose pseudoinverse. In the isotropic case $\Sigma_w \propto I$, this reduces to $\bar{\ell}_w = \mu_w$.

Since the representation dimension is larger than the number of descendant samples for many concepts, the empirical covariance is poorly conditioned. We therefore estimate Σ_w using the Ledoit-Wolf shrinkage estimator [29] before applying the pseudoinverse. This covariance regularization is used only for estimating the Park et al. concept vectors; it is distinct from the full-vocabulary whitening used for Gemma unembeddings in Appendix C.7. The diagnostics in Figure 5 additionally use $\Sigma_w + 10^{-3} I$. Three shuffled co-occurrence estimates have zero mean; their undefined normalized projections are omitted.

Synthetic co-occurrence embeddings. To test whether the same diagnostic can arise without an LLM-specific hierarchical mechanism, we repeat the concept-vector estimation procedure on theoretical co-occurrence embeddings. We construct a Gram matrix from the fitted distance kernel and take its top $d = 2048$ positive eigenmodes, as in eq. (2). Applying the same estimator to these embeddings yields the same qualitative parent-child orthogonality pattern as Gemma unembeddings: true parent-child innovations concentrate near zero, while shuffled-parent baselines are displaced.

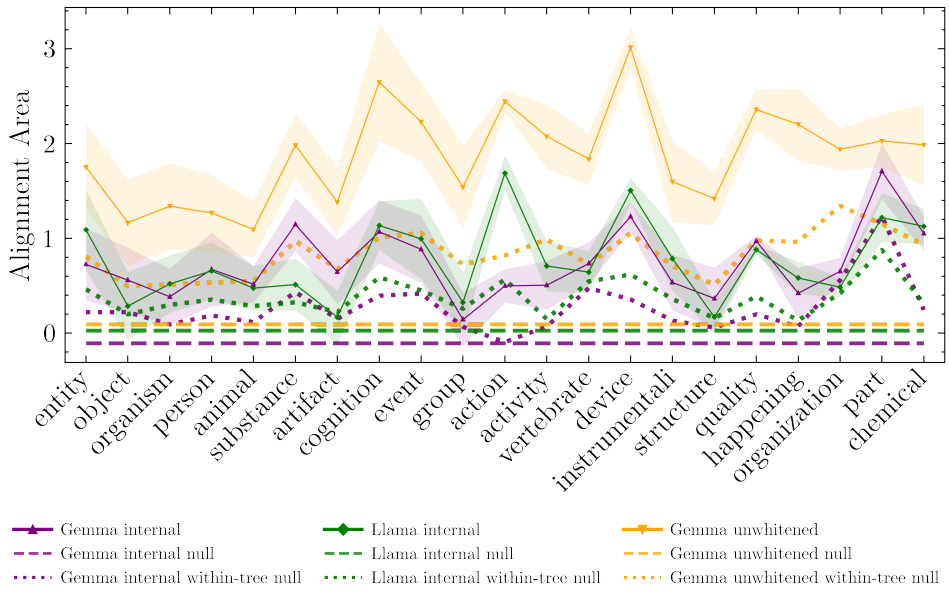

Figure 14: **Centered unwhitened and internal-activation representation controls.** We evaluate alignment using globally centered unwhitened Gemma embeddings, globally centered Gemma middle-layer residual-stream activations, and globally centered Llama middle-layer residual-stream activations. For each representation, the global mean vector is subtracted before computing the eigenspace alignment. Centering removes the constant anisotropic component that can otherwise increase the within-tree shuffle baseline. The centered representations remain substantially above both the global shuffle and within-tree shuffle baselines, showing that the hierarchical alignment is not explained by a shared global mode.

G Corpus statistics and LLM output geometry

The architectural link. An LLM with an affine softmax output head produces its next-token log-probabilities from inner products between a context representation and the unembedding vectors. Writing these factors as H and E,

$$L := \log P_\theta = HE^\top + \mathbf{1}b^\top - z\mathbf{1}^\top, \tag{97}$$

where b contains word biases and z the context-dependent log normalizers. After removing per-context and per-word offsets (softmax normalization and biases), its predictions therefore split exactly into a product of two factors, a context side and a word side,

$$\widetilde{L} := C_\mathcal{C} L C_\mathcal{V} = (C_\mathcal{C} H)(C_\mathcal{V} E)^\top = AB^\top. \tag{98}$$

Here the centering matrices subtract the row- and column-index means. This is the same structure as word2vec, whose two factors are the context and word embeddings. For LLMs it is the classical softmax-bottleneck property [25], and it holds by architecture, with no assumption about training.

The statistics being predicted contain the hierarchy. We work at the bigram level for simplicity. When the conditioning context is a single word, the corpus prediction target is described by one directed word-word co-occurrence matrix, with no averaging over longer contexts. Applying our coarse-to-fine alignment procedure to the relevant low-rank part of this corpus matrix reveals a strong hierarchy-splitting signal across 21 WordNet roots, already at the bigram level (Figure 16). The measured operator is double-centered empirical PMI with unobserved entries set to zero before centering; it is therefore a finite empirical surrogate, not the exact population log-conditional.

Two hypotheses complete the argument. (a) The trained model predicts bigram statistics well, so that its centered log-probability target approximately matches the corpus one. We do not claim direct verification of this hypothesis here.

(b) The two learned factors share the target's structure evenly (a balance property). For two-factor embedding models trained from small initialization, balance is supported by a conservation law of

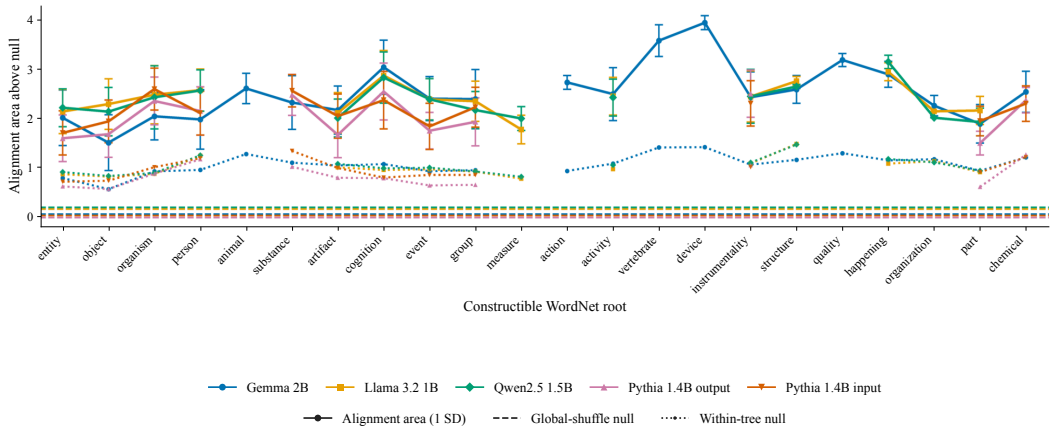

Figure 15: **Hierarchical alignment across language-model token representations.** Rootwise alignment areas are shown for Gemma-2B, Llama-3.2-1B, Qwen2.5-1.5B, and Pythia-1.4B output embeddings, together with Pythia input embeddings. Each representation uses its own eligible vocabulary and depth-3 trees: 21, 15, 15, 12, and 12 roots, respectively. Solid curves show mean alignment area above the random-subspace null; bars indicate one SD across sampled trees. Dashed lines show the mean global-shuffle null across roots; dotted curves show the within-tree shuffle null.

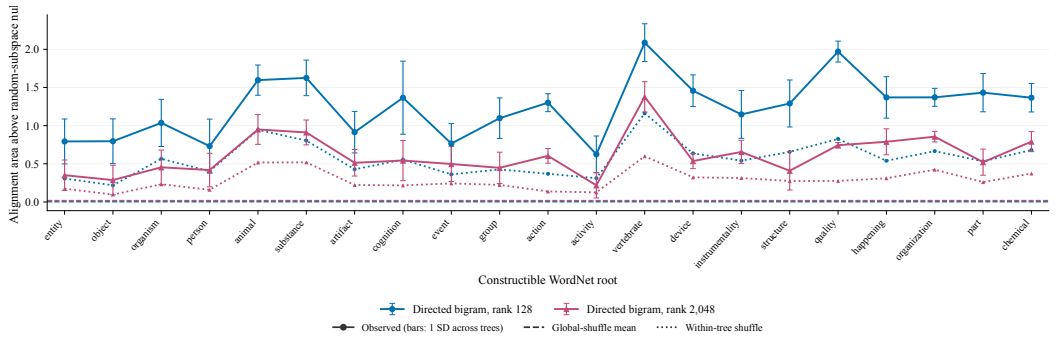

Figure 16: **Hierarchical alignment of directed-bigram PMI embeddings.** Left-singular embeddings $U_r\Sigma_r^{1/2}$ of the double-centered, zero-filled empirical PMI matrix for literal next-token bigrams are evaluated at ranks $r = 128$ and 2048. Both ranks use the same 21 roots and 5,000 depth-3 trees per root. Solid curves show mean alignment area above the random-subspace null; bars indicate one SD across sampled trees. Dashed lines show mean global-shuffle nulls across roots; dotted curves show within-tree shuffle nulls.

gradient flow [30, 31]. Nothing enforces it in an LLM, where the context side is computed by the deep network, so it is an idealization here. For an exact factorization, the balance condition

$$A^\top A = B^\top B \tag{99}$$

implies

$$AA^\top = (\widetilde{L}\widetilde{L}^\top)^{1/2}, \qquad BB^\top = (\widetilde{L}^\top\widetilde{L})^{1/2}. \tag{100}$$

Under (a) and (b), the corpus spectral structure transfers to both factors, the unembeddings and the context representations, through these identities. For a directed target, the relevant left and right singular structures must be checked separately.

Hierarchical alignment is observed in both learned factors. The unembeddings align across four models (Figure 15), and so do globally centered Gemma final-layer context representations after $[\mathrm{BOS}, j]$. For the latter, mean alignment area is 1.733 ± 0.102, versus -0.056 ± 0.055 under global shuffle and 0.800 ± 0.066 under within-tree shuffle (means $\pm$ SE across 21 roots). Direct support for (b) itself is, however, poor in our first probe with single-token contexts. Both bridging hypotheses remain open; bypassing hypothesis (b) should be a target of future work.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the paper's main claims: a co-occurrence-based spectral theory of hypernymy, hierarchical splitting geometry in word2vec embeddings, and corresponding empirical signatures in Gemma-2B unembeddings. These claims are addressed by the theoretical results and empirical tests in the main text.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper includes a dedicated Limitations section and states the main assumptions used in the theory and proofs, including the distance-kernel/co-occurrence model, idealized tree structure, and the role of perturbations around the noiseless setting. It also discusses the scope of the claims: the theory applies most directly to co-occurrence-driven word2vec-style embeddings, while explaining the same geometry in modern transformers requires additional assumptions about training dynamics and internal representation formation.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.

- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The main text states the assumptions for the theoretical results, and Appendix A provides formal proofs for the main spectral claims, including the distance-kernel reduction, hierarchy-aligned eigenstructure, coarse-to-fine ordering, exponential-kernel consequences, and perturbative stability statement. The appendix also records how the decomposition extends to symmetric non-binary trees.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper and appendix disclose the main experimental pipeline, including vocabulary filtering, token matching, polysemy filtering, WordNet arborescence construction, binary subtree sampling, co-occurrence estimation, word2vec-style spectral construction, LLM representation processing, baselines, and alignment metrics. The supplemental code submitted with the paper reproduces the figure generation and data-processing pipeline.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may

be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The submission includes code that reproduces all figures and data processing. The experiments use public resources, including WordNet, the November 2023 English Wikipedia dump, pretrained Gemma/Llama models, and standard Python libraries, with data access and preprocessing described in the appendix.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies the eligible vocabulary construction, WordNet arborescence construction, binary tree sampling procedure, word2vec spectral construction, Gemma unembedding preprocessing, and the top-k eigenspace alignment metric used in the experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The empirical results are aggregated across sampled WordNet subtrees and compared against shuffled baselines, with error bands reported as standard deviations or standard errors in the relevant figures.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Appendix C.8 reports the compute environment and runtimes.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

Answer: [Yes]

Justification: The research analyzes existing lexical resources, co-occurrence statistics, and pretrained representation vectors, and does not involve human subjects, private data, or deployment of a system.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [N/A]

Justification: This is foundational work on the geometry of language representations and does not introduce a deployed system, dataset, model, or capability with a direct path to societal impact. We therefore do not identify paper-specific broader impacts beyond the general scientific value of improving understanding of representation geometry.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate

to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release a high-risk model, scraped dataset, or other artifact requiring safeguards against misuse.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Appendix C.9 reports the existing assets used, including WordNet, the November 2023 English Wikipedia dump, pretrained Gemma/Llama model resources, TransformerLens, and standard scientific Python libraries. It states that these assets are cited, that the accompanying anonymized code/supplement lists versions, sources, and available license or terms-of-use information, and that the work does not redistribute pretrained model weights or the Wikipedia dump.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not introduce or release a new dataset, model, benchmark, or standalone software asset.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or human-subject experiments.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

Justification: The paper does not involve human-subject research, so IRB approval or equivalent review is not applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs are analyzed as objects of study, but LLM use is not an important, original, or non-standard component of the core research methodology. Any use of LLMs only for writing, editing, or formatting would not affect the scientific methodology.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.