CIRCLES OR ORDERINGS? A SYMMETRY TEST FOR CYCLIC CONCEPTS IN LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

The linear representation hypothesis treats a concept as a direction. Cyclic concepts such as days of the week and months look like an exception: several studies report that language models place them on a circle (Engels et al., 2025; Karkada et al., 2026; Feucht et al., 2026). We show that the tests behind such reports cannot tell a cycle from an ordering. A fitted successor operator has its spectrum at the Kth roots of unity for any tokens in a fixed order, random nouns included. Concentration of variation in the fundamental Fourier mode is as high for number words, ordinals, and temperature adjectives as it is for months. A short step from the last element back to the first also appears for graded scales whose two extremes lie close together. We derive a test from the symmetry itself. Advancing a concept by one step acts as an isometry on its K states exactly when their Gram matrix is circulant, so we measure the departure from circulance. Across nine models from three families (Pythia 160M to 12B, Llama-3.1-8B base and instruct, Qwen2.5-7B-Instruct), days and months are more circulant than every same-length ordered control in every model, measured at the concept token. Clock hours cannot be distinguished from a temperature scale, and compass directions are organized by cardinal versus intercardinal type rather than as an eight-cycle. Behavior does not follow automatically from the representation. With the same closed representation of months, Llama shows a wrap-around penalty consistent with the base-10 mechanism reported by Feucht et al. (2026), while Qwen does not. On the compass, Llama never answers a 45-degree step correctly, which matches its two-axis representation.

1 INTRODUCTION

The linear representation hypothesis (LRH) says that a concept is a direction in activation space, and that changing the concept means adding a vector, $h \mapsto h + t\,v$ (Park et al., 2024; Marks & Tegmark, 2024; Turner et al., 2024; Li et al., 2023; Zou et al., 2025). A direction suits a concept that can increase or decrease. It does not suit a concept that returns to where it started. Sunday is followed by Monday, and December by January. A representation that respects this cannot lay the seven days on a line without tearing the cycle. Consistent with this, Engels et al. (2025) find circular features for days and months, and a growing literature explains such circles through the statistics of language (Karkada et al., 2026; Modell et al., 2025; Prieto et al., 2026).

This paper asks a different question: what evidence shows that a representation is a cycle rather than an ordering? The usual evidence is a circular fit. One projects the states onto a plane, or fits the sine and cosine of $2\pi d/K$, or fits the linear map that sends each element to its successor and reads its eigenvalues. We show that all three respond to order, not to closure. Any K tokens visited in a fixed order give a successor map whose eigenvalues are the Kth roots of unity. A straight line traversed in order puts most of its variation into the fundamental Fourier mode, so it passes a circular fit. We confirm both effects on real models, using control sets that have an order but no wrap: number words, ordinals, sizes, and temperature adjectives. On the Fourier test, months are indistinguishable from ordinals.

Our test comes from the symmetry a cycle is supposed to have. If advancing a concept by one step is a rigid motion of its K states, every distance between two states can depend only on how far apart they are around the cycle. Equivalently, the Gram matrix of the states is circulant (Proposition 1). A

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.

line fails this badly. A random set of points passes it, but has no preferred frequency, and the Fourier test already rejects random points. Together the two tests separate cycles from both orderings and noise.

Applied to nine models, the test gives a clear and partly negative picture. Days and months are more circulant than every same-length ordered control in all nine models, at the token that names the concept. Clock hours, which a Fourier test ranks as the most circular concept, cannot be told apart from a temperature scale. Compass directions are organized by direction type, not as a ring. The closed geometry is also not the whole story of computation. Feucht et al. (2026) show causally that Llama-3.1-8B solves "six months after August" with a base-10 addition circuit. We replicate their behavioral numbers, find that a closed representation coexists with a wrap-around penalty in Llama and with none in Qwen, and find that Llama's compass arithmetic follows its two-axis compass representation.

We summarize the contributions:

- We show that three standard signals of circular structure (successor-operator spectra, held-out successor fits, and fundamental Fourier concentration) are produced by any consistent ordering, and that a wrap-gap test is fooled by graded scales (Section 3).
- We derive a symmetry test, non-circulance of the Gram matrix, from the requirement that one step act as an isometry, and pair it with the Fourier test (Section 2).
- Across nine models, days and months pass against every ordered control; clock, compass, and seasons do not (Section 4).
- Behavior is not fixed by the representation: the wrap-around penalty differs by model and by prompt, and compass arithmetic follows the compass representation (Section 5).

2 SETUP

2.1 A CYCLIC CONCEPT AS A SYMMETRY

Write $\Phi(x) \in \mathbb{R}^k$ for a model's state on input x, and let g advance a concept by one step (Monday to Tuesday). A cyclic concept of period K comes with the cyclic group $\mathbb{Z}_K = \langle g \mid g^K = 1 \rangle$. The natural hypothesis is approximate equivariance, $\Phi(g \cdot x) \approx \rho(g)\,\Phi(x)$ for a linear map $\rho(g)$, the assumption behind equivariant deep learning (Cohen & Welling, 2016; Kondor & Trivedi, 2018; Bronstein et al., 2021). Here it is a hypothesis about a trained model, not a property of the architecture.

Fitting such a $\rho(g)$ does not show the concept is cyclic. Such a map exists whether or not the states form a cycle. Take the centered states $z_0, \ldots, z_{K-1}$ of one sentence template, one per element of the cycle. If they are linearly independent apart from $\sum_d z_d = 0$, a linear map on their span sending each z_d to z_{d+1} always exists, and its eigenvalues are the Kth roots of unity other than 1 (Serre, 1977). States of distinct tokens are nearly independent in high dimensions, so this holds for random nouns as well as for days. What a genuine cycle adds is that the step is rigid: it preserves the geometry of the states.

Proposition 1. *Let $z_0, \ldots, z_{K-1} \in \mathbb{R}^k$ and let $G_{de} = \langle z_d, z_e \rangle$ be their Gram matrix. There is an orthogonal Q with $Qz_d = z_{d+1 \bmod K}$ for all d if and only if G is circulant, that is, G_{de} depends only on $(e - d) \bmod K$. In that case the eigenvalues of G are the Fourier energies $E_j = \frac{1}{K} \| \sum_d z_d\, e^{-2\pi i j d/K} \|^2$.*

The forward direction is one line: $\langle z_{d+1}, z_{e+1} \rangle = \langle Qz_d, Qz_e \rangle = \langle z_d, z_e \rangle$. For the converse, the sequences (z_d) and (z_{d+1}) have the same Gram matrix, so an orthogonal map carries one to the other (Horn & Johnson, 1985). The spectral statement holds because the discrete Fourier transform diagonalizes every circulant matrix (Gray, 2005). Appendix B gives details.

Two consequences organize the paper. First, whether a concept acts as a rigid cycle is a property of G alone, which we measure by the non-circulance

$$\nu = \frac{\|G - \mathrm{circ}(G)\|_F}{\|G\|_F}, \tag{1}$$

where $\mathrm{circ}(G)$ replaces each wrapped diagonal of G by its average. This is the circulant matrix closest to G in Frobenius norm (Chan, 1988), and it equals the average of G over all K cyclic shifts (Du et al., 2026). So ν is the share of G that changes when every element moves one step along the cycle. A perfect circle has $\nu = 0$. A line $z_d \propto d$ has $\nu \approx 0.87$. Second, circulance alone does not make a circle. K random points in high dimensions are nearly equidistant, so G is close to a multiple of the identity, which is circulant. What distinguishes a circle is that its energy sits in the fundamental pair $j = \pm 1$. When both hold, the step is the restriction of a rotation by $2\pi/K$ to the states, a rotation of $SO(2)$ sampled at K points.

The same picture applies to the LRH. For a rotation e^{tL} with generator L, one small step is $e^{tL}h \approx h + t\,Lh$: a translation along the local direction Lh (Hall, 2015). A direction is the first-order description of the action at one state. Around a closed cycle the local directions turn, and their average over the K states is zero, so no single vector advances the concept everywhere. We use this only as an argument about when a direction must fail; we do not test it separately.

2.2 Datasets, models, and how we read out states

A family is one sentence template with one slot, filled with each of the K elements in order, for example "The meeting is scheduled for Monday." through "... for Sunday." Every set has 100 families, formed from 10 templates and 10 subject nouns, and the data are built by string substitution with no model involved. Besides days (K=7), months (12), clock hours (12), compass directions (8), and seasons (4), we build controls with an order and no wrap: number words and ordinals (K=7 and 12), a seven-step size scale, and temperature scales (K=7 and 12, freezing to scorching). Random nouns in a fixed order (K=7 and 12, same templates as days and months) have neither order nor wrap. Appendix A gives every set in full, with its sentence frames and subject nouns.

We study Pythia at six sizes from 160M to 12B (Biderman et al., 2023), Llama-3.1-8B and Llama-3.1-8B-Instruct (Grattafiori et al., 2024), and Qwen2.5-7B-Instruct (Yang et al., 2025), nine models from three families, through TransformerLens (Nanda & Bloom, 2022). We read the normalized residual stream entering each block (`ln1.hook_normalized`) at two positions: the last token of the concept ("Monday", or "clock" in "one o'clock") and the final token of the sentence. We center each family over its K states, which removes the template, and project onto the top $4K$ principal components of the pooled centered states. The projection does not use the cyclic order.

The main results use one layer per model fixed before this study: for Pythia, a layer chosen earlier by an unrelated criterion on sentiment data, and 0.65 of the depth for the other models. We check robustness in two ways. We choose the layer on five templates and evaluate on the other five, with the same procedure applied to every control, and we report the result at every layer. Uncertainty is a 95% bootstrap over the 10 templates, which are the independent units.

2.3 Three measures of circularity

For one family, let E_j be the energy in Fourier mode j of $z_0, \ldots, z_{K-1}$ along the cyclic index. We report three quantities and average them over families.

- Fundamental share, $\mathrm{circ} = (E_1 + E_{K-1}) / \sum_{j \geq 1} E_j$, relative to its value $2/(K-1)$ for points with a flat spectrum. This is the logic of circular probes (Engels et al., 2025). Its ceiling is $(K-1)/2$, so we also report the normalized form $(\mathrm{circ} - b)/(1-b)$ with $b = 2/(K-1)$, which is 1 for a circle and 0 for random points at any K.
- Wrap gap, $\|z_{K-1} - z_0\|$ divided by the median of the other steps. It equals 1 for a circle and $K-1$ for a line.
- Non-circulance ν, Eq. (1).

Table 1 gives the values of each measure for a circle, a line, a half-circle arc, and random points, next to the measured ranges.

The decision rule follows from Proposition 1. A set counts as a cycle when its fundamental share is clearly above that of random points and its non-circulance is below that of ordered controls of the same length, measured in the same model. Neither condition is enough alone. Random points are

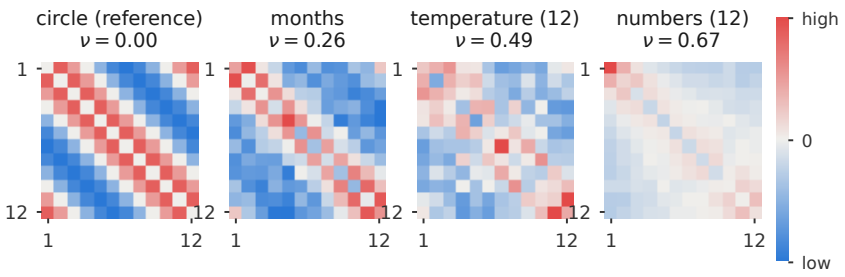


Figure 1: Mean Gram matrices of the 12 states (Pythia-12B, concept token, fixed layer), after subtracting the mean diagonal, which is circulant, and scaling each panel to its own maximum. A circulant matrix is constant along diagonals that wrap around the corners, as for the circle. Months show the same banded, wrapping pattern. The temperature scale is irregular, and the number words are dominated by one element. Titles give the non-circulance ν of Eq. (1).

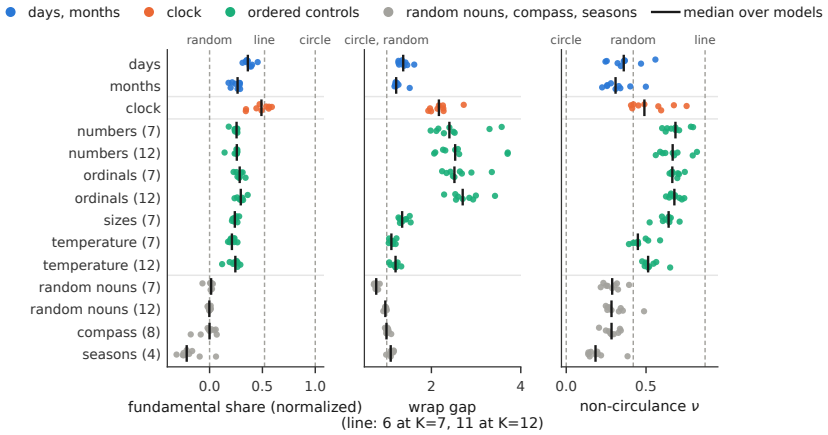


Figure 2: The three measures for every set, at the concept token and the fixed layer. Each dot is one of nine models; the black tick is the median. Dashed lines mark reference shapes. Left: the fundamental share puts months at the level of number words and ordinals. Middle: the wrap gap rejects numbers and ordinals but not the graded scales (sizes, temperature). Right: within each model, non-circulance is lower for days and months than for every ordered control (Figure 3); random nouns, compass, and seasons are also low, and are rejected by the fundamental share.

nearly circulant, and so is any set whose step is rigid but whose energy sits in a higher mode; a line has a high fundamental share but is far from circulant. Figure 1 shows what the test looks at.

3 Standard circle tests respond to order, not closure

We begin with the evidence a circle claim usually rests on, and show that ordered controls produce it too.

The successor spectrum is fixed by construction. The ridge-fitted successor map has its leading eigenvalue on the unit circle at the angle $2\pi/K$ for days, months, number words, and temperature alike ($51.4°$ or $30.0°$, modulus 0.99 to 1.00). For random nouns it sits at another Kth root of unity, with the same modulus. The fitted map also generalizes to held-out templates for random nouns, with R^2 between 0.75 and 0.96, in the range of days and months (Appendix C). None of these quantities separates a cycle from an arbitrary order.

The fundamental share does not separate cycles from orderings. A line traversed in order has Fourier energy falling as $1/\sin^2(\pi j/K)$, so about two thirds of it sits in the fundamental mode, and a line scores $\mathrm{circ}/\mathrm{base} = 1.99$ at $K{=}7$ and 3.44 at $K{=}12$. The measured controls come close: number words and ordinals reach $\mathrm{circ}/\mathrm{base}$ from 1.36 to 2.62, against 1.81 to 2.30 for months. Months

Table 1: Range of each measure over the nine models (concept token, fixed layer), with reference shapes at $K=7$ and $K=12$. Straightness is the share of within-family variance explained by a straight line in the order index.

set	circ/base	normalized circ	wrap gap	non-circulance	straightness
days	1.62–1.91	0.31–0.46	1.26–1.62	0.25–0.56	0.27–0.37
months	1.81–2.30	0.18–0.29	1.17–1.51	0.22–0.50	0.16–0.22
clock	2.54–3.65	0.34–0.59	1.93–2.72	0.41–0.75	0.31–0.45
numbers, ordinals	1.36–2.62	0.14–0.36	1.98–3.71	0.56–0.82	0.30–0.55
graded scales (sizes, temperature)	1.35–2.31	0.12–0.29	1.04–1.54	0.39–0.71	0.14–0.27
random nouns	0.87–1.06	−0.07–0.03	0.71–1.00	0.22–0.49	0.08–0.14
compass	0.55–1.17	−0.18–0.07	0.96–1.10	0.21–0.34	0.10–0.15
seasons	0.84–1.03	−0.31–0.06	0.99–1.16	0.14–0.38	0.31–0.39
reference: circle (K=7 / 12)	3.00 / 5.50	1.00 / 1.00	1.00 / 1.00	0.00 / 0.00	0.33 / 0.31
reference: line (K=7 / 12)	1.99 / 3.44	0.50 / 0.54	6.00 / 11.00	0.87 / 0.88	1.00 / 1.00
reference: arc (K=7 / 12)	2.50 / 4.36	0.75 / 0.75	3.86 / 7.03	0.69 / 0.74	0.78 / 0.80
reference: random (K=7 / 12)	0.99 / 1.00	0.00 / 0.00	1.00 / 1.00	0.41 / 0.43	0.17 / 0.09

Table 2: Number of models, out of nine, in which the cycle scores as more cyclic than a same-length ordered control (higher circ/base; lower wrap gap, straightness, and non-circulance). The last column counts models whose template confidence intervals for non-circulance do not overlap.

cycle	control	circ/base	wrap gap	straightness	non-circulance	CIs separated
days	numbers (7)	9/9	9/9	9/9	9/9	9/9
days	ordinals (7)	9/9	9/9	9/9	9/9	9/9
days	sizes (7)	9/9	5/9	0/9	9/9	9/9
days	temperature (7)	9/9	0/9	0/9	9/9	6/9
months	numbers (12)	6/9	9/9	9/9	9/9	9/9
months	ordinals (12)	2/9	9/9	9/9	9/9	9/9
months	temperature (12)	5/9	1/9	5/9	9/9	9/9

score above ordinals in only 2 of the 9 models (Figure 2, left; Table 2). Random nouns stay at 1, which is why the test looks informative when random nouns are the only control.

The wrap gap does not separate cycles from graded scales. It does reject number words and ordinals, whose last step back to the start is 2 to 3.7 times a typical step, while days and months stay between 1.2 and 1.6. The temperature and size scales, however, close up just as well (1.04 to 1.54). Their two extremes, freezing and scorching, lie close together, as in a horseshoe. Months have a smaller wrap gap than temperature12 in only 1 of 9 models, and days than temperature7 in none (Figure 2, middle).

4 Days and months pass the symmetry test

Within each model, non-circulance is lower for the cycles than for every ordered control. Across models the ranges overlap: days range from 0.25 to 0.56 and months from 0.22 to 0.50. Number words and ordinals range from 0.56 to 0.82 and the graded scales from 0.39 to 0.71. Model-level differences in scale make the within-model comparison the relevant one, and there the result holds without exception. In each of the seven same-length comparisons, the cycle is more circulant than the control in all nine models (Table 2, Figure 3). The template intervals do not overlap in all nine models for six of the seven comparisons, and in six of nine for days against temperature7, the hardest control. The gaps are large except against that control. The median difference in ν between control and cycle is 0.21 to 0.38 in the other six comparisons, with 0.09 in the closest model, but only 0.08 for days against temperature7, with 0.03 in the closest model. Nine of nine in the same direction has a one-sided sign-test probability of 0.002; the nine models share templates and, within Pythia, training data, so this is a guide rather than an exact p-value. The first condition of the decision rule is met by a wide margin: the normalized fundamental share of days and months is at least 0.18 in every model, against at most 0.07 for random nouns, compass, and seasons (Figure 5).

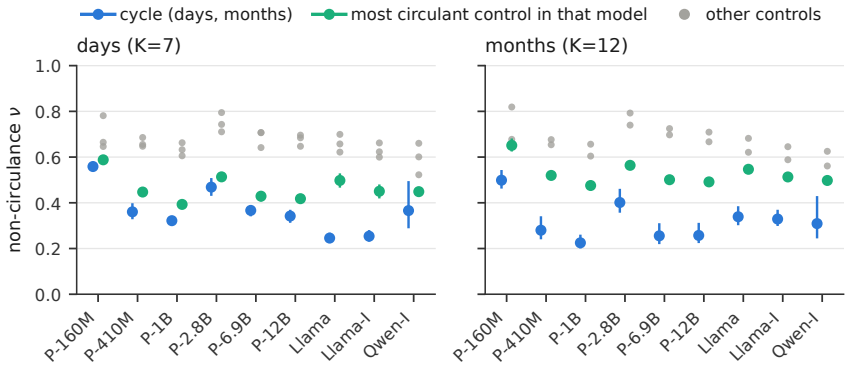


Figure 3: Non-circulance per model with 95% template bootstrap intervals. Blue: the cycle. Green: the same-length ordered control that is most circulant in that model, the hardest comparison. Gray: the other controls. Model labels: P = Pythia, Llama = Llama-3.1-8B, -I = instruct.

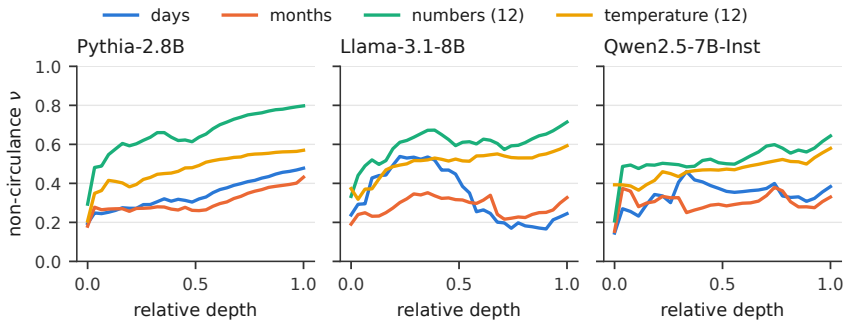


Figure 4: Non-circulance across depth at the concept token. In Pythia and Qwen the cycles stay below the controls at nearly every layer. In Llama-3.1-8B, days rise to the level of the temperature scale between about a quarter and a half of the depth and separate again later.

The result does not depend on the choice of layer. Months are more circulant than each control at 91% or more of the layers of every model. Days beat number words, ordinals, and sizes at every layer of every model, and beat temperature7 at 93% to 100% of layers in Pythia and Qwen but at only 55% to 58% in the two Llama models, where middle layers do not separate them (Figure 4). Selecting the layer on half of the templates and testing on the other half gives the same direction in seven to nine of the nine models for every comparison.

Two choices in the analysis could have produced this result as an artifact, and we checked that neither one did. First, word-specific variation spread evenly in all directions adds roughly a multiple of the identity to each Gram matrix. The identity is circulant, so such noise lowers ν for any set, and a set with more of it would look more cyclic. The identity-invariant version $\nu' = \|G - \mathrm{circ}(G)\|_F / \|G - \frac{\mathrm{tr}\, G}{K} I\|_F$ is unchanged by such noise. Second, the $4K$-dimensional projection is a choice. With ν or ν', and with $2K$, $4K$, or $8K$ components or no projection at all, every one of the seven comparisons holds in all nine models (Table 8, Appendix D).

The result is strongest at the concept token. At the final token of the sentence the cycles still beat number words and ordinals in all nine models, but they no longer separate from the graded scales (days beat temperature7 in 2 of 9 models, months beat temperature12 in 5 of 9). This is consistent with Feucht et al. (2026), who, in arithmetic prompts, find the circle at the concept token and not at the final position where Llama combines its inputs. The month cycle also survives inside a date such as "10 January", while the day of the month stays an ordering (Appendix G).

The other periodic sets, clock, compass, and seasons, fail the decision rule, each for a different reason. Clock hours have the highest $\mathrm{circ}/\mathrm{base}$ of all (2.54 to 3.65), yet their non-circulance (0.41 to 0.75) is below that of number words by a small margin in each model, level with temperature12

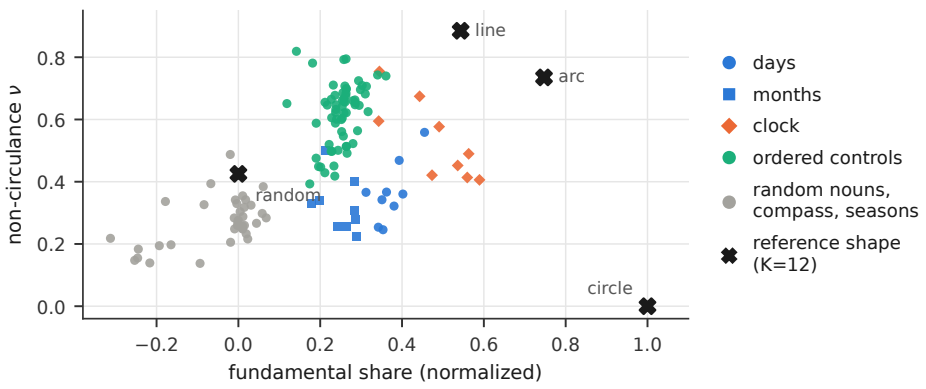


Figure 5: The decision rule. Each point is one set in one model (concept token, fixed layer); crosses are reference shapes at $K{=}12$. Days and months are the only sets with a fundamental share above random points and a non-circulance below the ordered controls. Clock sits between months and the controls; random nouns, compass, and seasons are circulant but have no fundamental energy.

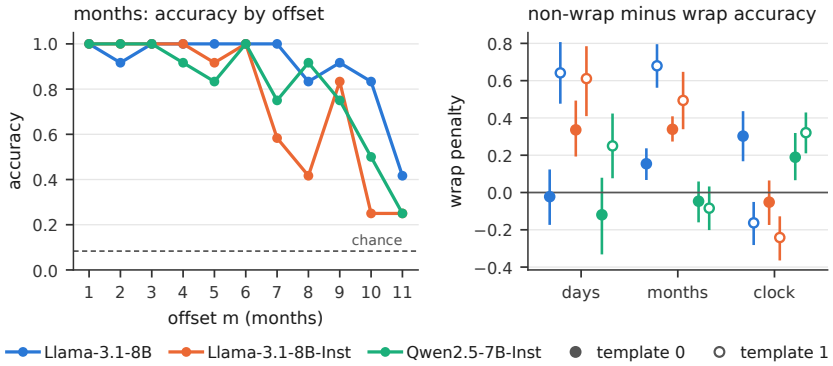


Figure 6: Left: accuracy by offset m on "What month is m months after X?", Feucht et al. template. Right: the wrap-around penalty, non-wrap minus wrap accuracy compared at the same offset, with 95% bootstrap intervals, for two prompt templates. A rigid cycle used as such predicts zero.

(lower in 5 of 9 models), and above that of months in all nine. Much of the clock's apparent circularity is the number line of its hour words. Compass directions have no excess energy in the fundamental mode (0.28 of their variance, against 0.29 for random points), and instead put 0.34 to 0.59 into mode $j{=}4$, which alternates between cardinal and intercardinal directions (random points: 0.14). They are organized by direction type, not as an eight-cycle (Figure 7, left). Their non-circulance is low (0.21 to 0.34), so the step itself is close to rigid, but it acts on a two-axis pattern rather than a ring. This is the case the decision rule is built for (Figure 5). Seasons score at or below the random baseline on the fundamental share (0.84 to 1.03), with excess energy in mode 2 in eight of nine models, which pairs spring with autumn and summer with winter.

We call the passing concepts cycles rather than $SO(2)$ rotations. The test establishes a rigid action of $\mathbb{Z}_K$ on K states. Showing a continuous rotation would require states between the K elements to land at intermediate angles. Our attempt with half hours is confounded, because the concept token of "half past one" is the word "one", and it is inconclusive (Appendix E).

5 The representation does not fix the computation

If a model computed "m steps after X" by stepping around a closed representation, the answer should not depend on whether the steps pass the end of the cycle. We test this from the output alone. We use the prompt format of Feucht et al. (2026), "Q: What month is six months after August?\nA:", plus one paraphrase. We score the K candidate answers by their summed log-probability followed

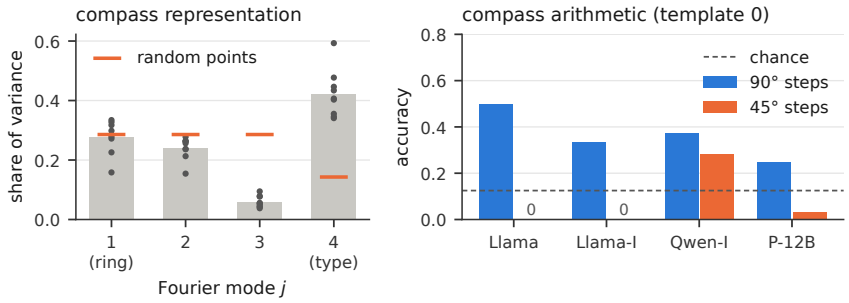


Figure 7: Left: Fourier mode energy of compass directions (bars: mean over nine models; dots: models; ticks: random points). Mode $j{=}4$ alternates cardinal and intercardinal directions. Right: accuracy on "What direction do you face after turning $45m$ degrees clockwise from X?" (template 0), split by 90-degree and 45-degree steps.

by a newline, and take the most likely one. Wrapped queries have larger offsets on average, and accuracy falls with the offset. We therefore compare wrapped and unwrapped queries at the same offset and average the differences (Appendix F).

With their template, base Llama-3.1-8B answers 0.90 of month queries and 0.93 of day queries correctly, close to the 82% and 92% that Feucht et al. (2026) report on a slightly larger range of offsets. Chance is 0.08 and 0.14. Base Pythia-12B stays near chance beyond one step (0.11 for months).

The wrap-around penalty depends on the model and on the prompt, although all three models represent months as a cycle by our test (Figure 6, right). Llama-3.1-8B-Instruct loses 0.34 in accuracy on wrapped queries at the same offset, for days and for months. Base Llama shows a penalty of 0.15 for months with the Feucht template (-0.02 for days) and 0.68 with the paraphrase (0.64 for days). Qwen shows none for months with either template (-0.05 and -0.08, both intervals containing zero), at an accuracy of 0.81. A penalty is what a sum computed in a number system with a privileged start predicts, the mechanism Feucht et al. (2026) identify in Llama. Its absence in Qwen shows that a closed representation neither forces nor rules out homogeneous behavior.

The compass shows the representation constraining behavior. Asked which direction one faces after turning a multiple of 45 degrees, both Llama models answer the 90-degree steps correctly 0.33 to 0.50 of the time and the 45-degree steps never, under both templates (Figure 7, right). On 45-degree steps, 97% to 100% of their answers keep the type of the starting direction: a cardinal start gives a cardinal answer. This is the structure of the representation, two interleaved four-cycles distinguished by mode 4, rather than one eight-cycle. Qwen, whose behavior is less extreme (0.28 on 45-degree steps), is also less tied to the type split in its answers (47% same type). This link is correlational; we did not intervene on the representation.

6 Scope and Negative Results

Beyond circulance, the symmetry view makes other predictions. We test two here. If concepts carried non-commuting generators, the commutator $[L_A, L_B]$ of two graded concepts, such as sentiment and formality, should predict how much steering one disturbs the other. It does not. In Pythia-410M, 2.8B, and 12B the correlation between the commutator norm and measured cross-talk is -0.11, -0.33, and -0.38, no better than for random operators with matched spectra. Ranking graded concepts by the effective rank of their operators gives an order that changes across Pythia sizes, so we claim no ranking. Every graded operator has a real spectrum, as expected for a concept that scales rather than turns. We do not test steering with these operators.

7 Related Work

Semantic relations as vector offsets predate transformers (Mikolov et al., 2013). Within language models the linear representation hypothesis was formalized by Park et al. (2024), and directions have

been reported for truth (Marks & Tegmark, 2024), relations (Hernandez et al., 2024), and world state (Nanda et al., 2023b), with steering methods built on them (Turner et al., 2024; Li et al., 2023; Zou et al., 2025). Park et al. (2025b) extend the picture to categorical and hierarchical concepts, which form simplices and orthogonal subspaces rather than single directions.

Circles are one case of a broader pattern: models encode several concepts along curved paths rather than straight lines. Engels et al. (2025) find circular day and month features and intervene on them; numbers are represented on helices and with base-10 Fourier features (Kantamneni & Tegmark, 2025; Zhou et al., 2024); counting positions lie on curved manifolds (Gurnee et al., 2026); and representations reorganize into the graph structure given in context (Park et al., 2025a). Several accounts derive such geometry from data statistics, in particular from translation symmetry in co-occurrence (Karkada et al., 2026; Modell et al., 2025; Prieto et al., 2026). Our contribution is methodological and complementary: a test that decides whether a measured curve is closed, validated against ordered controls that existing circular tests do not reject. The closest standard method is representational similarity analysis (Kriegeskorte et al., 2008), which correlates a model's distance matrix with that of a hypothesized shape, such as a circle. Our test asks a question that needs no shape: are the distances unchanged when the cycle shifts by one step? It therefore accepts any rigid cycle, which is why we pair it with the fundamental share. Measures of learned equivariance (Gruver et al., 2023) need the transformation to be known; here, fitting that transformation is the step that fails (Section 3). Du et al. (2026) prove that cyclic symmetry in next-token targets makes the optimal last-layer weights and context embeddings circulant, and they measure the distance to circulance for weekdays in four models. Their theory gives a reason to expect circulance. Our results show what the measure needs in order to be a test: random points are also nearly circulant, so a low value is informative only next to same-length ordered controls and together with the fundamental share.

On computation, small networks trained on modular arithmetic use Fourier features (Nanda et al., 2023a; Zhong et al., 2023). In a pretrained model, Feucht et al. (2026) show that Llama-3.1-8B solves cyclic arithmetic by base-10 addition and maps the sum back to the cycle in late layers. Our behavioral results replicate their accuracy and add that the wrap-around penalty varies across models with the same closed representation.

Equivariance is built into many vision and graph architectures (Cohen & Welling, 2016; Cohen et al., 2018; Kondor & Trivedi, 2018; Weiler & Cesa, 2021; Bronstein et al., 2021), and learned symmetry generators can be read out of trained networks (Moskalev et al., 2023). We treat equivariance as a hypothesis about a pretrained model and test the part of it that a fit cannot supply: that the group acts rigidly. Sparse autoencoders ask which directions exist (Elhage et al., 2022; Bricken et al., 2023; Cunningham et al., 2023; Templeton et al., 2026); a closed cycle is a set of such directions tied together by a symmetry, which the Gram test can check.

8 Discussion

Calling a concept circular should mean that stepping around it is a symmetry of the representation. The tests commonly used for that claim respond to any consistent order. A test derived from the symmetry separates days and months from number words, ordinals, and graded scales in every model we studied, and it downgrades the clock and the compass. The practical advice is simple: a circle claim should be checked against ordered controls of the same length, with a measure that a line and a horseshoe fail.

The limits are real. Each set has ten templates, so the templates, not the families, set the uncertainty. The result holds at the concept token and weakens at the end of the sentence. In the two Llama models, days separate from the seven-step temperature scale only in later layers. We establish a rigid cyclic action on K states, not a continuous rotation. The behavioral results are correlational and use two prompt templates. We also leave open why models close the cycle for months but not for clock hours, which name the same period.

The linear representation hypothesis fits naturally into this picture. A direction is the local, first-order description of a step. For a concept that closes on itself, the local directions turn and average to zero, so the cycle is exactly where a single direction cannot describe the concept, and it is also where a direct test of closure is needed.

AI USE STATEMENT

In this work, we used generative AI tools to help implement and debug parts of the experiment, analysis, and plotting code, to check reported numbers against the raw result files, and to edit parts of the manuscript for clarity and consistency. We did not use generative AI tools to generate the research idea, the evaluation protocol, or the main experiments, which the authors conceived and implemented. We have reviewed all AI-assisted work: the authors read and revised every AI-suggested edit, inspected all AI-assisted code, and confirmed that every reported number can be recomputed from the logged results. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The datasets are built by deterministic template substitution (Appendix A). All models are public and accessed through TransformerLens (Nanda & Bloom, 2022). The supplementary material contains the code, all datasets, and the saved outputs of every experiment. From those outputs, one script rebuilds every figure, table, and number in the paper, so no number is entered by hand; its README lists the commands that regenerate the outputs from the models.

REFERENCES

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar Van Der Wal. Pythia: a suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. https://transformer-circuits.pub/2023/monosemantic-features/index.html.

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

Tony F. Chan. An optimal circulant preconditioner for Toeplitz systems. *SIAM Journal on Scientific and Statistical Computing*, 9(4):766–771, 1988. doi: 10.1137/0909051.

Taco Cohen and Max Welling. Group equivariant convolutional networks. In Maria Florina Balcan and Kilian Q. Weinberger (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 2990–2999, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/cohenc16.html.

Taco S. Cohen, Mario Geiger, Jonas Koehler, and Max Welling. Spherical cnns, 2018. URL https://arxiv.org/abs/1801.10130.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models, 2023. URL https://arxiv.org/abs/2309.08600.

Zhehang Du, Hangfeng He, and Weijie Su. Uncovering symmetry transfer in large language models via layer-peeled optimization, 2026. URL https://arxiv.org/abs/2605.12756.

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, 2022. URL https://arxiv.org/abs/2209.10652.

Joshua Engels, Eric J. Michaud, Isaac Liao, Wes Gurnee, and Max Tegmark. Not all language model features are one-dimensionally linear, 2025. URL https://arxiv.org/abs/2405.14860.

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, 2026. URL https://arxiv.org/abs/2605.01148.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Robert M. Gray. Toeplitz and circulant matrices: a review. *Commun. Inf. Theory*, 2(3):155–239, August 2005. doi: 10.1561/0100000006. URL https://doi.org/10.1561/0100000006.

Nate Gruver, Marc Anton Finzi, Micah Goldblum, and Andrew Gordon Wilson. The lie derivative for measuring learned equivariance. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=JL7Va5Vy15J.

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, 2026. URL https://arxiv.org/abs/2601.04480.

Brian Hall. *Lie Groups, Lie Algebras, and Representations: An Elementary Introduction*, volume 222 of *Graduate Texts in Mathematics*. Springer, Cham, 2nd edition, 2015. ISBN 978-3-319-13466-6. doi: 10.1007/978-3-319-13467-3.

Evan Hernandez, Arnab Sen Sharma, Tal Haklay, Kevin Meng, Martin Wattenberg, Jacob Andreas, Yonatan Belinkov, and David Bau. Linearity of relation decoding in transformer language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=w7LU2s14kE.

Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 1985.

Subhash Kantamneni and Max Tegmark. Language models use trigonometry to do addition, 2025. URL https://arxiv.org/abs/2502.00873.

Dhruva Karkada, Daniel J. Korchinski, Andres Nava, Matthieu Wyart, and Yasaman Bahri. Symmetry in language statistics shapes the geometry of model representations, 2026. URL https://arxiv.org/abs/2602.15029.

Risi Kondor and Shubhendu Trivedi. On the generalization of equivariance and convolution in neural networks to the action of compact groups, 2018. URL https://arxiv.org/abs/1802.03690.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A. Bandettini. Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, Volume 2 - 2008, 2008. ISSN 1662-5137. doi: 10.3389/neuro.06.004.2008. URL https://www.frontiersin.org/journals/systems-neuroscience/articles/10.3389/neuro.06.004.2008.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=aLLuYpn83y.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets, 2024. URL https://arxiv.org/abs/2310.06824.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality, 2013. URL https://arxiv.org/abs/1310.4546.

Alexander Modell, Patrick Rubin-Delanchy, and Nick Whiteley. The origins of representation manifolds in large language models, 2025. URL https://arxiv.org/abs/2505.18235.

Artem Moskalev, Anna Sepliarskaia, Ivan Sosnovik, and Arnold Smeulders. Liegg: Studying learned lie group generators, 2023. URL https://arxiv.org/abs/2210.04345.

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

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability, 2023a. URL https://arxiv.org/abs/2301.05217.

Neel Nanda, Andrew Lee, and Martin Wattenberg. Emergent linear representations in world models of self-supervised sequence models, 2023b. URL https://arxiv.org/abs/2309.00941.

Core Francisco Park, Andrew Lee, Ekdeep Singh Lubana, Yongyi Yang, Maya Okawa, Kento Nishi, Martin Wattenberg, and Hidenori Tanaka. Iclr: In-context learning of representations, 2025a. URL https://arxiv.org/abs/2501.00070.

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

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*, 2025b. URL https://openreview.net/forum?id=bVTM2QKYuA.

Robert T. Powers and Erling Størmer. Free states of the canonical anticommutation relations. *Communications in Mathematical Physics*, 16(1):1–33, 1970. doi: 10.1007/BF01645492.

Lucas Prieto, Edward Stevinson, Melih Barsbey, Tolga Birdal, and Pedro A. M. Mediano. From data statistics to feature geometry: How correlations shape superposition. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=7akSRQS5Xh.

Jean-Pierre Serre. *Linear Representations of Finite Groups*, volume 42 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1977. ISBN 978-0-387-90190-6.

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, et al. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet, 2026. URL https://arxiv.org/abs/2605.29358.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering, 2024. URL https://arxiv.org/abs/2308.10248.

Maurice Weiler and Gabriele Cesa. General $e(2)$-equivariant steerable cnns, 2021. URL https://arxiv.org/abs/1911.08251.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Ziqian Zhong, Ziming Liu, Max Tegmark, and Jacob Andreas. The clock and the pizza: Two stories in mechanistic explanation of neural networks. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=S5wmbQc1We.

Tianyi Zhou, Deqing Fu, Vatsal Sharan, and Robin Jia. Pre-trained large language models use fourier features to compute addition. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=i4MutM2TZb.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to ai transparency, 2025. URL https://arxiv.org/abs/2310.01405.

A Datasets

Each cyclic and control set has 100 families: 10 sentence frames, each combined with 10 subject nouns. Within a family only the slot token changes, and the K sentences follow the order below. The data are built by string substitution with no model involved; Table 3 lists every frame and subject noun. Random nouns reuse the frames and subjects of days (K=7) and months (K=12), so that only the slot token differs. Number words appear in non-temporal frames, so that the number-12 set shares its words with the clock but not its context.

set	K	elements, in order
days	7	Monday, …, Sunday
months	12	January, …, December
clock	12	one o'clock, …, twelve o'clock
compass	8	north, northeast, east, southeast, south, southwest, west, northwest
seasons	4	spring, summer, autumn, winter
numbers (7, 12)	7, 12	one, …, seven / twelve
ordinals (7, 12)	7, 12	first, …, seventh / twelfth
sizes	7	microscopic, tiny, small, medium, large, huge, gigantic
temperature (7)	7	freezing, cold, cool, mild, warm, hot, scorching
temperature (12)	12	freezing, frigid, icy, cold, chilly, cool, mild, warm, balmy, hot, sweltering, scorching
random nouns (7)	7	river, table, cat, engine, cloud, pencil, harbor
random nouns (12)	12	the above plus violin, carpet, lantern, garden, bottle
clock, half hours	24	one o'clock, half past one, two o'clock, …, half past twelve
digits	28	1, …, 28 (number frames, as for number words)
dates	12×28	"10 January" and "January 10" for every month and day 1 to 28 (Appendix G)

The graded concepts of Section 6 (sentiment, arousal, tense, certainty, formality, politeness, complexity) share 477 seed sentences, sentences of 20 to 120 characters from the SST-2 training set. For each seed and concept, Llama-3.1-8B-Instruct receives a two-shot prompt that asks for rewrites at six signed levels $a \in \{-3, -2, -1, 1, 2, 3\}$ as a JSON object, changing only the target attribute and keeping the content; the seed itself is level 0. Sampling uses temperature 0.3 and top-p 0.9. A family is discarded if the output does not parse, if a rewrite repeats the seed, or if fewer than five of the six rewrites are distinct. This leaves between 431 and 477 families per concept. The level order is taken from the prompt and is not checked by a separate classifier.

B Proof of Proposition 1 and a stability bound

If $Qz_d = z_{d+1}$ with Q orthogonal, then $G_{d+1,e+1} = \langle Qz_d, Qz_e \rangle = G_{de}$ for all d, e, which is the definition of a circulant matrix. Conversely, suppose G is circulant, so the families $(z_0, \ldots, z_{K-1})$ and $(z_1, \ldots, z_{K-1}, z_0)$ have the same Gram matrix. Two finite families of vectors with the same Gram matrix are related by an isometry between their spans (Horn & Johnson, 1985, Ch. 7). That isometry sends z_d to z_{d+1}, and it extends to an orthogonal map Q of $\mathbb{R}^k$ by any orthogonal map between the orthogonal complements, which have equal dimension. For the spectrum, a circulant matrix is diagonalized by the Fourier vectors $f_j = (e^{2\pi ijd/K})_d$ (Gray, 2005), with eigenvalue $\lambda_j = \sum_s G_{0s} e^{2\pi ijs/K}$. Writing Z for the $K \times k$ matrix of states, $G = ZZ^\top$ and $\bar{f_j}^\top G f_j = \|Z^\top \bar{f}_j\|^2 = \|\sum_d z_d e^{-2\pi ijd/K}\|^2$, so $\lambda_j = \frac{1}{K}\|\sum_d z_d e^{-2\pi ijd/K}\|^2 = E_j$.

Table 3: Sentence frames and subject nouns. Each frame is combined with each subject noun to form one family, and the element slot [X] is filled with each of the K elements in order.

sets	frames ([S] = subject noun, [X] = element)	subject nouns
days; random nouns (7)	The [S] is scheduled for [X]. / She said the [S] would happen on [X]. / [X] is usually when the [S] takes place. / They moved the [S] to [X]. / Our [S] always falls on [X]. / He promised to finish the [S] by [X]. / The [S] opens next [X]. / I have a [S] every [X]. / We agreed to meet for the [S] on [X]. / The [S] was postponed until [X].	meeting, appointment, delivery, rehearsal, interview, class, checkup, call, review, shipment
months; random nouns (12)	The [S] is due in [X]. / She said the [S] would begin in [X]. / [X] is usually when the [S] starts. / They rescheduled the [S] to [X]. / Our [S] always happens in [X]. / He plans to launch the [S] in [X]. / The [S] opens in [X]. / The [S] was moved to [X]. / We expect the [S] to finish by [X]. / The annual [S] is held every [X].	report, festival, conference, harvest, audit, campaign, release, budget, election, renewal
clock; half hours	The [S] starts at [X]. / They scheduled the [S] for [X]. / Our [S] always begins at [X]. / He set the [S] for [X]. / The [S] was moved to [X]. / We hold the [S] at [X]. / [X] is when the [S] begins. / The [S] is due at [X]. / She booked the [S] for [X]. / The [S] runs until [X].	meeting, call, appointment, class, rehearsal, briefing, interview, shift, session, checkup
compass	The [S] was heading [X]. / They turned [X]. / The [S] faces [X]. / We steered the [S] [X]. / The [S] drifted [X]. / He pointed the [S] [X]. / The [S] kept moving [X]. / Our [S] is oriented [X]. / The [S] sailed [X]. / The compass on the [S] read [X].	ship, boat, convoy, hiker, drone, yacht, caravan, glider, rover, ferry
seasons	The [S] usually happens in [X]. / They scheduled the [S] for [X]. / Our [S] always takes place in [X]. / The [S] is best enjoyed in [X]. / He plans the [S] every [X]. / The [S] was moved to [X]. / We hold the [S] in [X]. / [X] is when the [S] begins. / The annual [S] returns each [X]. / She looks forward to the [S] in [X].	festival, harvest, cleanup, retreat, market, migration, planting, reunion, sale, carnival
numbers (7, 12)	The [S] wrote the number [X]. / The [S] picked number [X]. / The [S] said the answer was [X]. / The [S] counted up to [X]. / The [S] circled the number [X]. / The [S] called out [X]. / The [S] rated it [X] out of twenty. / The [S] guessed [X]. / The [S] stopped counting at [X]. / The [S] chose the number [X].	teacher, student, child, player, judge, clerk, coach, host, doctor, referee
ordinals (7, 12)	The [S] finished [X]. / The [S] came in [X]. / The [S] was ranked [X]. / The [S] took [X] place. / The [S] placed [X] overall. / The [S] ended up [X] in the race. / The [S] was listed [X]. / The [S] crossed the line [X]. / Officials said the [S] finished [X]. / The [S] qualified [X].	runner, team, horse, car, swimmer, driver, cyclist, skier, rower, jockey
sizes	The [S] was [X]. / They said the [S] is [X]. / Everyone agreed the [S] looked [X]. / The new [S] is [X]. / She described the [S] as [X]. / The [S] turned out to be [X]. / He thought the [S] seemed [X]. / Their [S] is quite [X]. / The [S] in the photo is [X]. / We measured the [S] and it was [X].	box, stone, house, fish, tree, rock, crate, boat, statue, crystal
temperature (7, 12)	The [S] was [X]. / She said the [S] felt [X]. / Everyone agreed the [S] was [X]. / The [S] is [X] today. / He complained that the [S] was [X]. / By noon the [S] had become [X]. / The [S] seemed [X] to us. / They found the [S] [X]. / The [S] stayed [X] all day. / We noticed the [S] was [X].	water, room, soup, air, coffee, bath, wind, sand, tea, pool

Proposition 1 is exact: the step is rigid when $\nu = 0$. Measured values of ν are never zero, so we also bound how close to rigid the step is when ν is small. Let S be the $K \times K$ matrix that shifts rows by one step, so that SZ has rows $z_1, \dots, z_{K-1}, z_0$, and let

$$\epsilon = \min_{Q \text{ orthogonal}} \frac{\|ZQ - SZ\|_F}{\|Z\|_F}$$

be the relative error of the best rigid one-step map.

Proposition 2 (Stability)**.** *For centered states with $k \geq K$,*

$$\epsilon^2 \;\leq\; 2\sqrt{K}\,\frac{\|G\|_F}{\operatorname{tr} G}\,\nu \;\leq\; 2\sqrt{K}\,\nu, \qquad \nu \;\leq\; \frac{K}{4}\,(2\epsilon + \epsilon^2)\,\frac{\operatorname{tr} G}{\|G\|_F} \;\leq\; \frac{K^{3/2}}{4}\,(2\epsilon + \epsilon^2).$$

In particular, $\nu \to 0$ if and only if $\epsilon \to 0$.

For the first bound, write both configurations in polar form, $Z = PU$ and $SZ = P'U'$, where $P = G^{1/2}$, $P' = (SGS^\top)^{1/2}$, and U, U' have orthonormal rows. Centering makes Z rank-deficient, so U is not unique, but such a U always exists: from the singular value decomposition $Z = W\Sigma V^\top$, take $U = WV^\top$, and likewise for U'. The argument uses nothing else. An orthogonal Q with $UQ = U'$ exists, and then $ZQ - SZ = (P - P')U'$, so $\|ZQ - SZ\|_F = \|P - P'\|_F$. The Powers–Størmer inequality (Powers & Størmer, 1970) gives $\|P - P'\|_F^2 \le \|G - SGS^\top\|_* \le \sqrt{K}\,\|G - SGS^\top\|_F$. Because $\mathrm{circ}(G)$ is unchanged by the shift, $G - SGS^\top = (G - \mathrm{circ}(G)) - S(G - \mathrm{circ}(G))S^\top$, whose norm is at most $2\nu\|G\|_F$. Dividing by $\|Z\|_F^2 = \mathrm{tr}\, G$ gives the bound. For the second, let Q attain ϵ and write $SZ = ZQ + E$. Expanding $SGS^\top = (ZQ + E)(ZQ + E)^\top$ gives $\|G - SGS^\top\|_F \le (2\epsilon + \epsilon^2)\,\mathrm{tr}\, G$. A shift by j steps changes G by at most $\min(j, K - j)$ single steps, since each step preserves the Frobenius norm, and $\mathrm{circ}(G)$ is the average of G over all shifts. Averaging $\min(j, K - j)$ over j gives at most $K/4$. The last inequalities in each chain use $\mathrm{tr}\, G \le \sqrt{K}\,\|G\|_F$.

The constants depend on K and are not tight. Across all sets and models, the first bound exceeds the measured error by a factor of 1.9 to 4.8, so the proposition guarantees that ν controls how rigid the step is, but it does not predict the error. Measured directly, the error of the best rigid step orders the sets as ν does. Within each model it is lower for the cycle than for every same-length ordered control, in all nine models for all seven comparisons. Outside Pythia-160M, it is 0.19 to 0.36 for days and 0.17 to 0.33 for months, against 0.34 to 0.51 for the temperature scales and 0.41 to 0.80 for number words. Both inequalities hold for all $12{,}600$ families we measured. The error is also low for random nouns (0.17 to 0.39), which is why, as in the main text, a small non-circulance is read together with the fundamental share.

C Spectrum and operator-fit sanity checks

Table 4 shows that the quantities a circle claim based on a fitted successor map would report are the same for cycles, orderings, and random nouns. The leading eigenvalue with a non-trivial angle sits at a Kth root of unity with modulus near one in every case. The ridge fit (regularization 10^{-2}, top $4K$ components) generalizes to held-out templates for random nouns as well as for days and months. An orthogonal fit (Procrustes) is closer to a test, but random nouns pass it too, because random points are nearly equidistant; this is the case the fundamental share rules out.

Table 4: Successor-map diagnostics at the concept token and fixed layer, range over the nine models. Held-out R^2 uses five-fold cross-validation over templates and is split into interior pairs and the wrap pair $K{-}1 \to 0$.

set	leading angle (°)	$\lvert\lambda\rvert$	ridge R^2 (interior / wrap)	orthogonal R^2 (interior / wrap)
days	51.4–51.5	1.00–1.00	0.52–0.98 / 0.08–0.97	0.34–0.95 / −0.98–0.87
months	30.0–30.0	1.00–1.00	0.66–0.97 / 0.67–0.98	0.38–0.94 / 0.41–0.93
numbers (7)	51.4–51.5	0.99–1.00	0.21–0.96 / 0.56–0.94	−0.61–0.80 / 0.30–0.66
numbers (12)	30.0–30.0	0.99–1.00	0.47–0.94 / 0.60–0.92	−0.26–0.83 / 0.25–0.58
temperature (12)	30.0–30.1	0.99–1.00	0.76–0.93 / 0.77–0.93	0.43–0.79 / 0.48–0.76
random nouns (7)	102.9–154.3	0.99–1.00	0.81–0.96 / 0.84–0.97	0.68–0.91 / 0.52–0.86
random nouns (12)	60.0–150.0	0.99–1.00	0.75–0.94 / 0.83–0.96	0.54–0.88 / 0.53–0.89

D Per-model results and layer robustness

Tables 5 and 6 give non-circulance and the wrap gap for every set and model. Table 7 gives the template intervals behind Table 2, and Table 8 the robustness checks. Figure 4 in Section 4 traces non-circulance across depth.

E Continuity

To ask whether the clock is a continuous rotation rather than twelve discrete states, we interleave half hours with hours (K=24), fit the hour circle per family, and locate each "half past h" between

Table 5: Non-circulance at the concept token and fixed layer. P = Pythia, -I = instruct.

set	P-160M	P-410M	P-1B	P-2.8B	P-6.9B	P-12B	Llama-3.1-8B	Llama-3.1-8B-I	Qwen2.5-7B-I
days	0.56	0.36	0.32	0.47	0.37	0.34	0.25	0.25	0.37
months	0.50	0.28	0.22	0.40	0.26	0.26	0.34	0.33	0.31
clock	0.75	0.67	0.59	0.58	0.45	0.49	0.41	0.41	0.42
numbers (7)	0.78	0.69	0.63	0.79	0.71	0.68	0.66	0.62	0.60
numbers (12)	0.82	0.68	0.60	0.79	0.70	0.67	0.62	0.59	0.56
ordinals (7)	0.66	0.65	0.66	0.74	0.71	0.70	0.70	0.66	0.66
ordinals (12)	0.68	0.65	0.66	0.74	0.72	0.71	0.68	0.65	0.63
sizes (7)	0.65	0.66	0.61	0.71	0.64	0.65	0.62	0.60	0.52
temperature (7)	0.59	0.45	0.39	0.51	0.43	0.42	0.50	0.45	0.45
temperature (12)	0.65	0.52	0.48	0.56	0.50	0.49	0.55	0.51	0.50
random nouns (7)	0.39	0.30	0.25	0.32	0.29	0.26	0.22	0.23	0.32
random nouns (12)	0.49	0.34	0.26	0.33	0.28	0.27	0.25	0.26	0.35
compass (8)	0.34	0.25	0.27	0.33	0.30	0.28	0.34	0.28	0.21
seasons (4)	0.38	0.20	0.14	0.22	0.19	0.14	0.15	0.15	0.18

Table 6: Wrap gap at the concept token and fixed layer.

set	P-160M	P-410M	P-1B	P-2.8B	P-6.9B	P-12B	Llama-3.1-8B	Llama-3.1-8B-I	Qwen2.5-7B-I
days	1.34	1.46	1.39	1.62	1.42	1.37	1.26	1.26	1.28
months	1.21	1.21	1.19	1.20	1.17	1.22	1.26	1.23	1.51
clock	2.27	2.72	2.27	2.26	1.97	2.17	1.99	2.14	1.93
numbers (7)	3.30	2.49	2.16	3.57	2.52	2.40	2.28	2.11	1.98
numbers (12)	3.71	2.59	2.31	3.70	2.62	2.53	2.26	2.09	2.06
ordinals (7)	2.24	2.48	2.52	3.35	2.90	2.68	2.64	2.43	2.34
ordinals (12)	2.28	2.70	2.55	3.42	3.00	2.93	2.85	2.53	2.60
sizes (7)	1.29	1.35	1.43	1.43	1.54	1.51	1.32	1.30	1.26
temperature (7)	1.22	1.10	1.04	1.20	1.10	1.14	1.15	1.09	1.05
temperature (12)	1.20	1.21	1.06	1.32	1.16	1.20	1.24	1.21	1.08
random nouns (7)	0.78	0.75	0.76	0.71	0.72	0.75	0.81	0.79	0.84
random nouns (12)	0.99	0.96	0.92	0.97	1.00	0.98	0.98	0.97	0.96
compass (8)	1.10	0.99	0.99	0.98	1.04	1.01	0.99	0.96	1.03
seasons (4)	1.08	1.16	1.11	1.15	1.16	1.09	1.09	1.06	0.99

its measured neighbors h and $h{+}1$, where 0.5 is the midpoint. At the concept token the half hours sit close to h (median relative position -0.04 to 0.20 in the larger models), which looks like snapping, but the concept token of "half past one" is the word "one" itself, so the test cannot separate snapping from token identity. At the final token the hour circle itself is poorly fit (angle errors of 12° to 25°), so the test is uninformative there. We therefore make no claim about continuity.

F Behavioral probe

The templates are "Q: What month is $\{m\}$ months after $\{X\}$?\nA:" (template 0, from Feucht et al. (2026)) and "Q: If it is $\{X\}$ now, what month will it be in $\{m\}$ months?\nA:" (template 1), with the analogous forms for days and clock hours. For the compass, template 0 is "Q: What direction do you face after turning $\{45m\}$ degrees clockwise from $\{X\}$?\nA:" and template 1 asks for m steps clockwise on an eight-point compass. Offsets run from 1 to $2K-1$; in-cycle results use $m < K$. A query wraps if the answer lies past the last element in the canonical order. The wrap penalty averages, over offsets that have both wrapped and unwrapped queries, the difference in accuracy, weighted by the number of queries at that offset; its interval resamples queries within each offset and wrap cell. The instruct models receive the same raw text, without a chat template. Table 9 gives all values.

G Nested concepts: calendar dates

We ask whether the month cycle survives when a month is combined with a second, ordered concept. Each family is one sentence frame filled with every date in a grid of 12 months and days 1 to 28, which every month has. We write the dates in two orders, "10 January" and "January 10", and build 50 families for each order (10 frames and 5 subject nouns). A causal model's state at the first word

Table 7: Non-circulance with 95% bootstrap intervals over templates, for the cycles and their hardest controls.

model	days	temperature (7)	sizes (7)	months	temperature (12)	numbers (12)
Pythia-160M	0.56 [0.54, 0.58]	0.59 [0.57, 0.61]	0.65 [0.62, 0.67]	0.50 [0.46, 0.54]	0.65 [0.62, 0.68]	0.82 [0.78, 0.86]
Pythia-410M	0.36 [0.33, 0.40]	0.45 [0.44, 0.46]	0.66 [0.62, 0.69]	0.28 [0.24, 0.34]	0.52 [0.52, 0.52]	0.68 [0.65, 0.70]
Pythia-1B	0.32 [0.30, 0.34]	0.39 [0.38, 0.41]	0.61 [0.58, 0.63]	0.22 [0.20, 0.26]	0.48 [0.47, 0.48]	0.60 [0.59, 0.62]
Pythia-2.8B	0.47 [0.43, 0.51]	0.51 [0.49, 0.53]	0.71 [0.68, 0.74]	0.40 [0.36, 0.46]	0.56 [0.55, 0.58]	0.79 [0.77, 0.82]
Pythia-6.9B	0.37 [0.34, 0.39]	0.43 [0.41, 0.45]	0.64 [0.61, 0.68]	0.26 [0.22, 0.31]	0.50 [0.49, 0.51]	0.70 [0.67, 0.72]
Pythia-12B	0.34 [0.31, 0.37]	0.42 [0.40, 0.44]	0.65 [0.62, 0.68]	0.26 [0.22, 0.31]	0.49 [0.48, 0.50]	0.67 [0.64, 0.69]
Llama-3.1-8B	0.25 [0.22, 0.27]	0.50 [0.47, 0.53]	0.62 [0.59, 0.66]	0.34 [0.30, 0.38]	0.55 [0.53, 0.56]	0.62 [0.59, 0.65]
Llama-3.1-8B-Inst	0.25 [0.23, 0.28]	0.45 [0.42, 0.48]	0.60 [0.58, 0.63]	0.33 [0.30, 0.37]	0.51 [0.50, 0.53]	0.59 [0.57, 0.61]
Qwen2.5-7B-Inst	0.37 [0.29, 0.49]	0.45 [0.43, 0.47]	0.52 [0.51, 0.54]	0.31 [0.24, 0.43]	0.50 [0.49, 0.51]	0.56 [0.54, 0.59]

Table 8: Robustness of the paired comparison: number of models, out of nine, in which the cycle is more circulant than the control, for the original ν and the identity-invariant ν', with $2K$, $4K$, or $8K$ principal components or none. Reference values for random points depend on the dimension; the "all" column uses 512 dimensions.

		ν				ν' (identity-invariant)			
cycle	control	$2K$	$4K$	$8K$	all	$2K$	$4K$	$8K$	all
days	numbers (7)	9/9	9/9	9/9	9/9	9/9	9/9	9/9	9/9
days	ordinals (7)	9/9	9/9	9/9	9/9	9/9	9/9	9/9	9/9
days	sizes (7)	9/9	9/9	9/9	9/9	9/9	9/9	9/9	9/9
days	temperature (7)	9/9	9/9	9/9	9/9	9/9	9/9	9/9	9/9
months	numbers (12)	9/9	9/9	9/9	9/9	9/9	9/9	9/9	9/9
months	ordinals (12)	9/9	9/9	9/9	9/9	9/9	9/9	9/9	9/9
months	temperature (12)	9/9	9/9	9/9	9/9	9/9	9/9	9/9	9/9
reference: circle (K=12)		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
reference: line (K=12)		0.88	0.88	0.88	0.88	0.92	0.92	0.92	0.92
reference: random (K=12)		0.55	0.42	0.32	0.14	0.89	0.83	0.74	0.44

of a date has not seen the second word, so we read each state at the second word, where it has seen both, and at the final token. We split each family's grid into a month part (the mean over days), a day part (the mean over months), and the remainder, and compute the measures of Section 2.3 on each part, with no principal-component projection. As a control for the day part we use the digits 1 to 28 in the number frames of Appendix A.

Table 10 gives the results at the month token of "10 January" and the fixed layer. The month part is as circulant as months alone: non-circulance 0.21 to 0.33 in eight models and 0.63 in Pythia-160M, with a median difference from months alone of 0.03. The month cycle also holds within each single day (0.23 to 0.64). The day part is not a cycle. Its non-circulance (0.74 to 0.85) lies in the range of the ordered controls, and it is less circulant than the digits alone in all nine models (intervals separate in eight). Its wrap gap (2.72 to 5.08) is larger than for the digits alone in all nine models. The month part carries 72% to 89% of the variance, the day part 10% to 21%, and the remainder at most 9%, so at this token the model stores month and day nearly separately.

The month cycle does not travel well to later tokens. At the day token of "January 10", the month part carries only 7% to 22% of the variance and is less circulant than months alone in all nine models (0.40 to 0.71). At the final token of either order the remainder grows to as much as 52%. Both results agree with the main finding that the cycle is held at the concept token.

H Formal tests of the decision rule

The decision rule has two conditions, and we test both in every model. For the first, the alternative hypothesis is that the cycle's normalized fundamental share is higher than that of the random nouns of the same length. For the second, it is that the cycle's non-circulance is lower than that of each same-length ordered control. Families that share a sentence frame are not independent, so each set is

Table 9: Behavioral results. Accuracy is in-cycle ($m < K$); non-wrap and wrap accuracy use template 0; the wrap penalty is offset-stratified with 95% bootstrap intervals.

model	concept	acc. t0 / t1	non-wrap / wrap (t0)	wrap penalty t0	wrap penalty t1
Llama-3.1-8B	days	0.93 / 0.62	0.95 / 0.90	−0.02 [−0.14, +0.12]	+0.64 [+0.48, +0.80]
Llama-3.1-8B	months	0.90 / 0.58	1.00 / 0.80	+0.15 [+0.08, +0.23]	+0.68 [+0.56, +0.79]
Llama-3.1-8B	clock	0.55 / 0.14	0.86 / 0.24	+0.30 [+0.17, +0.43]	−0.16 [−0.27, −0.06]
Llama-3.1-8B	compass (8)	0.21 / 0.18	0.21 / 0.21	−0.10 [−0.20, +0.01]	+0.01 [−0.05, +0.08]
Llama-3.1-8B-Inst	days	0.76 / 0.50	1.00 / 0.52	+0.34 [+0.20, +0.49]	+0.61 [+0.42, +0.78]
Llama-3.1-8B-Inst	months	0.75 / 0.48	1.00 / 0.50	+0.34 [+0.28, +0.40]	+0.49 [+0.35, +0.64]
Llama-3.1-8B-Inst	clock	0.21 / 0.11	0.30 / 0.12	−0.05 [−0.16, +0.06]	−0.24 [−0.36, −0.14]
Llama-3.1-8B-Inst	compass (8)	0.14 / 0.16	0.14 / 0.14	+0.00 [+0.00, +0.00]	+0.05 [−0.11, +0.19]
Qwen2.5-7B-Inst	days	0.48 / 0.45	0.57 / 0.38	−0.12 [−0.32, +0.07]	+0.25 [+0.06, +0.39]
Qwen2.5-7B-Inst	months	0.81 / 0.30	0.89 / 0.73	−0.05 [−0.16, +0.05]	−0.08 [−0.20, +0.03]
Qwen2.5-7B-Inst	clock	0.40 / 0.37	0.65 / 0.15	+0.19 [+0.07, +0.30]	+0.32 [+0.21, +0.42]
Qwen2.5-7B-Inst	compass (8)	0.32 / 0.20	0.39 / 0.25	−0.01 [−0.21, +0.21]	−0.14 [−0.30, +0.02]
Pythia-12B	days	0.19 / 0.02	0.33 / 0.05	+0.31 [+0.22, +0.40]	+0.17 [+0.17, +0.17]
Pythia-12B	months	0.11 / 0.06	0.18 / 0.03	−0.02 [−0.08, +0.03]	−0.05 [−0.08, −0.02]
Pythia-12B	clock	0.08 / 0.05	0.03 / 0.12	−0.11 [−0.20, −0.03]	−0.03 [−0.08, +0.02]
Pythia-12B	compass (8)	0.12 / 0.02	0.07 / 0.18	−0.12 [−0.26, +0.04]	−0.04 [−0.11, +0.00]

Table 10: Calendar dates, read at the month token of "10 January" (fixed layer). Month and day are the month and day parts of each family's 12×28 grid, compared with months alone and with the digits 1 to 28 alone. The variance share is the fraction of the grid's variance carried by the month part, the day part, and the remainder (interaction).

	month: ν		day: ν		day: wrap gap		variance share	
model	in date	alone	in date	digits alone	in date	digits alone	month / day	interaction
Pythia-160M	0.63	0.49	0.85	0.77	2.72	2.66	0.75 / 0.21	0.04
Pythia-410M	0.27	0.28	0.79	0.65	3.48	2.27	0.89 / 0.10	0.01
Pythia-1B	0.23	0.22	0.75	0.59	3.05	2.00	0.82 / 0.16	0.03
Pythia-2.8B	0.33	0.38	0.74	0.72	3.06	2.65	0.76 / 0.17	0.07
Pythia-6.9B	0.22	0.25	0.76	0.65	3.44	2.17	0.87 / 0.11	0.02
Pythia-12B	0.24	0.25	0.75	0.63	3.08	2.06	0.79 / 0.15	0.06
Llama-3.1-8B	0.32	0.33	0.78	0.66	4.03	2.27	0.85 / 0.11	0.04
Llama-3.1-8B-Inst	0.28	0.32	0.78	0.63	3.85	2.14	0.81 / 0.11	0.08
Qwen2.5-7B-Inst	0.21	0.30	0.79	0.61	5.08	1.88	0.72 / 0.19	0.09

summarized by its ten frame means. Each test compares two sets of ten frame means with an exact permutation test of the difference in means, over all 184,756 ways to split the twenty values into two groups of ten; the smallest possible p-value is 5×10^{-6}. The claim that days and months pass rests on 81 tests (two conditions; four controls for days and three for months; nine models), which we correct jointly with the Holm method.

Table 11 gives the corrected p-values. Of the 81 tests, 78 reject at 0.05, each with a corrected p-value of at most 0.022. The three that do not are days against temperature7 in Pythia-160M, Pythia-2.8B, and Qwen2.5-7B-Instruct (uncorrected $p = 0.035$, 0.040, 0.083; corrected 0.105). Months pass every test in every model. For clock, tested separately, the first condition holds in all nine models, but the second holds against temperature12 in only three.

Table 11: Holm-corrected one-sided p-values of the exact permutation tests (concept token, fixed layer). Fundamental share: the set against random nouns of the same length. Non-circulance: the set against each same-length ordered control. Days and months form one family of 81 tests; clock is corrected separately. An asterisk marks a test that does not reject at 0.05. P = Pythia, -I = instruct.

set	test	against	P-160M	P-410M	P-1B	P-2.8B	P-6.9B	P-12B	Llama	Llama-I	Qwen-I
days	fundamental share	random nouns (7)	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
days	non-circulance	numbers (7)	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	0.006
days	non-circulance	ordinals (7)	0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
days	non-circulance	sizes (7)	0.002	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	0.022
days	non-circulance	temperature (7)	0.105*	0.002	<0.001	0.105*	0.009	0.002	<0.001	<0.001	0.105*
months	fundamental share	random nouns (12)	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
months	non-circulance	numbers (12)	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	0.005
months	non-circulance	ordinals (12)	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
months	non-circulance	temperature (12)	<0.001	<0.001	<0.001	0.001	<0.001	<0.001	<0.001	<0.001	0.011
clock	fundamental share	random nouns (12)	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
clock	non-circulance	numbers (12)	0.145*	1.000*	1.000*	<0.001	<0.001	<0.001	<0.001	<0.001	0.035
clock	non-circulance	ordinals (12)	1.000*	1.000*	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	0.003
clock	non-circulance	temperature (12)	1.000*	1.000*	1.000*	1.000*	<0.001	1.000*	<0.001	0.002	0.394*