FROM LANGUAGE STATISTICS TO SEMANTIC CONTROL: SPECTRAL GEOMETRY AND GROUP ACTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Language models represent ordered concepts in structured activation patterns that can span many directions. An intervention that looks correct in a low-dimensional projection can therefore leave other coordinates inconsistent. We connect language statistics, spectral geometry and semantic control to understand how this structure arises and how to change it consistently. Starting from a latent variable data generating process, we relate latent semantic structure to word co-occurrence and, under explicit embedding assumptions, to Fourier coordinates. Group-based month and weekday steering transfers to held-out prompts, while an open-interval extension transports a fitted neural year component bidirectionally. To steer richer concepts, we must also understand which geometries arise beyond the circle and what transformations they support. Drawing on lattice normal modes, we interpret spectral modes as semantic phonons: coordinated patterns of variation across semantic positions. Sturm–Liouville theory explains how boundary conditions shape these patterns for second-order operators, yielding a broader family of curved geometries even when the interior rule is unchanged. This work connects understanding neural geometry to reliable control: for safety-relevant manifolds such as refusal, interventions must intervene across all relevant subspaces.

1 INTRODUCTION

As neural networks become more capable, understanding how to control their behavior becomes increasingly important. A common approach is to identify structure in their internal activations and use it to construct targeted interventions. A single direction can sometimes provide a useful control, as in refusal interventions (Arditi et al., 2024), but structured concepts can occupy curves and multiple activation directions. Months and weekdays form approximately circular arrangements (Karkada et al., 2026; Engels et al., 2025); space and time preserve meaningful coordinates (Gurnee & Tegmark, 2024); and hierarchical and categorical relationships organize multiple subspaces (Park et al., 2025; Nava & Wyart, 2026). Such structure appears even in simple word embeddings (Korchinski et al., 2025; Karkada et al., 2026), suggesting that it can originate in the statistics of language itself. These observations motivate a basic question: *how should an intervention act when a semantic variable is represented across a structured manifold rather than a single direction?* Figure 1 outlines our answer: language statistics fix the spectral coordinates, and a consistent semantic shift acts on all of them at once.

We connect this problem to the geometrical structure of the *data generating process*. Latent semantic states affect which words occur together: in latent-discourse models, for example, a shared discourse state induces statistical relationships among nearby words (Arora et al., 2016). For a cyclic semantic variable, translation-symmetric co-occurrence is diagonalized by Fourier modes (Karkada et al., 2026). More generally, symmetry constrains an operator's eigenspaces through the representations of its symmetry group (Serre, 1977; Wigner, 1959). Under an explicit link between these statistics and representation geometry, the resulting spectral modes provide coordinates for the semantic variable. Importantly, one variable can occupy several such modes, so a low-dimensional projection can reveal only part of its representation.

Spectral coordinates provide a basis for semantic control, but geometry alone does not specify a reusable operation. Advancing one month should follow the same rule from any starting month; successive advances should compose, and reversal should undo them. For cyclic concepts, group

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.

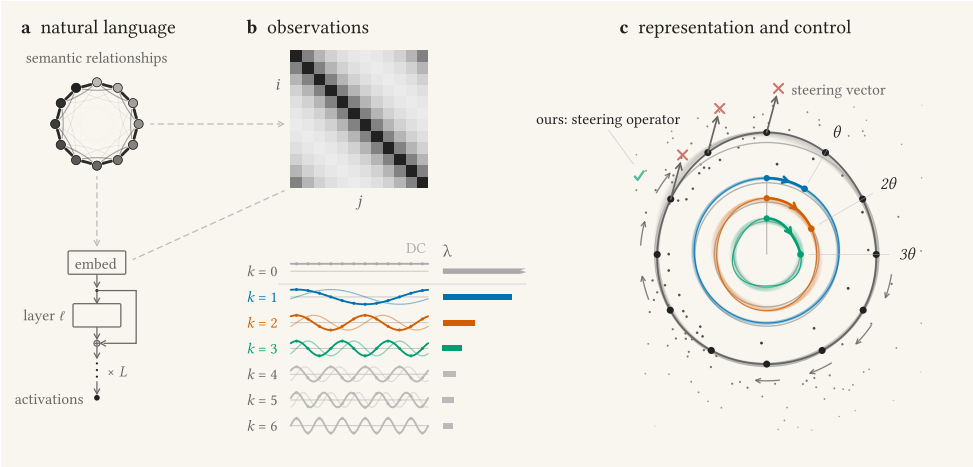


Figure 1: **From language statistics to semantic control.** (a) A contextual representation is read at one layer and token site. (b) Translation-symmetric co-occurrence produces spectral modes over semantic states. (c) A consistent semantic shift acts on all harmonic planes: plane k rotates by $k\theta$. A fixed displacement generally does not preserve the orbit. Panels are schematics.

actions implement this consistency through frequency-specific rotations. For ordered intervals, such as a range of historical years, forward and backward shifts instead operate without wrap-around and remain restricted to admissible states. Both constructions read coordinates from activations, transform them, and map the result back. Their differences motivate a broader account of how boundaries shape both representation geometry and the operations it supports.

We first connect an illustrative data generating process to co-occurrence and spectral geometry under explicit embedding assumptions (Section 3). We then construct cyclic activation operators and quantify errors from omitted harmonics (Section 4). Experiments demonstrate month and weekday steering across held-out prompt families with language-specific mean calibration, outperforming wrong-rate controls, alongside bidirectional transport of a fitted year component on an interval (Section 5). Finally, we interpret spectral modes as *semantic phonons*, drawing on lattice dynamics, and use Sturm–Liouville theory to explain how boundary conditions yield different mode families and geometries for second-order operators (Section 6). Together, these contributions connect the statistical origins of concept manifolds to the structure of the semantic operations they support.

2 Related Work

Representation geometry and language statistics. Linear directions can encode binary (Park et al., 2024; Marks & Tegmark, 2024) and continuous variables (Klindt et al., 2026), while circular concepts (Engels et al., 2025), space and time (Gurnee & Tegmark, 2024), categorical hierarchies (Park et al., 2025), and word analogies (Korchinski et al., 2025) exhibit richer relational geometry. This complements the study of capacity sharing across features through superposition (Elhage et al., 2022). Skip-gram and `GloVe` link embeddings to word–context statistics without uniquely fixing within-set distances (Mikolov et al., 2013; Levy & Goldberg, 2014; Pennington et al., 2014). Cagnetta & Wyart (2024) relate hierarchical latent correlations to recovery sample complexity, Nava & Wyart (2026) connect hierarchical geometry to co-occurrence, and Rende et al. (2024) study transformers' distributional simplicity bias. Most directly, Karkada et al. (2026) connect translation-symmetric language statistics to Fourier geometry under specified embedding assumptions. We study the transformations this geometry supports.

Groups and spectral structure. Fourier features support learned modular arithmetic (Nanda et al., 2023; Gromov, 2023; Zhong et al., 2023); related studies examine group multiplication (Chughtai et al., 2023; Stander et al., 2024) and sequential composition (Marchetti et al., 2026). Equivariant architectures impose transformation laws (Cohen & Welling, 2016; Bronstein et al., 2021), whereas we construct actions in pretrained activation spaces. Feucht et al. (2026) find evi-

dence for shared base-10 addition across cyclic tasks and intervene on Fourier phases. Most closely related, Appendix G of Gurnee et al. (2026) constructs Fourier embeddings of circulant similarity matrices and cyclic shifts on retained invariant subspaces. We connect this action to a semantic generator and test a fixed relative-shift operator across held-out prompt families.

Activation and manifold steering. Activation addition and contrastive steering use fixed displacements (Turner et al., 2024; Rimsky et al., 2024), with refusal as a one-dimensional example (Arditi et al., 2024). Circular-feature interventions exploit multidimensional geometry (Engels et al., 2025). The Goodfire manifold-steering work and Wurgaft et al. (2026) construct interventions along fitted manifolds. Rather than supplying a target coordinate, we read the current spectral coordinate from activations and apply frequency-specific relative shifts.

Modes and boundary conditions. Lattice normal modes describe coordinated displacement (Born & von Kármán, 1912), with symmetry organizing degeneracies (Dresselhaus et al., 2008, Chap. 11). For $-d^2/du^2$ on an interval, Neumann–Neumann and Neumann–Dirichlet endpoints yield integer and half-integer cosine families, respectively (Strang, 1999). Lexical scale structure motivates candidate semantic domains (Kennedy & McNally, 2005). Section 6 develops these connections beyond cycles; Appendix N distinguishes descriptive geometric fits from evidence for particular boundary conditions.

3 From a semantic generator to co-occurrence

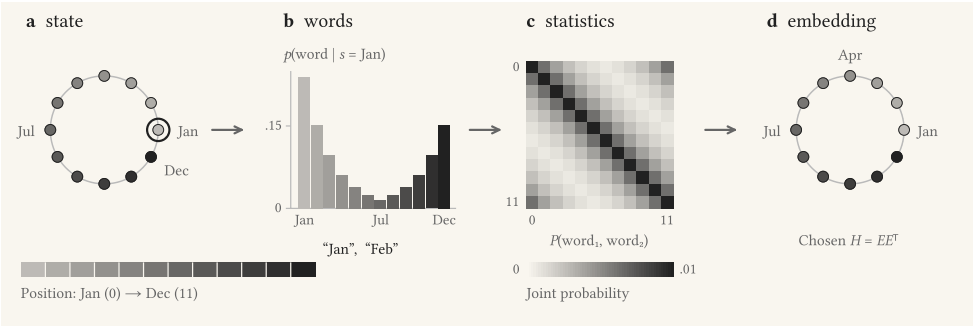


Figure 2: **From semantic states to embedding coordinates.** A latent calendar position determines the probabilities of two conditionally independent month names. Averaging over positions gives pair probabilities and a circular embedding. Gray intensity indicates joint probability.

3.1 Semantic states and the words they generate

We illustrate how latent semantic structure generates co-occurrence with a simple calendar model (Figure 2). Each pair is generated by sampling a latent position $s \in \{0, \ldots, 11\}$ uniformly, then drawing two month names independently conditional on s. For month angles $\theta_i = 2\pi i/12$, let the conditional probability be

$$f(i-s) = \frac{1 + \rho\cos(\theta_i - \theta_s)}{12}, \qquad 0 < \rho \leq 1. \tag{1}$$

The parameter ρ controls the preference for nearby months. Moving s shifts this profile, while uniform sampling makes all names equally frequent. Averaging the product of the conditional probabilities over s gives

$$P_{ij} = \frac{1}{12}\sum_s f(i-s)f(j-s) = \frac{1 + \frac{\rho^2}{2}\cos(\theta_i - \theta_j)}{12^2}. \tag{2}$$

Every row is a cyclic shift of the previous one: P is circulant. Importantly, note that this symmetry is imposed by the model, not established from corpus data. Polysemy (e.g., "May", Karkada et al., 2026), holidays, topic mixtures and geography can break it in language.

3.2 A STATISTICAL MATRIX AND ITS SPECTRAL COORDINATES

More generally, translation symmetry on a cycle means that co-occurrence depends only on separation, $C_{ij} = C(i-j)$, with indices taken modulo the number of states. Such matrices admit Fourier eigenvectors, or *modes*: profiles across semantic positions that retain their shape under matrix multiplication, changing only in scale. Their frequencies count oscillations around the cycle.

These statistical modes become embedding coordinates only after specifying what the embedding reproduces. For our generator, subtract the constant background and factor the resulting positive-semidefinite matrix:

$$H = 12^2 P - \mathbf{1}\mathbf{1}^\top = \frac{\rho^2}{2}(cc^\top + ss^\top) = EE^\top, \quad E = \frac{\rho}{\sqrt{2}}[c\ s], \tag{3}$$

where $c_i = \cos\theta_i$ and $s_i = \sin\theta_i$. The rows of E lie on a circle. Applying the statistical picture to neural activations therefore requires an explicit link to their Gram matrix. Thus, this geometry follows from the chosen representation relation, not from co-occurrence alone. PMI can have negative eigenvalues, and word–context factorizations leave within-set distances undetermined (Appendix J).

3.3 FROM MODES TO OBSERVED GEOMETRY

The centered embedding above uses only the first harmonic pair, but a representation can encode one semantic position in several pairs. For example, $(\cos\theta, \sin\theta, a\cos 2\theta, a\sin 2\theta)$ has one parameter but spans four linear directions when $a \neq 0$. Harmonic order counts oscillations, whereas PCA ranks empirical variance; their subspaces align only under additional metric and sampling conditions.

For twelve months in `Gemma-2b` at hidden-state index 12, the selected PC1/PC2 ellipse score is 0.9344, while those components retain 45.43% of full-space variance (Appendix N). Thus a strong geometric fit (PCA) in a projection can capture less than half of the activation variance. Thus, PCA may fall short of disentangling the harmonics of latent geometry. To change a represented value consistently, we must now specify how the semantic transformation acts across its coordinates.

3.4 MODES AS COORDINATED VARIATION

In a chain of masses and springs, normal modes are coordinated displacement patterns whose relative amplitudes remain fixed as they evolve (Born & von Kármán, 1912; Dresselhaus et al., 2008). We call analogous profiles across semantic positions *semantic phonons*: their frequency describes variation across concept values, not oscillation during inference.

A statistical kernel and a dynamical matrix can share these profiles through a common symmetry. On a cycle, translation-invariant statistics give a circulant matrix, diagonalized by Fourier modes regardless of the dependence on separation. Symmetry supplies a Fourier basis, while the kernel determines its eigenvalues. These paired modes are the coordinates a semantic shift must act on; cyclic indexing alone does not establish the required statistical symmetry.

4 FROM A REPRESENTED STATE TO A REUSABLE OPERATION

4.1 A SHIFT MUST WORK FROM EVERY STARTING MONTH

Advancing one month should send January to February and December to January. Two advances should equal a two-month shift, and reversal should restore the start. These requirements define an action of the cyclic group $\mathbb{Z}_D$, with $i \mapsto i + m \pmod{D}$. The Fourier modes of Section 3 provide coordinates adapted to this action: conjugate character pairs form real sine–cosine planes (Appendix C). A spline that is injective on the circle can likewise implement $s(\theta) \mapsto s(\theta + \Delta)$; harmonic coordinates make this action block-diagonal like the group's irreducible representation.

In contrast, one fixed nonzero displacement cannot implement a finite cycle. For activations y_i indexed by cyclic states,

$$\frac{1}{D}\sum_i (y_{i+m} - y_i) = 0, \tag{4}$$

because shifting the indices permutes the same terms. Thus the desired displacements average to zero, rather than defining a common nonzero steering direction. An anchor-specific displacement can reach its target without implementing the same relative change at every state.

4.2 Why each harmonic needs its own rotation rate

The first harmonic turns once around the calendar; the second turns twice. January to March changes the semantic angle by $\Delta = 2 \cdot 2\pi/12 = \pi/3$, requiring rotations of $\pi/3$ and $2\pi/3$, respectively. Applying the same angle to both is correct only in the first plane (Figure 3).

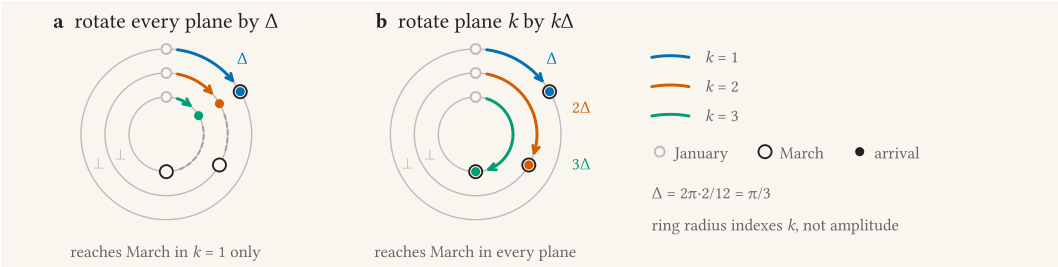


Figure 3: **Frequency-specific shifts.** For January to March, $\Delta = \pi/3$. (a) Rotating every plane by Δ reaches March only in the fundamental plane. (b) Rotating plane k by $k\Delta$ reaches March in all three planes. Open circles mark January, thick-outlined circles March and filled points the arrival. Ring radii indicate harmonic order, not amplitude. This is a schematic multi-harmonic orbit; the embedding in Figure 2 contains only the first pair.

On the retained harmonic pairs, a shift by $\Delta = 2\pi m/D$ acts on row vectors as

$$\mathbf{R}^m = \bigoplus_k R\left(\frac{2\pi k m}{D}\right), \qquad R(\alpha) = \begin{pmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{pmatrix}. \tag{5}$$

Each block rotates at its harmonic rate. Complete sine–cosine pairs remain invariant under shifts, as required by the construction in Appendix G of Gurnee et al. (2026). For even D, the Nyquist coordinate contributes an additional scalar block $(-1)^m$ when retained. Twelve months have five pairs and this scalar; only the paired part admits continuous rotation interpolation.

4.3 Reading coordinates and writing back the change

Fix a model of width M, a layer and a token site. Let $\mathbf{Y} \in \mathbb{R}^{D\times M}$ contain one activation per concept value, with mean μ and centered matrix $\bar{\mathbf{Y}} = \mathbf{Y} - \mathbf{1}\mu$. Let $\mathbf{Z} \in \mathbb{R}^{D\times P}$ contain orthonormal nonconstant Fourier columns in label order, retaining complete pairs. The least-squares decoder and its induced metric are

$$\mathbf{W} = \mathbf{Z}^\top \bar{\mathbf{Y}}, \qquad G = \mathbf{W}\mathbf{W}^\top. \tag{6}$$

Assume $\mathbf{W}$ full row rank, define the readout B, orthogonal projector Π and represented shift S^m:

$$B = \mathbf{W}^\top G^{-1}, \quad \Pi = B\mathbf{W}, \quad S^m = B\mathbf{R}^m\mathbf{W}. \tag{7}$$

For a new activation y_0, read $z_0 = (y_0 - \mu)B$ and retain the residual $r_0 = (y_0 - \mu)(I - \Pi)$. The intervention is

$$y_m = \mu + r_0 + z_0\mathbf{R}^m\mathbf{W}. \tag{8}$$

Equivalently, $y_m = y_0 + z_0(\mathbf{R}^m - I)\mathbf{W}$. The map is linear on centered activations and affine on raw activations: the operator is fixed, but its displacement depends on the input. The component outside the fitted span is unchanged.

Since $\mathbf{W}B = I$, the full centered maps $T^m = (I - \Pi) + S^m$ satisfy $T^aT^b = T^{a+b}$ and $T^D = I$, whereas $S^D = \Pi$. If $\bar{\mathbf{Y}} = \mathbf{Z}\mathbf{W}$, restoring the mean gives $y_i \mapsto y_{i+m}$: the fitted activations form an orbit. Algebraic closure alone does not guarantee this empirical correspondence. Moreover, all $D-1$ nonconstant Fourier columns reconstruct any centered D-row matrix, so reconstruction alone does not show that a few harmonics dominate.

Distance preservation additionally requires G to commute with every $\mathbf{R}^m$: distinct harmonic blocks must be orthogonal, with equal-length orthogonal directions within each pair (Condition C). Appendix C gives context-transfer conditions and cached alternatives for approximate readouts. Open-domain extensions are discussed in Section 6 and Appendix L.

4.4 What a low-variance plane can change

For an orbit satisfying Condition C across all harmonic blocks, let a_k be the activation-space radius of pair k and $\mathcal{K}$ the retained pairs. With the residual fixed, leaving other pairs unrotated gives

$$E^2_{\text{trunc}} = 4 \sum_{k \notin \mathcal{K}} a_k^2 \sin^2(k\Delta/2).$$

For small continuous paired-plane shifts, this is approximately $\Delta^2 \sum_{k \notin \mathcal{K}} k^2 a_k^2$. High-frequency modes can therefore contribute appreciably despite small amplitudes, although an omitted mode causes no error when the intended shift fixes its phase. Appendix E derives these expressions. Geometric error need not predict answer accuracy; the frequency sweep in Section 5 therefore evaluates control of model outputs.

5 Cyclic steering and open-interval transport

5.1 Cyclic steering across prompt families

We test month and weekday steering in `Llama-3-8B`, `Llama-3.1-8B` (team, 2024b) and `Gemma-2-2b-it` (team, 2024a). Each map is fitted on one English prompt family and evaluated on two held-out families per language in English, German, Spanish and French. All concept values appear in fitting; target-language means are calibrated separately. Validation selects $K = 6$ for months (five pairs plus Nyquist) and $K = 3$ for weekdays.

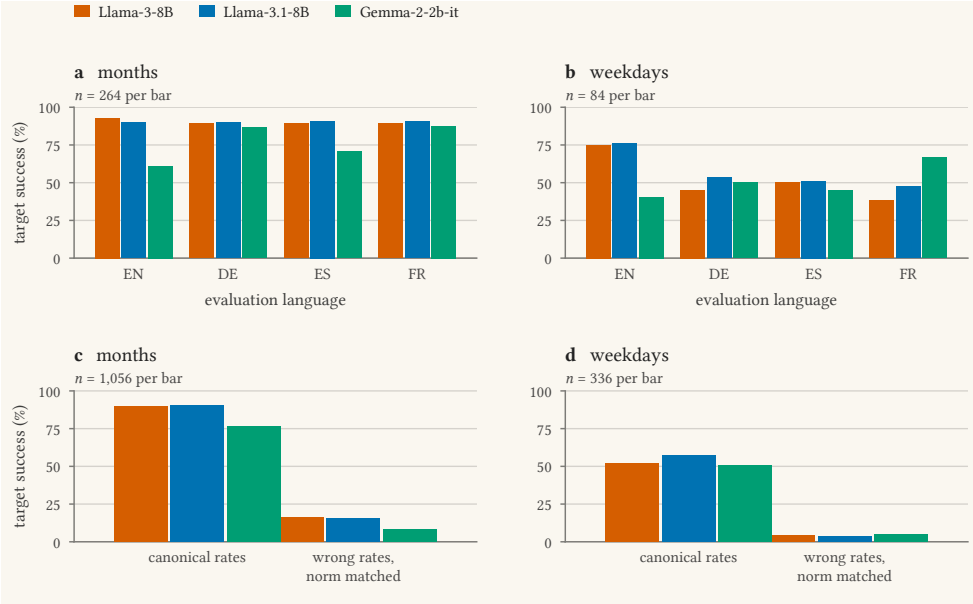


Figure 4: **Cyclic steering across prompt families.** Strict shifted-answer accuracy across three models. (a,b) Correct-rate rotations by evaluation language for months and weekdays. (c,d) Pooled comparisons with wrong-rate rotations, matched in displacement norm.

For each held-out prompt, we apply every nonidentity shift, giving 1056 month and 336 weekday interventions per model. A wrong-rate control permutes the paired rotation rates, preserves Nyquist and matches displacement norms. Strict accuracy measures whether the first answer line begins with the requested target under fixed alias and boundary rules. Intervention sites, selection procedures and decoding details are given in Appendix F.

In model order, month steering achieves 951, 953 and 806 successes out of 1056, versus 169, 163 and 89 for wrong rates. Weekday counts are 175, 192 and 170 out of 336, versus 15, 13 and 16 (Figure 4). Thus the coordinated harmonic rates improve control beyond the fitted prompts, with more consistent performance for months than weekdays. Frequency sweeps, full-copy controls and fitted-prompt cycle checks appear in Appendix F.

5.2 Additional tests

Candidate-likelihood scoring gives similar pooled results but can change individual predictions (Appendix H). Clock steering and full copying in `Llama-3-8B` achieve only $0/1056$ and $1/1056$, respectively, at the selected setting (Appendix F). In a separate `Gemma-2-2b-it` date protocol, full-fit changed-target writes succeed on $27/192$ cases, close to oracle-target copying at $28/192$, despite matching-answer reconstruction reaching $364/365$. These date scores use a post-hoc first-answer-line parser (Appendix I). Accurate reconstruction therefore need not imply successful control.

5.3 Open-interval transport of historical-year representations

Years provide an ordered domain without cyclic wrap-around. We therefore replace rotations with forward and backward shifts restricted to the observed interval. For each year $y \in \{1700, \ldots, 2020\}$, let h_y be the `Gemma-2-2b` activation for `In the year {year}` at the final prompt token, after decoder block 25 (zero-based). We use the base model without a chat template to capture activations. The site is selected for its unsteered chronological geometry. We construct semantic coordinates independently of the neural responses. For $t_y = 2(y - 1700)/320 - 1$, define

$$K_{ij} = \exp(-|t_i - t_j|), \qquad P = I - \frac{1}{321}\mathbf{1}\mathbf{1}^{\mathsf{T}}, \qquad PKP = U\Lambda U^{\mathsf{T}}.$$

We retain all 320 positive eigenmodes and set $Z = U_+\Lambda_+^{1/2}$. An affine one-year operator S_+ is fitted by minimum-norm least squares to adjacent coordinates $[z_y, 1] \mapsto [z_{y+1}, 1]$, keeping the homogeneous coordinate fixed. Reflection of the year grid defines R, giving the backward map $S_- = RS_+R$. No neural activations enter these shift fits.

The connection to neural representations is fitted on the 161 even years. Let X, Y be the training-centered semantic coordinates and activations, with means μ_Z, μ_H. We fit a ridge decoder

$$W_\alpha = \arg\min_W \|XW - Y\|_F^2 + \lambda\|W\|_F^2, \qquad \lambda = \alpha\,\sigma_{\max}(X)^2,$$

with fixed $\alpha = 10^{-2}$, and a minimum-norm reverse encoder $E_0 = Y^\dagger X$. Starting from the activation for year 1804, we encode once as $\widehat{z}_0 = (h_{1804} - \mu_H)E_0 + \mu_Z$. Independent trajectories apply up to 216 forward or 104 backward shifts. Removing the mean and decoding gives

$$\widetilde{h}_k = \mu_H + (\widehat{z}_k - \mu_Z)W_\alpha.$$

We do not restore the reconstruction residual, so this trajectory follows the fitted year component.

Figure 5 compares the semantic trajectory with its neural image. For each mode group I, reduced QR decomposition of $W_{\alpha,I,:}^{\mathsf{T}}$ defines the neural projection. The spans for modes 1–3 and 4–6 capture 70.36% and 15.45% of training activation variance; they overlap, so these fractions are not additive. Repeated shifts agree with direct decoding of the corresponding semantic grid coordinates to within 1.146×10^{-10} in full-neural-space L_2 norm. Mean squared error against raw activations is 2.484 forward and 2.373 backward, averaged over hidden coordinates and nonzero steps. Because the source year is fitted and all semantic transitions are supplied, this demonstrates consistency of fitted-component transport rather than held-out generalization or generated-year control. Unlike cyclic rotations, these shifts are restricted by the interval endpoints.

6 Beyond the cycle: modes and boundary conditions

The year experiment (Section 5.3) uses an ordered interval rather than a cycle. Temperature, size and expertise likewise have an order without wrap-around. We now ask how endpoints change the spectral modes of Section 3 and the operations they support.

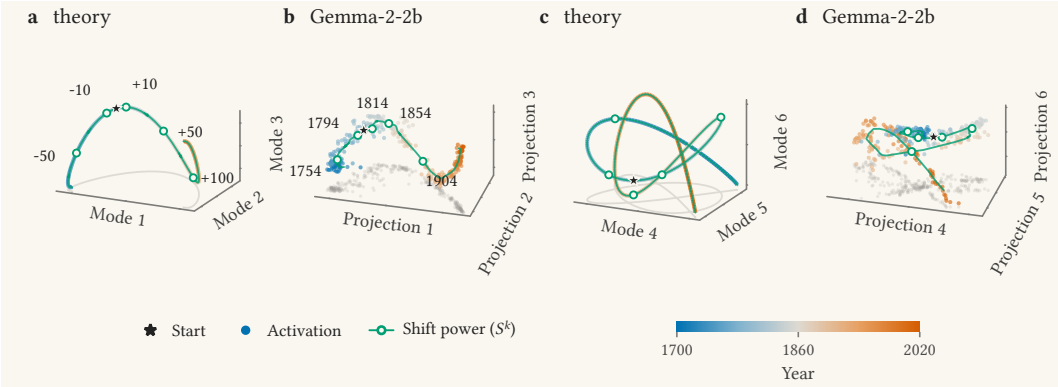


Figure 5: **Bidirectional year transport.** (a,c) Semantic kernel modes 1–3 and 4–6. (b,d) Recorded `Gemma-2-2b` activations and the transported fitted component in the corresponding decoder-row spans. The star marks source year 1804; circles mark selected shifts, and paths extend to the interval boundaries. Colors indicate year; gray shadows are geometric projections. S^k denotes k applications of the steering operator. Neural spans overlap and plotting ranges differ.

On a uniformly sampled interval, a separation-dependent kernel gives a Toeplitz matrix, but truncation at the endpoints breaks cyclic shift symmetry. Its modes depend on the kernel and its boundaries. A local chain offers an analytic description: in a suitable continuum limit, neighboring differences become derivatives. A regular self-adjoint second-order eigenproblem then takes Sturm–Liouville form

$$-(p(u)\phi'(u))' + q(u)\phi(u) = \nu\, w(u)\phi(u), \qquad p, w > 0. \tag{9}$$

Here u is position, ϕ a mode, ν its eigenvalue, and p, q, w specify the interior rule and weighting. Endpoint conditions complete the problem. For a regular finite interval, self-adjoint endpoint conditions form a four-parameter $U(2)$ family, including Neumann, Dirichlet, Robin and periodic cases (Zettl, 2005). The second-order description is an additional modeling assumption, not a consequence of semantic ordering alone.

The exponential kernel used for years makes this connection concrete. Its uncentered continuum version shares eigenfunctions with a second-order differential operator under Robin endpoint conditions; kernel eigenvalues are proportional to the reciprocals of differential eigenvalues. Appendix K derives this relation and explains how centering modifies the eigenproblem.

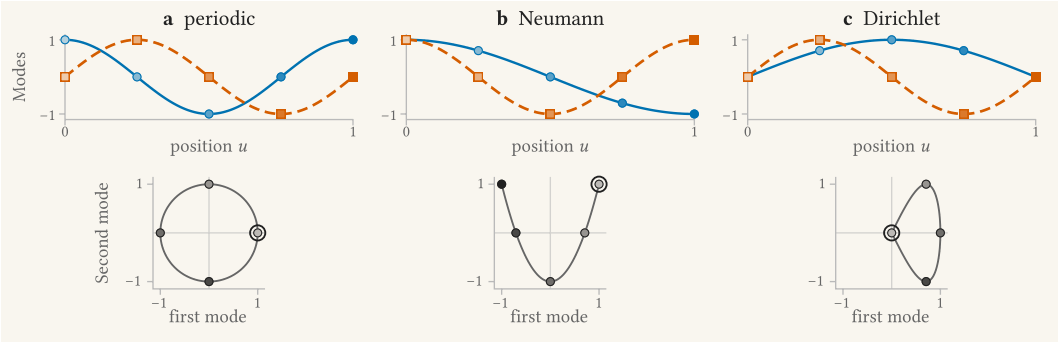


Figure 6: **Boundary conditions shape spectral geometry.** For $-d^2/du^2$, periodic, Neumann and Dirichlet endpoints yield different mode profiles (top) and two-mode embeddings (bottom). Coincident embedded endpoints do not make an interval a cycle.

Holding $-d^2/du^2$ fixed on $[0, 1]$ isolates the endpoint choice (Figure 6). Periodic conditions give the pair $(\cos 2\pi u, \sin 2\pi u)$, tracing a circle. Free (Neumann) ends give $(\cos \pi u, \cos 2\pi u)$, satisfying $y = 2x^2 - 1$. Pinned (Dirichlet) ends give $(\sin \pi u, \sin 2\pi u)$, mapping both endpoints to the origin while the domain remains an interval (Strang, 1999). These curves are exact for the stated

operator and conditions. The uncentered continuum exponential kernel instead yields Robin modes, approaching the Neumann or Dirichlet families in the limits described in Appendix K.

Lexical scale structure motivates candidate domains: *completely full*, unlike *completely tall*, indicates a bounded scale (Kennedy & McNally, 2005). This does not by itself select a differential boundary condition. Appendix N compares ordinal, logarithmic and affect representations with candidate mode shapes; these in-sample fits describe geometry rather than identify neural endpoint conditions.

Boundaries also change the steering problem. The year construction fits forward and backward shifts on admissible grid states, not an unrestricted translation group. A periodic cover offers another construction for reflecting boundaries, but folding can lose the orientation needed for composition (Figure 8, Appendix L). Thus interval transport need not use the paired rotations available on a cycle.

7 Discussion and conclusion

Knowing where a concept is represented does not yet tell us how to change it. A month can be readable from one harmonic plane even when shifting its full representation requires coordinated changes across several planes. Our construction makes this coordination explicit: a fixed operator applies the same relative change from different starting states, with composition and reversal determined by the semantic transformation. The month and weekday results show that this structure can guide behavioral interventions across prompt families. This suggests a useful goal for interpretability: understanding a concept through the operations it supports, as well as the coordinates that encode it.

The statistical perspective asks why those coordinates should arise in the first place. Our generator illustrates how relations among semantic states can produce co-occurrence patterns and, under an explicit embedding assumption, spectral geometry. The connection to lattice normal modes and Sturm–Liouville theory places circles within a broader account in which both the operator and its boundaries matter (Section 6). The year experiment provides a concrete extension: we can construct shifts in semantic coordinates and transport a fitted neural component along an interval without imposing wrap-around. This opens a route to studying ordered concepts whose extremes should remain distinct.

There is still a gap between these constructions and a general account of semantic control. We have not tested the generator against corpus statistics. Cyclic evaluations hold out prompt families rather than concept values and use language-specific mean calibration; year transport starts from a fitted state with all semantic transitions supplied. Date and clock failures show that geometric consistency alone is insufficient for changing answers, while splines remain competitive under their separate protocol (Appendix G). Successful interventions also do not establish that the network uses our constructed operations internally (Vig et al., 2020; Geiger et al., 2021; Wang et al., 2023).

The next steps should test both ends of this connection. Measured corpus kernels could test whether statistical modes predict neural geometry beyond our idealized generator. Interventions should be evaluated from unseen starting states, through successive compositions, and on generated answers rather than reconstruction alone. Extending the approach to interacting semantic variables would also let us ask when their operations can be combined without interference. The longer-term aim is to make semantic control predictable: to specify what should change, what should remain unchanged, and how several changes should combine. Explicit transformation laws give us a way to formulate these expectations and test where a model follows them.

AI USE STATEMENT

We used different LLMs to assist with manuscript drafting and revision, including structural suggestions, prose editing, LaTeX revisions, and checks of citation keys and cross-references. The authors take responsibility for the final manuscript.

REFERENCES

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in Language Models Is Mediated by a Single Direction. In *Advances in Neural Information Processing Systems*, volume 37, pp. 136037–136083, 2024. doi: 10.52202/079017-4322. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/f545448535dfde4f9786555403ab7c49-Abstract-Conference.html.

Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. A latent variable model approach to PMI-based word embeddings. *Transactions of the Association for Computational Linguistics*, 4:385–399, 2016. doi: 10.1162/tacl_a_00106. URL https://aclanthology.org/Q16-1028/.

Neil W Ashcroft and N David Mermin. *Solid State Physics*. Holt, Rinehart and Winston, 1976.

Usha Bhalla, Thomas Fel, Can Rager, Sheridan Feucht, Tal Haklay, Daniel Wurgaft, Siddharth Boppana, Matthew Kowal, Vasudev Shyam, Owen Lewis, Thomas McGrath, Jack Merullo, Atticus Geiger, and Ekdeep Singh Lubana. Do sparse autoencoders capture concept manifolds? *arXiv preprint arXiv:2604.28119*, 2026. doi: 10.48550/arXiv.2604.28119. URL https://arxiv.org/abs/2604.28119v1.

Max Born and Theodore von Kármán. Über schwingungen in raumgittern. *Physikalische Zeitschrift*, 13(8):297–309, 1912. URL https://archive.org/details/per_physikalische-zeitschrift_physikalische-zeitschrift_1912-04-15_8.

Michael M. Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. arXiv preprint arXiv:2104.13478v2, 2021. URL https://arxiv.org/abs/2104.13478v2.

Francesco Cagnetta and Matthieu Wyart. Towards a theory of how the structure of language is acquired by deep neural networks. In *Advances in Neural Information Processing Systems*, volume 37, pp. 83119–83163. Curran Associates, Inc., 2024. doi: 10.52202/079017-2645. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/9740da1c07c7b451af14e11523f95271-Abstract-Conference.html.

Bilal Chughtai, Lawrence Chan, and Neel Nanda. A toy model of universality: Reverse engineering how networks learn group operations. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 6243–6267. PMLR, 2023. URL https://proceedings.mlr.press/v202/chughtai23a.html.

Taco S. Cohen and Max Welling. Group equivariant convolutional networks. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 2990–2999. PMLR, 2016. URL https://proceedings.mlr.press/v48/cohenc16.html.

Mildred S. Dresselhaus, Gene Dresselhaus, and Ado Jorio. *Group Theory: Application to the Physics of Condensed Matter*. Springer, Berlin, 1 edition, 2008. doi: 10.1007/978-3-540-32899-5. URL https://link.springer.com/book/10.1007/978-3-540-32899-5.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy Models of Superposition. *Transformer Circuits Thread*, 2022. URL https://transformer-circuits.pub/2022/toy_model/index.html.

Joshua Engels, Eric J. Michaud, Isaac Liao, Wes Gurnee, and Max Tegmark. Not all language model features are one-dimensionally linear. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=d63a4AM4hb.

Pavel Etingof, Oleg Golberg, Sebastian Hensel, Tiankai Liu, Alex Schwendner, Dmitry Vaintrob, and Elena Yudovina. *Introduction to Representation Theory*, volume 59 of *Student Mathematical Library*. American Mathematical Society, 2011. doi: 10.1090/stml/059. URL https://pubs.ams.org/ebooks/stml/059/.

Sheridan Feucht, Tal Haklay, Usha Bhalla, Daniel Wurgaft, Can Rager, Raphaël Sarfati, Jack Merullo, Thomas McGrath, Owen Lewis, Ekdeep Singh Lubana, Thomas Fel, and Atticus Geiger. Arithmetic in the wild: Llama uses base-10 addition to reason about cyclic concepts. *arXiv preprint arXiv:2605.01148*, 2026. doi: 10.48550/arXiv.2605.01148. URL https://arxiv.org/abs/2605.01148v1.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal Abstractions of Neural Networks. In *Advances in Neural Information Processing Systems*, volume 34, pp. 9574–9586, 2021. URL https://proceedings.neurips.cc/paper/2021/file/4f5c422f4d49a5a807eda27434231040-Paper.pdf.

Andrey Gromov. Grokking modular arithmetic. *arXiv preprint arXiv:2301.02679*, 2023. doi: 10.48550/arXiv.2301.02679. URL https://arxiv.org/abs/2301.02679v1.

Wes Gurnee and Max Tegmark. Language models represent space and time. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=jE8xbmvFin.

Wes Gurnee, Emmanuel Ameisen, Isaac Kauvar, Julius Tarng, Adam Pearce, Chris Olah, and Joshua Batson. When models manipulate manifolds: The geometry of a counting task. *arXiv preprint arXiv:2601.04480*, 2026. doi: 10.48550/arXiv.2601.04480. URL https://arxiv.org/abs/2601.04480v1.

Dhruva Karkada, Daniel J. Korchinski, Andres Nava, Matthieu Wyart, and Yasaman Bahri. Symmetry in language statistics shapes the geometry of model representations. In *International Conference on Machine Learning*, 2026. doi: 10.48550/arXiv.2602.15029. URL https://arxiv.org/abs/2602.15029v3.

Christopher Kennedy and Louise McNally. Scale structure, degree modification, and the semantics of gradable predicates. *Language*, 81(2):345–381, 2005. doi: 10.1353/lan.2005.0071. URL https://doi.org/10.1353/lan.2005.0071.

David Klindt, Charles O'Neill, Patrik Reizinger, Harald Maurer, and Nina Miolane. A unifying framework from neural superposition to sparse interpretable codes. *Nature Machine Intelligence*, pp. 1–13, 2026.

Daniel J. Korchinski, Dhruva Karkada, Yasaman Bahri, and Matthieu Wyart. On the emergence of linear analogies in word embeddings. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 62069–62097. Curran Associates, Inc., 2025. doi: 10.52202/085713-2075. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/59997f550daa5880f50939927181177c-Abstract-Conference.html.

Omer Levy and Yoav Goldberg. Neural word embedding as implicit matrix factorization. In *Advances in Neural Information Processing Systems*, volume 27, pp. 2177–2185. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper_files/paper/2014/hash/b78666971ceae55a8e87efb7cbfd9ad4-Abstract.html.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-Time Intervention: Eliciting Truthful Answers from a Language Model. In *Advances in Neural Information Processing Systems*, volume 36, pp. 41451–41530, 2023. doi: 10.52202/075280-1797. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/81b8390039b7302c909cb769f8b6cd93-Abstract-Conference.html.

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*, 2026. doi: 10.48550/arXiv.2602.03655. URL https://arxiv.org/abs/2602.03655v2.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=aajyHYjjsk.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and Editing Factual Associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17359–17372, 2022. doi: 10.52202/068431-1262. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/6f1d43d5a82a37e89b0665b33bf3a182-Abstract-Conference.html.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. In *International Conference on Learning Representations, Workshop Track*, 2013. doi: 10.48550/arXiv.1301.3781. URL https://arxiv.org/abs/1301.3781v3.

Saif M. Mohammad. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 174–184. Association for Computational Linguistics, jul 2018. doi: 10.18653/v1/P18-1017. URL https://aclanthology.org/P18-1017/.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2301.05217v3.

Andres Nava and Matthieu Wyart. Hierarchical concept geometry in language models emerges from word co-occurrence. *arXiv preprint arXiv:2605.23821*, 2026. doi: 10.48550/arXiv.2605.23821. URL https://arxiv.org/abs/2605.23821v1.

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. URL https://proceedings.mlr.press/v235/park24c.html.

Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. The geometry of categorical and hierarchical concepts in large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=bVTM2QKYuA.

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

Riccardo Rende, Federica Gerace, Alessandro Laio, and Sebastian Goldt. A distributional simplicity bias in the learning dynamics of transformers. In *Advances in Neural Information Processing Systems*, volume 37, pp. 96207–96228. Curran Associates, Inc., 2024. doi: 10.52202/079017-3049. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/ae6c81a39079ddeb88b034b6ef18c7fe-Abstract-Conference.html.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering Llama 2 via Contrastive Activation Addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, 2024. doi: 10.18653/v1/2024.acl-long.828. URL https://aclanthology.org/2024.acl-long.828/.

James A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178, 1980. doi: 10.1037/h0077714. URL https://doi.org/10.1037/h0077714.

Jean-Pierre Serre. *Linear Representations of Finite Groups*. Springer, 1977.

Dashiell Stander, Qinan Yu, Honglu Fan, and Stella Biderman. Grokking group multiplication with cosets. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 46441–46467. PMLR, 2024. URL https://proceedings.mlr.press/v235/stander24a.html.

Gilbert Strang. The discrete cosine transform. *SIAM Review*, 41(1):135–147, 1999. doi: 10.1137/S0036144598336745. URL https://doi.org/10.1137/S0036144598336745.

Gemma team. Gemma 2: Improving open language models at a practical size. arXiv preprint arXiv:2408.00118, 2024a. URL https://arxiv.org/abs/2408.00118v3. Version 3.

Llama team. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*, 2024b. doi: 10.48550/arXiv.2407.21783. URL https://arxiv.org/abs/2407.21783v3.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, Alex Tamkin, Esin Durmus, Tristan Hume, Francesco Mosconi, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet. *Transformer Circuits Thread*, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering Language Models With Activation Engineering. arXiv preprint arXiv:2308.10248, 2024. URL https://arxiv.org/abs/2308.10248v5. Version 5, revised 10 October 2024; first submitted in 2023.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating Gender Bias in Language Models Using Causal Mediation Analysis. In *Advances in Neural Information Processing Systems*, volume 33, pp. 12388–12401, 2020. URL https://proceedings.neurips.cc/paper/2020/file/92650b2e92217715fe312e6fa7b90d82-Paper.pdf.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the Wild: a Circuit for Indirect Object Identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations*, 2023. doi: 10.48550/arXiv.2211.00593. URL https://openreview.net/forum?id=NpsVSN6o4ul.

Eugene P. Wigner. *Group Theory and Its Application to the Quantum Mechanics of Atomic Spectra*. Academic Press, New York, 1959.

Daniel Wurgaft, Can Rager, Matthew Kowal, Vasudev Shyam, Sheridan Feucht, Usha Bhalla, Tal Haklay, Eric Bigelow, Raphael Sarfati, Thomas McGrath, Owen Lewis, Jack Merullo, Noah D. Goodman, Thomas Fel, Atticus Geiger, and Ekdeep Singh Lubana. Manifold Steering Reveals the Shared Geometry of Neural Network Representation and Behavior. arXiv preprint arXiv:2605.05115, 2026. URL https://arxiv.org/abs/2605.05115v1. Version 1.

Bo Xiong. The lattice representation hypothesis of large language models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=5K1FG92m5s.

Anton Zettl. *Sturm–Liouville Theory*. American Mathematical Society, 2005.

Ziqian Zhong, Ziming Liu, Max Tegmark, and Jacob Andreas. The clock and the pizza: Two stories in mechanistic explanation of neural networks. In *Advances in Neural Information Processing Systems*, volume 36, pp. 27223–27250. Curran Associates, Inc., 2023. doi: 10.52202/075280-1186. URL https://papers.nips.cc/paper_files/paper/2023/hash/56cbfbf49937a0873d451343ddc8c57d-Abstract-Conference.html.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation Engineering: A Top-Down Approach to AI Transparency. arXiv preprint arXiv:2310.01405, 2023. URL https://arxiv.org/abs/2310.01405v3. Version 3, revised 10 October 2023.

A NOTATION

Conventions. Activations, orbit coordinates and displacements are row vectors; matrices act from the right, and $R(\alpha)$ rotates a row by $+\alpha$. The identity is I at whatever size the context requires, and $\mathbf{1}$ is the all-ones column used to broadcast a mean. Subscripts i, j index concept values, k harmonics, m steps. Several letters carry a second meaning elsewhere, and Section 5.3 reuses a number of them for its own construction; each such row says so.

Symbol	Type	Meaning
The calendar generator and its statistics (§3)		
D	scalar	concept values: 12 months, 7 weekdays, 365 dates
θ_i, θ_s, c_i, s_i	scalars	state angle $2\pi i/D$ and latent position; $c_i = \cos\theta_i$, $s_i = \sin\theta_i$. Not the cover profiles $c_k(i)$ of App. L
ρ	$(0, 1]$	concentration of the preference profile, Eq. (1)
P	$\mathbb{R}^{D\times D}$	joint name probabilities, Eq. (2); circulant. Not the coordinate count P, nor the projectors P of §5.3 and App. K
H, E	$\mathbb{R}^{D\times D}$, $\mathbb{R}^{D\times 2}$	$H = 12^2 P - \mathbf{1}\mathbf{1}^\top = EE^\top$, Eq. (3); rows of E are embedding coordinates. Not the reflection H of App. L.3
C	$\mathbb{R}^{D\times D}$	co-occurrence on one concept's tokens, $C_{ij} = C(i-j)$
Group action and harmonic coordinates (§4, App. C)		
$\mathbb{Z}_D$, m, k, Δ	—	the cyclic group $i \mapsto i + m \pmod D$; step count, harmonic index, and step angle $\Delta = 2\pi m/D$. Δ is never a position angle
$R(\alpha)$, $\mathbf{R}^m$	$\mathbb{R}^{2\times 2}$, $\mathbb{R}^{P\times P}$	planar rotation by $+\alpha$; $\mathbf{R}^m = \bigoplus_k R(2\pi km/D)$, Eq. (5). Not the reflection R of §5.3
$\mathbf{J}$	paired blocks	generator $\bigoplus_k k \left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)$, with $\mathbf{R}(\Delta) = \exp(\Delta\mathbf{J})$; scalar blocks excluded
$U(g)$, V_α, B_α	—	a representation, its irreducible components, and the blocks Schur's lemma permits. Here α labels irreps, and B_α is not the ridge readout of App. F
U, P_V, V	—	site shift; projector onto the retained space V. In App. J, U, V factor $M \approx UV^\top$
The steering operator (§4.3, App. C)		
M, ℓ	scalars	model width; layer index. In App. J, M is a product matrix and $\boldsymbol{\ell}$ a lattice vector
$\mathbf{Y}$, μ, $\bar{\mathbf{Y}}$	$\mathbb{R}^{D\times M}$, $\mathbb{R}^{1\times M}$, $\mathbb{R}^{D\times M}$	one activation per concept value; their mean; $\bar{\mathbf{Y}} = \mathbf{Y} - \mathbf{1}\mu$
K, P	scalars	highest retained harmonic; coordinate count $P = 2K$, or $2K-1$ with Nyquist
$\mathbf{Z}$, z_i	$\mathbb{R}^{D\times P}$, $\mathbb{R}^{1\times P}$	orthonormal Fourier columns, $Z^\top Z = I$; row i the ideal coordinates of value i
$\mathbf{W}$, W_k, G	$\mathbb{R}^{P\times M}$, $\mathbb{R}^{2\times M}$, $\mathbb{R}^{P\times P}$	$\mathbf{W} = Z^\top\bar{\mathbf{Y}}$; its harmonic-$k$ block; $G = \mathbf{W}\mathbf{W}^\top$, Eq. (6)
B, B_η	$\mathbb{R}^{M\times P}$	readout $\mathbf{W}^\top G^{-1}$; a regularized substitute
Π, S^m, T^m	$\mathbb{R}^{M\times M}$	$B\mathbf{W}$; $B\mathbf{R}^m\mathbf{W}$; $(I-\Pi)+S^m$, Eq. (7). Note $S^D = \Pi$, not I
y, $\bar{y}_i$, y_0, z_0, r_0	$\mathbb{R}^{1\times M}$	an activation; row i of $\bar{\mathbf{Y}}$, $= z_i\mathbf{W}$; the steered activation with its coordinates and residual, Eq. (8). §5.3 writes activations h, since y is a year there
Orbit amplitudes and error laws (§4, App. D, App. E)		
a_k, u_k, v_k, b	—	orbit radius in plane k and its orthonormal pair; Nyquist amplitude $\|W_{D/2}\|/\sqrt{D}$
$\mathcal{K}$	—	the retained pairs. Distinct from the count K
Condition C	—	G scalar on each harmonic block, i.e. $[G, \mathbf{R}^m] = 0$
j	$0, \ldots, D-1$	anchor of a stored displacement. In App. L, j indexes the cover
Year transport (§5.3)		
y, t_y, h_y	—	a calendar year; its rescaled position $2(y-1700)/320 - 1$; the activation there. Activations are h here because y is the year
K, P, U, Λ, Z	—	year kernel $K_{ij} = \exp(-\lvert t_i - t_j\rvert)$, centering projector, and $PKP = U\Lambda U^\mathsf{T}$ giving $Z = U_+\Lambda_+^{1/2}$. None is the K, P, U or Z above
S_+, S_-, R	—	one-year forward and backward shifts, with $S_- = RS_+R$ for the grid reflection R. Here R reflects rather than rotates

Symbol	Type	Meaning
X, Y, μ_Z, μ_H	—	centered coordinates and activations on the 161 training years, with their means
W_α, E_0	—	ridge decoder with penalty $\lambda = \alpha\,\sigma_{\max}(X)^2$; minimum-norm encoder $Y^\dagger X$. Here λ is a penalty, not a kernel eigenvalue
$\widehat{z}_k, \widetilde{h}_k$	—	coordinates and decoded activation after k shifts
Experiments and reporting (§5, App. F)		
K	scalar	retained harmonics; month $K = 6$, weekday $K = 3$
α, R^2	scalars	ridge coefficient; coefficient of determination. Neither relates to $R(\alpha)$
The boundary-value problem (§6, App. K)		
ϕ, n	—	mode profile on the interval; its index, by spectral order or by wavenumber in the mode families
u, x	scalars	position along the range, u in Eq. (9) and x in App. K. App. N uses u, v for valence and arousal
p, q, w	functions	stiffness, potential, density. w is neither $\mathbf{W}$ nor the mixture direction w of App. N; in App. J, q counts atoms per cell
ν, λ, κ	scalars	eigenvalue of Eq. (9) and of the kernel K, with $\lambda = 2\xi/(1 + \kappa^2\xi^2)$ and $-\phi'' = \kappa^2\phi$. λ is not the ridge penalty of §5.3
L, ξ	scalars	length of the range; correlation length. The differential operator is written $\mathcal{L}$
K	operator	the kernel $(Kf)(x) = \int_0^L e^{-\lvert x-y\rvert/\xi} f(y)\,dy$. Not the harmonic count K
c	scalar	the mean of $K\phi$, the forcing that centering introduces
$U(2)$	—	the self-adjoint endpoint conditions, a four-parameter family
Q	function	Liouville-transformed potential. Not the eigenvector matrix Q of App. J
The periodic cover (App. L)		
N	scalar	cover sites: $2D$ for Neumann, $2D + 2$ for Dirichlet. Not the word count N of App. N
f	map	fold onto $\{0, \ldots, D - 1\}$. Not the profile f of Eq. (1), nor a generic function
H	$\mathbb{R}^{2\times 2}$	the reflection $\mathrm{diag}(1, -1)$, with $HR(\alpha)H = R(-\alpha)$

B SUPPLEMENTARY CONNECTIONS

The object and purpose of an intervention. Our relative-shift construction belongs to a wider set of methods for changing model behavior. Representation engineering seeks high-level controls (Zou et al., 2023), inference-time interventions can elicit truthful answers (Li et al., 2023), and model editing changes stored factual associations through parameter updates (Meng et al., 2022). Their different intervention objects imply different evaluation questions. In particular, Equation (4) concerns the average displacement around a complete cycle and leaves these methods' demonstrated applications intact. A meaningful comparison must preserve each method's readout, write and generation procedure.

Overwriting a position and reading one. Engels et al. (2025) overwrite circular features from known token positions, supplying the target coordinate externally where the operator of Section 4.3 reads it from the activation. Where the orbit is exactly represented the two agree at fitted states, since the displacement carrying $\bar{y}_j$ to $\bar{y}_{j+m}$ is then $z_j(\mathbf{R}^m - I)\mathbf{W}$; stored displacements share this single-operator form only because the states lie on one orbit, and without exact representation they do not share it at all. A stored table is therefore not a weaker method where it has an entry, and a table re-read at each step composes as well. What the operator adds is a map defined at states no entry covers, and one needing no external report of the current value.

Settings of the fitted-manifold comparison. The fitted-manifold results we cite record procedures a matched comparison would have to preserve. Wurgaft et al. (2026) fit cubic splines in 64 PCA dimensions while showing three for visualization (§2.2), interpolate intrinsic coordinates (§3.1), study density and behavior-derived metrics together with the geodesics they induce (§3.4),

and use thin-plate splines for factored grid and cylinder control (§4). The last is a product domain construction rather than a single ordered coordinate, so it is not the setting our cyclic operator addresses.

Finding coordinates for geometric analysis. Applying an operation also requires locating the relevant representation. Sparse features can provide useful starting points (Templeton et al., 2024), although recovery of a curved concept manifold remains a measurement question (Bhalla et al., 2026). We therefore distinguish the structure of a recovered dictionary from that of the original activations. Concept-lattice proposals (Xiong, 2026) suggest richer relational structures that could extend this analysis; the proposed geometry of harm and its use for alignment are not established by the present study.

Extensions of the structural account. For finite-dimensional complex representations of a finite group, Maschke's theorem gives a decomposition into irreducible invariant subspaces (Etingof et al., 2011). We use this language in Appendices C and L to derive the cyclic action and determine what additional state is needed on an open interval. For affect, circumplex accounts and valence–arousal ratings motivate a two-coordinate example (Russell, 1980; Mohammad, 2018). Treating these coordinates as a product domain with Neumann boundaries is an additional modeling assumption, whose target functions can be fitted without establishing a neural endpoint mechanism.

C Cyclic actions in activation space

A retained embedding must be closed under the shift. To construct an action on retained coordinates, we first require that shifts remain within their span. The construction of Gurnee et al. (2026), Appendix G, provides the relevant example: a positive-semidefinite circulant similarity matrix has a square-root embedding reproducing its dot products, while spectral truncation approximates them in fewer dimensions and can produce ringing. Let U shift site indices and P_V project onto the retained space V. The condition $UP_V = P_V U$ ensures that $P_V U P_V$ maps projected points to projected successors and that its powers agree with repeated shifts within V.

The invariance condition concerns the retained subspace, not just the original similarity matrix. A cutoff through a degenerate eigenspace can select vectors whose shifts leave their span even when the full matrix commutes with U. Retaining complete real sine–cosine pairs avoids this problem for the cyclic case, with constant and Nyquist scalars treated separately. More generally, the retained space must be invariant, which explains both the restriction in the cited construction and the complete-pair assumption in our activation map.

How symmetry organizes an operator. We can understand this requirement through the general relationship between an operator and its symmetries. If A commutes with every $U(g)$, then $Av = \lambda v$ implies $AU(g)v = \lambda U(g)v$, so the transformed eigenvector remains in the same eigenspace. A subspace mapped into itself by every group transformation is invariant; it is irreducible when it has no smaller nonzero invariant subspace. Equivalent copies share the same transformation rule. For a finite group over the complex numbers, write the representation as $\bigoplus_\alpha V_\alpha \otimes \mathbb{C}^{m_\alpha}$, with irreducible component V_α and m_α copies. The tensor product records coordinates within a component and across its copies. Schur's lemma then gives the commuting-operator form $\bigoplus_\alpha I_{V_\alpha} \otimes B_\alpha$: the operator is scalar on each irreducible factor but can mix equivalent copies. This additional freedom explains why repeated components require more care than a single harmonic plane.

The cyclic case in real coordinates. For a cycle of D sites, each complex character $e^{2\pi \mathrm{i} ki/D}$ occurs once, making a translation-invariant site operator diagonal in the discrete Fourier basis. Real symmetry combines conjugate characters into two-dimensional scalar blocks; dropping symmetry permits an additional skew rotation term in a real commuting operator. With our right-acting row convention and $R(\alpha)$ as in Equation (5), the effect of a shift is explicit:

$$(\cos k\theta,\ \sin k\theta)\, R(k\Delta) = \big(\cos k(\theta+\Delta),\ \sin k(\theta+\Delta)\big).$$

This is why the k-th pair rotates at rate k. Constant coordinates instead form a scalar block, as does Nyquist for even D, with action $(-1)^m$. Nyquist has no second quadrature because its sampled sine

vanishes. On the paired blocks the shift extends to a one-parameter flow

$$\mathbf{R}(\Delta) = \exp(\Delta \mathbf{J}) \quad \text{with} \quad \mathbf{J} = \bigoplus_k k\left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right),$$

the translation generator in the Fourier basis. An invertible real scalar path from $+1$ to -1 would have to cross zero, so continuous interpolation must exclude this block or enlarge the representation.

Reading back coordinates. The next step is to determine whether the fitted activation link preserves composition. Let $Z^\top Z = I$, $\mathbf{W} = Z^\top \bar{\mathbf{Y}}$, and $G = \mathbf{W}\mathbf{W}^\top$, assuming full row rank of $\mathbf{W}$. The readout, projection, represented shift and full-space extension are

$$B = \mathbf{W}^\top G^{-1}, \quad \Pi = B\mathbf{W}, \quad S^m = B\mathbf{R}^m\mathbf{W}, \quad T^m = (I - \Pi) + S^m.$$

For an exactly represented row $\bar{y}_i = z_i\mathbf{W}$, multiplication by B recovers its coordinates, and the shift gives $\bar{y}_i S^m = z_{i+m}\mathbf{W}$. Since $\mathbf{W}B = I$, products of these maps satisfy

$$S^a S^b = S^{a+b}, \qquad S^0 = S^D = \Pi, \qquad (I - \Pi)S^m = S^m(I - \Pi) = 0.$$

These identities imply $T^a T^b = T^{a+b}$ and $T^0 = T^D = I$. Here the superscript on S^0 labels the zero shift rather than the ordinary zeroth matrix power. Preserving the complementary subspace is what makes T an invertible group action on the full activation space, with the same mean restored after each centered operation.

What exact transfer requires. Composition within one fitted representation does not by itself establish transfer between prompt contexts. A sufficient transferable change is a fixed phase and scale in each harmonic, together with a compatible mean and residual. We can express such a second family as

$$y_i^B = \mu_B + r + \sum_k a_k z_{i,k} R(\varphi_k) W_k, \qquad r\Pi = 0.$$

After subtracting μ_B, planar rotations commute with the fixed phase offsets, so the same action transports these states. Transport therefore cannot fail from phase or amplitude differences alone, which narrows a failure to the harmonic subspaces themselves, or to the mean and residual. This argument requires amplitudes, phases, mean and residual to remain fixed around the orbit, with scalar blocks handled separately. Merely sharing subspaces also permits distortions within a plane that do not commute with the action. Mean changes, approximate readout and downstream decoding introduce further obstacles to behavioral transfer. Successful answers in two languages therefore leave congruence of the activation orbits unestablished.

Why caching matters with approximate readout. Approximate readout changes the calculation because rank deficiency gives $\mathbf{W}\mathbf{W}^+ \neq I$, while an independently regularized inverse may also fail to recover the original coordinates. For any chosen B_η, we can instead cache $z_0 = (y_0 - \mu)B_\eta$ and $r_0 = y_0 - \mu - z_0\mathbf{W}$. Decoding the zero shift then restores the original activation exactly, and rotating the saved coordinates closes after a cycle even if the remainder is not orthogonal. Freshly encoding each intermediate activation adds another approximation at each step and need not reproduce this cached procedure. The reciprocal forward-ridge scaling recorded in Appendix F provides a specific cancellation, rather than a general guarantee for regularization.

D When rotations preserve activation distances

The fitted link can transform a circle in Fourier coordinates into an ellipse in activation space. Although the action still closes, neural norms may change. Condition C characterizes the additional metric symmetry needed to preserve them.

Lemma 1. *Suppose $\bar{\mathbf{Y}} = Z\mathbf{W}$ is an exact represented cyclic orbit, Z contains distinct complete real Fourier blocks with orthonormal columns, and $\mathbf{W}$ has full row rank. The following are equivalent: (i) G is scalar on each two-dimensional harmonic block, with no cross-block terms; (ii) $[G, \mathbf{R}^m] = 0$ for every group element; (iii) $\bar{\mathbf{Y}}\bar{\mathbf{Y}}^\top$ is circulant; (iv) $(S^m)^\top S^m = \Pi$ for every group element. Exceptional scalar blocks are included separately.*

Proof. A symmetric matrix commuting with every cyclic rotation is scalar on each distinct real planar block; multiplicity-freeness excludes cross-block intertwiners. Since Z intertwines cyclic site permutations and $\mathbf{R}$, invariance of $ZGZ^\top$ is equivalent to commutation, using $Z^\top Z = I$. Finally,

$$(S^m)^\top S^m = \mathbf{W}^\top (\mathbf{R}^m)^\top G^{-1} \mathbf{R}^m \mathbf{W}.$$

Full row rank makes equality to $\mathbf{W}^\top G^{-1} \mathbf{W}$ equivalent to rotation-invariance of G^{-1}, hence of G. □

The equivalence concerns every shift; checking the identity or half-turn alone is insufficient. For truncated fits it constrains the reconstructed orbit while leaving discarded responses unrestricted, and repeated components require the more general commuting-operator form in Appendix C. An ellipse shows that exact closure can hold without Condition C. Conversely, isotropic independent random activation columns yield scalar expected G, and sufficient width with appropriate moment and concentration assumptions can make an individual draw approximately scalar, without implying exact symmetry. A regular simplex also has a circulant Gram matrix. Distinguishing these possibilities from a spectrally concentrated circle therefore requires amplitude spectra and suitable controls.

E Errors from omitted harmonics and anchored addition

To relate an approximate operation to the ideal shift, we quantify its geometric error under Condition C. Express the paired component in mutually orthogonal unit planes as

$$\bar{y}_i = \sum_k a_k \{\cos(k\theta_i) u_k + \sin(k\theta_i) v_k\}, \qquad \theta_i = 2\pi i / D.$$

Here a_k is the physical radius of the orbit in plane k; a fixed orthogonal residual cancels when the activations are compared.

Leaving a plane unrotated. Truncation applies the intended rotation $k\Delta$ to retained planes while leaving the others unchanged. In an omitted plane this produces mismatch $a_k e^{\mathrm{i}k\theta_i}(1 - e^{ik\Delta})$. Orthogonality makes the total squared error a sum of plane-wise contributions:

$$E^2_{\mathrm{trunc}} = 4 \sum_{k \notin \mathcal{K}} a_k^2 \sin^2(k\Delta/2).$$

The resulting error is independent of the starting state. For a specified step it is zero exactly when every populated omitted term has $k\Delta \in 2\pi\mathbb{Z}$, so a nonzero omitted mode incurs no error if that shift fixes it. Along continuous paired-plane interpolation, the small-step expansion is

$$E_{\mathrm{trunc}} = |\Delta| \left(\sum_{k \notin \mathcal{K}} k^2 a_k^2 \right)^{1/2} + o(|\Delta|).$$

The weighted spectral tail therefore controls a geometric error rate. Translating it into answer errors additionally requires a model of downstream decoding.

Reusing a displacement from one anchor. An anchored additive intervention instead applies $\bar{y}_{j+m} - \bar{y}_j$ at state i, with $\Delta = 2\pi m / D$. The mismatch in plane k is proportional to $a_k(e^{\mathrm{i}k\theta_i} - e^{\mathrm{i}k\theta_j})(1 - e^{\mathrm{i}k\Delta})$, giving

$$E^2_{\mathrm{anchor}} = 16 \sum_k a_k^2 \sin^2(k\Delta/2) \sin^2(k(\theta_i - \theta_j)/2).$$

This expression recovers zero error at the anchor and also at other states whenever each populated product vanishes. It thus specifies the geometric comparison for this particular additive intervention.

Adding the scalar Nyquist term. For even D, the remaining scalar contribution has coordinate $b(-1)^i$ along a unit direction. Omitting it contributes $b^2[1-(-1)^m]^2$ to truncation error, while the anchored contribution is

$$b^2[(-1)^i-(-1)^j]^2[1-(-1)^m]^2 = 16b^2\sin^2\frac{\pi(i-j)}{2}\,\sin^2\frac{\pi m}{2}.$$

The scalar amplitude b is discrete. In orthonormal site coordinates, $b = \|W_{D/2}\|/\sqrt{D}$, and a paired radius under Condition C is $a_k = \sqrt{2/D}\|W_{k,c}\|$. A decoder confined to a retained fundamental plane recovers its transported coordinates exactly, so a concept can be linearly readable while an intervention on it still misses: a probe sees one plane, whereas a shift must carry every populated harmonic. That projection property does not guarantee full-activation fidelity or token generation. These expressions are analytic derivations, with no numerical verification claimed.

F Steering protocol and the evidence behind the counts

F.1 Models, intervention sites and selection

The base Llama checkpoints (team, 2024b) are `meta-llama/Meta-Llama-3-8B`, revision

`8cde5ca8380496c9a6cc7ef3a8b46a0372a1d920`,

and `meta-llama/Llama-3.1-8B`, revision

`d04e592bb4f6aa9cfee91e2e20afa771667e1d4b`.

The additional checkpoint is instruction-tuned `google/gemma-2-2b-it` (team, 2024a), revision

`299a8560bedf22ed1c72a8a11e7dce4a7f9f51f8`.

All three receive raw few-shot prompts with special tokens but no chat wrapper, so the comparison does not isolate architecture, parameter count or tuning. Inference uses BF16 with eager attention on an H100 NVL, with Python 3.12.3, PyTorch 2.13.0+cu130 and Transformers 4.57.6.

The receiving site is the post-decoder-block/post-MLP residual at the final original prompt token. Selected zero-based layers are 26 for both Llamas and 20 for Gemma. The `Llama-3-8B` month site is fixed at layer 26 rather than validation-selected; all other month and weekday sites use English-validation full-state-copy accuracy. Frequency count is selected separately by English-validation accuracy, ties favoring smaller K, yielding month $K = 6$ and weekday $K = 3$ for every model.

F.2 What is held out

Each domain has thirteen balanced families containing every state. There is one English fitting family, one English validation family, two held-out families per language in four languages, and three target-language calibration families. Months have 156 rows: 12 fitting, 12 validation, 36 calibration and 96 held out. Weekdays have 91: 7, 7, 21 and 56, respectively. All states are fitted, while only prompt families are held out; calibrated target-language means accompany the unchanged English map.

Applying each nonidentity shift, including the half-turn when applicable, gives $96 \times 11 = 1056$ month and $56 \times 6 = 336$ weekday interventions per model. Targets shift the requested answer independently of the unsteered prediction. Repeated shifts are dependent observations, not IID samples. Unsteered rows instead use unique prompts and original targets (Table 4); they are not pooled with shifted-target counts.

The English month fitting template is the complete raw string

```
Q: What day comes after Monday?
A: Tuesday
Q: What day comes after Friday?
A: Saturday
```

```
Q: What month comes after {cue}?
A:
```

The held-out English advance template is

```
Advance one month from May.
Answer:  June
Advance one month from {cue}.
Answer:
```

The cue precedes the requested answer in both templates, and the multilingual templates follow the same structure.

F.3 Fitting, writing and scoring

The fitting procedure maps real orthonormal Fourier coordinates to centered activations:

$$W_\alpha = (Z^\top Z + \alpha I)^{-1} Z^\top \bar{\mathbf{Y}}, \qquad B_\alpha = \mathrm{pinv}(W_\alpha),$$

with pseudoinverse cutoff 10^{-12} and $\alpha = 0$ for the main comparison. For a receiving activation y, it reads $z = (y - \mu_{\text{language}})B$, saves $r = y - zW$, and writes $r + z\mathbf{R}^m W$. Full-state copy uses an oracle donor from the same language and family, not a learned cross-family map or a guaranteed ceiling. The wrong-rate control cyclically permutes paired rates, retains Nyquist's action and matches displacement norms in float64, which need not remain exactly matched after BF16 conversion.

Greedy decoding repeats the fixed write at the original prompt position on each growing prefix with `use_cache=False`, stopping at EOS, a complete canonical value plus delimiter, or 32 tokens. The strict scorer parses the first physical answer line under fixed prefix, alias and boundary rules, allowing delimited trailing prose. It is not whole-continuation exact match. Malformed, truncated and wrong-language fields remain distinct; shared spellings alone do not establish language. First-token scores answer different questions and are not the primary measure. The unsteered shifted-target reference reuses baseline outputs rather than adding samples.

F.4 Counts and frequency curves

Intervention rows use held-out families, all nonidentity shifts, $\alpha = 0$ and the stated K; copy rows have no K or α, and unsteered rows have shift zero. Language rows are reported separately from their pooled views and are never summed with them. Tables 2 and 3 report the resulting counts.

Table 2: Strict shifted-answer counts at selected K: months 6, weekdays 3. Wrong rates are norm-matched before BF16 conversion.

Model	Domain	Canonical	Wrong rates	Full copy
`Llama-3-8B`	Months	951/1056	169/1056	1047/1056
`Llama-3-8B`	Weekdays	175/336	15/336	327/336
`Llama-3.1-8B`	Months	953/1056	163/1056	1045/1056
`Llama-3.1-8B`	Weekdays	192/336	13/336	329/336
`Gemma-2-2b-it`	Months	806/1056	89/1056	908/1056
`Gemma-2-2b-it`	Weekdays	170/336	16/336	234/336

Increasing retained frequencies improves canonical accuracy on these panels (Figure 7), without isolating the necessity of individual low-variance directions. All language-resolved curves remain in the source-linked export; the figure pools the eight families per domain and model without mixing models.

The preserved original `Llama-3-8B` forward-ridge table reports a different comparison on the same panel (Table 5).

Table 3: Strict canonical numerators by language; the last column gives each language denominator. Two held-out families contribute to each cell.

Model	Domain	EN	DE	ES	FR	Each n
`Llama-3-8B`	Months	245	235	236	235	264
`Llama-3-8B`	Weekdays	63	38	42	32	84
`Llama-3.1-8B`	Months	237	237	240	239	264
`Llama-3.1-8B`	Weekdays	64	45	43	40	84
`Gemma-2-2b-it`	Months	161	228	186	231	264
`Gemma-2-2b-it`	Weekdays	34	42	38	56	84

Table 4: Unsteered original-task accuracy on unique held-out prompts; targets remain unshifted.

Model	Months	Weekdays
`Llama-3-8B`	96/96	56/56
`Llama-3.1-8B`	96/96	56/56
`Gemma-2-2b-it`	85/96	39/56

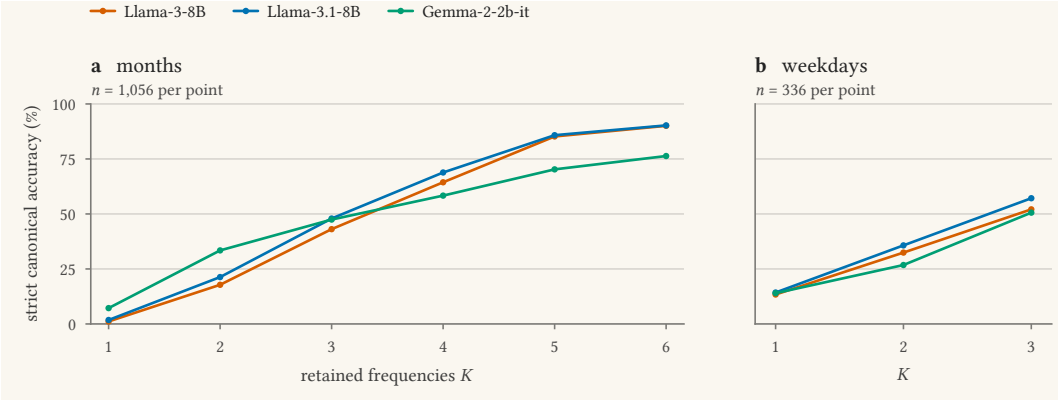


Figure 7: **Frequency sweeps.** Strict canonical accuracy at every tested frequency count, pooled over the four languages separately for each model; axes start at zero.

Table 5: Full-K forward-ridge counts across coefficients on the same held-out prompts.

Domain	K	$\alpha = 0$	0.1	1	10
Months	6	951/1056	951/1056	951/1056	951/1056
Weekdays	3	175/336	175/336	175/336	175/336
Clocks	6	0/1056	0/1056	0/1056	0/1056

With orthonormal Z, $W_\alpha = W_0/(1+\alpha)$, and its pseudoinverse scales reciprocally. These factors cancel in $B_\alpha \mathbf{R}^m W_\alpha$ if numerical rank is retained, explaining unchanged full-K counts. An independently regularized inverse is a different calculation, not evaluated by that sweep.

F.5 Gemma formatting sensitivity

Only Gemma's raw trials are complete enough for a post-hoc normalization analysis. The normalizer unwraps one eligible balanced bold alias on the first physical line while preserving prefix, suffix, later lines and truncation state; it neither searches explanations nor broadly strips Markdown. Selected canonical month accuracy changes from $806/1056$ to $822/1056$, and weekdays from $170/336$ to $171/336$. Matched wrong rates change from 89 to 98 and from 16 to 17, with the same respective denominators. English sequence counts change from 42 to 55 out of 132 months and from 0 to 1

out of 42 weekdays; English advance remains $119/132$ and $34/42$. Formatting explains part, not all, of the deficit. Strict scoring remains primary for all models; the corresponding Llama trials are unavailable, so we do not report a normalized control.

F.6 Clock diagnostics and checks answering different questions

The earlier `Llama-3-8B` clock study uses twelve hours without AM/PM. All six English-validation K settings score zero, selecting $K = 1$; held-out selected steering gives $0/1056$, full copy $1/1056$, and $K = 5$ has one success. Its original-task held-out baseline is $96/96$. The generated-answer panel contains no clock or date results; the date study below uses a different protocol.

A write-free replay check finds that BF16 activation differences are stable under repetition but vary with input shape, which limits exact reproduction of these runs without explaining the failure.

The restored month check instead tests the fitted orbit using the original Llama revision, post-MLP layer 26 and an SDPA/position-ID path. Fitting and evaluating the same twelve English prompts with $W = Z^\top \bar{\mathbf{Y}}$ and the literal Gram inverse gives $120/120$ restricted and unrestricted next-token accuracy for both $K = 5, 6$, over ten nonidentity shifts excluding the half-turn. The $K = 6$ walk reuses these cases; neither adds held-out continuation evidence.

A separate `Llama-3.1-8B` arithmetic follow-up compares direct and projected bases at post-attention layer 16/$K = 3$ for start state and layer 31/$K = 1$ for answer state. Its 72 reused selection-validation prompts yield 576 interventions at shifts 1 and 6, with no held-out rows. The further comparisons and date records below use their own scoring conventions.

G Historical controls and transfer

The historical notebook records and the modern generated-answer panel answer related but different questions. Historical month comparisons use restricted-token argmax, answer-indexed states, and language-specific mean calibration; their percentages are not additional observations to pool with generated answers.

Rotation, lookup, spline and copy. Lookup stores the measured displacement for each state and step and replays it by known position, so it has an entry for every request in this panel. The saved percentages compare methods within each row:

Prompt context	Rotation	Lookup	Spline	Copy
Fitted English	100	100	99	100
Plain English	96	92	97	100
Two-after English	77	100	88	100
German	98	100	98	100

The spline retains 10 or 11 PCs to reach 95% variance; rotation retains $K = 6$. Target-language means are calibrated. They rule out a blanket claim that splines cannot transfer across languages. They leave a matched comparison of capacity, calibration and semantic parameterization unresolved. A known coordinate can supply a composable action with either a spline or a harmonic representation.

Geometry is not a localization criterion. The historical layer scan covers blocks 2–30. At layer 15, the fundamental's share is 0.52, yet rotation has zero control in the saved comparison. At layer 26 the share is 0.32, with English/German rotation scores of $100\%/98\%$. The diagnostic selection identifies the late band 22–30 by normalized circular-displacement gain at least 0.99, then chooses its midpoint 26. This is not selection by exact donor accuracy and is not the modern English-validation selection rule.

Nulls, rates and truncation. Permuting the rows of an activation matrix leaves its PCA spectrum invariant. A fixed-semantic-order Fourier shuffle instead changes alignment with the ordered harmonic basis. These are different controls. Historical own-rate, common-rate, shuffled-rate and

fundamental-only comparisons are also distinct from the modern displacement-norm-matched permutation. The saved association of a truncation-tail quantity with output KL uses 20 condition means; it is not a per-example causal law.

Composition and cross-context state construction. In the saved half-month comparison, circular movement is 0.57 for rotation and 0.47 for spline. The scalar Nyquist component collapses at half steps in that implementation, so this does not establish exact continuous composition. Coordinate matrices satisfy their integer group identities by construction. An externally carried coordinate walk makes independent writes from that trajectory; it does not show autonomous neural rollout or repeated native computation. Historical cross-style overlap and phase comparisons describe fitted harmonic planes, while transport additionally depends on means, within-plane distortions, readout and decoding. A separate final-block `Gemma-2-2b-it` German demonstration rotates an English-derived activation and adds a January bilingual offset. It does not establish congruence with the native German orbit or calibration-free transfer.

Arithmetic remains exploratory. The `Llama-3.1-8B` follow-up fits start-state and answer-state interfaces at different post-attention sites, using layers $16/K = 3$ and $31/K = 1$. It reuses 72 selection-validation prompts for 576 interventions at shifts 1 and 6, and contains zero test-intervention rows. Direct and projected basis comparisons can change intervention norm; apparent improvements therefore do not isolate the intended geometric change. The original inverse-ridge candidates and the modern forward-ridge cancellation are different procedures. No held-out arithmetic intervention success is claimed.

H CANDIDATE LIKELIHOOD IS A SEPARATE READOUT

The completed `Llama-3.1-8B` panel ranks supplied complete answer strings under fixed interventions. It sums natural-log probabilities over contextual answer tokens, retains the leading space and applies no length normalization. The answer-only convention removes only a standalone terminal newline token, ID 198. It is applied post hoc to the saved token probabilities, with no change to model parameters, site or frequency selection; we report both conventions.

Answer-only canonical accuracy is $953/1056$ for months and $192/336$ for weekdays, compared with $947/1056$ and $183/336$ when the newline is included. There are 19 changed month choices and 21 changed weekday choices, more than the net correctness improvement. Equal pooled totals between answer-only ranking and generation do not imply equal predictions: English weekday counts change from 64 to 65, while Spanish changes from 43 to 42. Norm-matched wrong rates score $161/1056$ and $13/336$; real-target copying scores $1044/1056$ and $329/336$.

Mean shortest circular error, including correct cases, is $270/1056 = 0.2557$ months and $235/336 = 0.6994$ weekday steps. Corresponding unsteered shifted-target errors are 3.273 months and 2 days. All candidates are valid labels; this closed-set measure cannot estimate how often free generation leaves the candidate set. A softmax over candidate scores is likewise not unrestricted model probability mass. The full panel has 15,848 condition records on 152 distinct held-out prompts. Repeated conditions are not independent prompts, and the exported records omit write vectors.

I THE 365-DAY CALENDAR: PRESERVATION AND CHANGED TARGETS

I.1 HISTORICAL GEOMETRY AND HELD-OUT-STATE FAILURE

Historical `Llama-3-8B` analyses write at block 26 on two-shot next-day prompts. Their full-calendar $K = 60$ fit uses 120 harmonic columns, not the 364 columns of a complete centered basis. Its reconstruction R^2 is 0.893. The annual component accounts for 17.61% of retained fit energy, frequency 12 for 8.56%, and frequency 52 for 0.10%; these denominators exclude omitted activation variance. Month/day-of-month alignment measures reduction in projected variance after grouping, not classification or causal ablation.

The first raw similarity lag peaks outside the initial neighborhood occur at 10, 20 and 31 rather than monthly multiples. Adding a day-of-month kernel raises in-sample R^2 from 0.610 to 0.660 over 66,430 unordered date pairs. These pairs share dates and are not independent or held-out evidence.

Only twelve frequency planes have saved numerical group-alignment values; the missing full series is not reconstructed.

For a $+30$-day request, real-target copy obtains the correct month on $364/365$ cases but the full-date prefix on only $131/365$. Rotation at $K = 60$ gives $362/365$ and $129/365$, respectively. For $+1$, full-date-prefix scores fall to $11/365$ for copy and $10/365$ for rotation; all 353 off-boundary rotation cases retain the original expected date. Prefix acceptance differs from exact equality of the entire continuation.

The half-calendar experiment tests a train-fitted operator on real held-out neural states. Fitting 182 dates yields training/test $R^2 = 0.877/0.007$ on 183 held-out dates. Predicted-state writes obtain $8/183$ full-date-prefix successes, compared with $183/183$ self-copy and $1/183$ mean write. For $+30$, the train-half operator gets only $2/183$ held-out month answers, versus $182/183$ for the full-year operator and $183/183$ real-target copy. Its $7/182$ on its own half also indicates interface instability beyond off-support generalization.

I.2 Completed external Gemma test

The external `Gemma-2-2b-it` study uses the final residual block 25, FP32 TransformerLens conversion and a prefill-only write with cached decoding. It replaces activations using known calendar coordinates, rather than learning an encoder from a new neural state. Its full-calendar $K = 60$ fit has $R^2 = 0.9564$. A half-year $K = 91$ ridge fit, with selected strength 12.33, has held-out state-space $R^2 = -0.6202$. Under the saved corrected parser, matching-answer baseline, full reconstruction and half reconstruction score $365/365$, $364/365$ and $181/183$. Even a constant mean write preserves $43/48$ original answers. These are preservation tests: the prompt still specifies the correct date.

A historical periodic spline with 64 PCs gets $30/183$ matching-prompt successes compared with $181/183$ ridge prediction. It forcibly closes the partial-year curve, so it is not a boundary- or capacity-matched test of general spline extrapolation.

The saved parser correction. The original strict grammar rejects valid dates followed by a newline explanation. It scores baseline and full reconstruction as zero, half reconstruction as $180/183$, oracle and full changed-target writes as zero, and ridge changed-target writes as $1/192$. A post-hoc first-answer-line parser, cross-checked against an independent datetime parser on all 1,585 scored generations, retains impossible dates as failures. We report both the original and corrected scores.

Changing the target remains difficult. The saved panel uses 48 selected dates spanning months and shifts $-30, -1, +1, +30$, giving 192 requested targets per condition. Corrected exact-date scores use *all* 192 cases: oracle target $28/192$, full fit $27/192$, ridge $16/192$, and unsteered $0/192$. Invalid outputs number 7, 6 and 5 for the three interventions. Only circular-error means exclude invalid answers, giving $1701/185 = 9.195$, $1704/186 = 9.161$ and $2218/187 = 11.86$ days, versus 15.50 days unsteered. On matched valid subsets the unsteered means are approximately 15.50–15.58 days.

All 86 within-month oracle requests retain the original date, and neither fitted method succeeds on within-month targets. For -1, full-fit mean error worsens from 1 to 4.844 days. From an original answer of January 1, requesting January 2 can yield January 1, whereas requesting December 31 yields December 1. Impossible answers include February 29 and April 31 in the non-leap calendar. A final-block prefill write changes first-token logits without rewriting earlier keys/values or future decode states; later changes can propagate through the generated prefix. This helps separate month and day control in this protocol, without establishing absence of day information elsewhere. The earlier historical Llama site does not admit the same final-block explanation.

J Operator modes and lattice dynamics

Our use of modes rests on the idea that relationships among positions constrain how their values can vary together. Lattice dynamics provides a developed example of this reasoning: coupled objects have coordinated displacement patterns determined by an operator. We use that operator mathematics to describe semantic variation. The correspondence concerns patterns and symmetry, rather than mechanical displacement of activations, physical vibration or quantum dynamics.

Setting the correspondence out: in a harmonic crystal the force constants depend only on the separation between cells, so the force-constant operator commutes with lattice translations, and Fourier transformation into that group's characters $e^{\mathrm{i}\mathbf{k}\cdot\boldsymbol{\ell}}$ splits it into independent blocks whose normal modes are the phonons (Ashcroft & Mermin, 1976). Tokens ordered along a concept take the role of the cells and the co-occurrence kernel that of the force constants. With one token per site each block is one dimensional, so here the Fourier modes are the normal modes themselves; several channels per semantic position would restore a branch index, as in the multi-atom case next.

When a frequency has more than one kind of motion. The lattice example also shows why specifying a frequency may leave further structure to determine. With q atoms per primitive cell, a crystal has $3q$ displacement coordinates at each wavevector. Fourier transformation separates wavevectors, after which a $3q$-dimensional block must be diagonalized to obtain polarization vectors and phonon branches. Our scalar cyclic operator has one complex coordinate per wavevector and therefore no additional branch index; multiple channels per semantic position would introduce analogous internal blocks. The rows of $\mathbf{W}$ attach activation-space amplitudes to the site modes. Symmetry fixes their transformation law, while their lengths and angles depend on the metric conditions in Appendix D.

Why a statistical product need not fix distances. Transferring the analogy to embeddings requires distinguishing an energy from a statistical product. A physical quadratic energy approximates behavior near equilibrium, whereas an embedding objective may constrain a bilinear product $M \approx UV^\top$. For any invertible A, replacing the factors by UA and $VA^{-\top}$ preserves that product but can change word distances. A symmetric PMI matrix also need not be positive-semidefinite: if $M = Q\Lambda Q^\top$, then $E = Q|\Lambda|^{1/2}$ gives $EE^\top = Q|\Lambda|Q^\top$, differing from M whenever an eigenvalue is negative. Equation (3) specifies a positive-semidefinite Gram matrix to make the geometric relation exact. Approximating it at lower rank still requires choosing the retained eigenvectors.

Periodicity and allowed frequencies. Frequency indexing follows from two related properties of a periodic domain. For lattice vectors $\boldsymbol{\ell}$ and reciprocal vectors G_{rec}, the identity $e^{\mathrm{i}G_{\mathrm{rec}}\cdot\boldsymbol{\ell}} = 1$ makes wavevectors separated by G_{rec} represent the same lattice character. Finite periodic boundary conditions further restrict the wavevectors to a discrete set. Neither property determines the amplitudes with which a neural representation expresses those frequencies.

K Deriving endpoint conditions

What an oscillation count tells us. Endpoint conditions matter because an interior differential rule alone does not specify its modes. The main-text example belongs to the Sturm–Liouville family,

$$-(p\phi')' + q\phi = \nu w\phi, \qquad p, w > 0.$$

Here ϕ is the mode profile, ν its eigenvalue, p, q position-dependent coefficients, and w a positive weight; primes indicate position derivatives. With regular coefficients and self-adjoint separated endpoint conditions, the spectrum is simple and ordered, and eigenfunction n, counted from zero, has n interior zeros. This orders spatial oscillations but need not order empirical PC variance: kernel eigenvalues can rank modes differently, centering alters the operator, and periodic problems have paired degeneracies. Similarly, exact Fourier diagonalization of a circulant matrix does not extend to exact sine/cosine diagonalization of a finite Toeplitz matrix. An asymptotic eigenvalue distribution leaves individual eigenvectors unspecified. The same inversion is exact whenever the kernel's Fourier symbol is a rational function of frequency, as in the Matérn family, and holds only in the long-wavelength limit otherwise; higher-order rational symbols require correspondingly higher-order differential equations with their own endpoint domains. What does survive the loss of the shift is a reflection: a symmetric Toeplitz matrix is centrosymmetric, commuting with the reversal permutation, so every eigenvector is exactly even or odd about the midpoint. That parity selects no endpoint condition, since the Neumann and Dirichlet families both alternate even and odd modes, which is what the derivation below supplies.

What the coefficients mean. The energy form $\int_0^L (p\phi'^2 + q\phi^2)\,dx$, with norm $\int_0^L w\phi^2\,dx$, gives the coefficients their reading: p is a stiffness, the cost of varying along the token axis; w is a density, the weight defining orthogonality and hence how densely tokens sample the underlying axis; and q

is a potential penalizing amplitude, inert when constant. For the exponential kernel below, $p = \xi^2$ and $q = 1$, so stiffness is the squared correlation length. The three are not separately identifiable: the Liouville transformation $t = \int_0^x \sqrt{w/p}\, dx'$, $\psi = (pw)^{1/4}\phi$ reduces any such problem to $-\psi'' + Q\psi = \nu\psi$, leaving only Q and $\int_0^L \sqrt{w/p}\, dx$ recoverable from data.

What constant coefficients assume. Within the second-order differential model, constant p, q, w describe a homogeneous interior. The eigenproblem reduces to $-\phi'' = \kappa^2\phi$, with $\kappa^2 = (\nu w - q)/p$. The endpoint conditions determine the admissible profiles. Relating this differential model to a statistical kernel requires a separate derivation, as in the exponential example below.

A kernel whose endpoint rule can be derived. An exponential kernel illustrates how the integral operator itself can determine the endpoint rule, and is the decay reported for cyclic concepts in corpora (Karkada et al., 2026). For $L, \xi > 0$, consider

$$(Kf)(x) = \int_0^L e^{-|x-y|/\xi} f(y)\, dy.$$

Nearby positions receive greater weight through this kernel. Splitting the integral at x separates two smooth pieces, whose derivatives give

$$(Kf)' = -\xi^{-1}\int_0^x e^{-(x-y)/\xi} f(y)\, dy + \xi^{-1}\int_x^L e^{-(y-x)/\xi} f(y)\, dy, \quad (Kf)'' = \xi^{-2}Kf - 2\xi^{-1}f.$$

At either endpoint one piece vanishes, yielding $(Kf)'(0) = (Kf)(0)/\xi$ and $(Kf)'(L) = -(Kf)(L)/\xi$. Together with the second derivative, this identifies

$$K = 2\xi(1 - \xi^2\partial_x^2)^{-1}$$

on the Robin domain $u'(0) = u(0)/\xi$, $u'(L) = -u(L)/\xi$. The normalization and domain are both necessary for this identity. The eigenvalue equation $K\phi = \lambda\phi$ then becomes $-\phi'' = \kappa^2\phi$, with

$$\lambda = \frac{2\xi}{1 + \kappa^2\xi^2}, \qquad \phi'(0) = \phi(0)/\xi, \quad \phi'(L) = -\phi(L)/\xi. \tag{10}$$

Equation (10) inverts the ordering: λ decreases in κ, so the kernel's leading eigenfunctions are its slowest-varying ones. An endpoint condition fixes the admissible κ and therefore fixes which profiles lead.

Separating even and odd functions about $L/2$ gives the respective root conditions

$$\tan(\kappa L/2) = \frac{1}{\kappa\xi}, \qquad \tan(\kappa L/2) = -\kappa\xi.$$

Together, the positive roots give one root in each $(n\pi/L, (n+1)\pi/L)$, $n \geq 0$. Increasing ξ/L sends the lowest wavenumber to zero and the remaining roots toward the Neumann ladder. For small ξ/L, the fixed-index asymptote $\kappa_n \sim (n+1)\pi/(L + 2\xi)$ approaches Dirichlet with extrapolation length ξ. These continuum limits explain this kernel's behavior; they provide neither a discrete approximation-error estimate nor a universal rule for open vocabularies.

Why centering changes the answer. Connecting this calculation to PCA requires accounting for mean removal. Let P remove the constant function in $L^2[0, L]$. A mean-zero eigenfunction of PKP satisfies $K\phi = \lambda\phi + c$, with c the mean of $K\phi$. Differentiation therefore yields

$$-\phi'' = \kappa^2\phi - \frac{c}{\lambda\xi^2}, \qquad \lambda\phi'(0) = \frac{\lambda\phi(0) + c}{\xi}, \quad \lambda\phi'(L) = -\frac{\lambda\phi(L) + c}{\xi}.$$

The constant introduces forcing and modifies the endpoint equations unless $c = 0$. As a result, removing the lowest uncentered Robin mode is insufficient to derive the centered PCs.

Which endpoint conditions are admissible. The interval restricts the operator before it restricts the modes. On the whole line $-\partial_x^2$ is self-adjoint on its own; on $[0, L]$ the boundary terms in $\int(\mathcal{L}u)\bar{v} - \int u\overline{(\mathcal{L}v)}$ do not vanish, and making them vanish requires a linear constraint on ϕ, ϕ' at

the two ends. The admissible constraints form a four-real-parameter family in bijection with $U(2)$ (Zettl, 2005): a Robin angle at each end, a mixing angle setting how strongly the two ends are tied together, and a relative phase. Separated conditions are those of zero mixing angle, with Dirichlet, Neumann and Robin as special cases; periodic conditions are the non-separated branch. This is a finite-dimensional classification, not a finite list. It also separates the two spectral regimes used here: separated conditions give simple eigenvalues and standing waves, whereas the periodic branch gives the degenerate $(\cos, \sin)$ pairs that Section 4 rotates.

Ideal mode families and their curves. Holding $-\phi'' = \kappa^2\phi$ fixed isolates familiar endpoint families, indexed here by wavenumber rather than by spectral order. Neumann ends give $\cos(n\pi x/L)$, $n \geq 0$, while Dirichlet ends give $\sin(n\pi x/L)$, $n \geq 1$. Pinned–free conditions give $\sin((n+\frac{1}{2})\pi x/L)$, $n \geq 0$, with cosine for the reversed endpoints. Periodic value and derivative conditions give a constant and harmonic pairs at $2\pi n/L$. In outward-normal notation, Robin conditions are $-\phi'(0) + \beta_0\phi(0) = 0$ and $\phi'(L) + \beta_L\phi(L) = 0$.

The geometric consequences follow by pairing mode values. Normalized Neumann modes satisfy $\cos 2\theta = 2\cos^2\theta - 1$, giving a parabola whose plotted equation changes with scales, offsets and signs. A Dirichlet pair $(\sin\theta, \sin 2\theta)$ joins the two endpoints in its display. Neither curve specifies which PCs an empirical fit selects. On a Neumann rectangle, modes are products $\cos(n_x\pi x/L_x)\cos(n_y\pi y/L_y)$, with $(1,1)$ vanishing on both midlines. The affect example introduces an additional modeling choice by interpreting its semantic coordinates as this rectangular domain.

L Extending operations to open semantic domains

The paired-rotation construction of Section 4 does not directly apply to a cosine-only standing-wave basis. A generic continuous translation introduces a sine component:

$$\cos n(\theta + \Delta) = \cos(n\Delta)\cos(n\theta) - \sin(n\Delta)\sin(n\theta).$$

This obstruction concerns closure of the cosine-only function space, not the existence of the in-range affine grid shifts constructed in Section 5.3. Here we study periodic covers that supply paired coordinates; folding back to the interval can discard the orientation needed to compose shifts.

We make the additional state explicit by extending particular discrete sine and cosine bases to a periodic cover. General Robin conditions require a separate construction. Mixed endpoints yield an extension that is antiperiodic over twice the interval and periodic over four times its length.

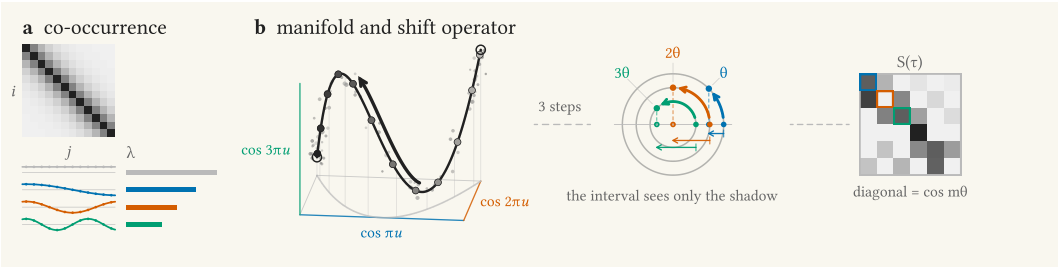


Figure 8: **Standing waves and the missing quadrature.** (a) On an open range the kernel's leading modes are the smoothest standing waves its endpoints permit: $\cos\pi u$, $\cos 2\pi u$, $\cos 3\pi u$. (b) Their joint values trace a curve, along which a shift moves the state. Each mode is one quadrature of a plane rotating at θ, 2θ or 3θ in the cover, and the interval sees only the projection, so the shift matrix in the cosine basis is dense rather than a block rotation and does not compose. Panels are a construction, not measured activations.

Figure 8 places the obstruction in context: the standing-wave modes of an open range are each one quadrature of a rotating plane in the cover, so a shift expressed in the cosine basis alone is a dense operator rather than a block rotation.

L.1 Mirroring a Neumann chain

For a Neumann chain, mirroring D rows at half-sample endpoints produces $N = 2D$ rows:

$$Y_{\text{double}} = [Y_0; \ldots; Y_{D-1}; Y_{D-1}; \ldots; Y_0], \qquad f(i) = \begin{cases} i, & 0 \le i < D, \\ 2D - 1 - i, & D \le i < 2D. \end{cases}$$

Duplicating each response preserves the mean, while reflection about $-1/2$ selects the type-II discrete cosine transform (DCT-II) profiles

$$c_k(i) = \cos \frac{\pi k(i + 1/2)}{D}, \qquad k = 0, \ldots, D - 1.$$

The shifted sine coefficients vanish, and the shifted Nyquist cosine is identically zero at the sampled sites; centering removes the constant. Mirroring whole-sample endpoints instead gives $2D - 2$ sites, unequal endpoint multiplicities and DCT-I modes. The endpoint sampling convention therefore changes the discrete problem.

With a complete cover basis and $W = Z^\top \bar{Y}_{\text{double}}$, the row $z_i \mathbf{R}^m W$ reconstructs the extended centered response at $i + m$, using block angles $\pi km/D$. Knowing the cover index permits exact reconstruction of the folded sequence, while truncation reconstructs its fit. However, vanished sine rows make the full paired $WW^\top$ singular, preventing use of the full-row inverse in Section 4.3. The coefficients can instead be obtained directly by a DCT without materializing the cover.

L.2 An odd Dirichlet extension

For Dirichlet endpoints, we choose an origin μ, write $X_i = Y_i - \mu$, $i = 1, \ldots, D$, and introduce two zero ghost sites. An odd extension then gives

$$X_{\text{double}} = [0; X_1; \ldots; X_D; 0; -X_D; \ldots; -X_1], \qquad N = 2D + 2.$$

This cover has zero mean regardless of the original mean. Its type-I discrete sine transform (DST-I) modes are $\sin(\pi ki/(D + 1))$, $k = 1, \ldots, D$, with vanishing cosine coefficients again preventing a full paired inverse. Shifts have block angles $\pi km/(D + 1)$. Decoding a ghost site gives μ, and decoding a mirrored site gives $2\mu - Y_i$, which may not correspond to an observed concept. Thus the pinned boundary is imposed by the extension itself.

L.3 Why folding can lose composition

Retaining the cover also makes the reflection operation explicit. With row convention $H = \text{diag}(1, -1)$, rotations $R(2\pi km/N)$ and reflections H form the dihedral action on plane k, satisfying

$$H^2 = I, \qquad HR(\alpha)H = R(-\alpha).$$

Reflection exchanges conjugate harmonics, so restricting to reflections separates cosine-even and sine-odd components. The zero-frequency and Nyquist blocks, when present, remain scalar exceptions.

Figure 9 collects the covers of the two preceding subsections and makes the cost of discarding one explicit. The pairing $j \leftrightarrow N - 1 - j$ has no fixed point for even N, so every interval site has exactly two preimages, travelling in opposite senses. One step of the ring action sends them to $f(j + m)$ and $f(j - m)$, and these agree only when $2m \equiv 0 \pmod N$, which for $N = 2D$ leaves the identity and the half-turn. At every other step the same interval position therefore admits two successors, and a map defined on position alone cannot choose between them. Repeated lift–rotate–fold steps need not equal one composed shift, whereas composition remains exact while the cover index or an equivalent orientation is retained. This construction identifies the required state without establishing a closed group of arbitrary interval-activation maps or a behavioral open-chain result. Appendix M records exploratory open-domain and dihedral checks.

M Supporting open-domain and dihedral records

An imposed reflecting number window. In historical `Llama-3-8B` number-word successor prompts, the one-through-twelve window has endpoint similarity -0.17 and neighboring similarity 0.20, using a mean-diagonal-normalized centered Gram statistic. This supports a non-closing

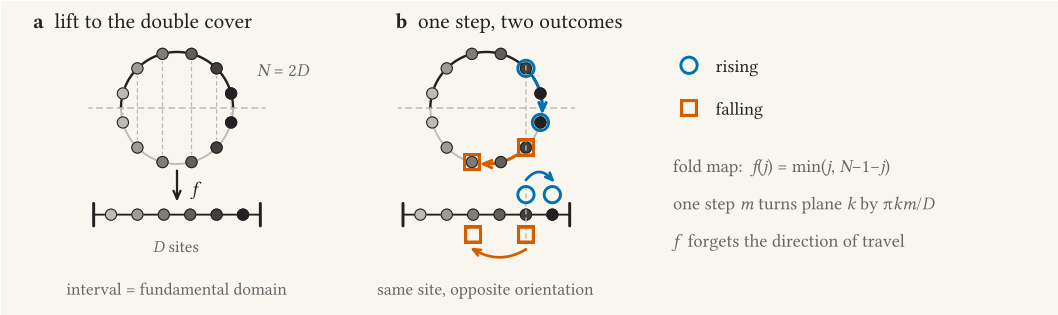


Figure 9: **Folding loses orientation.** (a) An open chain of D sites lifts to a ring of $N = 2D$ in which the interval is a fundamental domain of the reflection, with fold map $f(j) = \min(j, N-1-j)$. (b) Two cover points fold to the same interval site while travelling in opposite directions, so one step of the ring action sends them to different successors. A step of m turns plane k by $\pi km/D$. Composition is preserved while the cover index is retained and can fail once it is discarded. The panels are a construction, not measured activations.

ordered window but does not select a unique boundary condition. Half-step cosine coordinates impose repeated-endpoint reflection. At last-token block 24, $K = 5$ gives $9/9$ fitted-style in-window shifts; the boundary interface uses $K = 10$ with reconstruction $R^2 = 0.948$. The fold matches $12/12$ predictions, or $10/10$ after no-op-equivalent cases are excluded. Screw continuation gives $0/12$ despite $12/12$ real continuation-state copying. Refitting a longer window moves the wall, identifying a property of the interface rather than an intrinsic neural boundary.

An externally carried 25-point trajectory yields $25/25$ independent writes; repeated stateless folding instead absorbs at the wall. At an early concept-token site, screw continuation reaches $9/12$. A shared teen-residual direction, estimated from thirteen-through-twenty states, rescues $8/8$ tuned cases at selected block 20, where screw alone, direction alone and fold-plus-direction are each $0/8$. Fresh starts yield $7/12$ despite $12/12$ copy. This is tuned sufficiency with partial transfer, not window-only extrapolation or necessity. The lower-boundary fold gives $6/6$, or $5/5$ informative cases; a $2/3$ screw result uses a declared readable subset excluding some zero targets despite zero being in the vocabulary. A separate in-sample log-distance improvement has no behavioral continuation test.

Compass: complete anchors, permissive scoring. The `Gemma-2-2b-it` compass demonstration fits eight bearings on opposite-direction prompts at final block 25. Full-rank $K = 4$ gives $56/56$ rotations, $64/64$ reflections and $960/960$ mixed products by case-insensitive substring matching. Cardinal names can match within intercardinal outputs. Without the generated strings, these cannot be promoted to exact-label accuracy. Low-rank results remain near one-half. The 960 product evaluations collapse to eight reflection actions and 64 unique source/action pairs already covered by the reflection panel. Products are composed before one synthetic write, not by successive native operations. Full-rank anchor replacement can explain these task scores; no unseen-bearing or held-out-context result is saved.

Valence: forced choices and graded logits. A historical `Llama-3-8B` task uses nine ordered feeling words at last-token block 24 and forced pleasant/unpleasant choice against each landed word's baseline. Six of twelve shifted pairs already share the source label, giving no intervention $6/12$. On the six informative pairs, $K = 1$ gets $4/6$ and $K = 7$ gets $6/6$, while mean logit-difference error falls from 3.80 to 0.24. These are binary equivalence and graded-logit tests, not unconstrained sentiment generation. A direct sign flip matches $8/8$ informative mirrored labels. The $K = 1$ covering write differs from it on real anchors by up to 0.739 of mean edit size; equality requires the on-curve limit or an explicit implementation substitution. Separate early concept-token Gemma curves show rank effects without strict monotonicity.

Separate early-site Gemma demonstrations. Base `Gemma-2-2b` concept-token checks at block 1 give month counts $1, 7, 9, 9, 12, 12$ out of 12 at $K = 1, \ldots, 6$, versus $1/12$ for an April-

minus-January vector and $0/12$ no-op. Number counts at $K = 2, 4, 6, 8, 10, 11$ are $0, 1, 7, 6, 9, 9$ of 9, and valence counts at $K = 1, \ldots, 8$ are $6, 9, 11, 12, 12, 12, 12, 12$ of 12. These are prefix-matched full-vocabulary top-token checks, not the modern generated-answer panel. Rank effects are not uniformly monotonic.

N READING THE SAVED-DATA GALLERY

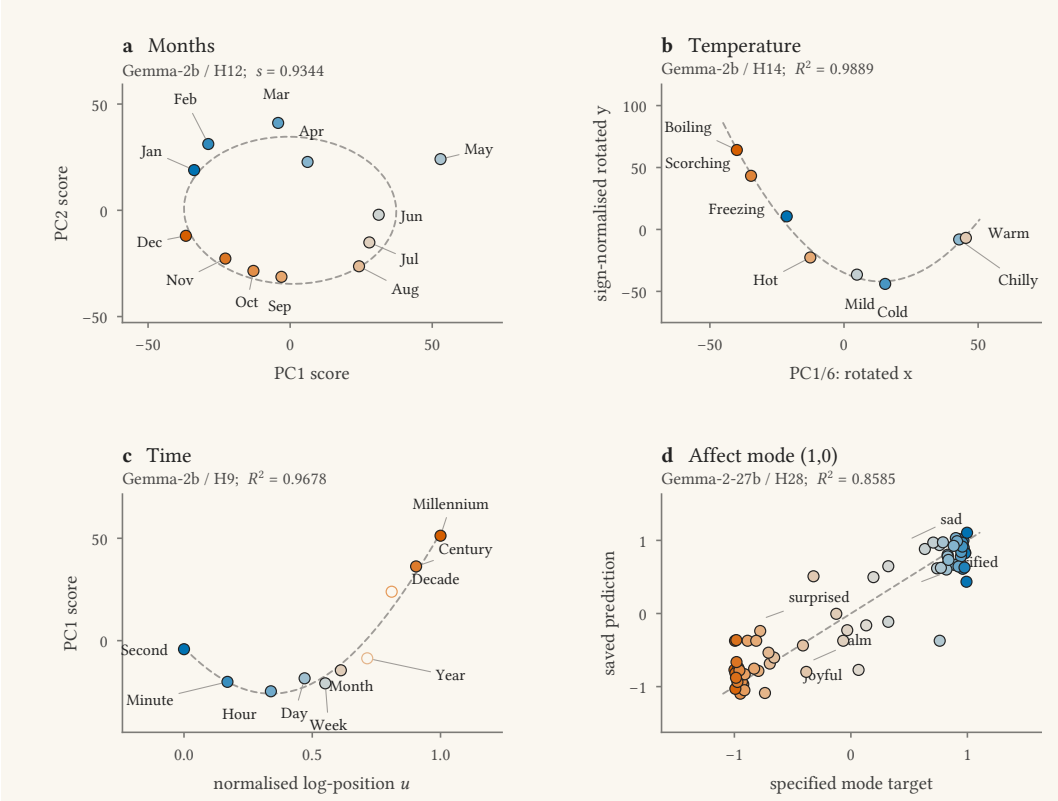


Figure 10: **Saved representation gallery.** Selected representation coordinates for months, temperature, time and affect with saved fits or specified targets. Hollow points were excluded from displayed coefficient fitting, but remain in reported scores. Selection and fitting use the same observations.

We use the gallery to examine how selected representation coordinates relate to the curves and target functions motivated by the theory (Figure 10). All 20 panels are included: nineteen agree with their source captions, while storage retains the discrepancy discussed below. The comparisons use the saved observations and fits, without new activations, PCA calculations or fitting.

Which representations are shown. The source implementation is an anonymized repository pinned to a single commit. Its numeric archive records the observations, candidate and displayed fits, masks, hidden-state indices and component identities. The one-dimensional comparisons use 300-dimensional `GloVe-Wiki-Gigaword` vectors and the `Gemma-2b` analysis with 18 decoder layers and width 2048. Historical checkpoint revisions are unknown, and these captures differ from the earlier `Gemma-2-2b-it` intervention setting. We retain the source's hidden-state indices, which it calls layers, without equating them to intervention hook stages.

How a display is selected. The source selects a display after mean-centering cached inputs. PCA retains at most $\min(N - 1, 6)$ components for periodic and logarithmic fits and $\min(N - 1, 10)$ for ordinal fits. Contextual examples search indices 1–18, excluding index 0, using the same words for selection and fitting. Periodic fits search PC pairs and harmonics one and two; ordinal fits search PC pairs and parabola orientation. Clipping is based on median residuals with a minimum retained

fraction, rather than a two-standard-deviation rule. Final saved ellipse/parabola coefficients and reported scores use all points. Both intermediate clipping masks and final coefficient-fit masks are retained, and all words remain visible. These in-sample comparisons provide neither held-out scores nor sampling error bars.

Two questions behind the numbers. Interpreting the resulting display requires distinguishing geometric fit from explained variation. Fit quality measures how well the selected coordinates follow a target curve, whereas retained variance measures how much activation variation those coordinates contain. The selected-PC retained-spectrum share divides their eigenvalue sum by the retained-PC sum; the full-space share uses the full centered spectrum. For a fitted mixture direction w, directional variance uses the corresponding denominator and differs from the sum of variances over all PCs used in the regression.

For an ellipse, the source combines the all-point coordinate scores as $s_{\text{ell}} = \sqrt{\max(0, R_x^2)\max(0, R_y^2)}$. This geometric mean clips negative coordinate scores and differs from a pooled multivariate R^2. Neither it nor the other target R^2 scores measures either variance fraction. Table 6 reports display scores alongside full-space fractions; exact retained fractions, selections and masks will be given in the supplementary material.

Following the ordered examples. For periodic concepts, the supplied month, weekday and compass order determines the progression around the fitted curves. We retain the saved parametric fits, including the source's choice to hide a fit in particular `GloVe` panels. Ordinal words also have a supplied order, and their coordinate signs follow the original display, including the second rotated coordinate's sign flip. A parabola can therefore summarize these selected points without identifying Neumann boundaries for a neural operator. Figures 11 and 12 show the periodic and ordinal panels.

Concept	Source/index	PCs or mode	Fit score	Full var. (%)
Months	GloVe	2/3	0.5481	21.83
Months	G2b/12	1/2	0.9344	45.43
Weekdays	GloVe	1/3	0.6264	84.30
Weekdays	G2b/5	1/2	0.9618	53.48
Compass	GloVe	3/4	0.9485	23.42
Compass	G2b/7	3/4	0.9870	31.14
Quality	GloVe	2/8	0.8746	23.94
Quality	G2b/18	1/9	0.9230	29.11
Temperature	GloVe	4/7	0.7656	21.74
Temperature	G2b/14	1/6	0.9889	38.68
Valence	GloVe	1/2	0.9564	46.14
Valence	G2b/1	1/3	0.8949	36.91
Storage	GloVe	2	0.9645	23.64
Storage	G2b/4	2	-0.2812	23.73
Time	GloVe	1	0.7578	27.50
Time	G2b/9	1	0.9678	29.57
Money	GloVe	1	0.7836	34.85
Money	G2b/1	1	0.9915	37.62
Affect	G27b/28	(1,0)	0.8585	7.26
Affect	G27b/28	(1,1)	0.6741	3.56

Table 6: All 20 notebook fit panels in the saved replay. G2b and G27b abbreviate the source `Gemma-2b` and `Gemma-2-27b`; suffixes are hidden-state indices. PCs are one-based. The full-space fraction belongs to the displayed PCs or fitted affect direction, not to fit quality. Periodic scores use the source's geometric mean of nonnegative coordinatewise R^2; other rows report target R^2. The storage G2b/4 row reports the actual PC2 display, not the selected PC6 score. All selections and scores are in-sample.

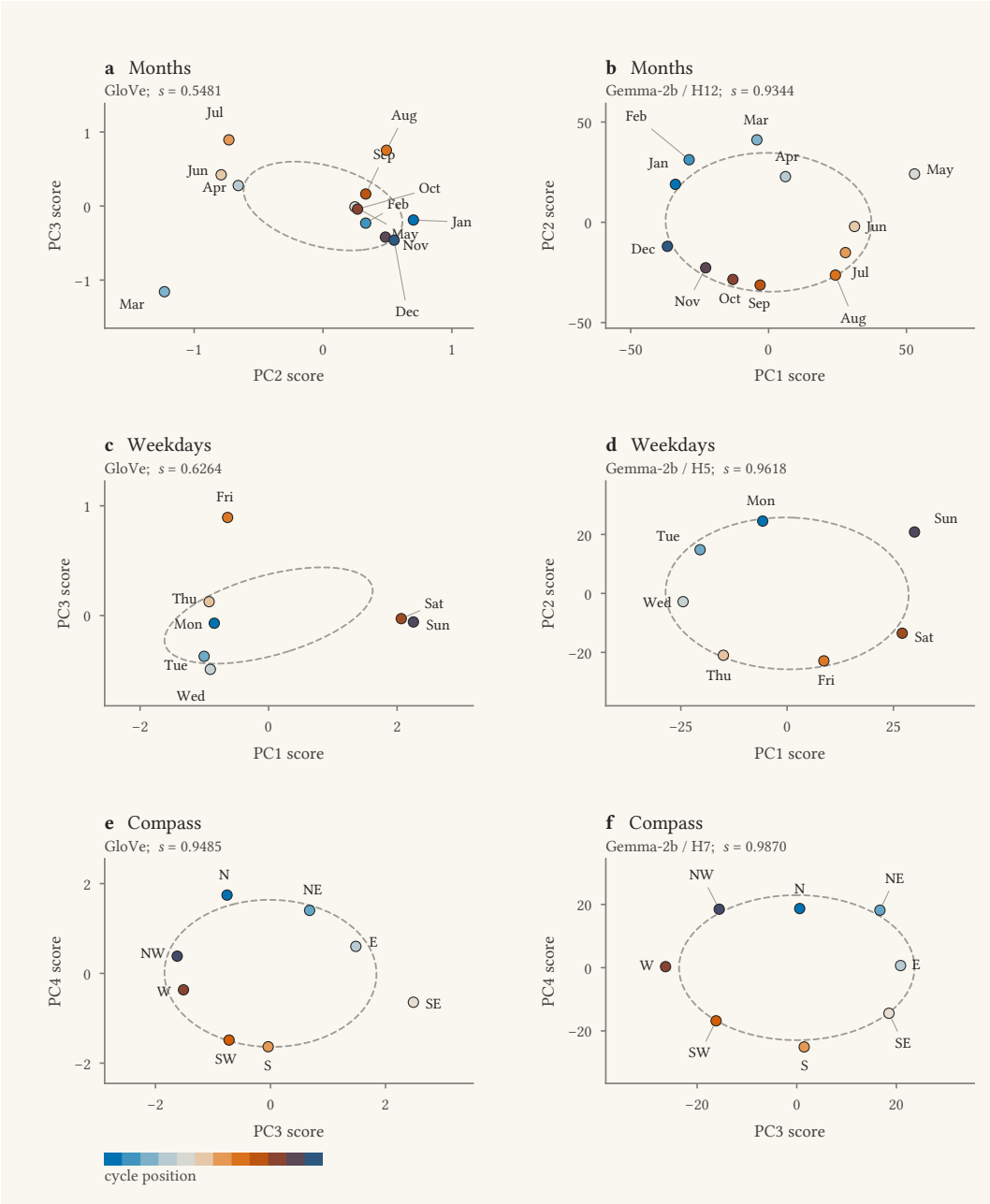


Figure 11: **Periodic representations.** Rows compare months, weekdays and compass directions; columns compare `GloVe` and `Gemma-2b`. Colors and word labels show how semantic order relates to each curve. Gray lines are the saved displayed fits, and titles report hidden-state indices where applicable and all-point ellipse scores s_{ell}. Axes identify the selected PCs; selection and fitting use the same points.

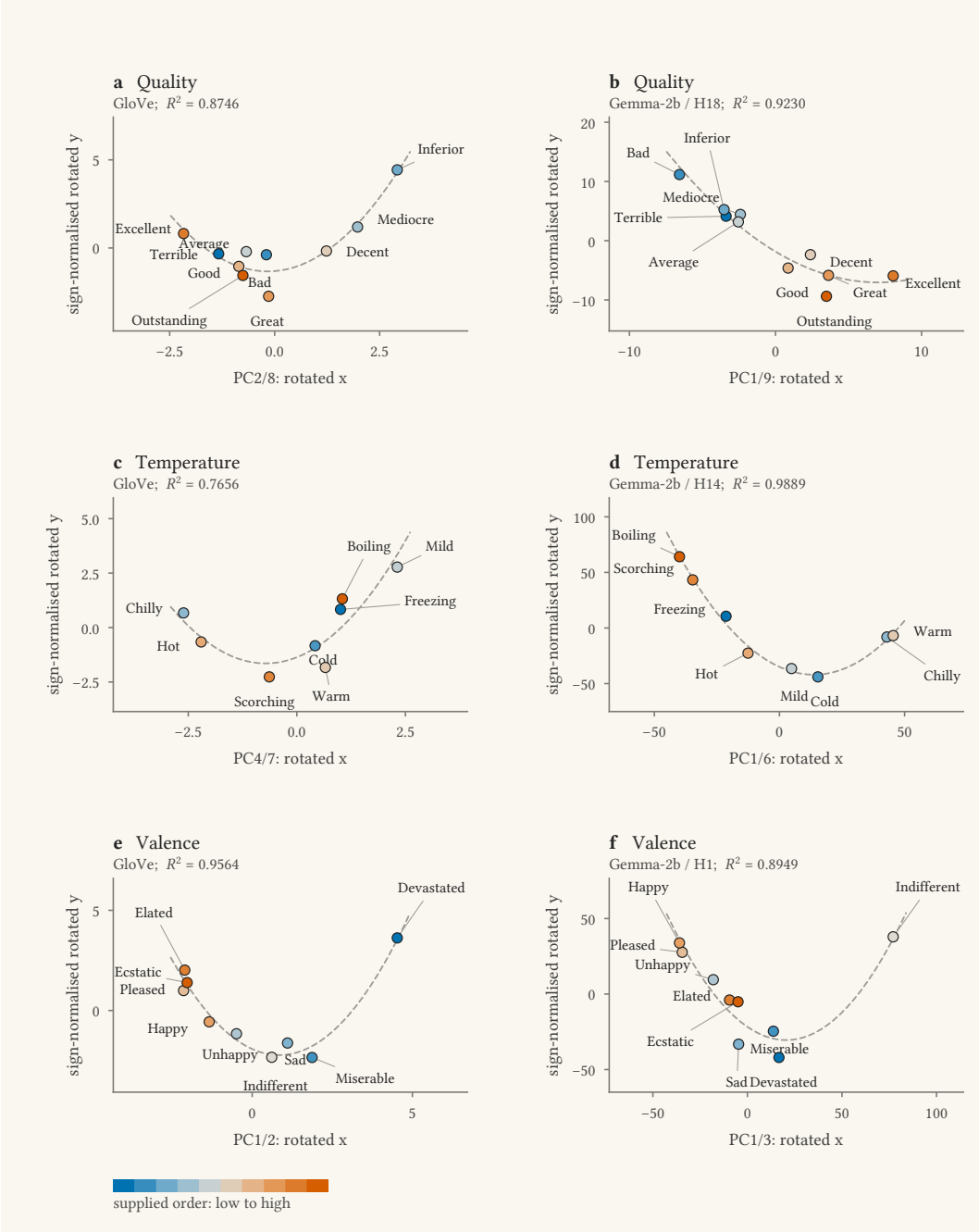


Figure 12: **Ordinal representations.** Rows show quality, temperature and emotion valence for `GloVe` and `Gemma-2b`. Colors follow the supplied low-to-high order, allowing the coordinate progression to be compared with the saved parabola. Each panel retains all observations, its selected PC plane and source sign convention. Titles give the selected hidden-state index and all-point fit R^2.

N.1 WHY LOGARITHMIC STORAGE IS A DIAGNOSTIC

Logarithmic quantities expose a consequential difference between two fitting procedures. The notebook fits $A\sin(\pi u/2) + B\cos(\pi u/2) + C$ to one PC, with normalized logarithmic coordinate u. Fixing the frequency at $\pi/2$ while varying A, B, C permits phase, amplitude and offset changes. The library instead regresses a capped PC mixture onto fixed targets $\sin(n\pi u/2)$. The notebook's initial PC/layer search scores unclipped all-point fits; subsequent clipping changes displayed coefficients, so selection and display scores can differ.

This fitting freedom matters when interpreting endpoints. A free-phase curve need not satisfy either endpoint condition, and the library's $n = 2$ target has nonzero slope at the upper endpoint. Consequently, the observed fits describe logarithmic-coordinate relationships without confirming pinned–free modes.

The storage example makes the distinction visible. At `Gemma-2b` index 4, selection chooses PC6 with score 0.9296, while the float64 display helper reselects PC2. The latter gives all-point $R^2 = -0.2812$, below a constant-mean baseline, and is retained in Figure 14. The source's zero-based component indices are respectively 5 and 1. We report both alternatives.

N.2 AFFECT TARGETS FROM MIXTURES OF PCS

The affect comparison extends the construction to two coordinates motivated by a circumplex account (Russell, 1980). The `Gemma-2-27b` analysis uses 69 words with valence/arousal coordinates attributed to `NRC-VAD 2018` (Mohammad, 2018); agreement with an independent official copy remains unverified. Each mode target is fitted by least squares from ten PCs and an intercept on these same words. Index 28 maximizes the summed $(1, 0)$ and $(0, 1)$ target scores, while the displayed targets are $(1, 0)$ and $(1, 1)$. Thus each prediction is a fitted linear mixture of PCs, shown in Figure 13.

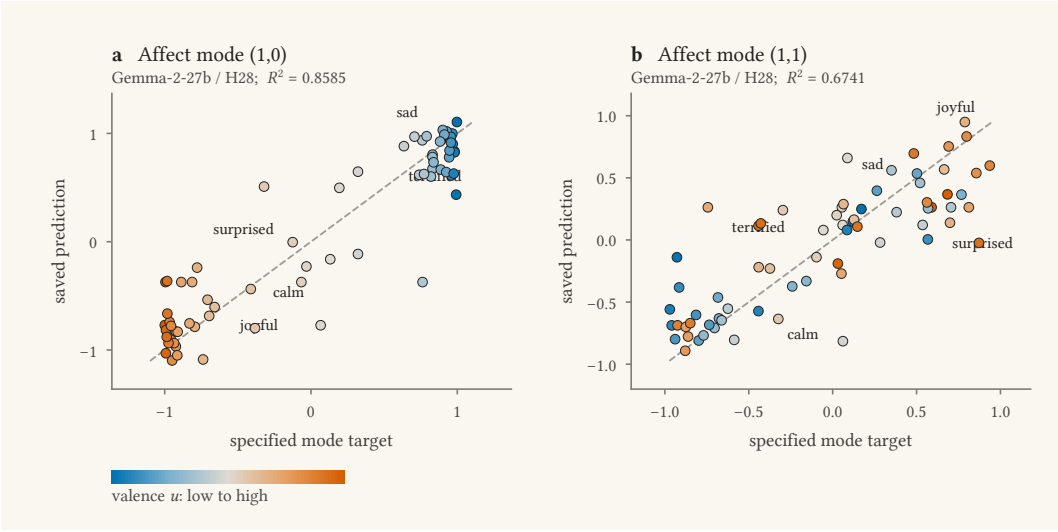


Figure 13: **Affect representations.** PC-mixture predictions are compared with the saved targets (a) $\cos(\pi u)$ and (b) $\cos(\pi u)\cos(\pi v)$. The diagonal shows exact agreement, color indicates valence u, and selected labels identify words. All 69 words at source index 28 enter both in-sample scores; the saved data retain every target and label.

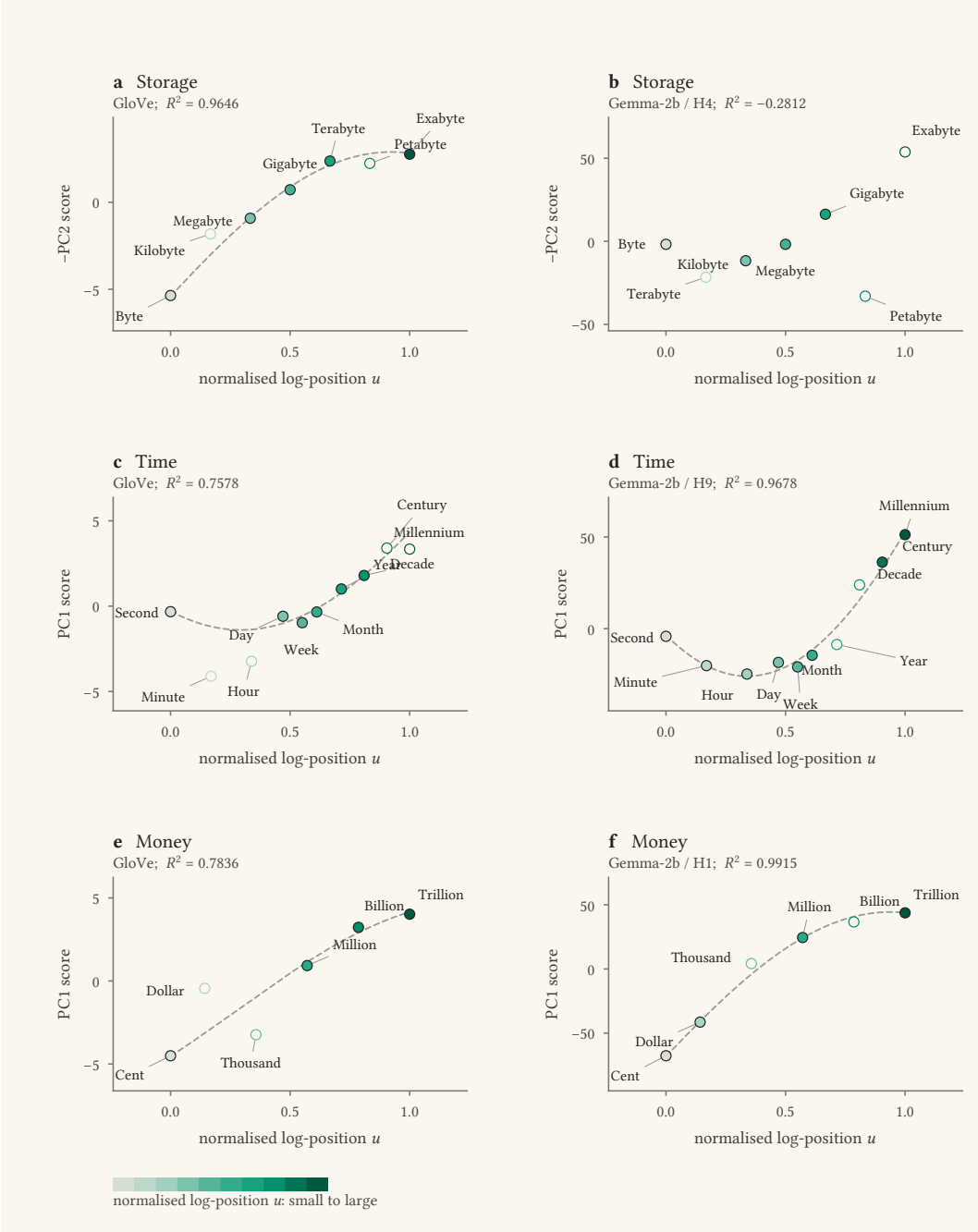


Figure 14: **Logarithmic representations and the storage discrepancy.** Rows compare storage, time and money for `GloVe` and `Gemma-2b`, using normalized logarithmic position. Curves and all-point R^2 values correspond to the saved display components and coefficients. The `Gemma-2b` storage panel shows PC2 even though selection chose PC6 at that index. Hollow points were excluded from fitting the displayed coefficients but remain in the reported score.