GROUP-INVARIANT STATISTICS DETERMINE EMBEDDING GEOMETRY: HARMONIC ANALYSIS OF REPRESENTATIONS FROM BACH TO THE NIGHT SKY

Anonymous authors
Paper under double-blind review

ABSTRACT

The representations that language models learn for concepts such as months, weekdays, and places display consistent geometric structure: circles and saddle-shaped "Pringle" manifolds. Recent work traced these structures to *translation symmetry* in word co-occurrence statistics, deriving the observed Fourier geometry when co-occurrence depends only on distance on an abelian lattice of concepts. We demonstrate that more general notions of symmetry lead to equally structured predictions. Considering symmetries defined by arbitrary finite groups, compact groups, and homogeneous spaces, we prove that whenever the co-occurrence statistics of a word family are invariant under a group G, the learned word embeddings consist of matrix elements of the irreducible representations (irreps) of G. Circles and Pringles arise when G is cyclic, in which case the irreps are Fourier modes. We verify the irrep structure in three experimental settings. (i) The cyclic group $\mathbb{Z}_{12}$: for the months of the year we recover the known circular geometry. (ii) A dihedral group acting on the major and minor triads: we unify two classical observations—that transposition and chord inversion form a group (T/I) acting on chords (music theory), which *implies* that the well-known "circle of fifths" emerges in learned chord embeddings (machine learning). (iii) We explain and reproduce a recently discovered spherical representation of celestial objects in large language models (LLMs) as a spherical-harmonic embedding derived from our theory. Our results demonstrate that the geometry of learned representations is often a consequence of the statistical symmetry of underlying data.

1 INTRODUCTION

Large language models (LLMs) organize concepts geometrically in their activations. Months of the year lie on circles (Engels et al., 2025), latitude and longitude of cities can be linearly decoded from LLM activations (Gurnee & Tegmark, 2024), and celestial objects reside on a sphere inside a LLM's residual stream (Berdnikov & Liokumovich, 2026). A compelling explanation for the ubiquitous circular geometry was given by Karkada et al. (2026): when the co-occurrence of words in text is only a function of their distance on some "cyclic latent lattice", the optimal embedding under a quadratic word-embedding model (Karkada et al., 2025) consists of Fourier modes which in turn generates circles. For example, they found that the measured co-occurrence of January and February is about the same as November and December in general text, which explains why language models embed the 12 months on a circle.

Contributions. Cyclic symmetry is just one example of structure that can exist in data (Figure 1). Here, we ask and answer the general question: *what does an arbitrary symmetry of co-occurrence statistics imply for embedding geometry?* The answer is a complete picture of the components used to construct word embedding geometry, assembled from group harmonic analysis. Specifically, our contributions are:

- **A general theory of word embedding geometry under group symmetry**: if the words form a torsor for a group G (*i.e.*, if we can form a bijection between the words and group elements,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

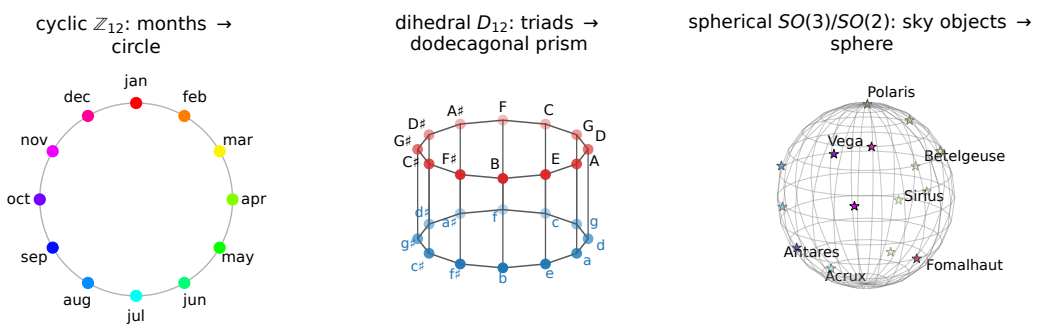


Figure 1: **Months, musical chords, and celestial objects exhibit distinct latent geometries**, dictated by the symmetry of their word co-occurrence statistics in text corpora. *Left*: month statistics are approximately invariant under cyclic shift ($\mathbb{Z}_{12}$), which places them on a circle. *Middle*: musical chord statistics are approximately invariant under transposition and inversion (D_{12}), which places them on a dodecagonal prism. *Right*: celestial object statistics are approximately invariant under 3D rotations ($SO(3)$), which places them on a sphere.

Theorem 4.1) or a homogeneous space G/K (Theorem 4.2), the embedding's entries are matrix elements of the irreps of G that can be expressed analytically.

- **A non-Abelian validation of the theory** *that unifies existing, but disconnected, observations from music theory*: we use our theory to reconcile two classical observations about music: the existence of transposition symmetry (T/I) in music theory (Crans et al., 2009), and the so-called circle of fifths that emerges in learned chord embeddings (Lahnala et al., 2021). Specifically, we show that Bach's chorales have approximately T/I-invariant chord statistics (invariance captures $R^2 = 0.822$ of the co-occurrence structure), and that this statistical symmetry *forces* the observed geometry. The empirical circle of fifths matches our theoretical predictions at a per-chord cosine of 0.98. We observe the same structure in ChordBERT, a masked language model for symbolic chord sequences.
- **A continuous validation of the theory** *that challenges existing results on LLM activations*: We reanalyze the sky-sphere LLM activations of Berdnikov & Liokumovich (2026) and show that the representation of celestial objects is not a sphere, as the authors initially reported, but is actually *spherical-harmonics*. Specifically the eigenmodes of the LLM activations are degree-ordered spherical-harmonic blocks (ℓ=1, ℓ=2, ℓ=3), in five of the seven models examined. We reproduce these harmonic results using four public-domain astronomy books (131k tokens, selected by a "star guide" criterion) and examining classical word embeddings. We show that co-occurrence statistics place the objects on the true celestial sphere at a statistically significant per-object cosine 0.55 (median 0.68), which is within the range of accuracy that an idealized spherical model predicts for a sample of this size. The same symmetry fingerprint, present in a 131k-token field guide and in a 123B-parameter LLM's activations, sharpens in the model, as expected if training supplies vastly more such text.

Together, these results show that word embeddings admit analytical expressions determined by the symmetry of their co-occurrence statistics in quadratic embedding models, and suggest that LLM activations might be described by the same rule.

2 Related Work

We review work linking word statistics to their embeddings, focusing on symmetry-driven statistics, and survey the role of group theory in interpretability research.

From word statistics to their embeddings. Early work has linked the structure of word embeddings to their co-occurrence statistics, but did not connect that structure to any symmetry of these statistics. Skip-gram with negative sampling implicitly factorizes a shifted Pointwise Mutual Information (PMI) matrix (Levy & Goldberg, 2014). GloVe (Global Vectors for Word Representation)

fits log co-occurrence directly (Pennington et al., 2014). Karkada et al. (2025) derive a quadratic word-embedding model in which trained embeddings realize a low-rank factorization of the normalized statistics $M^\star$ (Eq. 1), the premise we adopt (Assumption A.1).

Word statistics described by a symmetric group and their embeddings. Karkada et al. (2026) prove that translation-invariant statistics on $\mathbb{Z}_L^D$ yield sinusoidal embeddings and validate this for periodic concepts; our Appendix B.6 recovers their Proposition 1 as the cyclic case. Engels et al. (2025) found the day/month circles empirically. Gurnee & Tegmark (2024) linearly decoded measurements of space and time from activations. Berdnikov & Liokumovich (2026) discovered the sky sphere, asking explicitly whether the statistics-based paradigm explains it, which our Section 7 addresses. The linear representation hypothesis and the geometry of categorical concepts (Park et al., 2024; 2025) concern complementary, non-manifold structure.

Group theory in interpretability. Beyond word embeddings, group theory has been successfully used to describe representations emerging across neural network architectures. Networks trained on group operations learn irrep-based algorithms: Fourier features for modular addition (Nanda et al., 2023) and general irreducible representations for finite groups (Chughtai et al., 2023; Marchetti et al., 2026). Those results concern models trained *on a group operation*. By contrast, our work concerns models trained on *natural corpora whose statistics happen to be group-invariant*, with the group acting on concepts rather than being the task.

3 Background on Group Theory

We briefly review the concepts from group theory that our results are built from. We refer to Diaconis (1988); Folland (2016) for details.

Groups. A *group* is a set G equipped with an associative binary operation $(g, h) \mapsto gh$, an identity element $1 \in G$ with $1g = g1 = g$, and an inverse $g^{-1} \in G$ for each $g \in G$ with $g^{-1}g = gg^{-1} = 1$. A group is *Abelian* if $gh = hg$ for all $g, h \in G$, and non-Abelian otherwise. The cyclic group $\mathbb{Z}_n$ (integers modulo n under addition) is Abelian. The dihedral group D_n of rotations and reflections of a regular n-gon is not, since rotating and reflecting do not generally commute. $\mathbb{Z}_n$ and D_n are finite, but groups can also be continuous, i.e., the group $SO(3)$ of all possible 3D rotations. We will use $\mathbb{Z}_{12}$, D_{12} and $SO(3)$ to characterize the word embeddings of the months of the year, musical chords, and celestial objects, respectively (Figure 1).

Group actions. A (left) *action* of G on a set L is a map $(g, x) \mapsto g \cdot x$ with $1 \cdot x = x$ and $g \cdot (h \cdot x) = (gh) \cdot x \; \forall g \in G, \forall x \in L$. In our theory, L is a set of words: for example, the months of the year, on which the cyclic group $G = C_{12}$ acts. The action is *transitive* if every $x \in L$ can be moved to every $y \in L$ by some $g \in G$. The action is *free* if $g \cdot x = x$ implies $g = 1$, i.e, only the identity can fix a $x \in L$. When the action is both free and transitive, L is a *G-torsor*: choosing any base point $x_0 \in L$ yields a bijection $G \to L$, $g \mapsto g \cdot x_0$. For example, the action of $G = C_{12}$ on the months is free and transitive, such that the months themselves can be thought of as elements of C_{12} after choosing a base point (Figure 1 left). When the action is transitive but not free, each point x has a nontrivial *stabilizer* $K = \{g \in G : g \cdot x = x\}$, and L is identified with the *homogeneous space* G/K of left cosets gK. The sphere $L = S^2 = SO(3)/SO(2)$ of the celestial objects is our example of a homogeneous space (Figure 1 right): $SO(3)$ acts transitively on S^2, the stabilizer of a point is the group $SO(2)$ of rotations about the axis through the point.

Irreducible representations. Our theory shows that word embeddings can be computed analytically using irreducible representations of group, which we define now. A *unitary representation* of G of dimension d_ρ is a map $\rho : G \to U(d_\rho)$ into the unitary group such that $\rho(gh) = \rho(g)\rho(h)$ for all $g, h \in G$. It is *irreducible* (an *irrep*) if no proper nonzero subspace of $\mathbb{C}^{d_\rho}$ is mapped to itself by every $\rho(g)$. Any unitary representation decomposes into a direct sum of irreps. A finite group has finitely many irreps up to unitary equivalence. All irreps of Abelian groups (e.g., C_{12}) are one-dimensional: their $\rho(g)$ are complex scalars for any irrep ρ and group element g. Higher dimensional irreps may exist for non-Abelian groups (e.g., D_{12} and $SO(3)$): their $\rho(g)$ are generally complex matrices of size $d_\rho \times d_\rho$. Our theory will often refer to the entries of those matrices.

Lastly, irreps are used to define the *group Fourier transform* of any function f over a finite group G as $\hat{f}(\rho) = \sum_g f(g)\rho(g)^\dagger$. The sum becomes an integral if G is compact.

4 THEORY: FROM INVARIANT STATISTICS TO HARMONIC GEOMETRY

We state the framework and main results. Complete statements and proofs are in Appendices A–D.

Word embeddings. For words x, y in a n-word family, we write P_x, P_y their *unigram probabilities* –how often word x, resp. y, appears in text – and P_{xy} their *co-occurrence probability* –how often x, y appear together (formal definitions in Appendix A). We define the $n \times n$ *normalized co-occurrence matrix* element-wise by

$$M^\star_{xy} = \frac{P_{xy} - P_x P_y}{\frac{1}{2}(P_{xy} + P_x P_y)}, \tag{1}$$

and $PM^\star P$ its *doubly centered version*, where $P = \mathrm{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$ is the centering projector.

Recent work shows that the matrix $PM^\star P$ specifies which word embeddings will be learned from text corpora. Specifically, Karkada et al. (2025) established that the r leading PCA scores of a set of n `word2vec`-style embeddings in d dimensions $\overline{W} \in \mathbb{R}^{n\times d}$ can be approximated with the positive eigenpairs $\{(\lambda_i, v_i)\}_{i\le r}$ of $PM^\star P$, taken in decreasing order. Namely, up to an orthogonal transform and realification, the projection of $\overline{W}_x \in \mathbb{R}^d$ along the first r principal components of $\overline{W}$ is approximately $\left(\sqrt{\lambda_i} v_i(x)\right)_{i\le r}$. We use this characterization of empirical word embeddings throughout our paper, as also done in Karkada et al. (2026).

Assumption: Statistical invariance. Our starting point is the assumption that a given n-word family carries a symmetry, as in the real-world examples of Figure 1. Specifically, we assume that a symmetry group G acts on a word set L and that the co-occurrence matrix verifies:

$$M^\star_{g\cdot x,\, g\cdot y} = M^\star_{x,y}, \quad \forall g \in G,\ \forall x, y \in L. \tag{2}$$

We say that the word statistics are *invariant* to the group G. If the words L form a *G-torsor*, fixing any word as the group identity[1] creates a bijection between the word set and the group: $L \simeq G$. Abusing notation, Assumption 2 implies that $M^\star_{g,h} = C(g^{-1}h)$ for a kernel $C : G \to \mathbb{R}$ with $C(g^{-1}) = C(g)\ \forall g, h \in G$, the latter property forced by the symmetry of $M^\star$. If instead the action is transitive with stabilizer K, the word set is the homogeneous space $L \simeq G/K$ and C is bi-K-invariant.

Theory: Structure of word embeddings Using the characterization of word embeddings and the statistical invariance assumption above, we derive two theorems giving the analytical expression of word embeddings under group symmetry. Their formal versions are in Thm. B.3 and Thm. D.1.

Theorem 4.1 (Finite group, informal)**.** *Let G be a finite group of order n, and let the word set L be a G-torsor satisfying Assumption 2. For each nontrivial real irrep ρ of G, let $\widehat{C}(\rho)$ be the group Fourier coefficient at ρ of the kernel C associated with $M^\star$, with eigenvalues $\lambda_\rho^{(1)}, \dots, \lambda_\rho^{(d_\rho)}$. Then, denoting pc_r the top-r score map $\mathrm{pc}_r : \mathbb{R}^d \to \mathbb{R}^r$, the word embedding W_g of any $g \in G$ verifies:*

$$\mathrm{pc}_r(W_g) = \left[\sqrt{\frac{\lambda_\rho^{(s)}\, d_\rho}{n}}\ \rho(g)_{as}\right]_{(\rho,s,a)}, \tag{3}$$

for some choices of (ρ, s, a) sorted in decreasing order of $\lambda_\rho^{(s)} > 0$, and up to orthogonal transform and realificaction.

A version of Theorem 4.1 also holds for compact groups, see Thm. C.5. These theorems demonstrate that word embeddings of a G-torsor word set appear like the matrix elements of the irreps of G. The text source determines, through its kernel C, which columns of the irreps will be most pronounced

[1] The base word choice is inconsequential as shown in Lemma B.4.

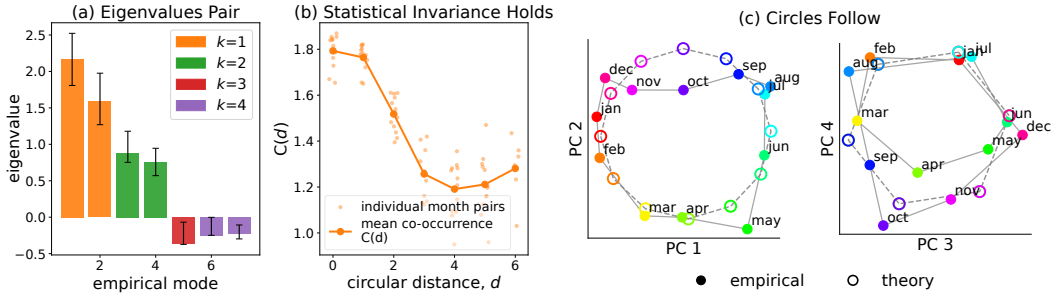


Figure 2: **Month statistics are $\mathbb{Z}_{12}$-invariant $\rightarrow$ embed as circles (known result, re-derived).** $M^\star$ is formed using month co-occurrences in the corpus `text8`, using month-sense disambiguation for "may." The approximate $\mathbb{Z}_{12}$-invariance of $M^\star$ causes circles to emerge in the months' word embeddings. See App. E for all details. (a) Spectrum of $PM^\star P$ (corpus bootstrap error bars, $2.5\% - 97.5\%$), wavenumber assigned by dominant Fourier mode. The first 4 eigenvalues split into approximate pairs with the same wavenumber, a hint that $\mathbb{Z}_{12}$-invariance could be present in the data. (b) Averaged kernel $C(d)$ of the individual month pair co-occurrences. The data follow the kernel rather than just averaging to it (average predicts raw data with $R^2 = 0.87$), asserting an approximate invariance to Z_{12}. (c) The two circular planes. Left: $k = 1$ plane with median $31°$/month. Right: $k = 2$ plane with median $66°$/month.

(largest $\lambda_\rho^{(s)}$). Furthermore, if the dimension prescribed to the word embeddings is too small, irrep elements with sufficiently small $\lambda_\rho^{(s)}$ are truncated.

Next, we consider homogeneous spaces. We save the general discussion for the Appendix and focus on the particular case of the sphere S^2 for clarity, which is the case relevant for our experiments.

Theorem 4.2 (Homogeneous space S^2)**.** *Let L be a finite word set in which each word x is assigned a distinct point $p_x \in S^2$, and satisfying Assumption 2. For each degree $\ell \geq 1$, let $c_\ell = \frac{1}{2}\int_{-1}^{1} C(t)\, P_\ell(t)\, dt$ for P_ℓ the degree-ℓ Legendre polynomial and C a kernel corresponding to $M^\star$. Then, denoting pc_r the top-r score map $\mathrm{pc}_r \colon \mathbb{R}^d \to \mathbb{R}^r$, the word embedding W_x of any $x \in L$ satisfies:*

$$\mathrm{pc}_r(W_x) = \left[\sqrt{c_\ell}\, Y_\ell^m(p_x)\right]_{(\ell,m)}, \tag{4}$$

for some choice of (ℓ, m), sorted in decreasing order of $c_\ell > 0$, where Y_ℓ^m denotes a real spherical harmonic of degree ℓ, and and up to orthogonal transform and realificaction. Since there are $2\ell+1$ spherical harmonics of degree ℓ, the embedding contains subspaces of dimension $2\ell + 1$.

Realification. Embeddings are real, while the eigenvectors above are in general complex. Hence, we must convert them to real coordinates, using a methodology that depends on the type of irrep (Proposition B.6). In our experiments, irreps can be realified (if needed) in a simple, formulaic way.

5 The Cyclic Case: Months and the Pringle

We test the theory on a known case: the cyclic group $G = \mathbb{Z}_{12}$ acting on the twelve months. We utilize the `text8` corpus (Wikipedia dump, 17M tokens, (Mahoney, 2011)). We recover the same results as (Karkada et al., 2026), but using group theory instead of their Fourier analysis.

Predictions All irreps of $\mathbb{Z}_{12}$ are 1-dimensional complex Fourier modes, indexed by a wavenumber $k \in \{0, \ldots, 11\}$. The theory predicts that the embedding of month $m \in \{0, \ldots, 11\}$ belongs to a direct product of circles of the form $\big(\cos(2\pi km/12), \sin(2\pi km/12)\big)$ with nonzero k. The radii of these circles is obtained through the discrete Fourier transform of the kernel $C(d)$, the average co-occurrence of months $d \in \{0, \ldots, 6\}$ months apart. Specifically, $\sqrt{\widehat{C}(d)}$ is the radius of the circle with $k = d$. These radii are the corpus supplied scalars for ranking the irreps.

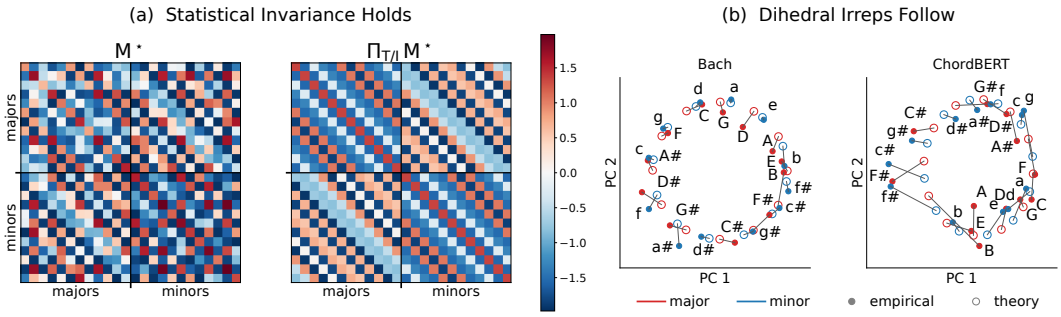


Figure 3: **Chord statistics are T/I-invariant $\rightarrow$ embed as dihedral irreps.** See Appendix F for all details. (a) Left: Empirical $M^\star$ over the 24 triads from 371 Bach chorales. Majors and minors are listed from C to B in chromatic order. Right: $\Pi_{T/I}M^\star$ is the projection of $M^\star$ onto matrices that are T/I-invariant. Importantly, $\Pi_{T/I}M^\star$ captures much of the structure of $M^\star$ ($R^2 = 0.822$), demonstrating an approximate T/I-invariance in $M^\star$. (b) Uppercase and lowercase refer to major and minor chords respectively. An orthogonal transform is used to align theoretical predictions with empirical embeddings in each case. Left: due to the statistical invariance in (a), a dihedral irrep, in the form of interleaved circles of majors and minors, emerges in the Bach chord embeddings. Our predictions obtain a per-chord cosine similarity of 0.98 with the real embeddings. Right: Our empirical embeddings are the average activation of each triad in the layer 3 (selected for high isotypic energy, Figure 7) residual stream of ChordBERT across all 371 chorales. The same dihedral structure emerges. We use the Gram matrix of these centered embeddings (normalized by the root mean square of its diagonal) as a proxy for $M^\star$. This gives us theoretical predictions with a per-chord cosine sim. of 0.91.

Results: The embeddings realize the cyclic irreps As we see in Figure 2 B, the averaged kernel is faithful to the individual months' co-occurrences. Although the statistics will never be perfectly invariant, they are close enough that our theoretical circular predictions line up nicely with the actual embeddings in Figure 2 C. The first 4 empirical PCs of the embeddings match the theoretical predictions at a per-month cosine of 0.96 (median 0.97), under one aligning orthogonal transform per plane, compared to a label-shuffle null of mean 0.35. The classic Pringle lives in the first 3 dimensions of these embeddings, a sinusoidal on top of a circular base.

6 The Nonabelian Case: Dihedral Chord Geometry in Bach

We test the theory on a novel case: the dihedral group $G = D_{12}$ acting on the 24 major and minor triads, which form a D_{12}-torsor under the transposition–inversion group $T/I \cong D_{12}$. We utilize the chord sequences from the 371 Bach chorales, `music21` Cuthbert & Ariza (2010): 24k chord tokens, full sonority vocabulary of 92 types with the 24 triads as the analyzed subsystem. See App. F for details.

Predictions. The 24 major and minor triads form a D_{12}-torsor under the transposition-inversion group $T/I \cong D_{12}$. Hence, Theorem 4.1 applies. D_{12} has four 1-dimensional embeddings and five 2-dimensional irreps $(E_1 \ldots, E_5)$. From foresight, none of the 1-dimensional irreps are strongly present in the empirical embeddings, so we will neglect them. We predict that the embeddings of the majors $m \in \{0, \ldots, 11\}$ are a direct product of circles of the form $\big(\cos(2\pi hm/12 + \phi), \sin(2\pi hm/12 + \phi)\big)$, while the embeddings of the minors $n \in \{0, \ldots, 11\}$ are instead a direct product of circles of the form $\big(\cos(2\pi hn/12 - \phi), \sin(2\pi hn/12 - \phi)\big)$, existing in the same set of planes, where ϕ is some constant (dependent on the corpus) and $h \in \{1, \ldots, 5\}$. Furthermore, we predict that these interleaved circles of majors and minors share the same radius, a corpus determined value computed as in Thm. 4.1.

Statistical assumption is verified: Chord co-occurrence statistics exhibit dihedral symmetry We form $M^\star$ with a 3-token window from our collected 24k chord tokens. We can find the projection of $M^\star$ onto matrices that are invariant with respect to some group G, $\Pi_G M^\star$, by the element-wise

computation $\Pi_G M^\star_{x,y} = \frac{1}{|G|}\sum_{g\in G} M^\star_{gx,gy}$, i.e., replacing entries that would be equivalent under invariance by their average. Transposition invariance only, $\Pi_{\mathbb{Z}_{12}} M^\star_{x,y}$ (26 independent parameters), explains $R^2 = 0.830$ of $M^\star$. Full T/I invariance, $\Pi_{T/I} M^\star$ (19 independent parameters), still explains $R^2 = 0.822$ with 7 less parameters. It would be impossible for T/I invariance to outperform $\mathbb{Z}_{12}$ as it strictly adds restrictions on which elements must be equal, yet it has near identical performance. Furthermore, it is not the case that T/I and $\mathbb{Z}_{12}$ score highly due to some trivial reason. A chord-relabeling null gives $p = 5 \times 10^{-4}$: none of 2000 random relabelings yields a T/I-invariance R^2 approaching the observed 0.822.

Result I: The embeddings realize the dihedral irreps We factorize $M^\star$ and analyze its leading plane. Our prediction of interleaved circles of major and minors emerges (Figure 3 B, left). The plane's isotypic energy (loosely speaking, value from 0 to 1 on how much a *subspace* corresponds to a given irrep), is very high at 0.958 (E_5), far above a chance level of $d_\rho^2/24 = 4/24 \approx 0.17$. (App. Fig. 6). Most importantly, our theoretical predictions for the 24 chords give a *per-chord* cosine of 0.98. In fact, our theoretical predictions for the first 3 planes of embeddings (6 dimensions) are all high at $0.98/0.94/0.80$, respectively (Appendx F).

Result II: The circle of fifths In particular, the leading plane is *the circle of fifths*, a classic organizing structure of tonal harmony, realized as interleaved circles at the corpus dictated radius. The circle of fifths has been found *empirically* in chord embeddings before, on a far larger and entirely different music source of ∼89k popular songs (Lahnala et al., 2021); the present result *derives* it. It is a column of the 2-dimensional E_5 irrep of D_{12}.

Result III: Same structure in chord representation model Obtaining dihedral irreps in our chord embeddings is theoretically guaranteed once statistical invariance is established. However, in ChordBERT He et al. (2026), a 4.71M parameter transformer, the circle of fifths also emerges (Figure 3 B, right). The structures we predict for word embeddings potentially have wide spread appearances across many architectures; if a structure is useful in one setting, it is likely helpful elsewhere.

Competing symmetry models and data efficiency Although the chord statistics appears to follow a dihedral (T/I) structure, other conjectures for the underlying symmetry could perform even better. We compare seven other natural hypotheses about the structure of the co-occurrence matrix in F. Some notable examples include $\mathbb{Z}_{12}$ (transposition-only invariance as before, 26 independent parameters), and the raw empirical $M^\star$ itself (300 independent parameters). The T/I symmetry has (at least) comparable predictive performance on held-out chorales to each of other models, even those with more parameters. Furthermore, the T/I symmetry is the most efficient hypothesis at gaining predictive power when fit on a fewer number of chorales. Both of these observations point to T/I being the most stringent form of symmetry the data reliably follows.

7 THE CONTINUOUS CASE: THE SKY SPHERE AS A HARMONIC EMBEDDING

Berdnikov & Liokumovich (2026) discovered that large language models, prompted about *celestial proximity*, represent 188 stars and constellations on a decodable sphere in the residual stream, and posed the question of this geometry's origin. Our theory answers this question.

Predictions The sphere is the homogeneous space $S^2 = SO(3)/SO(2)$. Although an LLM's residual stream is not strictly in our theory, which considers word embeddings, we use Theorem 4.2 to predict what to expect in its structure. If the "effective" statistics are invariant (co-occurrence of celestial objects is a function only of angular separation), each dimension of the representation takes the form of a spherical harmonic.

The dimensions corresponding to $\ell = 1$, first order spherical harmonics, are precisely a sphere when plotted, as has been seen in Berdnikov & Liokumovich (2026). However, our theory further predicts the presence of spherical harmonics of higher orders (e.g. $\ell = 2, 3, \ldots$).

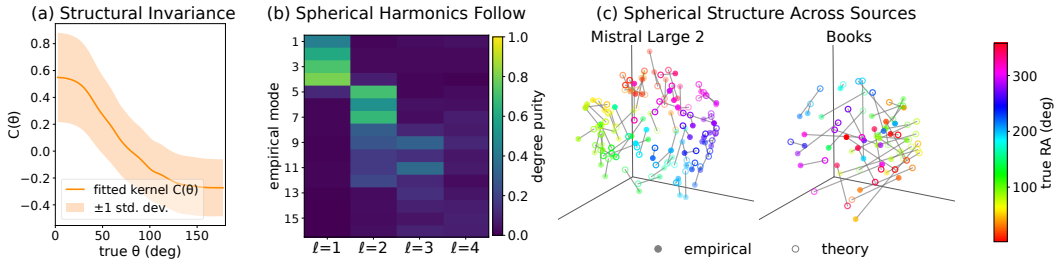


Figure 4: **LLM's structure is function of true celestial angles → spherical harmonics emerge, same structure also in text.** Mistral Large 2 (123B) residual stream for 188 sky objects (Berdnikov & Liokumovich, 2026) at layer 72 (where positions $\in \mathbb{R}^3$ on the celestial sphere can be best linearly decoded). See Appendix G for all details. (a) Weighted Nadaraya–Watson regression kernel predicts Gram matrix of LLM activations over objects from actual angular separation. Standard deviation follows the curves shape, suggesting structure is largely dependent on the true angle, our proxy for statistical invariance of $M^\star$. (b) Spherical harmonic purity of the model's PCs: blocks of spherical harmonics appear. $\ell = 0$ mode is removed by centering. (c) Comparison with $\ell = 1$ harmonics (sphere) evaluated at true celestial positions after a fitting rotation (and global scaling for the LLM). Color bar shows right ascension Left: LLM residual steam with a randomly selected 75/188 objects shown. Best-aligned 3-dimensional subspace is found within the first four empirical modes (same search repeated for permutation null). Per-object cosine sim. of 0.92 (median 0.97) against an object permutation null of 0.12 (95% 0.17). Right: Top 3 PCs of book corpus's centered $M^\star$. Per-object cosine sim. of 0.55 (median 0.68) against a null 0.22 (95% 0.32).

A small technical detail: our theory assumes a uniform sampling on S^2, which celestial objects do not exactly obey. To prevent any of our estimators being strongly influenced by over-sampled regions of the sphere, we employ inverse density weighting (see Appendix G).

(Effective) statistical assumption is verified: Gram matrix of activations is function of true angular separation. Figure 4 analyzes the released activations for Mistral Large 2, $X \in \mathbb{R}^{188\times 16}$ (mean across prompts for each object, top 16 PCs, App. G). As a proxy for $M^\star$, we consider the doubly centered, Gram matrix $PXX^TP \in \mathbb{R}^{188\times 188}$ (normalized by the root mean square of its diagonal). In place of statistical invariance, we ask for *structural invariance*: off diagonal-elements of this matrix are a function of only the true angular separation between the corresponding celestial objects. In Figure 4 A, we see this approximately holds. The standard deviation of a regressed kernel, predicting the upper triangle of the Gram matrix from objects' true angular separation, hugs the curve's shape. As further evidence, an averaging of all the upper triangular elements, grouped in bins of $10°$, gives an $R^2 = 0.49$ for predicting the original matrix (against a object permutation null of 0.002).

Results: The embeddings realize the spherical harmonics As seen in Figure 4 B the eigenmodes form *degree-ordered harmonic blocks*: modes 1–4 are ℓ=1, 5–8 are ℓ=2, and then ℓ=3 emerges. A completely random direction would have a purity on the order of 10^{-2} so the values in this heatmap are far above chance. Five of seven models in Berdnikov & Liokumovich (2026) demonstrate the same spherical harmonic ordering (Table 2). Notably, the ℓ=1 energy occupies four modes rather than three because PC 1 is largely occupied by a star/constellation type direction, as noted by Berdnikov & Liokumovich (2026).

Results: The same structure appears in an independent corpus. The harmonic structure should be visible in text embeddings themselves, specifically in the *genre* whose discourse is organized by sky position: star observing guides. However, many astronomy books are organized by topics such as mythology or famous constellations rather than by complete "tours of the sky". Hence, from 24 open-source general astronomy books, we utilize the metric $D = \mathrm{corr}(M^\star_{ij}, \cos\theta_{ij})$ with a one-sided location-permutation $p < 0.05$ (discussed in App. G) to select our corpus. Although passing this criterion would certainly be required for developing spherical harmonics in the embeddings, linear correlation with the true cosines does not by itself ensure the structure will be present. Four

guides pass (D from 0.22 to 0.45, 131k tokens in total). Volume is not enough to score highly on this metric, *genre* is critical: `text8` (17M tokens, 130× the guides) gives $D = 0.01$ ($p = 0.48$). `text8`'s counts are dominated by non-sky versions of the terminology. (Table 3; App. G). Notably, the guides' $M^\star$ over 39 well-counted objects yields a kernel of similar form as the model's activation kernel (correlation 0.92 with Figure 4 A).

Figure 4 C compares the LLM activations in layer 72 (where celestial positions are best linearly decoded) and the book corpus's embeddings with degree-1 spherical harmonics evaluated at the celestial object's true positions. For the books, we view the first 3 PCs of its embeddings, while for the model, we look at the most $\ell = 1$ aligned 3D-subspace in its first 4 PCs, hypothesizing that the structure could be *smeared* across dimensions. In both settings the spherical structure emerges, although in a much sharper fashion for the model. We perform the same analysis for degree-2 harmonics. Locating the most $\ell = 2$ aligned subspace in PCs 5-12 of the model (based on Figure 4 B), we obtain a per-object cosine similarity of 0.73 (median 0.82) compared to a object permutation null of 0.17 (95%: 0.21). Similarly, looking at PCs 4-8 of the books' centered $M^\star$, we see a per-object cosine sim. of 0.39 (median 0.44) against a null of 0.27 (95% 0.35). In either case, the $\ell = 2$ structure is weaker than that of $\ell = 1$, but it is certainly still present.

Book corpus's weaker harmonic structure results from less data: It may at first seem surprising that the books exhibit a lesser strength of spherical harmonic structure than the LLM residual stream. We argue that the book corpus's performance is typical for its size. We simulate new data assuming the zonal kernel fit to the books exactly specifies the true underlying statistics, and see how well dipole harmonics tend to align with the synthetic data (details in G). Figure 5 demonstrates that our book corpus is actually a typical (if not good) draw at its size (≈ 93% from 1000 simulations). At around roughly ∼32–64× the corpus size, we would expect the dipole structure in the textual embeddings to be just as present as in the model. The idea that a 123B parameter LLM could have consumed this amount of relevant stargazing material is very plausible. Hence, a possible explanation for the difference in structure quality between the model and textual embeddings is simply data volume.

8 Discussion

A unifying theory of structured embeddings. A Pringle traced by the months, the circle of fifths in Bach's chords, and a globe of spherical harmonics in a language model's activations initially appear unrelated. However, all three are predictions from the same theory. Underlying symmetry in the statistics of data, when present, is a powerful restraint on permissible embeddings.

Limitations The theory's scope is set by its premise: it considers only families of words whose co-occurrence statistics are, at least approximately, invariant under some group—a strong structural property that many word families will simply not have. Months, chords, and celestial objects were chosen because their statistics demonstrably carry such an invariance. Yet, for most of language such structure might not be immediately clear. For LLMs there is the further caveat that our theorems concern trained word embeddings, and their application to activations rests on an empirical premise.

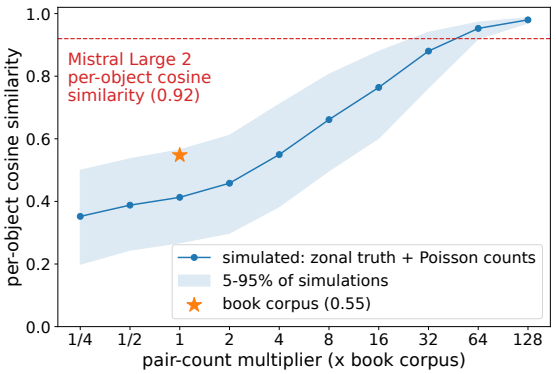


Figure 5: **Weaker harmonic structure in book corpus could be a sample size effect.** We take the zonal kernel fit to the books as ground truth and simulate new synthetic $M^\star$'s, using Poisson counts for off-diagonal elements at various multiples of the corpus's size. (full methodology in App. G). We measure how well the simulated data aligns with dipole spherical harmonics, using the same analysis pipeline as in Figure 4. Per-object cos. sim. (mean with 5–95% band) captures the observed corpus (star) at 1× and crosses the LLM's performance (0.92, dashed) at ∼32–64× the books' counts: an *effective corpus size* for the model.

REFERENCES

Aleksandr Berdnikov and Yevgeny Liokumovich. Sky sphere representation in language models, 2026. URL https://arxiv.org/abs/2607.27092.

Bilal Chughtai, Lawrence Chan, and Neel Nanda. A toy model of universality: Reverse engineering how networks learn group operations. In *International Conference on Machine Learning*, pp. 6243–6267. PMLR, 2023.

Alissa S Crans, Thomas M Fiore, and Ramon Satyendra. Musical actions of dihedral groups. *The American Mathematical Monthly*, 116(6):479–495, 2009.

Michael Scott Cuthbert and Christopher Ariza. Music21: A toolkit for computer-aided musicology and symbolic music data. In J. Stephen Downie and Remco C. Veltkamp (eds.), *ISMIR*, pp. 637–642. International Society for Music Information Retrieval, 2010. ISBN 978-90-393-53813. URL http://dblp.uni-trier.de/db/conf/ismir/ismir2010.html#CuthbertA10.

Feng Dai and Yuan Xu. *Approximation theory and harmonic analysis on spheres and balls*. Springer, 2013.

Persi Diaconis. *Group representations in probability and statistics*. Institute of Mathematical Statistics, 1988.

Josh Engels, Eric Michaud, Isaac Liao, Wes Gurnee, and Max Tegmark. Not all language model features are one-dimensionally linear. In *International Conference on Learning Representations*, volume 2025, pp. 84591–84622, 2025.

G.B. Folland. *A Course in Abstract Harmonic Analysis, Second Edition*. Textbooks in Mathematics. Taylor & Francis Group, 2016. ISBN 9781498727136.

Wes Gurnee and Max Tegmark. Language models represent space and time. In *International Conference on Learning Representations*, volume 2024, pp. 2483–2503, 2024.

Yulong He, Ivan Smirnov, and Yanming Li. Modeling stylistic co-evolution in symbolic music heritage collections, 2026. URL https://arxiv.org/abs/2607.23957.

Dhruva Karkada, James Simon, Yasaman Bahri, and Michael DeWeese. Closed-form training dynamics reveal learned features and linear structure in word2vec-like models. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 3967–4000. Curran Associates, Inc., 2025. doi: 10.52202/085713-0140.

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.

Allison Lahnala, Gauri Kambhatla, Jiajun Peng, Matthew Whitehead, Gillian Minnehan, Eric Guldan, Jonathan K. Kummerfeld, Anıl Çamcı, and Rada Mihalcea. Chord embeddings: Analyzing what they capture and their role for next chord prediction and artist attribute prediction. In *Artificial Intelligence in Music, Sound, Art and Design*, pp. 171–186, Cham, 2021. Springer International Publishing. ISBN 978-3-030-72914-1.

Omer Levy and Yoav Goldberg. Neural word embedding as implicit matrix factorization. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

Matt Mahoney. Large text compression benchmark (text8). http://mattmahoney.net/dc/textdata.html, 2011.

Giovanni Luca Marchetti, Daniel Kunin, Adele Myers, Francisco Acosta, and Nina Miolane. Sequential group composition: A window into the mechanics of deep learning. In *International Conference on Machine Learning (ICML)*, 2026.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. In *International Conference on Learning Representations*, 2023.

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*, volume 235 of *Proceedings of Machine Learning Research*, pp. 39643–39666. PMLR, 2024.

Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. The geometry of categorical and hierarchical concepts in large language models. In *International Conference on Learning Representations*, volume 2025, pp. 76441–76463, 2025.

Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162.

Jean-Pierre Serre. *Linear Representations of Finite Groups*, volume 42 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1977. ISBN 978-0-387-90190-9.

APPENDIX OVERVIEW

Appendices A–D contain the complete theory while Appendices E–G contain all experimental details. Specifically, the appendix is organized as follows.

Appendix A discusses how `word2vec`-style embeddings can be approximated by an eigendecomposition of a co-occurrence matrix of the words, and states the "statistical invariance under a group" assumption necessary for our theory to apply. Appendix B considers embedding structure when the words are invariant under a finite group (Theorem B.3), and considers two special cases: the cyclic and dihedral cases, which are relevant to our "months" and "chords" experiments. Appendix C generalizes the theory to compact groups. Appendix D further treats homogeneous spaces G/K, considering the special case of Gelfand Pairs and in particular S^2, which is relevant to our sky-sphere experiments.

Appendices E, F, G give the experimental details for the month, music, and sky-sphere experiments, respectively. Finally, Appendix H discusses reproducibility, while Appendix I gives our AI use statement.

A WORD EMBEDDINGS AND STATISTICAL INVARIANCE

A.1 FROM STATISTICS TO EMBEDDINGS

Let L be a finite set of n words. Following Karkada et al. (2026), for arbitrary words $x, y \in L$ we consider the *normalized co-occurrence matrix* defined by

$$M^{\star}_{xy} \;=\; \frac{P_{xy} - P_x P_y}{\frac{1}{2}\left(P_{xy} + P_x P_y\right)}, \tag{5}$$

where P_{xy} is the probability that words x, y co-occur within a context window and P_x is the unigram probability (both defined below). Since $P_{xy} = P_{yx}$, the matrix $M^{\star}$ is symmetric.

Estimation conventions. Throughout the paper, $P_{xy}\ \forall x, y \in L$ and $P_x\ \forall x \in L$ are estimated from a respective corpus $\mathcal{C}$ of N tokens as follows. Fix a function $f(\delta) \geq 0$ supported on token gaps $\delta = 1, \ldots, B$, with total mass $F := \sum_{\delta=1}^{B} f(\delta)$. We define a weighted count matrix

$$W_{x,y} = \sum_i \delta_{\mathcal{C}[i],x} \sum_{d=1}^{B} f(d)(\delta_{\mathcal{C}[i+d],y} + \delta_{\mathcal{C}[i-d],y})$$

where $\mathcal{C}[i]$ is the i^{th} token in the corpus. For multiple corpora, we can form a single W_{xy} through summation. Note that this definition enforces $W_{xy} = W_{yx}$. Then, with additive smoothing $\alpha \geq 0$, and the number of mentions of word x being n_x we define

$$P_{xy} \;=\; \frac{W_{xy} + \alpha}{2FN}, \qquad P_x \;=\; \frac{n_x}{N}, \tag{6}$$

so that $2FN$ is the total window mass (N tokens, each opening a two-sided window of mass $2F$, neglecting edge effects). We can further define an *opportunity normalized* co-occurrence rate:

$$R_{xy} \;=\; \frac{P_{xy}}{P_x P_y} \;=\; \frac{W_{xy}+\alpha}{n_x n_y \cdot 2F/N} \;=\; \frac{W_{xy}+\alpha}{\mathbb{E}_0[W_{xy}]}, \tag{7}$$

where $\mathbb{E}_0[W_{xy}] = n_x n_y \cdot 2F/N$ is the expected value of $W_{x,y}$ when mentions are placed independently (neglecting edge effects). We have that $M^\star_{xy} = 2(R_{xy}-1)/(R_{xy}+1) = 2\tanh(\frac{1}{2}\ln R_{xy})$, so $M^\star$ is strictly increasing in R_{xy} and saturates at ± 2.

Per-system choices of (B, f, α) are stated in the experimental appendices (months: $B = 16$, $f(\delta) = B+1-\delta$, $\alpha = 0$; chords: $B = 3$, flat: $f(\delta) = 1$, $\alpha = 0.5$; sky: $B = 30$, flat: $f(\delta) = 1$, $\alpha = 0.25$).

Word Embeddings Karkada et al. (2025) shows that the matrix $PM^\star P$ can be used to generate word embeddings. Specifically, the r leading PCA scores of a set of n `word2vec`-style embeddings in d dimensions $\overline{W} \in \mathbb{R}^{n\times d}$ can be approximated with the positive eigenpairs $\{(\lambda_i, v_i)\}_{i\le r}$ of $PM^\star P$, taken in decreasing λ_i order. Up to an orthogonal transform and realification, the projection of $\overline{W}_x \in \mathbb{R}^d$ (for word x) along the first r principal components of $\overline{W}$ is approximately $\left(\sqrt{\lambda_i} v_i(x)\right)_{i\le r}$. This is how we generate all empirical word embeddings in this paper.

Furthermore, transforming the generation of word embeddings into an eigendecomposition is incredibly fruitful for theory. Whenever $M^\star$ obeys certain symmetries, we can theoretically predict its eigenpairs, which we discuss in the upcoming sections.

A.2 Statistical Invariance under a Group

Definition A.1 (Torsor)**.** A (left) *G-torsor* is a set L with a (left) action of G that is free and transitive: for every $x, y \in L$ there is exactly one $g \in G$ with $g \cdot x = y$. A torsor is "a group that has forgotten its identity element": it becomes isomorphic to G once a base point is chosen, but no choice is canonical.

Example A.2. The 24 major and minor triads form a torsor under the transposition–inversion group $T/I \cong D_{12}$ and the 12 months form a $\mathbb{Z}_{12}$-torsor.

Definition A.3 (Statistical invariance under a group)**.** We say that a G-torsor word set L is statistically invariant under the group G if its $M^\star$ obeys

$$M^\star_{g\cdot x,\, g\cdot y} \;=\; M^\star_{x,y} \qquad \text{for all } g \in G,\; x, y \in L.$$

Lemma A.4 (Kernel)**.** *Consider a G-torsor word set L. Fix a base point $x_0 \in L$ and identify $L \cong G$ via $g \mapsto g \cdot x_0$. Then, abusing notation, Definition A.3 is equivalent to the existence of some (kernel) $C : G \to \mathbb{R}$ with*

$$M^\star_{g,h} \;=\; C(g^{-1}h)\;\; \forall g, h \in G, \tag{8}$$

and symmetry of $M^\star$ forces $C(g^{-1}) = C(g)\;\forall g \in G$ (note we change from the main text convention of denoting 1 as identity to the more standard e).

Proof. First, we assume that $M^\star$ is invariant under the group G. Then we define $C : G \to \mathbb{R}$ by $C(g) = M^\star_{e,g}\;\forall g \in G$. Hence, from statistical invariance, for any $g, h \in G$, $M^\star_{g,h} = M^\star_{e,g^{-1}h} = C(g^{-1}h)$.

Now, we assume that $M^\star_{g,h} = C(g^{-1}h)\;\forall g, h \in G$. Hence, for any $a \in G$, we have that $M^\star_{a\cdot g, a\cdot h} = C(g^{-1}a^{-1}ah) = C(g^{-1}h) = M^\star_{g,h}$, which is precisely statistical invariance under G. Finally, $C(g^{-1}h) = M^\star_{g,h} = M^\star_{h,g} = C(h^{-1}g) = C((g^{-1}h)^{-1})\;\forall g, h \in G$, which gives us that $C(g) = C(g^{-1})\;\forall g \in G$. □

B Finite groups

Throughout this section, G is a finite group with $|G| = n$. Furthermore, we assume we have a G-torsor word set L whose $M^\star$ in statistically invariant under G. Under some arbitrary choice of base point, this allows us to associate a real kernel $C : G \to \mathbb{R}$ with $M^\star$ as in Lemma A.4.

B.1 HARMONIC ANALYSIS PRELIMINARIES

Let $\widehat{G}$ denote a complete set of inequivalent *unitary* irreducible representations $\rho : G \to U(d_\rho)$, so that $\sum_{\rho \in \widehat{G}} d_\rho^2 = n$. We note the Schur orthogonality relations for matrix coefficients: for $\rho, \sigma \in \widehat{G}$,

$$\sum_{g \in G} \rho_{ab}(g)\, \overline{\sigma_{ce}(g)} \;=\; \frac{n}{d_\rho}\, \delta_{\rho\sigma}\, \delta_{ac}\, \delta_{be}, \tag{9}$$

and the generalized Fourier transform of C,

$$\widehat{C}(\rho) \;=\; \sum_{g \in G} C(g)\, \rho(g) \;\in\; \mathbb{C}^{d_\rho \times d_\rho}. \tag{10}$$

where d_ρ is the dimension of the irrep ρ.

Lemma B.1. *Under C real with $C(g^{-1}) = C(g)$, each $\widehat{C}(\rho)$ is Hermitian.*

Proof. $\widehat{C}(\rho)^\dagger = \sum_g \overline{C(g)} \rho(g)^\dagger = \sum_g C(g) \rho(g^{-1}) = \sum_g C(g^{-1}) \rho(g) = \sum_g C(g) \rho(g) = \widehat{C}(\rho)$. □

Accordingly each $\widehat{C}(\rho)$ admits an orthonormal eigenbasis $\{u_\rho^{(s)}\}_{s=1}^{d_\rho} \subset \mathbb{C}^{d_\rho}$ with real eigenvalues,

$$\widehat{C}(\rho)\, u_\rho^{(s)} \;=\; \lambda_\rho^{(s)}\, u_\rho^{(s)}, \qquad \lambda_\rho^{(s)} \in \mathbb{R}.$$

B.2 CENTERING REMOVES EXACTLY THE TRIVIAL REPRESENTATION

Lemma B.2 (Centering)**.** *The constant vector $\mathbf{1}$ is an eigenvector of $M^\star$ with eigenvalue $\mu_0 = \sum_g C(g)$. $M^\star$ is real symmetric so we may construct an orthogonal eigenbasis for its eigenspace with $\mathbf{1}$ explicitly included. $PM^\star P$ (with $P = \mathrm{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$) has these same eigenpairs as $M^\star$ except that the pair $(\mu_0, \mathbf{1})$ is replaced by $(0, \mathbf{1})$; no other eigenvector or eigenvalue is altered.*

Proof. $M^\star \mathbf{1} = \big(\sum_g C(g)\big)\mathbf{1}$ from equation 8, so $\mathbf{1}$ is indeed an eigenvector. Furthermore, if $M^\star v = \lambda v$ with $v \perp \mathbf{1}$ then $Pv = v$. Hence, $PM^\star Pv = PM^\star v = \lambda Pv = \lambda v$, while $PM^\star P\mathbf{1} = 0$. □

For obtaining our word embeddings from $PM^\star P$, we can diagonalize $M^\star$ and discard the constant mode. This constant mode corresponds to the trivial representation of the group, $\rho(g) = 1\ \forall g \in G$.

B.3 EMBEDDING THEOREM FOR FINITE GROUPS

Theorem B.3 (Diagonalization of invariant statistics; finite case)**.** *Let G be a finite group with n elements. Assume we have a G-torsor (Definition A.1) word set L whose $M^\star$ (Equation 5) is statistically invariant under G. Under an arbitrary choice of base point (arbitrary from Lemma B.4), we associate with $M^\star$ a kernel $C : G \to \mathbb{R}$ as in Lemma A.4. From Lemma B.1, $\widehat{C}(\rho)$ admits an orthonormal eigenbasis $\{u_\rho^s\}_{s=1}^{d_\rho} \subset \mathbb{C}^d$ with real eigenvalues $\lambda_\rho^{(s)}$. Let $\widehat{G}$ denote a complete set of irreducible unitary representations of G as usual. For each $\rho \in \widehat{G}$, each eigenvector index $s \in \{1, \ldots, d_\rho\}$ and each row index $a \in \{1, \ldots, d_\rho\}$, we define*

$$\psi^{\rho,a,s}(g) \;:=\; \sqrt{\tfrac{d_\rho}{n}}\, \big[\rho(g)\, u_\rho^{(s)}\big]_a. \tag{11}$$

Then:

- *(i) Each $\psi^{\rho,a,s}$ is an eigenvector of $M^\star$ with eigenvalue $\lambda_\rho^{(s)}$: $M^\star\, \psi^{\rho,a,s} = \lambda_\rho^{(s)}\, \psi^{\rho,a,s}$. Furthermore, there exists a choice of equivalent irreducible unitary representations, ρ', such that we have eigenvectors $\sqrt{\frac{d_\rho}{n}} \rho'(g)_{as}$.*
- *(ii) The family $\{\psi^{\rho,a,s} : \rho \in \widehat{G},\ 1 \le a, s \le d_\rho\}$ is an orthonormal basis of $\ell^2(G) \cong \mathbb{C}^n$ ($\sum_\rho d_\rho^2 = n$ vectors).*

(iii) Hence, the spectrum of $M^\star$ is $\{\lambda_\rho^{(s)} : \rho \in \widehat{G}, s \in [d_\rho]\}$, each occurring with multiplicity (at least) d_ρ and

$$PM^\star P = \sum_{\rho \neq \rho_{\text{triv}}} \sum_{s=1}^{d_\rho} \lambda_\rho^{(s)} \sum_{a=1}^{d_\rho} \psi^{\rho,a,s} \left(\psi^{\rho,a,s}\right)^\dagger.$$

where $P = I - \frac{1}{n}\mathbf{1}\mathbf{1}^T$.

Proof. (i) Substituting $h = gk$,

$$\begin{aligned}(M^\star \psi^{\rho,a,s})(g) &= \sum_h C(g^{-1}h)\, \psi^{\rho,a,s}(h) = \sqrt{\tfrac{d_\rho}{n}} \sum_k C(k) \left[\rho(g)\rho(k)\, u_\rho^{(s)}\right]_a \\ &= \sqrt{\tfrac{d_\rho}{n}} \left[\rho(g)\, \widehat{C}(\rho)\, u_\rho^{(s)}\right]_a = \lambda_\rho^{(s)}\, \psi^{\rho,a,s}(g).\end{aligned}$$

This proves the first claim. Now, for some irrep ρ, let $U_\rho = \left[u_\rho^{(1)} | \cdots | u_\rho^{(d_\rho)}\right]$. Note that U_ρ is unitary since $\widehat{C}(\rho)$ is Hermitian. Let us define the equivalent unitary irrep $\rho' := U_\rho^\dagger\, \rho\, U_\rho$. Furthermore, $\widehat{C}(\rho)$ = $U^\dagger \Lambda U$ where $\Lambda = \operatorname{diag}\left(\lambda_\rho^{(1)}, \ldots, \lambda_\rho^{(d_\rho)}\right)$. Hence,

$$\widehat{C}(\rho') = \sum_{g \in G} C(g) U_\rho^\dagger\, \rho\, U_\rho = U_\rho^\dagger\, \widehat{C(\rho)}\, U_\rho = \Lambda.$$

We can choose the basis vectors $\{e_i\}_{i=1}^{d_\rho}$ as the eigenbasis for $\widehat{C}(\rho')$. Plugging into our formula for the eigenvectors above, we obtain

$$\sqrt{\frac{d_\rho}{n}}[\rho(g)e_s]_a = \sqrt{\frac{d_\rho}{n}}\rho(g)_{as}$$

(ii) Orthonormality: expanding $\psi^{\rho,a,s}(g) = \sqrt{d_\rho/n}\sum_b \rho_{ab}(g)\,(u_\rho^{(s)})_b$ and applying (9),

$$\langle \psi^{\rho,a,s}, \psi^{\sigma,c,t}\rangle = \tfrac{\sqrt{d_\rho d_\sigma}}{n} \sum_{b,e} (u_\rho^{(s)})_b \overline{(u_\sigma^{(t)})_e} \sum_g \rho_{ab}(g)\overline{\sigma_{ce}(g)} = \delta_{\rho\sigma}\,\delta_{ac}\,\langle u_\rho^{(t)}, u_\rho^{(s)}\rangle_{\mathbb{C}^{d_\rho}} = \delta_{\rho\sigma}\delta_{ac}\delta_{st}.$$

Completeness follows by counting: there are $\sum_\rho d_\rho^2 = n$ orthonormal vectors in an n-dimensional space. (iii) is immediate from (i)–(ii) and Lemma B.2. □

Two freedoms are apparent in Theorem B.3: the choice of equivalent irreps, and the choice of eigenbasis of $\widehat{C}(\rho)$. We show the third potential source of freedom, choosing the base point, produces no permissible $\psi^{\rho,a,s}$ outside of the previous two freedoms in Lemma B.4. Both of these freedoms do not amount to "structural" changes. We spare the small derivations, but one can see the first freedom amounts to a unitary rotation of our eigenbasis in the a coordinate; i.e., the basis vectors $\{\psi^{\rho,a,s}\}_{a=1}^{d_\rho}$ are unitarily rotated within their subspace. Additionally, the choice of eigenbasis amounts to a unitary rotation across s in $\psi^{\rho,a,s}$ (for fixed ρ, a) for which all $\psi^{\rho,a,s}$ posses the same eigenvalue.

Lemma B.4 (Choice of base point)**.** *Consider a G-torsor word set L. We can fix a base point $x_0 \in L$ and identify $L \cong G$ via $g \mapsto g \cdot x_0$. Under the alternative choice of base point $x_0' = b \cdot x_0$, the group elements assigned to the word x, g_x, becomes $g_x\, b^{-1}$. Firstly, this relabeling has no effect on $M^\star$ nor its eigendecomposition (what we denote as $M^\star$'s indices has no effect on its numerical values). Hence, we can freely choose any base point without any effect on the word embeddings.*

What may change is our eigenvectors in equation 11. Fortunately, the proposed change of base point produces the same set of permissible $\psi^{\rho,a,s}$'s as in Theorem B.3, with the usual freedoms of choosing equivalent irreps and an eigenbasis of $\widehat{C}(\rho)$.

Proof. $M^\star_{x,y} = C(g_x^{-1}g_y)$ where g_x, g_y are the elements of G assigned to words x, y before the base point change. Since the matrix elements do not change, after selecting the new base point, $g \to gb^{-1}$, and we have that

$$M^\star_{x,y} = C'(g_x'^{-1}g_y') = C'(bg_x^{-1}g_y b^{-1}).$$

Hence, $C'(bkb^{-1}) = C(k) \ \ \forall k \in G$, so $C'(k) = C(b^{-1}kb)$. where C' is the new kernel. We can see that

$$\widehat{C}'(\rho) = \rho(b)\widehat{C}(\rho)\rho(b)^{-1}.$$

Therefore, we can choose our eigenbasis of $\widehat{C}'(\rho)$ to be $\{\rho(b)u_\rho^{(s)}\}_{s=1}^{d_\rho}$. The formula for the a modified ψ then becomes

$$\sqrt{\frac{d_\rho}{n}}[\rho(g')\rho(b)u_\rho^{(s)}]_a = \sqrt{\frac{d_\rho}{n}}[\rho(g)u_\rho^{(s)}]_a,$$

unchanged from before. Furthermore, if we used our freedom of choice of equivalent irreps or eigenbasis of $\widehat{C}'(\rho)$ when constructing the modified ψ, those same freedoms are present in the original set of permissible ψ's. Hence, the sets of permissible ψ before and after the base change are the same. □

Similar results to Lemma B.4 can be shown in the cases that G is replaced with a compact group (Appendix C), or homogeneous space (Appendix D).

B.4 Special case: kernel is class function

The natural analogue of the Karkada et al. (2026) assumption that statistics depend only on *distance* in some cyclic latent space is that our kernel C, as in Lemma A.4, is a *class function*. I.e., it is constant on conjugacy classes of some respective group G. An explanation as to why follows.

Let's assume $M^\star_{g,h} = D(g,h)\ \forall g,h \in G$ for some "distance function" $D : G \times G \to \mathbb{R}$. Intuitively, distances between points shouldn't change as we move them together through space. This suggests the condition that D is bi-invariant under G. In other words,

$$D(ag, ah) = D(ga, ha) = D(g,h)\ \forall a, g, h \in G.$$

Hence,

$$C(g) = D(e,g) = D(a, ag) = D(e, aga^{-1}) = C(aga^{-1}) \quad \forall a, g \in G,$$

so C being a class function is indeed the natural analogue.

Proposition B.5 (Fourier transform of class function C is scaled identity)**.** *If C is a class function, then for every $\rho \in \widehat{G}$,*

$$\widehat{C}(\rho) \ = \ \hat{c}_\rho \, \mathrm{I}_{d_\rho}, \qquad \hat{c}_\rho \ = \ \frac{1}{d_\rho} \sum_{g \in G} C(g)\, \chi_\rho(g) \ \in \ \mathbb{R},$$

where $\chi_\rho = \operatorname{tr} \rho$. Consequently in Theorem B.3 the eigenvalue $\hat{c}_\rho$ has multiplicity d_ρ^2, one for each matrix coefficient of ρ, and we can take $\psi^{\rho,a,b}(g) = \sqrt{d_\rho/n}\ \rho(g)_{ab}$.

Proof. For any $a \in G$,

$$\rho(a)^{-1}\, \widehat{C}(\rho)\, \rho(a) = \sum_{g \in G} C(g)\, \rho(a^{-1}ga) = \sum_{g \in G} C(aga^{-1})\, \rho(g) = \sum_{g \in G} C(g)\, \rho(g) = \widehat{C}(\rho),$$

so $\widehat{C}(\rho)$ commutes with ρ. Hence, by Schur's lemma $\widehat{C}(\rho)$ must be a scalar (which we will denote $\hat{c}_\rho$) times the identity matrix. Taking traces of $\widehat{C}(\rho) = \hat{c}_\rho \mathrm{I}_{d_\rho}$ gives $d_\rho \hat{c}_\rho = \sum_g C(g)\chi_\rho(g)$, the desired relation.

The consequences for Theorem B.3 follow as every unit vector of $\mathbb{C}^{d_\rho}$ is now an eigenvector of $\widehat{C}(\rho)$. Additionally, note that our eigenvalues of $\widehat{C}(\rho)$, $\hat{c}_\rho$, are still real as before since

$$\overline{\hat{c}_\rho} = \frac{1}{d_\rho} \sum_{g \in G} \overline{C(g)\chi_\rho(g)} = \frac{1}{d_\rho} \sum_{g \in G} C(g)\chi_\rho(g^{-1}) = \frac{1}{d_\rho} \sum_{g \in G} C(g^{-1})\chi_\rho(g) = \frac{1}{d} \sum_{g \in G} C(g)\chi_\rho(g) = \hat{c}_\rho$$

□

B.5 REALIFICATION OF EIGENFUNCTIONS

Embeddings are real, while the eigenvectors (11) are in general complex. $M^\star$ is a real symmetric matrix, so it always possible to find a real orthonormal basis for its eigenspace. The only question is whether it is somehow possible to modify our theory's outputted irrep-based eigenvectors in a formulaic manner as to obtain real embeddings.

The answer depends on the "type" of irrep we are considering. Namely, we consider its Frobenius-Schur indicator

$$\nu(\rho) \;=\; \frac{1}{n}\sum_{g\in G}\chi_\rho(g^2) \;\in\; \{1,0,-1\},$$

with $\chi_\rho(g^2)$ denoting the trace of $\rho(g^2)$. $\nu = 1, 0, -1$ are denoted as *real* type, *complex* type, and *quaternionic* type, respectively.

Proposition B.6 (Realification Protocols)**.** *Transforming the eigenvectors* (11) *into real form depends on the type (Frobenius-Schur indicator) of the given irrep. The different methodologies are as follows.*

- *(a)* ***Real type*** *($\nu(\rho) = 1$). This is the simplest case. Since ρ is real type, it is guaranteed to be expressible as real matrices upon some appropriate change of basis. Furthermore, after enforcing the reality of ρ, the eigenvectors of $\widehat{C}(\rho)$,$\{u_\rho^{(s)}\}_{s=1}^d$, can be chosen real as $\widehat{C}$ is now a real symmetric matrix. Together, these enforce the reality of the eigenvectors in* (11)*.*

- *(b)* ***Complex type*** *($\nu(\rho) = 0$). Since ρ is of complex type, no change of basis can convert it to real matrices. Hence, a different approach is required. Namely, for irreps ρ and $\bar{\rho}$, and $a, s \in [d_\rho]$, we generate new real orthonormal eigenvectors through*

$$\psi^{\rho,a,s}, \psi^{\bar{\rho},a,s} \to \sqrt{2}\,\mathrm{Re}\,\psi^{\rho,a,s}, \sqrt{2}\,\mathrm{Im}\,\psi^{\rho,a,s} \tag{12}$$

- *(c)* ***Quaternionic type*** *($\nu(\rho) = -1$). Since ρ is of quaternionic type, it is impossible to express it as real matrices. Furthermore, unlike the complex case, there is generally no "simple" way to arrive at real eigenvectors which are clean functions of our eigenvectors in equation* (11)*. An approach instead is to take the real and imaginary components of each $\psi^{\rho,a,s}$ and just apply Gram-Schmidt to arrive at real, orthogonal eigenvectors.*

Proof. (a) Proof already discussed in proposition.

(b) Since ρ is of complex type, it is well known that $\rho \ncong \overline{\rho}$. ρ and $\overline{\rho}$ are genuinely nonequivalent irreps which correspond to different eigenvectors in equation (11). The reality of our proposed substitution is immediate. All that is left is to demonstrate that the proposed vectors are indeed eigenvectors of $M^\star$ and are orthonormal.

We begin by proving that $\mathrm{Re}\,\psi^{\rho,a,s}$ and $\mathrm{Im}\,\psi^{\rho,a,s}$ are eigenvectors with the same eigenvalue $\lambda_\rho^{(s)}$. Namely,

$$\mathrm{Re}\,\psi^{\rho,a,s} = \frac{1}{2}(\psi^{\rho,a,s} + \overline{\psi^{\rho,a,s}}), \quad \mathrm{Im}\,\psi^{\rho,a,s} = \frac{1}{2i}(\psi^{\rho,a,s} - \overline{\psi^{\rho,a,s}})$$

Note that

$$\overline{\psi^{\rho,a,s}} = \sqrt{\frac{d_\rho}{n}}\left[\overline{\rho}(g)\overline{u_\rho^{(s)}}\right]_a.$$

$\overline{u_\rho^{(s)}}$ is an eigenvector of $\overline{\widehat{C}}(\rho)$, which in turn equals $\widehat{C}(\overline{\rho})$ from the reality of C. Hence,

$$\overline{\psi^{\rho,a,s}} = \sqrt{\frac{d_\rho}{n}}\left[\overline{\rho}(g)u_{\overline{\rho}}^{(s)}\right]_a = \psi^{\overline{\rho},a,s},$$

so $\overline{\psi^{\rho,a,s}} = \psi^{\overline{\rho},a,s}$ is another eigenvector of $M^\star$. Furthermore, $\psi^{\overline{\rho},a,s}$ and $\psi^{\rho,a,s}$ share the same eigenvalue since

$$M^\star\overline{\psi^{\rho,a,s}} = \overline{M^\star\psi^{\rho,a,s}} = \overline{\lambda_\rho^{(s)}}\,\overline{\psi^{\rho,a,s}} = \lambda_\rho^{(s)}\overline{\psi^{\rho,a,s}},$$

taking advantage of $M^\star$ being real symmetric. Therefore, $\mathrm{Re}\,\psi^{\rho,a,s}$ and $\mathrm{Im}\,\psi^{\rho,a,s}$ are indeed eigenvectors with eigenvalue $\lambda_\rho^{(s)}$.

Finally, checking the orthonormality of $\sqrt{2}\,\mathrm{Re}\,\psi^{\rho,a,s}$ and $\sqrt{2}\,\mathrm{Im}\,\psi^{\rho,a,s}$ is relatively simple as $\psi^{\rho,a,s}$ and $\psi^{\overline{\rho},a,s} = \overline{\psi^{\rho,a,s}}$ are orthonormal with respect to any of the original eigenvectors.

(c) Gram-Schmidt returns real vectors if real vectors are first provided. Furthermore, define the eigen-subspace $W_\lambda = \{\phi : M^\star\phi = \lambda\phi\}$. W_λ and $W_{\lambda'}$ are orthogonal for $\lambda \neq \lambda'$ since $M^\star$ is real symmetric. Hence, if Gram-Schmidt is provided k vectors in W_λ it will return k vectors in W_λ, i.e., it will return other eigenvectors.

All that is left to show is that the real and imaginary components of the ψ's are all eigenvectors and span the space of all eigenvectors of $M^\star$. Firstly, note that since $M^\star$ is real with real eigenvalues, the real and imaginary components of any $\psi^{\rho,as}$ are automatically eigenvectors. Namely,

$$M^\star\psi^{\rho,a,s} = \lambda_\rho^{(s)}\psi^{\rho,a,s} \Rightarrow M^\star\,\mathrm{Re}\,\psi^{\rho,\mathrm{a,s}} + iM^\star\,\mathrm{Im}\,\psi^{\rho,a,s} = \lambda_\rho^{(s)}\,\mathrm{Re}\,\psi^{\rho,a,s} + i\lambda_\rho^{(s)}\,\mathrm{Im}\,\psi^{\rho,a,s}$$

splits into two separate eigenvalue-eigenvector equations. Finally, these real and imaginary components of the ψ span the space of all eigenvectors automatically since the ψ span the space of all eigenvectors. Hence, Gram-Schmidt will not be missing any eigenvectors.

□

Only quaternionic type irreps would provide real eigenvectors whose structure do not fall neatly from our equation 11. Fortunately, these types of irreps are arguably "rarer" in more typical groups, and do not appear in any of our experiments.

B.6 Special cases: cyclic and dihedral groups

We briefly state the structure Theorem B.3 predicts for the cyclic ($\mathbb{Z}_{12}$) and dihedral (D_{12}) cases relevant to our experiments.

Cyclic: $G = \mathbb{Z}_{12}$ (months). All irreps are one-dimensional: characters $\chi_k(m) = e^{2\pi ikm/12}$, indexed $k = 0, \ldots, 11$. The generalized Fourier transform is actually the discrete Fourier transform. Namely, $\widehat{C}(\rho)$ become scalars

$$\hat{c}_k = \sum_{d=0}^{11} C(d)\, e^{-2\pi ikd/12} = \sum_d C(d)\cos(2\pi kd/12),$$

which are real since $C(d) = C(-d)$. After realification, the conjugate pairs (χ_k, χ_{-k}), $k = 1, \ldots, 5$, fuse into two-dimensional cosine/sine planes sharing the eigenvalue $\hat{c}_k$. $k = 0$ is the constant mode (removed by centering) and $k = 6$ the one-dimensional alternating mode, which empirically tends to be neglected in natural data. The predicted embedding of month m is therefore, up to one orthogonal transform per plane,

$$m \;\longmapsto\; \Big(\sqrt{\hat{c}_k}\,\cos\big(\tfrac{2\pi km}{12}\big),\ \sqrt{\hat{c}_k}\,\sin\big(\tfrac{2\pi km}{12}\big)\Big)_{k\in\mathcal{K}}, \qquad \mathcal{K} = \{k \geq 1 : \hat{c}_k > 0,\ \text{decreasing } \hat{c}_k\}:$$

concentric circles traversed at $30k^\circ$ per month, radius $\sqrt{\hat{c}_k}$.

Dihedral: $G = D_{12}$ (major/minor triads). Let T_m and I_m denote the m^{th} major and minor triad respectively for $m \in \{0, \ldots, 11\}$ (with base chord C). $D_{12} = \mathbb{Z}_{12} \rtimes \mathbb{Z}_2$ has four one-dimensional irreps and five two-dimensional irreps E_h, $h = 1, \ldots, 5$, with $E_h(T_m) = R(2\pi hm/12)$ and $E_h(I_m) = R(2\pi hm/12)\,F$. $R(\theta)$ denotes typical 2D rotation matrix by θ and $F = \mathrm{diag}(1, -1)$. Note that all irreps are of real type. We will neglect the one-dimensional irreps as they do not tend to prevalent, as least for our experiments. For each E_h, $\widehat{C}(E_h)$ is a real symmetric 2×2 matrix with eigenpairs $(\lambda_h^{(s)}, u_h^{(s)})$, $s = 1, 2$. Writing $u_h^{(s)} = (\cos\varphi, \sin\varphi)$, each realized mode (h, s) with $\lambda_h^{(s)} > 0$ contributes a plane in which (up to one orthogonal transform per plane)

$$T_m \to \sqrt{\lambda_h^{(s)}}\big(\cos\big(\tfrac{2\pi hm}{12} + \varphi\big),\, \sin\big(\tfrac{2\pi hm}{12} + \varphi\big)\big), \quad I_m \to \sqrt{\lambda_h^{(s)}}\big(\cos\big(\tfrac{2\pi hm}{12} - \varphi\big),\, \sin\big(\tfrac{2\pi hm}{12} - \varphi\big)\big),$$

two interleaved twelve-point circles of *equal radius* (forced by orthogonality of E_h), stepping $30h^\circ$ per semitone, with relative offset 2φ.

C COMPACT GROUPS

The finite theory extends to compact groups with only minor adjustments: sums become (Haar) integrals, matrices become integral operators, and the spectral theorem for compact self-adjoint operators supplies the discrete spectrum that makes "top eigenvalues" (and hence low-rank embeddings) meaningful.

One practical detail: in this section, we are intrinsically assuming that we have a bijection between a continuous group G and some set of words. In reality, one has a finite set of words which sample the continuous group. Hence, for theorem C.5, one forms predicted word embeddings by taking the eigenfunctions and sampling them at the group elements corresponding to the words, forming vectors.

C.1 HARMONIC ANALYSIS PRELIMINARIES

We let G be a compact Hausdorff topological group with normalized Haar measure dg ($\int_G dg = 1$), and work in the Hilbert space $L^2(G)$ with $\langle f_1, f_2 \rangle = \int_G f_1 \overline{f_2}\, dg$. Let $\widehat{G}$ be the (now countably infinite, in general) set of equivalence classes of unitary irreps $\rho : G \to U(d_\rho)$. The Peter–Weyl theorem states that the matrix coefficients $\{\sqrt{d_\rho}\, \rho_{ab} : \rho \in \widehat{G},\ 1 \le a, b \le d_\rho\}$ form a complete orthonormal basis of $L^2(G)$, with orthogonality

$$\int_G \rho_{ab}(g)\, \overline{\sigma_{ce}(g)}\, dg = \frac{1}{d_\rho}\, \delta_{\rho\sigma}\, \delta_{ac}\, \delta_{be}. \tag{13}$$

The generalized Fourier transform of some $C \in L^2(G)$ now takes the form

$$\widehat{C}(\rho) \;=\; \int_G C(g)\, \rho(g)\, dg \;\in\; \mathbb{C}^{d_\rho \times d_\rho}.$$

C.2 INTEGRAL OPERATOR AND THE DISCRETE SPECTRUM

We naturally extend the notion of the previous real symmetric matrix $M^\star$ to an integral operator $\mathcal{M}$. Namely, we first define the integral operator $\mathcal{M}$ on some $f \in L^2(G)$ by

$$(\mathcal{M}f)(g) \;:=\; \int_G k(g,h)\, f(h)\, dh$$

where $k \in L^2(G \times G)$

The reality and symmetry of $M^\star$ translate to requiring that k is both real and symmetric in its inputs. Finally, the statistical invariance of $M^\star$ with respect to some finite group G becomes the property that $\mathcal{M}$ commutes with left translation $L_b\ \forall b \in G$, where $(L_b f)(g) = f(b^{-1}g)$. I.e., $\mathcal{M}L_b = L_b\mathcal{M}$ $\forall b \in G$.

From these properties, $\mathcal{M}$ takes the form of

$$(\mathcal{M}f)(g) \;:=\; \int_G C(g^{-1}h)\, f(h)\, dh \tag{14}$$

for some real kernel $C \in L^2(G)$ with $C(g^{-1}) = C(g)$.

Lemma C.1 (Hilbert–Schmidt)**.** *For $C \in L^2(G)$, $\mathcal{M}$ is a Hilbert–Schmidt (hence compact) operator, with $\|\mathcal{M}\|_{\mathrm{HS}} = \|C\|_{L^2}$.*

Proof. By left invariance of Haar measure,

$$\|\mathcal{M}\|_{\mathrm{HS}}^2 = \int_G \int_G |C(g^{-1}h)|^2\, dh\, dg = \int_G \Big(\int_G |C(u)|^2 du \Big) dg = \|C\|_{L^2}^2 < \infty.$$

□

Lemma C.2 (Self-adjointness)**.** *If C is real with $C(g^{-1}) = C(g)$, then $\mathcal{M} = \mathcal{M}^*$.*

Proof. $\overline{k(g,h)} = \overline{C(h^{-1}g)} = \overline{C\big((g^{-1}h)^{-1}\big)} = C(g^{-1}h) = k(g,h).$ □

The properties of $\mathcal{M}$ being Hilbert-Schmidt and self-adjoint allow us to sensibly discuss its top eigenvalues as the well known result in Corollary C.3 describes.

Corollary C.3 (Discrete real spectrum)**.** *Let G be a compact group with normalized Haar measure, and let $C \in L^2(G)$ be real with $C(g^{-1}) = C(g)$. Then the operator $(\mathcal{M}f)(x) = \int_G C(x^{-1}y)\, f(y)\, dy$ on $L^2(G)$ is Hilbert–Schmidt (Lemma C.1) and self-adjoint (Lemma C.2). Consequently: $L^2(G)$ admits an orthonormal basis of eigenfunctions of $\mathcal{M}$. The eigenvalues $\{\mu_k\}$ are real with $\sum_k \mu_k^2 = \|C\|_{L^2}^2 < \infty$ and every nonzero eigenvalue has finite multiplicity and $\mu_k \to 0$.*

Lemma C.4 (Centering)**.** *The constant function $\mathbf{1}$ (already unit-norm) is an eigenfunction: $\mathcal{M}\mathbf{1} = \mu_0\mathbf{1}$ with $\mu_0 = \int_G C$. With $P = \mathrm{I} - \Pi_0$, where $(\Pi_0 f)(g) = \int_G f\, dh$ is the projection onto constants, $P\mathcal{M}P$ retains every eigenpair of $\mathcal{M}$ except that $(\mu_0, \mathbf{1})$ becomes $(0, \mathbf{1})$.*

Proof. Identical to Lemma B.2, with sums replaced by integrals. □

C.3 Embedding Theorem for Compact Groups

Lemma B.1 itself holds verbatim. Only the proof changes slightly as the sums become Haar integrals and we utilize the inversion invariance of the Haar measure. Hence, $\widehat{C}(\rho)$ is Hermitian, giving orthonormal eigenpairs $\widehat{C}(\rho)u_\rho^{(s)} = \lambda_\rho^{(s)} u_\rho^{(s)}$, $\lambda_\rho^{(s)} \in \mathbb{R}$.

Our setup allows us to state a theorem that is largely equivalent to the finite case.

Theorem C.5 (Diagonalization; compact case)**.** *Let G be a compact group with normalized Haar measure dg. Let integral operator $\mathcal{M}$ take the form as in Equation 14, so $\mathcal{M}$ has a discrete real spectrum from Corollary C.3. Let the kernel of $\mathcal{M}$ be $C : G \to \mathbb{R}$. Then $\widehat{C}(\rho)$ admits an orthonormal eigenbasis $\{u_\rho^{(s)}\}_{s=1}^{d_\rho} \subset \mathbb{C}^d$ with real eigenvalues $\lambda_\rho^{(s)}$ as mentioned above.*

Furthermore, define functions

$$\psi^{\rho,a,s}(g) \;=\; \sqrt{d_\rho}\,\big[\rho(g)\,u_\rho^{(s)}\big]_a, \qquad \rho \in \widehat{G}, \quad 1 \le a, s \le d_\rho,$$

Then:

(i) Each $\psi^{\rho,a,s}$ is an eigenfunction of $\mathcal{M}$ with eigenvalue $\lambda_\rho^{(s)}$.

(ii) The family $\{\psi^{\rho,a,s} : \rho \in \widehat{G},\ 1 \le a, s \le d_\rho\}$ is an orthonormal basis of $L^2(G)$.

(iii) Hence, the spectrum of $\mathcal{M}$ is $\{\lambda_\rho^{(s)} : \rho \in \widehat{G}, s \in [d_\rho]\} \cup \{0\}$, each nonzero eigenvalue occurring with finite multiplicity, at least d_ρ, and

$$P\mathcal{M}P \;=\; \sum_{\rho \neq \rho_{\mathrm{triv}}} \sum_{s=1}^{d_\rho} \lambda_\rho^{(s)} \sum_{a=1}^{d_\rho} \psi^{\rho,a,s}\,\langle \psi^{\rho,a,s}, \cdot \rangle,$$

where P is defined in C.4.

Proof. We spare writing every detail as the proof is largely a copy of the finite case. (i) The eigenfunction computation is the integral version of Theorem B.3(i). (ii) Orthonormality comes from (13) while completeness results from Peter–Weyl. (iii) follows immediately from (i) and (ii). □

Remark C.6 (Kernel is class function, compact case). The discussion in Appendix B.4 is largely the same in the compact case. Eigenvalues still take on multiplicities of d_ρ^2, and we can express our eigenfunctions as $\psi^{\rho,a,s} = \sqrt{d_\rho}\,\rho(g)_{a,b}$.

Remark C.7 (Realification, compact case). The discussion in Appendix B.5 is largely analogous. The Frobenius-Schur indicator becomes $\nu(\rho) = \int_G \chi_\rho(g^2)\, dg \in \{1, 0, -1\}$, and Proposition B.6 holds nearly verbatim. (Note that in the quaternionic case, Gram-Schmidt can be applied to functions as it works on inner product spaces in general, not just vectors)

D HOMOGENEOUS SPACES, GELFAND PAIRS, AND THE SPHERE

Our final generalization considers words in bijection with a homogeneous space of G. As mentioned at the start of Appendix C, assuming we have a word set in bijection with a continuous set is not realistic. In practice, we have a finite word set which samples a homogeneous space. We in turn sample our predicted eigenfunctions in Theorem D.1 at these points when making actual predictions.

D.1 SETUP

Let $K \leq G$ be a closed subgroup of the compact group G, and consider $X = G/K$, a homogeneous space with G-invariant probability measure dx. Words are identified with points of X.

As in Appendix B, we consider an operator $\mathcal{M}$ of the form

$$(\mathcal{M}f)(g) \;:=\; \int_G C(g^{-1}h)\, f(h)\, dh$$

where C is real, $C(g) = C(g^{-1})$, and $f \in L^2(G)$. However, we now require that C is bi-K-invariant: $C(k_1 g k_2) = C(g)$ for $k_1, k_2 \in K$ (so its value is independent of the particular representative of some coset gK). $\mathcal{M}$ descends to operating on $f \in L^2(X)$ through

$$(\mathcal{M}f)(y) := \int_X \kappa(y,x) f(x) dx \tag{15}$$

where $\kappa \in L^2(X \times X)$ is real symmetric and G-invariant ($\kappa(gx, gy) = \kappa(x,y)\ \forall g \in G$). We could equivalently replace κ with a kernel C defined by $C(g) = \kappa(eK, gK)\ \forall g \in G$. Note that C is immediately bi-K-invariant, and has the usual properties of being real and obeying $C(g) = C(g^{-1})\ \forall g \in G$. This gives us the alternative form of $\mathcal{M}$

$$(\mathcal{M}f)(gK) \;=\; \int_G C(g^{-1}h)\, f(hK)\, dh, \tag{16}$$

expressed instead as an integral over the group.

Let $\widehat{G}$ be the set of equivalence classes of unitary irreps of G, $\rho : G \to U(V_\rho)$. Recall that

$$L^2(G/K) \;\cong\; \bigoplus_{\rho \in \widehat{G}} V_\rho^{\oplus m_\rho}, \qquad m_\rho \;=\; \dim V_\rho^K, \tag{17}$$

where $V_\rho^K = \{v \in V_\rho : \rho(k)v = v\ \forall k \in K\}$ is the subspace of K-fixed vectors. An irrep contributes exactly as many copies of V_ρ as it has K-fixed directions.

D.2 EMBEDDING THEOREM FOR HOMOGENEOUS SPACES

Theorem D.1 (Diagonalization; homogeneous space)**.** *Let $K \leq G$ be a closed subgroup of compact group G and let $X = G/K$. Assume we have an integral operator $\mathcal{M}$ as in Equation 16 with kernel $C : G \to \mathbb{R}$. Let $\widehat{G}$ be the set of equivalence classes of unitary irreps of G, $\rho : G \to U(V_\rho)$. We define*

$$B_\rho := P_K \hat{C}(\rho) P_K$$

where $P_K = \int_K \rho(k) dk$ is the orthogonal projection of V_ρ onto V_ρ^K. It can be seen that B_ρ is Hermitian and preserves V_ρ^K, so its restriction to V_ρ^K admits an orthonormal eigenbasis $\{u_\rho^s\}_{s=1}^{m_\rho} \subset V_\rho^k$ with real eigenvalues $\lambda_\rho^{(s)}$. Furthermore, define functions

$$\psi^{\rho,a,s}(gK) \;=\; \sqrt{d_\rho}\, \left[\rho(g)\, u_\rho^{(s)}\right]_a, \qquad \rho \in \widehat{G}, \quad 1 \leq a \leq d_\rho,\ 1 \leq s \leq m_\rho$$

where $m_\rho = \dim V_\rho^K$. Then:

(i) $\psi^{\rho,a,s}$ are well-defined on X.

(ii) Each $\psi^{\rho,a,s}$ is an eigenfunction of $\mathcal{M}$ with eigenvalue $\lambda_\rho^{(s)}$.

(iii) The family $\{\psi^{\rho,a,s} : \rho \in \widehat{G},\ 1 \le a \le d_\rho,\ 1 \le s \le m_\rho\}$ is an orthonormal basis of $L^2(X)$.

(iv) Hence, the spectrum of $\mathcal{M}$ on $L^2(X)$ is $\{\lambda_\rho^{(s)} : \rho \in \widehat{G},\ s \in [m_\rho]\} \cup \{0\}$; each nonzero eigenvalue occurs with finite multiplicity, at least d_ρ, and

$$P\mathcal{M}P \;=\; \sum_{\substack{\rho \neq \rho_{\mathrm{triv}} \\ m_\rho \neq 0}} \sum_{s=1}^{m_\rho} \lambda_\rho^{(s)} \sum_{a=1}^{d_\rho} \psi^{\rho,a,s}\, \langle \psi^{\rho,a,s}, \cdot \rangle,$$

where P is defined in C.4.

Proof. (i) For any $k \in K$,

$$\psi^{\rho,a,s}(gkK) = \sqrt{d_\rho}\,\big[\rho(g)\,\rho(k)\,u_\rho^{(s)}\big]_a = \sqrt{d_\rho}\,\big[\rho(g)\,u_\rho^{(s)}\big]_a = \psi^{\rho,a,s}(gK),$$

using $u_\rho^{(s)} \in V_\rho^K$.

(ii) The computation of Theorem B.3 applies in a similar fashion: substituting $h = gk$,

$$(\mathcal{M}\psi^{\rho,a,s})(gK) = \int_G C(g^{-1}h)\,\psi^{\rho,a,s}(hK)\,dh = \sqrt{d_\rho}\,\Big[\rho(g)\,\widehat{C}(\rho)\,u_\rho^{(s)}\Big]_a.$$

Since $u_\rho^{(s)} \in V_\rho^K$ and $\widehat{C}(\rho)$ agrees with B_ρ on V_ρ^K (as $P_K u_\rho^{(s)} = u_\rho^{(s)}$ and $P_K \widehat{C}(\rho) = \widehat{C}(\rho)$ by bi-K-invariance),

$$\widehat{C}(\rho)\,u_\rho^{(s)} = B_\rho\,u_\rho^{(s)} = \lambda_\rho^{(s)}\,u_\rho^{(s)},$$

hence $\mathcal{M}\psi^{\rho,a,s} = \lambda_\rho^{(s)}\,\psi^{\rho,a,s}$.

(iii) Orthonormality: under the identification of $L^2(X)$ with the right-K-invariant functions in $L^2(G)$, the inner products may be computed on G, where Schur orthogonality gives

$$\langle \psi^{\rho,a,s}, \psi^{\sigma,c,t} \rangle_{L^2(X)} = \langle \psi^{\rho,a,s}, \psi^{\sigma,c,t} \rangle_{L^2(G)} = \delta_{\rho\sigma}\,\delta_{ac}\,\langle u_\rho^{(t)}, u_\rho^{(s)} \rangle_{V_\rho} = \delta_{\rho\sigma}\,\delta_{ac}\,\delta_{st}.$$

Completeness: by (17), the ρ-isotypic component of $L^2(X)$ has dimension $d_\rho\, m_\rho$. The functions $\{\psi^{\rho,a,s} : 1 \le a \le d_\rho,\ 1 \le s \le m_\rho\}$ are $d_\rho m_\rho$ orthonormal elements of this component and therefore span it. Taking the union over $\rho \in \widehat{G}$ gives completeness.

(iv) This result follows directly from (i)-(iii). We note that the trivial irrep is dropped due to centering as usual. □

Note that our discussion of realification as in Appendix B.5 applies largely unchanged for homogeneous spaces, just as for the compact group case.

D.3 Gelfand pairs

Definition D.2. (G, K) for compact G is a *Gelfand pair* if $m_\rho = \dim V_\rho^K \le 1$ for every $\rho \in \widehat{G}$, i.e. the decomposition (17) is *multiplicity-free*.

Corollary D.3 (Diagonalization; Gelfand pair)**.** *Assume (G, K) is a Gelfand pair with $m_p = \dim V_\rho^K \le 1$ for every $\rho \in \widehat{G}$. Theorem D.1 can be simplified. For each ρ with $m_\rho = 1$, fix a unit $u_\rho \in V_\rho^K$. Then B_ρ becomes the 1×1 scalar*

$$\lambda_\rho \;=\; \int_G C(g)\,\varphi_\rho(g)\,dg, \qquad \varphi_\rho(g) := \langle \rho(g)\,u_\rho, u_\rho \rangle.$$

The spectrum of $\mathcal{M}$ is the scalar sequence $\{\lambda_\rho\}$, each with eigenspace of dimension at leaset d_ρ, spanned by $\psi^{\rho,a}(gK) = \sqrt{d_\rho}\,[\rho(g)u_\rho]_a$. The functions φ_ρ are often denoted as spherical functions.

Proof. Since $\dim V_\rho^K = 1$, the restriction of B_ρ to $V_\rho^K = \mathbb{C}\,u_\rho$ is multiplication by the scalar $\lambda_\rho = \langle B_\rho u_\rho, u_\rho \rangle$. As $P_K u_\rho = u_\rho$ and P_K is self-adjoint, $\langle B_\rho u_\rho, u_\rho \rangle = \langle \widehat{C}(\rho)\,u_\rho, u_\rho \rangle$. Hence,

$$\lambda_\rho = \int_G C(g) \langle \rho(g) u_\rho, u_\rho \rangle dg = \int_G C(g) \varphi_\rho(g) dg$$

□

D.4 Special Case: The sphere

We consider the particular case of the sphere which is relevant to our experiments. Take $G = SO(3)$, $K = SO(2)$, $X = G/K = S^2$. We state 3 well known facts about the sphere (Dai & Xu, 2013).

(1) The irreps of $SO(3)$: ρ_ℓ $\ell \geq 0$, are real and have dimension $2\ell + 1$, each with $\dim V_{\rho_\ell}^K = 1$. Hence, $(SO(3), SO(2))$ is a Gelfand pair.

(2) L^2 functions on the sphere have the isotypic decomposition $L^2(S^2) = \bigoplus_\ell \mathcal{H}_\ell$, where $\mathcal{H}_\ell$ are the degree-ℓ spherical harmonics. Alternatively, for each $\ell \geq 0$, let

$$\Psi_\ell \;:=\; \operatorname{span}\left\{\psi^{\rho_\ell, a} \;:\; 1 \leq a \leq 2\ell + 1\right\} \;\subset\; L^2(S^2)$$

span of the modes of Corollary D.3 for the irrep ρ_ℓ. We must have that $\mathcal{H}_\ell = \Psi_\ell$ as they correspond to the same isotypic component. Hence, instead of using $\psi^{\rho,a}(gK)$ as our orthonormal basis, we can instead use the real degree ℓ spherical harmonics $\{Y_\ell^m\}_{\ell \geq 0, |m| \leq \ell}$.

(3) The spherical functions in Corollary D.3 become Legendre polynomials, $\varphi_{\rho_\ell}(g) = P_\ell(\cos\theta_g)$ where $\cos\theta_g = \langle g x_0, x_0 \rangle$, $x_0 = eK \in S^2$. The inner product is done with respect to the points on the sphere realized in R^3.

Finally, note that the kernel $\kappa(x, y)$ on S^2 is invariant under $G = SO(3)$ iff $\kappa(x, y) = c(\langle x, y \rangle)$, i.e., the kernel is only some function c of the inner product between two points. Therefore, recall that our alternative form of the kernel C obeys $C(g) = \kappa(eK, gK)$. Hence $C(g) = c(\langle x_0, gK \rangle) = c(\cos\theta_g)$. Altogether, applying Corollary D.3 gives us that

$$\mathcal{M} Y = \lambda_\ell Y \quad \forall Y \in \mathcal{H}_\ell$$

where λ_ℓ is given by

$$\lambda_\ell = \int_G C(g)\varphi_\rho(g) dg = \int_G c(\cos\theta_g) P_\ell(\cos\theta_g) dg =$$
$$\frac{1}{4\pi} \int_0^{2\pi} \int_0^{\pi} c(\cos\theta) P_\ell(\cos\theta) \sin\theta \, d\theta \, d\phi = \frac{1}{2} \int_{-1}^{1} c(t) P_\ell(t) dt$$

and each λ_ℓ has multiplicity as least $2\ell + 1$. (Centering, i.e. considering $P\mathcal{M}P$, will set $\lambda_0 = 0$)

E Months: experimental details

Corpus. `text8` (Mahoney, 2011): the first 10^8 bytes of cleaned English Wikipedia (17,005,207 whitespace tokens; lowercase, punctuation stripped, digits spelled out). Counts of 2,771–3,650 per month after filtering "may" (see below). Using this corpus, we form the normalized co-occurrence matrix $M^\star \in \mathbb{R}^{12\times 12}$ as in equation 5 (settings are $B = 16$, $f(\delta) = B{+}1{-}\delta$). We factorize $M^\star$ to form our empirical word embeddings (centering removes the constant mode). Visually, The leading PCs are circular: PCs 1–2 are the $k{=}1$ circle ($\lambda = 2.17, 1.59$), while PCs 3–4 are $k{=}2$ circle ($0.88, 0.76$).

Sense disambiguation for "may". Raw counts: 19,115 for "may", $\sim 6\times$ the other months, dominated by the verb, which inflates the unigram P_x in Eq. 1. Fortunately, we can roughly select cases where "may" is used as a month. Namely, we keep an occurrence of "may" iff its two following tokens are both digits, or its preceding token is a digit, or an adjacent token is another month. This retains 3,129 occurrences (16.4%), in line with the other months. This improves our per-month cosine prediction from 0.91 to 0.96.

Wavenumber assignment/purity (the coloring of Figure 2 A). Define the vectors $c_k := \big(\cos(2\pi k t/12)\big)_{t=0}^{11}$ and $s_k := \big(\sin(2\pi k t/12)\big)_{t=0}^{11}$ in $\mathbb{R}^{12}$, and let $F_k := \operatorname{span}\{c_k, s_k\}$ be the "Fourier plane" of wavenumber $k = 1, \ldots, 5$ (for $k = 6$, the one-dimensional span of the alternating vector c_6). Denote P_k as the orthogonal projector onto F_k. These subspaces orthogonally decompose the centered space, so for any unit eigenvector v of $PM^\star P$ the "energies" $\|P_k v\|^2$ sum to 1 over $k = 1, \ldots, 6$. We assign v the wavenumber capturing the most energy, $k^*(v) = \arg\max_k \|P_k v\|^2$, and call the winning energy its *purity*. $\|P_{k^*} v\|^2 \in [0, 1]$: 1 is perfectly in one of these subspaces, while a uniformly random unit direction scores $2/12 \approx 0.17$ in expectation ($1/12$ for $k = 6$). Leading-mode purities: $0.99, 0.86$ ($k{=}1$ pair) and $0.92, 0.88$ ($k{=}2$ pair).

Corpus Bootstrap (errors bars of Figure 2 A). The corpus is divided into 40 contiguous segments ($\sim$425k tokens each). A resample draws 40 of these segments *with replacement*, forming a new $M^\star$ from which new eigenvalues can be computed. 300 resamples give percentile (2.5%–97.5%) intervals.

Theory vs. Empirical (Figure 2 C). Our theoretical predictions take the form as in Appendix B.6, using $M^\star$ to estimate the kernel C (by averaging over appropriate elements). We align our predictions to each empirical plane with by solving an orthogonal Procrustes problem. The first PCs align with our theoretical predictions at a per month cosine of 0.96 (median 0.97). A month label-shuffle null has mean per-month cosine of 0.35 (95% 0.53) across 500 shuffles.

F MUSIC: EXPERIMENTAL DETAILS

Corpus and chord extraction. All 371 Bach chorales from the `music21` corpus (Cuthbert & Ariza, 2010). Scores are "chordified" into beat-level sonorities. Each sonority is reduced to its pitch-class set and matched against the 24 major/minor triads (inversions and octave doublings identified; sevenths and other sonorities retained as their own types). This yields 24,374 chord tokens over a vocabulary of 92 types, of which the 24 triads (78% of tokens) are the subsystem under consideration. We form our $M^\star \in \mathbb{R}^{24\times 24}$ from this corpus as in equation 5 (settings are $B = 3$, $f(\delta) = 1$, $\alpha = 0.5$).

Accessing quality of statistical hypotheses. Suppose we have a group G' (which acts on the 24 triads) under which we suspect that $M^\star$ could be approximately statistically invariant. We can score how well this hypothesis holds by first constructing the projection of $M^\star$ onto matrices that are exactly statistically invariant, $\Pi_{G'} M^\star_{x,y} = \frac{1}{|G'|}\sum_{g\in G'} M^\star_{gx,\,gy}$, setting matrix elements which would be equal to their average. In general, a hypothesis need not be a group; it must only specify which elements of $M^\star$ would be equal in an ideal situation. We then compute our score

$$R^2 \;=\; \frac{\sum_c n_c \left(\overline{M^\star}_c - \overline{M^\star}\right)^2}{\sum_{x,y}\left(M^\star_{xy} - \overline{M^\star}\right)^2},$$

where c runs over the equivalence classes of pairs under the candidate invariance, n_c is the number of entries in class c, $\overline{M^\star}_c$ is the class mean of equivalence class c, and $\overline{M^\star}$ is the grand mean over all entries. In other words, this is the fraction of the entry-wise variance of $M^\star$ explained by the class-wide averaging.

The T/I group obtains an R^2 of 0.822. An illustration of its performance is shown in Figure 3 A. None of 2000 random relabeling of the chords are able to surpass this value. The Z_{12} hypothesis, which is the same as T/I but with only transposition-invariance, scores barely higher at 0.830. It is impossible for T/I to surpass $\mathbb{Z}_{12}$ as it has additional restrictions, yet it performs essentially the same. This is strong evidence that the T/I symmetry is valid. Other hypotheses, as defined in Table 1 score notably worse ($\mathbb{Z}_{24}$: 0.699, PLR: 0.442, $\mathbb{Z}_{12}\times\mathbb{Z}_{12}$: 0.440, latent $\mathbb{Z}_{12}$: 0.632, T/I conjugacy classes: 0.435). Notably the T/I conjugacy class hypothesis, which adds even more restrictions to the T/I hypothesis, scores much worse, suggesting that T/I is the sharpest symmetry that is reliably present in the data.

Isotypic energy. The T/I action on the 24 triads is a permutation representation $L(g) \in \mathbb{R}^{24\times 24}$ (the regular representation acting on $\mathbb{C}^{24}$). By the canonical Peter–Weyl decomposition of the regular representation, $\mathbb{C}^{24} \cong \bigoplus_\rho V_\rho^{\oplus d_\rho}$: every irrep appears, with multiplicity equal to its dimension. In other words, the space $\mathbb{C}^{24}$ can be canonically decomposed into "isotypic" subspaces corresponding to each given irrep of T/I. The orthogonal projector onto the ρ-component is given by

$$P_\rho \;=\; \frac{d_\rho}{24}\sum_{g\in T/I} \overline{\chi_\rho(g)}\, L(g).$$

See Serre (1977) for more details on orthogonal projectors onto isotypic components. For T/I acting on triads, the two-dimensional irreps E_h, $h \in [5]$, each have isotypic dimension $d_\rho^2 = 4$. The *isotypic energy* of an empirical plane (defined by orthonormal v_1, v_2) in irrep ρ is defined as

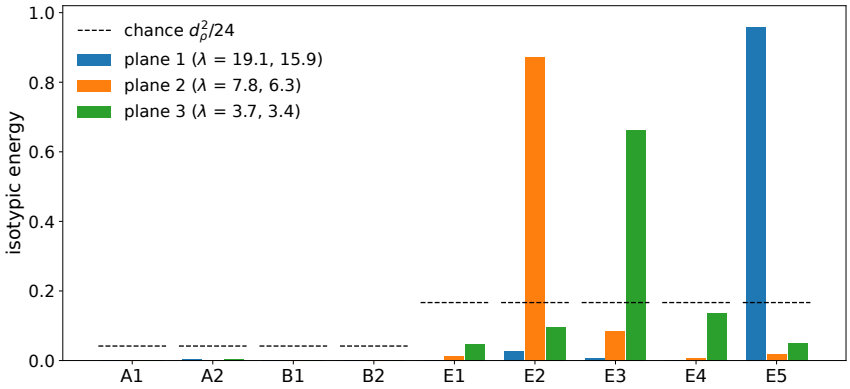


Figure 6: Irrep content of the top three embedding planes: we consider the first six modes of the centered $M^\star$. One-dimensional irreps are A_1, A_2, B_1, B_2 and two-dimensional irreps are $E_1, \dots E_5$. A_1 corresponds to the trivial irrep which centering automatically removes. We compute the isotypic energy of each plane for each irrep. One irrep dominates per plane (E_5, E_2, E_3; energies $0.958, 0.873, 0.662$ vs. chance $1/6$). Eigenvalues of centered $M^\star$ are listed in legend; notice they approximately pair.

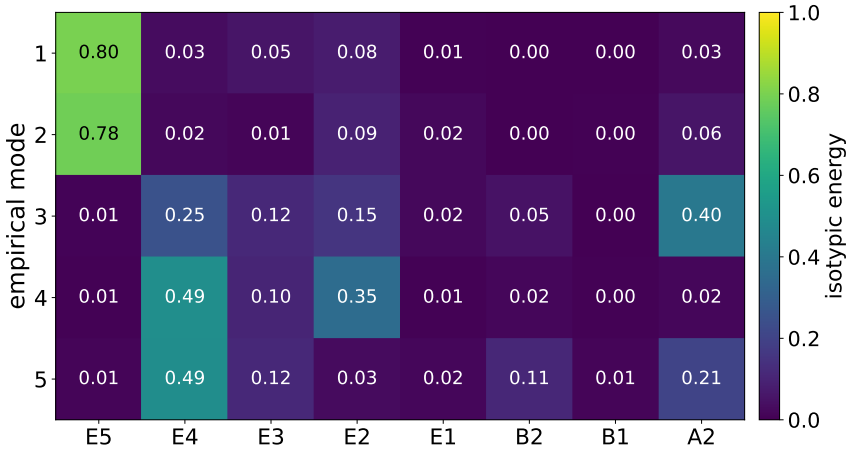


Figure 7: Irrep content of ChordBERT layer 3 PCs: one-dimensional irreps are B_1, B_2, A_2, and two-dimensional irreps are E_1, E_2, E_3, E_4, E_5. We do not show the remaining trivial irrep A_1 as it is removed automatically from centering. The first 2 PCs demonstrate high isotypic energy in E_5.

$\frac{1}{2}(\|P_\rho v_1\|^2 + \|P_\rho v_2\|^2) \in [0,1]$ with chance level for a random plane $d_\rho^2/24$ (1/24 and 1/6 for one and two-dimensional irreps respectively). Note that the isotypic energy is invariant to the choice of v_1, v_2. For computing the isotypic energy of a single vector v, we compute $\|P_\rho v\|^2 \in [0,1]$ instead. See Figure 6 and Figure 7 for the isotypic energies corresponding to Bach chorales and ChordBERT layer 3 respectively.

ChordBERT activations We input the 371 Bach Chorales separately into ChordBERT (He et al. (2026)) and in layer 3's residual stream average the activations corresponding to each of the 24 triads over every single one of its occurrences. We focus on layer 3 in particular as its first 2 modes have very high isotypic energy in irrep E_5 (Figure 7), suggesting the potential presence of a circle of fifths (which does indeed emerge, Figure 3 B).

Theory vs. Empirical (Figure 4 B) For the left subplot we use the $M^\star$ formed from the 371 Bach chorales to form empirical embeddings. On the right, we instead consider the averaged embeddings per triad in layer 3 of ChordBERT, as discussed above. In place of $M^\star$ we consider the Gram matrix across the centered embeddings ($\in \mathbb{R}^{24\times 24}$), normalized by the root mean square of its diagonal.

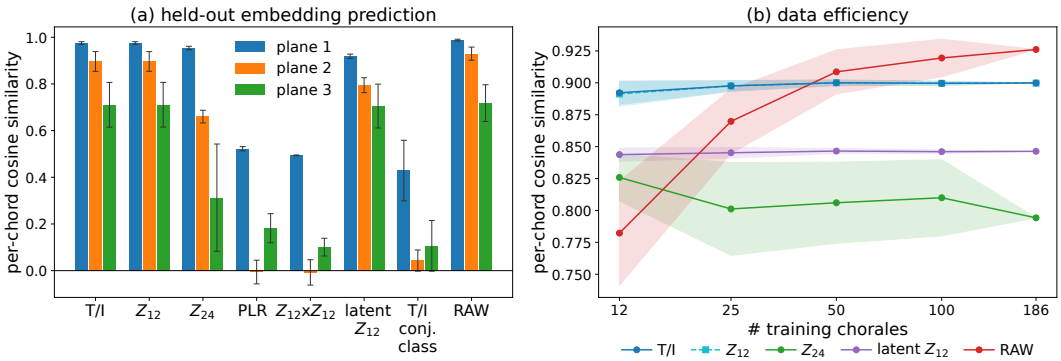


Figure 8: (a) Held-out embedding prediction per plane for the eight hypotheses of Table 1: mean $\pm$ sample standard deviation over 10 random splits of the chorales into disjoint training/test halves. (b) Data efficiency: Training sets are nested subsets drawn without replacement from one fixed half of the 371 chorales, the other half serving as a fixed test set. Each point averages 15 redraws (shaded: ± 1 sample standard deviation), except the rightmost, where the training set is the entire pool and no redraw is possible. The 19-parameter T/I hypothesis converges by $\sim$12 chorales. The 300-parameter unconstrained model (RAW) needs $\approx 4\times$ more data to match it.

Our theoretical predictions take the form as in Appendix $B.6$ in both subplots, with the kernel C being estimated from $M^\star$ or the Gram matrix respectively (by averaging over appropriate elements). A fitting orthogonal transformation aligns the theoretical predictions with the empirical embeddings in both cases. The per-chord cosine is 0.98 (median 0.99) for the chorales and 0.91 (median 0.95) for ChordBERT. A chord relabeling null with 500 permutations gives 0.40 (95% 0.56) and 0.29 (95% 0.48) respectively in each case.

Furthermore, recall from Figure 6 that the first three planes of the Bach chorale embeddings all have high isotypic energy in their respective 2D irreps. This suggests comparing our theoretical predictions against the first 6 dimensions of the empirical embeddings from $M^\star$, using one fitting orthogonal transformation per plane. We obtain a per-chord cosine similarity of 0.95 (median 0.96) overall with per-plane values of 0.98, 0.94, and 0.80, respectively. Our theory certainly holds predictive power beyond the first 2 dimensions, but weakens as we look as less prevalent modes.

Model comparison protocol. Chorales are split into disjoint halves (10 random splits). Each hypothesis is fit on the training half. Namely, the empirical $M^\star$ is projected onto a matrix which satisfies the hypothesis G', i.e., $\Pi_{G}^{'} M^\star$ as defined above (the RAW model uses the training $M^\star$ itself without projection). $\Pi_{G}^{'} M^\star$ then predicts the *held-out* half's embeddings. Namely, the first 6 modes of the centered $M^\star$ (of the held out data) are aligned with the modes of $\Pi_{G'} M^\star$ (from the training data) using one orthogonal transform per plane. Per-chord cosines are then computed. Results for various hypotheses (Table 1) are shown in Figure 8 A. The data itself without any imposed invariances,"RAW", performs the best by a small margin. However, $\mathbb{Z}_{12}$ and T/I don't do much worse. As mentioned before, the fact that T/I performs about as well as $\mathbb{Z}_{12}$, when it has more restrictions and less independent parameters, is a great indication that T/I is a natural symmetry of the data. Figure 8 B performs a similar analysis as in A, but instead refits on nested subsets with a fixed held out test set. T/I attains a cosine similarity with the held out data of 0.892 at 12 chorales, the best performance of all models at this point; RAW requires around 50 training chorales to match.

Changing window size for constructing $M^\star$. As seen in Figure 9, we construct $M^\star$ and consider the corresponding embeddings for a variety of window sizes. We see that the fitted rotation angle and isotypic energy remain largely consistent. Plane 3's values change somewhat, but this kind of behavior is understandable in weaker modes of the data. Finally, we also consider the eigenvalues of the centered $M^\star$ in Figure 9 C. While rotation angle and isotypic energy are largely functions of the statistical symmetries of the data, the eigenvalues are very dependent on the corpus; corpora that have the same symmetries can still have wildly different eigenvalues. Hence, it is not too surprising to see that the eigenvalues have noticeably more drift than the other considered quantities.

Model	Params	Indep.	Invariance assumed on $M^\star$ over the 24 triads
T/I	24	19	Statistical invariance under T/I.
$\mathbb{Z}_{12}$	48	26	Statistical invariance under only transposition of T/I.
$\mathbb{Z}_{24}$	24	13	Invariance under a "transposition" that moves along both majors and minors as in the circular order seen in Figure 3 B. This hypothesis predicts embeddings where the major and minor circles are perfectly inter-spaced.
PLR	24	19	Statistical invariance under the classic PLR group of music theory
$\mathbb{Z}_{12} \times \mathbb{Z}_{12}$	26	15	Statistical invariance under transposition, but separately for both majors and minors. (let $m, m+1, m+2$ be sequential major chord indices while $n, n+1, n+2$ correspond to minors) E.g., $M^\star_{m,m+1} = M^\star_{m+1,m+2}$ and $M^\star_{n,n+1} = M^\star_{n+1,n+2}$ but in general we have $M^\star_{m,m+1} \neq M^\star_{n+1,n+2}$) Major and minor cross terms of $M^\star$ are set to one constant. This hypothesis essentially assumes independent circular structures in majors and minors.
latent $\mathbb{Z}_{12}$	12	7	Majors are identified with their relative minors (music theory term, this identification looked reasonable from Figure 3 B), and invariance under transposition is assumed. This hypothesis corresponds to the embeddings forming a single circle with majors on top of minors.
T/I conjugacy class	9	9	We begin with statistical invariance under T/I. Each equivalence class corresponds to a group element. We add the restriction that if two or more of these group elements correspond to the same conjugacy class of T/I (different type of class, group theoretic concept), then the equivalence classes are combined.
RAW	576	300	No invariance, the raw matrix

Table 1: The eight statistical hypotheses of the music model comparison. Each hypothesis is applied to $M^\star$ by setting elements of $M^\star$ which would be equal under the hypothesis to their average (Except RAW which is $M^\star$ itself). *Params* counts the model's number of equivalence classes without considering the symmetry of $M^\star$. *Indep.* counts the number of free parameters once this symmetry merges equivalence classes together.

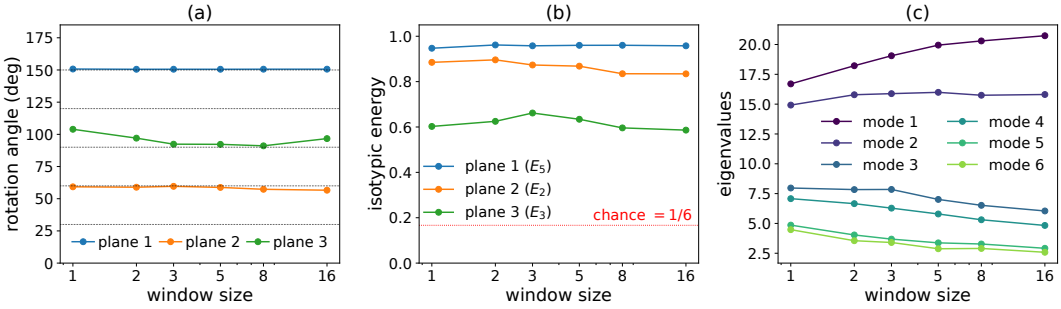


Figure 9: We use various window sizes for constructing the centered $M^\star$ (the original size was 3) which we then use to form 6-dimensional chord embeddings. (a) Rotation angles are computed by fitting a rotation matrix to each pair of dimensions (plane) of the embeddings; we consider the angle between chords in chromatic order separately for both the majors and minors. (b) Isotypic energy of each plane in its respective irrep. Chance level is always $d_\rho^2/24 = 1/6$ for the considered irreps. (c) eigenvalues of the centered $M^\star$.

G SKY: EXPERIMENTAL DETAILS

LLM activations Released data of Berdnikov & Liokumovich (2026): 188 named sky objects (stars, constellations, etc.) $\times$ 25 proximity-style prompts, residual-stream activations at all layers exported as the top-16 PCA coordinates per layer. We use mean activations across the prompts for

each layer. The layer from which the true celestial positions can be linearly decoded the best is selected as the working layer (layer 72 of 89 for Mistral Large 2). In place of $M^\star$, we consider Gram matrix ($\in \mathbb{R}^{188\times 188}$) of the sky objects' activations, normalized by the root mean square of its diagonal. We will denote this matrix as G_{Gram} for convenience. We will also use 188 and n interchangeably in this section.

Need for weighted estimators. Our theory in Appendix D intrinsically assumes a continuous setting in which an infinite set of words are in bijection with the sphere. Of course, this is not practical. In reality, we are discretizing the theory by sampling points on S^2 which correspond to sky objects on the celestial sphere. If all celestial objects were sampled uniformly on the sphere, no weighted estimators would be required. However, the celestial objects we consider sample the sphere in a "lightly" non-uniform fashion. Acceptable estimators could still be obtained by ignoring this small bias, but we decide to be proactive in counteracting it.

Inverse density weighting Essentially all estimators in this section incorporate inverse density-weighted. We fit a density $\hat{\rho}$ to our 188 objects using an on-sphere Gaussian kernel density estimator (KDE). Namely, let $x_1, \ldots, x_n \in S^2$ be the sky-object positions and $\theta_{ij} = \arccos(\langle x_i, x_j\rangle)$ the great-circle angle. With $h = 0.35$ rad, the on-sphere Gaussian KDE is

$$\hat{\rho}(x_i) \;=\; \sum_{j=1}^{n} \exp\!\left(-\frac{\theta_{ij}^2}{2h^2}\right).$$

A weight w_i is now associated with each object i. These weights inversely related to $\hat{\rho}$, and are renormalized to sum to 1:

$$\tilde{w}_i = \frac{1}{\hat{\rho}(x_i)}, \qquad w_i = \frac{\tilde{w}_i}{\sum_{k=1}^{n} \tilde{w}_k}.$$

Hence, areas of the sphere which were previously over or under-sampled are counterbalanced.

Empirical Word Embeddings The integral operator in Appendix D on some $f \in L^2(S)$ discretizes as $(\mathcal{M}f) \approx \sum_j M^\star_{ij} f(x_j)$, where the typical normalized co-occurence matrix $M^\star$ emerges. However, with reweighting, this is lightly modified to $\sum_j M^\star_{ij} w_j f(x_j)$. Something else affected by reweighting is centering. The previous $P = I - \mathbf{1}\mathbf{1}^T/n$ is modified to $P_w = I - \mathbf{1}w^T$.

Altogether, for computing empirical embeddings, using say the $M^\star$ from our book corpus, we solve generalized eigenproblem

$$P_w M^\star W P_w v = \lambda v \tag{18}$$

with diagonal $W \in R^{188\times 188}$ and $W_{ii} = w_i$.

Regressions All regressions in this section incorporate reweighting.

(i) Kernel estimation (LLMs would use G_{Gram} in place of the book corpus's $M^\star$) The kernel C (as in Theorem 4.2) is the weighted estimate of $\mathbb{E}[M^\star \mid t]$,

$$\tilde{C}(t) \;=\; \frac{\sum_{i<j} w_i w_j\, K_h(t - t_{ij})\, M^\star_{ij}}{\sum_{i<j} w_i w_j\, K_h(t - t_{ij})}, \qquad t_{ij} = \langle x_i, x_j\rangle,$$

where $t_{i,j}$ is the true cosine between two objects x_i, x_j on the celestial sphere. K_h is a 1D Gaussian with standard deviation $h = 0.08$. Note, the kernel as described here is a function of cosine similarity while in the main paper it is a function of angle. The two forms are interchangeable.

The local standard deviation of this fit is computed as follows. Let

$$\kappa_{ij}(t) \;=\; w_i w_j\, K_h(t - t_{ij})$$

For convenience, write

$$\hat{m}_1(t) = \tilde{C}(t), \qquad \hat{m}_2(t) \;=\; \frac{\sum_{i<j} \kappa_{ij}(t)\, (M^\star_{ij})^2}{\sum_{i<j} \kappa_{ij}(t)}.$$

The local standard deviation of the kernel fit is then

$$\hat{\sigma}(t) \;=\; \sqrt{\max\big(\hat{m}_2(t) - \hat{m}_1(t)^2,\, 0\big)},$$

i.e. the estimate of $\sqrt{\mathrm{Var}(M^\star_{ij} \mid t_{ij} = t)}$.

(ii) Decoding: Consider the unit vectors $x_i \in \mathbb{R}^3$ representing the n celestial objects. Suppose we are trying to linearly decode these positions from some $U \in \mathbb{R}^{n\times k}$, e.g., the residual stream of any of the LLMs. Let $\hat{B} = \arg\min_B \sum_i w_i \|x_i - B^\top U_i\|^2$ where $B \in \mathbb{R}^{k\times 3}$ and $U_i \in \mathbb{R}^k$. We report the quality of the fit of $\hat{B}$ as

$$R^2_{\mathrm{dec}} \;=\; 1 - \frac{\sum_{i=1}^n w_i \|x_i - \hat{B}^\top U_i\|^2}{\sum_{i=1}^n w_i \|x_i - \bar{x}_w\|^2},$$

where $\overline{x}_w = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i}$. This metric chooses the working layer of each model where U become the centered activations of each given LLM (Table 1).

(iii) Gram zonality: (books would use $M^\star$ in place of G_{Gram} in this section). Write $t_{ij} = \langle x_i, x_j\rangle$ and $\theta_{ij} = \arccos(t_{ij})$. Using only upper-triangle pairs $i < j$ with weights $\omega_{ij} = w_i w_j$, partition $\theta \in [0, \pi]$ into $B = 18$ equal bins. In bin b set

$$\hat{\mu}_b \;=\; \frac{\sum_{(i,j)\in b} \omega_{ij}\, M_{ij}}{\sum_{(i,j)\in b} \omega_{ij}}$$

With $\hat{M}_{ij} = \hat{\mu}_{b(ij)}$ and $\bar{M} = \sum_{i<j} \omega_{ij} M_{ij} \big/ \sum_{i<j} \omega_{ij}$,

$$R^2_{\mathrm{zonal}} \;=\; 1 - \frac{\sum_{i<j} \omega_{ij}(M_{ij} - \hat{M}_{ij})^2}{\sum_{i<j} \omega_{ij}(M_{ij} - \bar{M})^2}.$$

A permutation null (shuffle object labels, keep $\{M_{ij}\}$, refit the bin means) supplies a chance baseline for R^2_{zonal}.

Sampling Spherical Harmonics Sampled real spherical harmonics $(Y^m_\ell(x_i))_{i=1}^n$ are only approximately orthonormal at finite N. We orthonormalize them with Gram-Schmidt in the weighted inner product $\langle\cdot,\cdot\rangle_w$ ($\langle u, v\rangle_w = \sum_{i=1}^n w_i u_i v_j$ where $u, v \in \mathbb{R}^n$.) We conduct this degree by degree, beginning with $\ell = 1$ and moving upwards (ending for whatever degree is high enough for our purposes). This yields a matrix $Q_\ell \in \mathbb{R}^{n\times(2\ell+1)}$ for each ℓ, whose columns are an orthonormal basis of the sampled degree-ℓ subspace, with $Q_\ell^\top Q_{\ell'} = 0$ for $\ell \neq \ell'$ by construction. These Q_ℓ are the orthonormalized spherical harmonics we show as our theoretical predictions in Figure 4 C.

Degree energies (Figure 4 B): Consider a unit eigenvector v of the weighted eigenproblem in equation 18. The "degree-ℓ" energy of v, $E_\ell(v)$, is the squared orthogonal projection onto that subspace,

$$E_\ell(v) \;=\; \|Q_\ell^\top v\|^2 \;=\; \sum_{m=-\ell}^{\ell} \langle q_{\ell m}, v\rangle^2 \;\in\; [0, 1],$$

computed for $\ell = 0, \ldots, 4$. The vector's assigned degree is $\ell^*(\tilde{u}) = \arg\max_\ell E_\ell(\tilde{u})$, and its *purity* in that degree is $E_{\ell^*}(\tilde{u})$. This is the color bar of 4 B. Note that the considered degrees in this figure consider $\sum_{\ell\le 4}(2\ell+1) = 25$ of the $n = 188$ dimensions, so energies need not sum to 1. The chance energy level for a random direction is $(2\ell+1)/n$ ($\approx 0.016/0.027$ for $\ell = 1, 2$), so purities of 0.5–0.9 sit two orders of magnitude above chance. Finally, note that $E_0(\tilde{u}) = 0$ exactly due to centering.

Selection of Star-Guide Corpus Candidate pool: 24 public-domain astronomy books (Project Gutenberg, $\sim$1.6M words). Not every astronomy book carries the sky's geometry in its statistics. Many are organized by mythology, best telescope practices, or by a handful of famous constellations. A minority are written as a general *tour of the sky*, directing the reader from object to nearby object with co-mentions largely determined by geometrical positions. The metric below selects for this property.

Model	layer	R^2_{zonal}	R^2_{dec}	canon^2 (top-3)	degree sequence
Qwen3-235B	90	0.581	0.928	0.94	1 1 1 1 2 2 2 2 2
Mistral-Large-123B	72	0.490	0.905	0.91	1 1 1 1 2 2 2 2 3
GLM-4.5-Air	46	0.484	0.904	0.91	1 1 1 1 2 2 2 2 2
Qwen3-32B	62	0.440	0.865	0.87	1 1 1 2 1 2 2 2 3
Llama-3.3-70B	63	0.395	0.898	0.91	1 1 1 2 1 2 2 3 3
Mixtral-8x22B	56	0.238	0.836	0.84	1 1 1 2 1 2 2 4
GPT-OSS-120B	31	0.072	0.634	0.64	scrambled

Table 2: Mistral-Large-123B is the same as Mistral Large 2. Each model is evaluated at the layer where celestial positions on the sphere ($\in \mathbb{R}^3$) can be linearly decoded best from the activations (R^2_{dec}). R^2_{zonal}: the fraction of the activation Gram matrix structure explained by true angular separation alone. Canon^2 (top-3): average squared canonical cosines between the three leading activation modes and the dipole harmonics (subspace overlap; 1 = identical). Degree sequence: the harmonic degree (ℓ) of each successive eigenmode—a clean harmonic embedding reads $1\,1\,1\,2\,2\,2\,2\,2\,3\ldots$ Five of seven models show strongly harmonic structure (relatively high zonality R^2 with degree-ordered modes). Mixtral-8x22B is borderline, and GPT-OSS-120B is the true outlier, with weak spherical harmonic structure by every measure.

Tokenization maps every object alias to a canonical token. Each book's $M^\star$ (equation 5, $B = 30, f(\delta) = 1, \alpha = 0.25$) is formed using objects with ≥ 5 mentions (fallback 3, marked † in Table 3). Consider the normalized co-occurence rate $R_{ij} = \frac{P_{ij}}{P_i P_j}$ as in equation 7; this counts how much more or less often a pair of objects i and j co-occur relative to how much their individual prevalences would predict. We would ideally select books where pairs of objects have high R_{ij} if they are geometrically close and vis-versa. Furthermore, recall that $M^\star_{ij}$ is a strictly increasing function of R_{ij} while also being bounded, which prevents any issues with R_{ij}'s heavy tail.

This line of reasoning leads us to select our books by

$$D \;=\; \operatorname{corr}\left(M^\star_{ij},\, \cos\theta_{ij}\right) \quad \text{over unordered pairs,}\;\; p < 0.05$$

with p coming from a one-sided location-permutation null (500 trials). $\theta_{i,j}$ is the true angle between two astronomical objects i and j on the celestial sphere. Passing books are *A Field Book of the Stars* (Olcott 1907; $D = +0.34$, $p < .004$), *Round the Year with the Stars* (Olcott 1912; $+0.45$, $p < .004$), *Astronomy with an Opera-Glass* (Serviss 1888; $+0.22$, $p = .01$), *Astronomy for Young Australians* (Bonwick 1866; $+0.26$, $p = .01$). In total we obtain 131k tokens of star guide prose with 39 sufficiently counted objects. Notably, the selected 4-book corpus kernel correlates with the model's activation kernel at 0.92, so they have similar structure.

Simpler statistics fail in instructive ways: narrative *coherence* (median angular hop between successive distinct mentions vs. a shuffled null, Table 3) completely ignores whether successive mentions actually fall within our $M^\star$'s window. Alternatively, a metric completely based on co-occurrences within the window, something like the mean $\cos\theta$ of all co-mentions in the same window, also typically fails. These metrics are often confounded by jointly famous pairs (e.g. Castor–Pollux) that are often mentioned together, but ignores all the geometrically close pairs that are never mentioned. D avoids these problems by working with $M^\star$ and considering all pairs.

A caught negative shows the metric working: *Astronomy for Young Folks* (monthly chapters revisiting the same famous celestial objects) is strongly coherent ($z = 8.1$) yet fails on our selection metric ($D = -0.08$, $p = 0.83$). Similarly, a large generic corpus will generally not perform well at this metric. For example, `text8` (17M tokens of cleaned Wikipedia, $130\times$ the size of the selected guides) demonstrably does not. Its $D = +0.01$ ($p = 0.48$) and coherence $z = 4.6$ and are both far behind the selected corpus (Table 3). `text8` is dominated by celestial terms with alternative meanings (algol, peacock, mimosa, etc.). Any signal it does possess seems to arise from jointly famous pairs of celestial objects, a confounding that D removes as discussed before.

Theory vs. Empirical (Figure 4 C). We compare the centered Mistral Large 2 activations and the modes of the book corpus's centered $M^\star$ with the dipole (degree 1) spherical harmonics evaluated at the object's true positions. Note that spherical harmonics as orthonormalized as in the Sampling Spherical Harmonics paragraph above. We will utilize the same Q_ℓ.

Book	Author, year	ment.	D (p)	coher. (z)	λ_1 (p)	kernel corr
Round the Year with the Stars	Olcott, 1912	361	**+0.45 (0.00)**	30 (8.8)	+1.98 (0.00)	+0.86
A Field Book of the Stars	Olcott, 1907	345	**+0.34 (0.00)**	46 (14.3)	+1.69 (0.00)	+0.85
Astronomy for Young Australians	Bonwick, 1866	116	**+0.26 (0.01)**†	46 (9.6)	+1.16 (0.01)	+0.74†
Astronomy with an Opera-Glass	Serviss, 1888	364	**+0.22 (0.00)**	33 (7.4)	+0.86 (0.03)	+0.75
Half-Hours with the Telescope	Proctor, 1868	101	+0.14 (0.15)†	25 (3.5)	+0.34 (0.08)	+0.50†
Recreations in Astronomy	Warren, 1879	158	+0.11 (0.12)†	18 (3.8)	+0.19 (0.17)	+0.61†
Astronomy for Amateurs	Flammarion, 1904	136	+0.02 (0.38)†	25 (4.2)	−0.03 (0.50)	+0.11†
Pleasures of the Telescope	Serviss, 1901	156	+0.01 (0.45)†	21 (2.6)	+0.01 (0.51)	−0.12†
Astronomy for Young Folks	Lewis, 1922	301	−0.08 (0.83)	26 (8.2)	−0.21 (0.72)	−0.21
text8 (general Wikipedia, 17M tok)	–, –	792	+0.01 (.48)	79 (4.6)	+0.00 (.53)	+0.18

Table 3: Corpus selection (leading 9 candidates of 24). **Selection column** $D = \mathrm{corr}(M^\star, \cos\theta)$ with one-sided object celestial sphere location-permutation p; boldface: selected at $p < 0.05$ with 500 shuffles. Coherence is the median angular hop between successive distinct mentions in degrees. $z = (\bar{m}_{\mathrm{null}} - m_{\mathrm{obs}})/\sigma_{\mathrm{null}}$ is the hop's deficit below the object textual position order-shuffled null (500 shuffles), with $\bar{m}_{\mathrm{null}}$ and σ_{null}, the mean and standard deviation of the median hop across all shuffles. λ_1 is the empirical c_1 from Theorem 4.2 using the book's kernel; 500 shuffles with object celestial sphere location-permutation are used to compute p. Kernel corr is the correlation with the LLM activation kernel. All listed books contain ≥ 15 well-counted (≥ 5 mentions, †: fall back of 3 mentions) objects.

Subspace alignment: There are $2\ell + 1$ spherical harmonics of degree ℓ, but the number of modes of the LLM activations with high degree ℓ purity often exceed $2\ell + 1$ (Figure 4 B). Hence, for comparing our theoretical predictions with the empirical LLM activations, we first find the "best aligned subspace" (within some span of PC modes) in the activations with our theoretical predictions. For the dipole (ℓ =1) alignment, we use PCs 1-4, while for quadrupole ($\ell = 2$) alignment, we instead use PCs 5-12.

The actual alignment for degree ℓ is conducted with canonical angles. Let $U \in \mathbb{R}^{n\times k}$ have columns of PCA modes under consideration, where $k \geq 2\ell + 1$. Let the singular value decomposition (SVD) of $Q_\ell^T U$ take the usual form of $Q_\ell^T U = \Phi\Sigma\Psi^T$, where $\Phi \in \mathbb{R}^{(2\ell+1)\times(2\ell+1)}$, $\Sigma \in \mathbb{R}^{(2\ell+1)\times k}$, and $\Psi \in \mathbb{R}^{k\times k}$. Our best aligned subspace is then $U\Psi_{:,1:2\ell+1}$; this becomes our empirical data.

Note that this procedure is not conducted for the book corpus. We directly consider its $M^\star$ modes 1-3 for dipole and 4-8 for quadrupole.

Point by point alignment: After any potential subspace alignment, we solve an orthogonal Procrustes problem to align our theoretical predictions point wise with the empirical data for either the LLMs or the books. A global scaling is also fit for the LLMs, but not for the books. For the books, we utilize our theoretically predicted scaling as in Theorem 4.2.

Objection permutation null: Our null in Figure 4 (and for the quadrupole case) consists of randomly assigning objects to different positions and performing the exact same analysis. We utilize 300 permutations.

Size effect on books worse performance (Figure 5). We take the fitted zonal kernel to the books ($\tilde{C}$ in Regressions (i) section above) as ground truth and generate synthetic corpora that differ from the real one only in being exactly zonal. Concretely, for each pair the kernel value is converted back to count units by the following (cf. App. A for all definitions):

$$R_{ij}^{\mathrm{fit}} = \frac{2 + \tilde{C}(t_{ij})}{2 - \tilde{C}(t_{ij})}, \qquad \mathrm{rate}_{ij} = n_i n_j \, \frac{2F}{N} \times \; R_{ij}^{\mathrm{fit}}.$$

Off-diagonal counts are drawn as $W_{ij} \sim \mathrm{Poisson}(m \cdot \mathrm{rate}_{ij})$ at multipliers $m \in \{0.25, 0.5, 1, 2, 4, 8, 16, 32, 128\}$ (token count scaled from N to mN) and are used to generate a new $M^\star$. Each draw is then scored by the same pipeline as the observed corpus: per-object cosine is found for the top 3 modes of the centered $M^\star$ after an aligning orthogonal transform. 1000 draws per multiplier are used. Notably, one asymmetry is forced: a simulated draw has no measured diagonal (pair counts are drawn off-diagonal only), so its diagonal is set to the refit $\tilde{C}(1)$.

The observed value sits at the $\sim$93rd percentile of the matched simulations at $m = 1$ (star in Figure 5; band at $1\times$: $[0.27, 0.56]$, observed 0.55), and the curve rises monotonically toward 1, passing the level of the model's per object cosine sim. between $32\times$ and $64\times$ the corpus's counts. Hence, we can say that the "effective corpus size" of the LLM is roughly in this range. It is plausible that the model could have easily consumed this amount of stargazing material during its training.

H REPRODUCIBILITY

Code for all experiments will be available two weeks after submission at `https://github.com/group-invar-stats/group-invariant-statistics.git`.

I AI USE STATEMENT

In this work, we used generative AI tools for formulating mathematical claims, providing critical ingredients for proving mathematical claims, implementing methods in code, assisting the writing of proofs, refining hypotheses, providing feedback on computational experiments, and interpreting results. We have not used generative AI tools for generating synthetic data sets, helping develop theoretical models or conceptual frameworks, assisting with translation, cleaning and reformatting a dataset, or supporting qualitative and thematic data analysis. Additionally, we used generative AI tools for creating code, drafting parts of the paper, analyzing existing literature, proposing keywords for the paper, and helping to find relevant sources. We have reviewed all AI-assisted work. Any generated code or mathematics were reviewed and edited for correctness. An independent literature survey was conducted. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.