THE DEVIATION CARRIES THE CODE: CAUSAL SURGERY ON GEOMETRIC REPRESENTATIONS IN LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Causal analyses locate language-model arithmetic in geometric structures such as the month circle and the number helix. Each describes the representation through that idealized structure. We call the remainder orthogonal to that structure the deviation. Yet no work has manipulated its content. Concurrent theory predicts that such components exist, and we test whether models use them. We edit token embeddings, rewriting only the tested components. A matched-norm control substitutes random content, so behavior changes trace to content. We study four open-weight models and a grokked network on month, weekday, hour, letter, and two-digit addition arithmetic. The structure alone never matches the unedited model. In month, hour, and letter arithmetic, the deviation carries most or all of what models use. Weekday arithmetic varies by model, and addition needs both parts. We pit a circle rotated one step against the deviation at half its norm. In Llama-3.1-8B, the deviation wins on 0.95 of prompts and the circle on 0.01. Patched into three models' residual streams, the deviation wins the same month and weekday conflicts. This holds at every tested layer where the operand position still controls the answer. Strength-based readouts cannot certify such components and are uncalibrated. They certify structure in 93.5% of structureless inputs matched to the GPT-2 embedding spectrum. Our test certifies structure in 4.0% of those inputs.

1 INTRODUCTION

A central goal of interpretability research is to identify the representations through which language models compute. Interventions built on such identifications, steering, editing, ablation, are only as reliable as the identification itself. For arithmetic over structured concepts, a line of causal analyses has produced identifications with striking geometric content. The twelve months lie on a circle in the internal representations of language models, and interventions on that circle alter calendar arithmetic (Engels et al., 2025). Numbers form a generalized helix in the residual stream, and on addition, patching in a model's number representations as reconstructed from the helix recovers most of the effect of patching in the representations themselves (Kantamneni & Tegmark, 2025, see also Feucht et al., 2026; Gurnee et al., 2026). Kantamneni & Tegmark (2025, §4.5) themselves probe the limit of the idealized description. In GPT-J the helix remains causally relevant on five further numerical tasks but falls below a principal-component baseline on three of them (integer division, multiplication by 1.5, and solving $x - a = 0$ for x). They read this as a sign of potential additional structure that helical fits do not capture, and they remain confident that the helix is used for addition itself. This paper studies the part of each representation that an idealized structure leaves out, and asks whether the computation uses it.

These analyses share a structural property. Each addresses the representation through an idealized structure: the circle for cyclic concepts and the helix for numbers, to which we will add the line for ordinal concepts. Given the subspace the structure spans, each concept's representation decomposes uniquely into two orthogonal parts: the projection, the best approximation the structure can express, and the orthogonal remainder, everything it cannot. We name the remainder the *deviation*. Prior experiments differ in what they do to the deviation, and none manipulates its content. Steering the magnitude of an identified structure leaves it in place (Feucht et al., 2026). Patching the structure's reconstruction into the residual stream deletes it (Kantamneni & Tegmark, 2025). Setting the circle

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.

to a target value and mean-ablating the rest of the state replaces it with its average across prompts (Engels et al., 2025). In every case the deviation enters the experiment as background to be held, removed, or neutralized. We present the first direct causal study of it, and Appendix F shows that the structures we hold fixed coincide with the structures these prior analyses intervene on.

The deviation is a predicted object, not an incidental one. Karkada et al. (2026) prove that, for concepts whose co-occurrence statistics are translation symmetric, as holds for the months of the year, learned embeddings contain sinusoidal components at every frequency, a component at frequency k tracing k cycles across the ordered concepts, with amplitudes that decay as frequency rises. The circle on which prior work intervenes is the component at frequency one, and the deviation is the progressively fainter components above it. The theory settles that these components are present, but presence is not use: whether the computation uses them, in the sense that editing them changes the model's answers, requires intervention, and settling that is the experimental contribution of this paper.

Our method (Fig. 1) is a protocol of tightly controlled interventions on the operand tokens' embeddings, trained parameters that, unlike input-dependent hidden states, form a fixed matrix whose edits can be constructed and checked exactly in double precision. Each condition rewrites only the components under test, the components each constructed matrix is meant to protect are checked to a relative error below 10^{-9} before it is written back in the model's half-precision format (§2.2), and the central control substitutes a random deviation whose strength (squared norm) equals the true deviation's at every frequency, so that any behavioral difference between them is attributable to content alone (§2.3).

Across month, weekday, hour, letter, and two-digit addition arithmetic in four open-weight models and a network grokked on modular addition, the idealized structure alone never reproduces the unmodified model's accuracy on the same prompts, which is our baseline throughout. No single component is individually necessary. In month, hour, and letter arithmetic the deviation carries most to all of the information the computation uses, and in weekday arithmetic the split varies by model (Appendix A). Two-digit addition in Llama-3.1-8B needs both parts, with the structure alone at 0.48 and the deviation alone at 0.37 against a baseline of 1.00, and a quarter of the deviation restoring 0.83. The decisive condition places the components in direct conflict: with the circle rotated by one step, so that a row's structure asserts the next month, and the deviation retained at half its norm, so that it still asserts the original month, Llama-3.1-8B answers with the deviation's month on 0.95 of prompts and the circle's on 0.01, an override that a strength-matched random deviation in the same condition does not produce. The grokked network, whose algorithm is known exactly (Nanda et al., 2023; Morwani et al., 2024), anchors these verdicts against ground truth. The verdicts are not a fact about the input encoding alone. When the same constructions are patched into the residual stream of Qwen-2.5-7B, Mistral-7B-v0.3, and Llama-3.1-8B, the rotation conflict resolves for the deviation in month and weekday arithmetic in all three models, at every tested layer at which the operand position still controls the answer. At every such layer the deviation alone also stays ahead of the structure alone in five of these six pairs and in Qwen-2.5-7B letter arithmetic, with Llama-3.1-8B weekdays the exception. In two-digit addition the structure alone is ahead at every such layer, as it is at the embeddings, in line with the residual-stream helix of prior work (§3.3).

The deviation's content is arbitrary where training data underdetermine it (grokked seeds select frequency sets indistinguishable from random draws) and convergent where they decide it (independently trained models agree at $\bar{r} = 0.65$–0.88, correlating with co-occurrence statistics measured from text). The same amplitude decay suggests why the deviation went unmeasured. Its components sit below what strength-based readouts can certify, and such readouts are themselves uncalibrated. On null data generated at the GPT-2 embedding spectrum they report structure in 93.5% of trials, and a spectrum-conditioned significance test we construct in Appendix D reduces that rate to 4.0%.

Contributions. **(i) The first direct causal study of the deviation.** The idealized structure is never sufficient and no single component is individually necessary. In month, hour, and letter arithmetic the deviation carries most to all of the information the computation uses at the embeddings. Through the residual stream the rotation conflict resolves for the deviation in all six cyclic model–domain pairs patched at depth, at every tested layer at which the operand position still controls the answer (§3, §3.3). **(ii) Consistency with the mid-network structures of prior work:** in two-digit addition the helix alone stays ahead of the deviation alone at the embeddings and at every such layer of Llama-

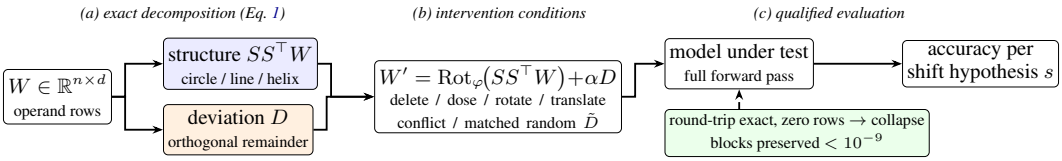

Figure 1: The intervention pipeline. (a) Operand representations, token embeddings or layer-ℓ operand-position states, decompose exactly into the idealized structure and the deviation. (b) Conditions delete, dose, rotate, or translate the components, place them in conflict, or substitute a strength-matched random deviation $\tilde{D}$ whose norm at every frequency equals the true deviation's. (c) The model under test, unmodified apart from the constructed rows, evaluates fixed prompts once the checks of §2.3 pass, and rotation and translation conditions are scored against every shifted answer hypothesis.

3.1-8B, in line with the residual-stream helix that Kantamneni & Tegmark (2025) found causally implicated in addition (§3.3). **(iii) A determination pattern for representational content:** arbitrary where training data underdetermine the solution, convergent and correlated with text statistics where they decide it (§3.4). **(iv) A calibrated significance test for claimed structure,** controlling the rate at which structure is reported in data that has none, with a closed-form detection floor (Appendix D).

Related work. Networks trained on finite-group tasks supply our ground-truth anchor. A network that groks modular addition (Power et al., 2022) computes it by composing multiple Fourier frequencies (Nanda et al., 2023; Zhong et al., 2023; Morwani et al., 2024), and the same reverse-engineering programme extends to composition in larger groups (Stander et al., 2024; Wu et al., 2025). Networks learn circuits from one fully characterized family, but which circuits a given network learns is arbitrary (Chughtai et al., 2023), and our seed experiments extend that arbitrariness to the content of the deviation. Pre-trained token embeddings carry Fourier features of integers that models use to add (Zhou et al., 2024). Language models of many architectures converge on the same periodic spectrum for number tokens, yet only some learn representations from which a number's residue class is linearly decodable (Fu et al., 2026). Their distinction separates the presence of structure from its decodability, and ours separates presence from causal use. Our significance test contributes to work on the statistical validity of interpretability claims (Sharkey et al., 2025; Bolukbasi et al., 2021; Méloux et al., 2025) by controlling the rate at which symmetric structure is reported in data that has none, calibrated where ground truth permits exact validation.

2 Method

Our protocol has three stages, summarized in Fig. 1. An exact decomposition separates the operand embeddings into the idealized structure and the deviation (§2.1). Experimental conditions construct new embeddings that alter only the components under test, and the model under test, unmodified apart from those embeddings, answers a fixed set of arithmetic questions (§2.2). Two validity checks qualify every language-model condition and one control condition licenses the causal claim about content (§2.3). The same constructions extend to the operand's hidden states through the residual stream (§2.4). Tasks, qualification checks, and scoring are specified in §2.5, with templates and further detail in Appendix G. The calibrated significance test we apply to every structure-existence claim is constructed in Appendix D.

2.1 Decomposing embeddings into structure and deviation. The decomposition assembles objects that prior work has already specified. Expressing the representations of an ordered concept set in the real Fourier basis over the concept index is the standard analysis of the modular-arithmetic literature (Nanda et al., 2023; Chughtai et al., 2023), applied directly to the number embeddings of language models by Zhou et al. (2024). The idealized structures are taken from the causal analyses that intervene on them. What this section contributes is the control regime built on these objects: an exact decomposition of the operand rows into named blocks, conditions that rewrite only the blocks under test, and verification in every run that every other block is preserved.

Let $\mathcal{T} = \{t_0, \ldots, t_{n-1}\}$ be an ordered set of n concept tokens, such as the twelve months, the seven weekdays, the twenty-six letters, or the integers zero through ninety-nine. Let $W \in \mathbb{R}^{n \times d}$ collect their input embedding rows in that order, where d is the model's embedding dimension. Row i of W is the embedding of concept t_i. Column j of W lists the values that embedding coordinate j takes across the n concepts in their order, so each column is a function on the ordered index set $\{0, \ldots, n-1\}$, and the decomposition below is a change of basis for such functions.

The basis is the real Fourier basis over the concept index: the constant vector, and at each frequency k from 1 to $\lfloor n/2 \rfloor$ the cosine–sine pair assigning concept i the values $\cos(2\pi ki/n)$ and $\sin(2\pi ki/n)$, tracing k full cycles across the ordered concepts. For even n the sine at $k = n/2$ vanishes at every index, so the top pair collapses into a single alternating vector. These n vectors are orthonormal (each has unit length, any two are orthogonal) and complete (any vector of length n is a unique weighted sum of them), so each column of W splits exactly into frequency components. We call the two-dimensional subspace spanned by the cosine–sine pair at frequency k the *plane* at frequency k, and a named group of basis vectors, such as the constant, one plane, or the ramp introduced below, a *block*. As in §1, the *strength* of a component is its squared norm.

The *idealized structure* is the subspace that prior work identifies and intervenes on, specified per domain. For cyclic domains (months, weekdays, hours) the structure is the constant together with the plane at k=1, which is the circle (Engels et al., 2025), and it spans $s = 3$ basis directions (the grokked network's structure is given in Appendix I). For the ordinal domain (letters) the structure is the constant together with the centered linear *ramp*, the vector whose entries increase in equal steps with the mean subtracted. This is the line, with $s = 2$. For the numerical domain the structure is the generalized helix of Kantamneni & Tegmark (2025): the constant, the ramp, and the Fourier blocks at periods $T \in \{2, 5, 10, 100\}$, with $s = 9$. A block at period T sits at frequency n/T, and its components complete one cycle every T concepts, so integers T apart receive identical values on it, which is why these periods express the base-ten regularities of two-digit numbers. At period 2 the sine vanishes at every integer, so that block is the single alternating vector.

One technical step is needed before each structure block can be edited independently of the others. The ramp is not a Fourier basis vector. It overlaps several frequencies at once, so a structure containing both the ramp and Fourier blocks is a list of blocks that are not mutually orthogonal, and deleting one block would partially delete another. We therefore orthonormalize the structure blocks by blockwise Gram–Schmidt in the order constant, ramp, then the blocks at periods 100, 10, 5, and 2, which leaves the total subspace they span unchanged while making the blocks mutually orthogonal, so each block can afterwards be deleted, retained, or transformed without touching the others (construction in Appendix F). The total structure subspace, and therefore the deviation, does not depend on this order. The content assigned to an individual block does, so single-block results are reported relative to it.

Collecting the resulting orthonormal basis vectors as the columns of $S \in \mathbb{R}^{n \times s}$, the matrix $SS^\top$ is the orthogonal projection onto the structure's subspace, and the decomposition is

$$W = SS^\top W + (I - SS^\top)\, W, \qquad D = (I - SS^\top)\, W, \tag{1}$$

where I is the $n \times n$ identity and the first term is the structure. The *deviation* D is every component of the learned embeddings that the idealized structure does not span. The two terms are orthogonal by construction, so the split is exact and unique, and D is the projection of W onto the orthogonal complement of the structure. For cyclic domains the Fourier blocks are mutually orthogonal, so the deviation is exactly the Fourier components outside the structure's plane. For the language-model domains these are the frequencies $k \geq 2$, the highest of which is a single alternating vector when n is even. For the ordinal and numerical domains the structure contains the ramp, so the deviation is the projection onto the orthogonal complement of the Gram–Schmidt-orthonormalized structure rather than a set of whole planes. All decompositions and checks run in double precision on copies of the model's half-precision rows. Appendix F relates these structures to the objects prior causal analyses intervene on.

2.2 Intervention conditions. Throughout, the *model under test* is one of the four open-weight language models of §3 or the network grokked on modular addition specified in Appendix I. An experimental condition replaces the n operand rows of that model's embedding matrix with the corresponding rows of a constructed matrix W', runs the model end to end on a fixed set of arithmetic questions for the domain (§2.5), and records accuracy. We write only the n operand rows of the input

embedding. The remaining vocabulary, all attention and MLP parameters, and the output embedding are unchanged, and models that tie input and output embeddings are untied by cloning before any edit, so that an edit to an input row cannot leak into the readout. Because W is a matrix of trained parameters rather than an input-dependent activation, every component that a condition is not meant to alter can be preserved exactly in the constructed rows. In both language-model campaigns we check this before the rows are written and the condition is scored, requiring the relative change of each checked component, computed in double precision, to be below 10^{-9}. Appendix G gives the components each campaign checks and the rounding that write-back to half precision adds.

The conditions fall into four families. *Deletion and dose* conditions keep the structure exactly and scale the deviation by a fraction $\alpha \in \{0, 0.25, 0.5, 0.75, 1\}$, so that $W' = SS^\top W + \alpha D$: the endpoint $\alpha = 0$ gives the structure alone, and $\alpha = 1$ returns the unmodified embeddings. *Component* conditions delete or retain a single named block, such as one plane of the helix or the ramp, leaving all other blocks intact. *Rotation and translation* conditions apply the transformation the structure's own geometry defines: a rotation of the circle by angle φ within its plane, or a shift of the ramp by m index steps, so that the structure asserts a displaced concept while everything else is unchanged. *Conflict* conditions combine a rotated or shifted structure with a retained fraction of the deviation, so that the two components assert different answers and the model's output reveals which it follows. Rotation and translation conditions are scored against every shifted answer hypothesis (Appendix G).

2.3 Validity checks and the matched-random control. Two checks qualify every language-model condition before it is scored, and one control condition licenses the causal claim about content, each ruling out a specific failure.

The *round-trip check* writes the original rows back after every edit and requires the restored weights to be bitwise identical, and before any condition it requires a write of the unedited rows to leave the model's predictions unchanged, which together rule out residue from the editing machinery itself. The *zero check* replaces all operand rows with zeros and requires accuracy to fall to at most the larger of twice chance and half the unedited accuracy, and the run is aborted otherwise. Surviving accuracy would mean the edited rows are not the rows the computation uses, a failure that token-boundary ambiguity can silently cause. In every accepted run of the hour, letter, and addition domains, zero-row accuracy in fact fell to chance level or below, up to sampling error. Zeros serve here rather than random rows because this check tests addressing, not content. Random content enters the design separately, with its strength matched, as the control condition below.

The *strength-matched random deviation* is the control that licenses causal language about content. The deviation decomposes into one component per frequency, and each component has a strength, its squared norm, and a content, the particular directions it occupies in the d-dimensional embedding space. For each frequency, we draw a random component and rescale it so that its norm equals the true deviation's norm at that frequency exactly, then substitute the assembled random deviation for the true one in any condition. By construction, no measurement of strength, at any frequency, can distinguish the random deviation from the true one, so any difference in the model's behavior between them is attributable to content alone. Five independent draws are run for each dose condition in the language models and three in the grokked network, and reported with their spread. Rotation conflicts and every activation condition use a single draw, and translation conditions have no random counterpart.

2.4 Extending the interventions through the residual stream. The same decomposition extends into the network under one design constraint: all activation experiments use operand-first prompts, in which the operand token precedes the offset and any variable content, optionally behind a constant few-shot prefix. Under causal attention, the hidden state at the operand position then depends only on the constant prefix and the operand token, so for each layer ℓ the operand-position states again form a fixed matrix $W^{(\ell)} \in \mathbb{R}^{n \times d}$, and every construction of §2.2 applies unchanged. We collect $W^{(\ell)}$ for the residual stream after each block and for each block's MLP output, construct conditions from the layer's own states, and write the constructed state back at the operand position through a forward hook while the rest of the model under test runs normally. The depth campaign covers month and weekday arithmetic in Qwen-2.5-7B, Mistral-7B-v0.3, and Llama-3.1-8B, letter arithmetic in Qwen-2.5-7B, and two-digit addition in Llama-3.1-8B. Templates, tested layers, and the condition list are in Appendix H.

Three adaptations are required, because a state written through a hook, unlike a parameter row, cannot be restored bitwise. First, the round-trip check becomes a check on the patching apparatus. Patching a state through the hook and patching nothing at all should give identical answers, but the two paths differ in floating-point detail, which can flip the argmax on prompts where two answer tokens are nearly tied. We therefore require the do-nothing patch to reproduce the unpatched accuracy on the same prompt set to within 0.04, a tolerance set in the campaign code. Across all accepted runs, do-nothing accuracy varied by at most 0.024 across layers for any model and domain. Second, the zero check becomes a per-layer diagnostic rather than a gate. Zeroing the operand position collapses accuracy only while the operand's information still resides at that position, so a failure to collapse at depth is informative rather than disqualifying. Collapse uses the criterion of the embedding zero check, zero-patch accuracy at most the larger of twice chance and half the unpatched accuracy. The set of depths where collapse occurs, which we call the *causal window*, is recorded per model and domain, and causal claims are interpreted within it. Beyond the window, attention has relocated the operand's information and position-level patching undertests. Third, for the numerical domain only the first operand is patched, because the second operand's state depends on the prefix containing the first. An embedding edit, by contrast, changes every occurrence of an operand token, so the two campaigns apply different interventions to two-digit addition, and this restriction is disclosed wherever those results are reported. The strength-matched random control carries over without modification, constructed from each layer's own deviation.

2.5 Tasks, qualification, and scoring. Each domain defines offset-varying prompts, which ask for the concept displaced from the operand by a stated offset, with the numerical domain adopting the Llama-3.1-8B addition prompt of Kantamneni & Tegmark (2025). Llama-3.1-8B is the only model to enter the numerical domain. A model enters a domain only when two checks pass: the *tokenization check*, that every operand appears as a single token and every answer begins with a distinct single token, verified programmatically (failures are reported as tokenizer limitations, not null results), and the *capability check*, that the best of the domain's prompt templates, three to five in the embedding campaign and one or two in the activation campaign, reaches accuracy 0.40. The threshold is a uniform floor for admission, set in the campaign code, and every comparison in the paper is within-model (Appendix G). Mistral-7B-v0.3 on letters is the one model–domain pair that did not enter surgery (Appendix G).

Answers are scored by restricting the final-position logits to the answer token set and taking the argmax, and each condition is reported with a binomial standard error computed from its prompt count. Templates, the embedding and activation campaigns' independent template sweeps, both campaigns' unmodified baselines, and archival details are in Appendix G.

3 Results

We report the embedding-level results first: deletion and restoration (§3.1), conflict (§3.2), and the helix dissection (Fig. 2c). Results through the residual stream follow (§3.3), then the ground-truth anchor, the determination of the deviation's content in summary (§3.4, with full results in Appendix A). Every value is accuracy on the fixed prompt set of the named domain and model, under the scoring rule of §2.5. The *baseline* is the unedited model's accuracy on that same prompt set. Every language-model condition passed the checks of §2.3, or through the residual stream those of §2.4.

3.1 The deviation carries the information the computation uses. Figure 2 shows the deletion results. Deleting the month circle leaves accuracy unchanged in Qwen-2.5-7B ($0.750 \rightarrow 0.750$) and Mistral-7B-v0.3 ($0.788 \rightarrow 0.788$), and nearly so in Llama-3.1-8B ($0.970 \rightarrow 0.917$) and Gemma-2-9b ($0.750 \rightarrow 0.720$), where in each pair the first value is the unedited model's accuracy and the second is its accuracy with the circle removed. The structure alone, the constant together with the circle, falls far below that unedited accuracy in every model, and the strength-matched random deviation scores with the structure rather than with the unedited model (Fig. 2a). Hour arithmetic, a third cyclic domain, shows the same pattern in all four models. The deviation alone scores 0.49 to 0.72 against baselines of 0.69 to 0.95, while the structure alone scores 0.09 to 0.33 (Appendix A).

The same dissociation holds in a domain with no period at all (Fig. 2b). In letter arithmetic the structure is the constant together with the line, and it carries little of the task: the structure alone scores 0.073 in Qwen-2.5-7B, 0.059 in Llama-3.1-8B, and 0.078 in Gemma-2-9b, where chance is

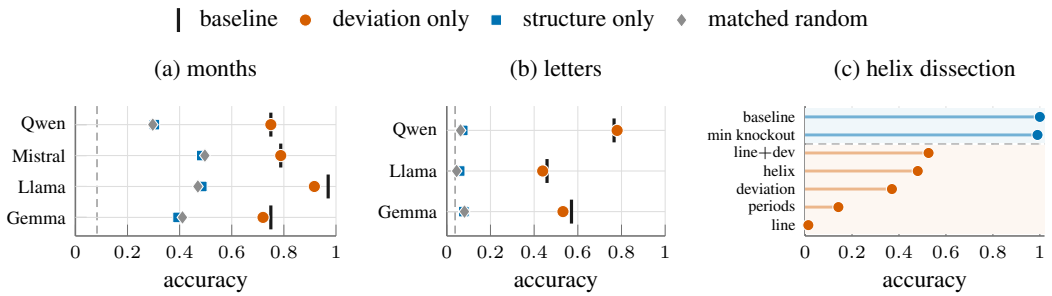

Figure 2: Component deletion at the embeddings. (a) Months and (b) letters. Black ticks mark the unedited baseline, the model's accuracy on the same prompts with no embedding edited, and dashed lines mark chance ($1/12$ for months, $1/26$ for letters). The deviation alone (circles) matches or nearly matches baseline in every model and both domains, while the structure alone (squares) and the strength-matched random deviation at full strength (diamonds, mean of five draws) score together, far below baseline for the month circle and near chance for the letter line. (c) Helix dissection (two-digit addition, Llama-3.1-8B), ordered by effect. Deleting any single helix block costs at most 0.010 (top band), and every isolated part or pair falls far below baseline (bottom band). *Helix* keeps the constant, line, and periods without the deviation. *Periods* and *line* each keep the constant and that part alone. *Line+dev* keeps the constant, line, and deviation without the periods. *Deviation* keeps the constant and the deviation.

0.038, while the deviation alone matches or nearly matches baseline in the same three models (0.780 against 0.766, 0.439 against 0.459, and 0.532 against 0.571). Mistral-7B-v0.3 did not enter surgery on letters and is not reported (Appendix G).

Two-digit addition in Llama-3.1-8B, the only model to enter that domain, needs both parts (Fig. 2c). The helix alone scores 0.480 and the deviation alone 0.370 against a baseline of 1.000, a quarter of the deviation restores 0.826, and deleting any single helix block leaves accuracy at 0.990 or above. Among partial structures, the line with the deviation scores 0.526, the periods alone 0.142, and the line alone 0.014.

Restoration depends on content rather than strength. Restoring the true deviation in increasing fractions α raises accuracy toward baseline in every hour, letter, and addition condition, while the strength-matched random deviation, averaged over five draws, stays within 0.05 of the structure-alone level at every α (Appendix A). In Qwen-2.5-7B months the random deviation scores 0.297 ± 0.036 at $\alpha{=}1$, the mean and sample standard deviation over five draws.

Weekday arithmetic does not reproduce the pattern uniformly. At the first residual site, after the first block, deviation sufficiency is full in Qwen-2.5-7B (0.714 against 0.714), partial in Mistral-7B-v0.3 (0.738 against 0.833), and weakest in Llama-3.1-8B, where the structure alone reaches 0.810 against a baseline of 0.905 and the deviation alone 0.595 (Appendix A.5). The embedding campaign shows the same variation, with the deviation alone ahead of the circle alone in Qwen-2.5-7B and Mistral-7B-v0.3 and behind it in Llama-3.1-8B and Gemma-2-9b (Appendix A.6).

3.2 Conflict separates content from strength. The conflict experiment rotates the circle by one step, so that the circle and the retained deviation assert different months, and scores each answer against both assertions: following the deviation means answering with the unrotated month, and following the circle means answering with the rotated one. The retained fraction α takes the values 0, 0.25, 0.5, 0.75, and 1, and answers cross from the circle to the deviation as it rises (Fig. 3a). We report the crossing point rather than the endpoint because the endpoint is the uninformative case: at $\alpha{=}1$ both components are present at full weight, whereas the override is already complete at $\alpha{=}0.5$, where the deviation's amplitude is halved and its strength, its squared norm, is therefore a quarter of its trained value, yet it overrides an intact, fully weighted rotated circle. With a strength-matched random deviation substituted, in a single draw at each of $\alpha{=}0.5$ and $\alpha{=}1$, the override does not occur.

The circle is nevertheless used as genuine geometry when it is all the model has: rotating the bare structure through its full orbit, accuracy follows the rotation-predicted month at the structure-

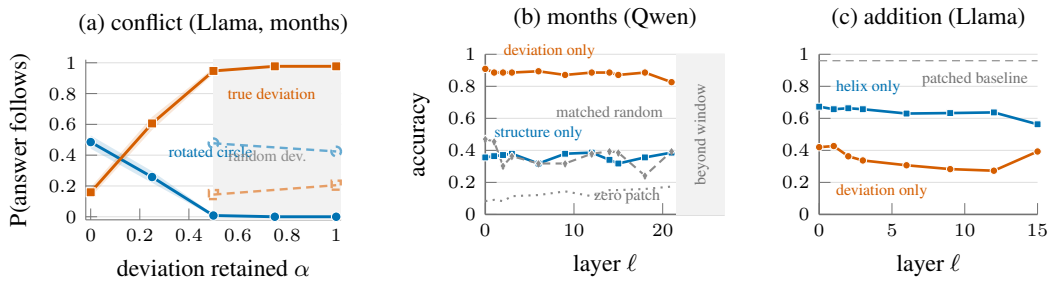

Figure 3: Conflict and depth. (a) Conflict (Llama-3.1-8B, months). As the retained fraction α rises, answers cross from the rotated circle to the true deviation, and the override is already complete at α=0.5 (shaded), where the deviation retains a quarter of its trained strength against a fully weighted circle. The strength-matched random deviation (dashed, one draw at each of α=0.5 and α=1) does not override. Bands show ± 1 binomial standard error. (b, c) The deviation through depth (residual stream, where ℓ=0 is the state after the first block). (b) Qwen-2.5-7B months. The deviation alone tracks baseline and the structure alone stays low at every layer of the causal window, whose edge the zero patch identifies, and the matched random stays with the structure rather than the deviation throughout. (c) Two-digit addition against the patched baseline (dashed), the accuracy obtained when the unmodified state is written back through the hook, with only the first operand patched. The helix alone stays ahead of the deviation alone at every layer of the window, and neither approaches the patched baseline.

only level at every whole-step angle while accuracy at the unrotated labels decays away from 0° (Appendix B), and those whole-step angles double as a validation that rotation is an exact relabeling.

3.3 The deviation through depth. Figure 3b,c shows the verdicts through the residual stream. The residual-stream experiments cover Qwen-2.5-7B (months, letters, weekdays), Mistral-7B-v0.3 (months, weekdays), and Llama-3.1-8B (months, weekdays, two-digit addition). We read each condition at every tested layer of the causal window, the depths at which zeroing the operand position still collapses accuracy (§2.4). With the full deviation retained, the one-step rotation conflict resolves for the deviation at every such layer in all six month and weekday pairs. The deviation alone also stays ahead of the structure alone at every such layer in five of those pairs and in Qwen-2.5-7B letters, with Llama-3.1-8B weekdays the exception. In Qwen-2.5-7B months the deviation alone sustains 0.83 to 0.91 against 0.32 to 0.39 for the structure alone at all eleven tested layers of the window, and the letter line stays at most 0.141 against a baseline of 0.58 to 0.60 (chance 0.038). Paired with the structure, a strength-matched random deviation never reaches the do-nothing baseline at any layer of the window in any of the eight pairs.

In two-digit addition the ordering reverses (Fig. 3c). With only the first operand patched, the helix alone stays ahead of the deviation alone at every layer of the window, 0.56 to 0.67 against 0.27 to 0.43, as it is at the embeddings, in line with the residual-stream helix that Kantamneni & Tegmark (2025) found causally implicated in addition. Neither part alone approaches the patched baseline of 0.96, and both end the window lower than they begin it. Appendix H sets out how these depth claims should be read: they concern the partition between the structure and its complement at each site, not a single content carried through the network.

3.4 Ground truth and determination. Two further results are summarized here and reported in full in Appendix A. The grokked network anchors the verdicts against ground truth: across five seeds the identical interventions yield 0.072 to 0.094 for the dominant frequency alone, at a chance level of 0.017, and 0.994 to 0.998 for the remainder. The deviation's content follows a determination pattern (Fig. 4, Appendix A): grokked seeds trained on identical, underdetermining data select frequency sets indistinguishable from random draws, while the four language models converge ($\bar{r} = 0.651$ on months, 0.876 on weekdays, all twelve pairs $p \leq 0.008$), and the convergent content correlates with co-occurrence statistics measured from text (pooled $r = 0.405$ on weekdays, 0.295 on months, Appendix E).

4 DISCUSSION

The four results fit one account, which we offer as a proposal rather than a finding. Karkada et al. (2026) prove that translation-symmetric co-occurrence statistics force learned embeddings to contain Fourier components at every frequency, with amplitudes that decay as frequency increases. Our determination experiments are consistent with the data deciding the content of those components, which is arbitrary where the training data underdetermine the solution and convergent, and correlated with measured text statistics, where they decide it. The same amplitude decay places the deviation's components below the detection floor of strength-based readouts (Appendix D), which would explain what the literature found: strength-based summaries could discover the circle and the helix, the strongest components, but could not have shown that in month, hour, and letter arithmetic most of the information the computation uses sits in components too faint for such readouts to distinguish from chance. Detecting a component and establishing what it does are different questions, and at the deviation's strengths only intervention on content can answer the second. Prior interventions established that the idealized structures are used. They left the content of the deviation untested, and testing it is the step this paper adds.

The results carry a caution and a corrective for intervention-based interpretability. Steering, editing, or ablating an idealized structure acts on a component that, on its own, falls short of the unedited model's accuracy in every domain we test, often by a wide margin, so effect sizes calibrated on the idealized geometry may misjudge what an intervention does to the computation. The corrective is procedural, and this paper supplies both parts: a calibrated test of whether a claimed structure is present at all, applied before any intervention on it is interpreted, and causal claims licensed by a control matched to the component's strength at every frequency.

In sum, the deviation is predicted by theory for translation-symmetric concepts, and it carries most to all of the information the computation uses in month, hour, and letter arithmetic at the embeddings and wins the rotation conflict through the residual stream in every month and weekday pair we patch. Its content converges across models where the training data can decide it, and on our proposed account it escaped measurement because amplitude decay places it below the detection floor of strength-based readouts. The representations through which language models compute are richer than their visible geometry, and this paper supplies instruments for studying the remainder.

5 LIMITATIONS

Our embedding-level interventions are checked in double precision before the edited rows are written back, and the write-back itself rounds each entry to the model's half-precision format (§2.2). The check is narrower in the month and weekday campaign, which covers the circle but not the deviation, and absent for the grokked network. The activation-level extension is looser still, and its claims are bounded accordingly. The do-nothing control loosens to a 0.04 tolerance (observed spread at most 0.024 across layers), the zero control becomes a diagnostic rather than a gate, all depth claims are confined to the causal window because position-level patching undertests beyond it, the numerical domain patches only the first operand, and each condition uses a single matched-random draw without a per-condition preservation check. The depth campaign covers three of the four models, excluding Gemma-2-9b, and runs letter arithmetic only in Qwen-2.5-7B. How attention and MLP components use the structure and the deviation at depth remains open.

The numerical domain runs on a single model, because only Llama-3.1-8B passes the single-token check among our four. Hour arithmetic uses number-word operands, whose embeddings also serve every other use of those words, and we do not adjudicate between this polysemy account and other causes of the partial hour results. The letter templates were revised after a first sweep, and Mistral-7B-v0.3 did not enter surgery on letters, with its accuracy on the revised templates unlogged (Appendix G). The exploratory content analyses postdate the causal results and are labeled as such, and the de-stationarization they use removes cyclic index distance even in the ordinal and numerical domains. The calibration of the existence test was performed at the GPT-2 embedding spectrum, and the reported false-certification rates are established at that spectrum. All findings concern arithmetic over structured token sets, and extension to other computations is a conjecture, not a claim.

AI USE STATEMENT

In this work, we did not use generative AI tools for any task on the required-disclosure list. Specifically, we did not use them to generate synthetic data sets, develop theoretical models or conceptual frameworks, formulate mathematical claims, supply critical ingredients for or assist in writing proofs, propose or refine hypotheses, design our research methodology or experiments, implement methods, clean or reformat datasets, support qualitative or thematic data analysis, interpret results, or assist with translation.

Additionally, we used generative AI tools for several tasks where disclosure is recommended. The primary use was editing the manuscript to improve readability and polishing prose the authors had already written. We also used these tools to draft parts of the manuscript, to suggest a structure for sections, to create and edit software code used for analysis, to format references, to search for and summarise existing literature, and to identify relevant literature. Separately, we asked a generative AI tool for reviewer-style feedback on a complete draft, in the manner of a colleague reading the paper before submission. That feedback concerned the clarity and presentation of work the authors had already completed.

We have reviewed all AI-assisted work. The research question, the intervention protocol, the control design, all experimental work, and all claims in this paper are the authors' own. Every numerical value reported here was reconciled against the experiment output files by the authors. Passages drafted or edited with AI assistance were reviewed and revised by the authors before inclusion, and the authors adopted them only where they independently judged the content correct and supported by the data. Feedback received was evaluated and acted on at the authors' discretion.

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 protocol is specified in full in Sections 2.1 to 2.5, with the correspondence to the structures of prior work in Appendix F, every prompt template and qualification check in Appendix G, the depth constructions in Appendix H, the grokked network in Appendix I, and the existence test in Appendix D. The supplementary material contains two notebooks and the result files they write, covering the embedding campaign for month and weekday arithmetic, the grokked anchor, and hour, letter and two-digit addition arithmetic. Both notebooks enforce their checks at run time rather than reporting them afterwards: operand token ids are detected per model and per template from the tokenized prompt, input and output embeddings are untied by cloning before any edit, the round-trip check requires the restored weights to be bitwise identical, the zero check aborts the run if accuracy fails to collapse, and every component a condition is not meant to alter is asserted to a relative change below 10^{-9} before the rows are written back. Seeds are derived deterministically from one master seed per condition and recorded with every row, and a run manifest records the master seed, the library versions and the device for each campaign.

Every accuracy in Tables 1 to 6, 9 and 10, and every accuracy quoted in Sections 3.1 and 3.2 and Appendix B, traces to a row of the archived result files through the key columns `model`, `task` or `domain`, `phrasing_idx`, `condition`, `block`, `alpha`, `phi_deg`, `mode_k`, `seed` and `shift`, and per-prompt predictions are archived for every condition, so any reported accuracy can be recomputed prompt by prompt. Tables 7, 11, 13 and 14 are archived as separate files with their permutation nulls, as are the grokked network's training curve, trained weights and per-seed statistics for all five seeds, and the per-model existence tests of Appendix D. Four limits apply. The models are public checkpoints, two of which are gated and require access to be granted before they will load, and no private or newly collected data is used anywhere in this work. Accuracies are argmax decisions over a restricted answer-token set, so a prompt whose top two answer logits are close can resolve differently across software environments, and re-running the pipeline on a later library release reproduced most reported values exactly while moving one or two such prompts. The letter templates were revised after a first sweep and the resume key records a template index rather than the template text, so three models' `phrasing_sweep` rows describe the first template set while their surgery rows describe the revised one, with the template used for surgery recorded in each control log. Finally, the activation campaign's per-layer results, the calibration of the existence

test on structureless inputs, the language-model ellipticities of Appendix E.3 and the fitted-probe alignment of Appendix F are not included in this archive.

REFERENCES

Tolga Bolukbasi, Adam Pearce, Ann Yuan, Andy Coenen, Emily Reif, Fernanda Viégas, and Martin Wattenberg. An interpretability illusion for BERT, 2021. URL https://arxiv.org/abs/2104.07143.

Bilal Chughtai, Lawrence Chan, and Neel Nanda. A toy model of universality: Reverse engineering how networks learn group operations, 2023. URL https://arxiv.org/abs/2302.03025.

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.

Deqing Fu, Tianyi Zhou, Mikhail Belkin, Vatsal Sharan, and Robin Jia. Convergent evolution: How different language models learn similar number representations, 2026. URL https://arxiv.org/abs/2604.20817.

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.

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.

Maxime Méloux, Giada Dirupo, François Portet, and Maxime Peyrard. The dead salmons of AI interpretability, 2025. URL https://arxiv.org/abs/2512.18792.

Depen Morwani, Benjamin L. Edelman, Costin-Andrei Oncescu, Rosie Zhao, and Sham Kakade. Feature emergence via margin maximization: case studies in algebraic tasks, 2024. URL https://arxiv.org/abs/2311.07568.

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

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets, 2022. URL https://arxiv.org/abs/2201.02177.

Lee Sharkey, Bilal Chughtai, Joshua Batson, Jack Lindsey, Jeff Wu, Lucius Bushnaq, Nicholas Goldowsky-Dill, Stefan Heimersheim, Alejandro Ortega, Joseph Bloom, Stella Biderman, Adrià Garriga-Alonso, Arthur Conmy, Neel Nanda, Jessica Rumbelow, Martin Wattenberg, Nandi Schoots, Joseph Miller, Eric J. Michaud, Stephen Casper, Max Tegmark, William Saunders, David Bau, Eric Todd, Atticus Geiger, Mor Geva, Jesse Hoogland, Daniel Murfet, and Tom McGrath. Open problems in mechanistic interpretability, 2025. URL https://arxiv.org/abs/2501.16496.

Dashiell Stander, Qinan Yu, Honglu Fan, and Stella Biderman. Grokking group multiplication with cosets, 2024. URL https://arxiv.org/abs/2312.06581.

Wilson Wu, Louis Jaburi, Jacob Drori, and Jason Gross. Towards a unified and verified understanding of group-operation networks, 2025. URL https://arxiv.org/abs/2410.07476.

Ziqian Zhong, Ziming Liu, Max Tegmark, and Jacob Andreas. The clock and the pizza: Two stories in mechanistic explanation of neural networks, 2023. URL https://arxiv.org/abs/2306.17844.

Tianyi Zhou, Deqing Fu, Vatsal Sharan, and Robin Jia. Pre-trained large language models use Fourier features to compute addition, 2024. URL https://arxiv.org/abs/2406.03445.

A Full results beyond months

This appendix reports the complete numerical results for the domains summarized in the main text. All values are accuracies from the embedding campaign unless a subsection states otherwise. Zero controls set every operand row to zero. Matched random replaces the true deviation with a strength-matched random deviation and is averaged over five draws. Zero checks fall to chance or below, up to sampling error, in every hour, letter, and addition run, and in Qwen-2.5-7B months to 0.083 against a chance level of 0.083.

A.1 Alphabet arithmetic

Table 1 gives the letters results behind Fig. 2b. In every model that entered surgery, the deviation alone matches or nearly matches baseline while the ordinal line alone falls to about the level of the matched random. Mistral-7B-v0.3 did not enter surgery on letters (Appendix G).

Model	Baseline	Deviation only	Line only	Matched random	Zero
Qwen-2.5-7B	0.766	0.780	0.073	0.063	0.010
Llama-3.1-8B	0.459	0.439	0.059	0.045	0.000
Gemma-2-9b	0.571	0.532	0.078	0.081	0.039

Table 1: Alphabet arithmetic under component deletion (n=205; chance $1/26 \approx 0.038$). The deviation alone matches or nearly matches baseline, and the line alone scores close to the matched random.

A.2 Hour arithmetic

Hours reproduce the cyclic pattern in partial form. Table 2 gives the deletion results: the deviation alone recovers most but not all of baseline in every model. Zero controls sit at chance. Restoration rises with α for the true deviation while the matched random stays near the circle-only level (Table 3), the conflict resolves for the deviation by α=0.5 in every model (Table 4), and the per-mode analysis shows the redundancy of the ensemble: no mode is individually necessary, and no single mode added to the circle recovers the full gap (Table 5).

Model	Baseline	Deviation only	Circle only	Zero
Qwen-2.5-7B	0.848	0.644	0.091	0.083
Mistral-7B-v0.3	0.689	0.553	0.273	0.083
Llama-3.1-8B	0.705	0.492	0.220	0.083
Gemma-2-9b	0.955	0.720	0.333	0.083

Table 2: Hour arithmetic under component deletion (n=132, chance $1/12 \approx 0.083$). Deviation sufficiency is partial in every model.

A.3 Two-digit addition: component dissection

Two-digit addition departs from the pattern of the other domains in the other direction. The helix alone outscores the deviation alone, 0.48 against 0.37, both are needed to reach the baseline of 1.00,

Model	Restoration	α=0	0.25	0.5	0.75	1
Qwen-2.5-7B	true deviation	0.091	0.568	0.803	0.833	0.848
	matched random	0.091	0.083	0.085	0.121	0.141
Mistral-7B-v0.3	true deviation	0.273	0.689	0.705	0.689	0.689
	matched random	0.273	0.277	0.276	0.288	0.297
Llama-3.1-8B	true deviation	0.220	0.455	0.591	0.652	0.705
	matched random	0.220	0.194	0.214	0.215	0.227
Gemma-2-9b	true deviation	0.333	0.773	0.947	0.947	0.955
	matched random	0.333	0.341	0.350	0.336	0.342

Table 3: Hours: restoration dose-response. True-deviation restoration rises toward baseline in every model, and the matched random, averaged over five draws, stays within 0.05 of the circle-only level.

Model	Answers follow	α=0	0.25	0.5	0.75	1
Qwen-2.5-7B	deviation	0.144	0.447	0.795	0.826	0.848
	circle	0.212	0.121	0.000	0.030	0.030
Mistral-7B-v0.3	deviation	0.144	0.485	0.712	0.712	0.697
	circle	0.227	0.098	0.000	0.000	0.000
Llama-3.1-8B	deviation	0.106	0.311	0.576	0.652	0.720
	circle	0.167	0.167	0.076	0.030	0.015
Gemma-2-9b	deviation	0.068	0.303	0.848	0.871	0.871
	circle	0.250	0.220	0.015	0.008	0.008

Table 4: Hours: conflict at one whole-step rotation of the circle. In every model the circle leads at α=0 and the deviation has overridden it by α=0.5. With the matched random substituted, in a single draw at each of α=0.5 and α=1, the override does not occur (deviation 0.061–0.152, circle 0.167–0.273).

and no single helix block is individually necessary. Table 6 gives the full dissection behind Fig. 2c. Deleting any single helix block costs at most 0.010, and the costliest single deletion is the line. Every isolated component or pair falls far below baseline. Adding a single periodic block to the deviation recovers much of the gap for the short periods and little for T=100. Replacing the true deviation with the strength-matched random adds 0.019 over the helix alone. Rotating a periodic plane or translating the line with the deviation retained leaves answers anchored at the original value, with accuracy at least 0.996 at shift 0 and at most 0.003 at every other shift, for every tested plane and step size.

A.4 Ground truth and determination

The grokked network (Appendix I) provides the anchor where ground truth is known: the identical interventions yield accuracy 0.072 to 0.094 for the dominant frequency alone and 0.994 to 0.998 for the remainder, across five training seeds that each reach a baseline of 1.000 at a chance level of 0.017. In a network whose algorithm is known, the deviation is the rest of its frequency ensemble. The language models show the same per-frequency redundancy in hour arithmetic, where knocking out any single frequency leaves most of baseline accuracy in all four models (Table 5).

The five grokked seeds, trained on identical data that underdetermine the solution, select frequency sets statistically indistinguishable from random draws: their mean Jaccard overlap, the size of the intersection of two seeds' frequency sets divided by the size of their union, is 0.142 against a random expectation of 0.171, and their mean pairwise content correlation is -0.022. The four language models, trained by different organizations on different corpora, converge instead (Fig. 4, Table 7): every one of the twelve pairs is individually significant at $p \leq 0.008$ against a permutation null. The weekday corpus match is led by workweek collocations. Monday–Friday and Tuesday–Thursday co-occur above the rate their calendar distance predicts and Saturday–Sunday below it, in the corpus and in all four models, a contrast internal to the weekday set that the de-stationarization of Appendix H cannot create. The month match is dominated by deficits in the pairs adjacent to *May*, whose lexical ambiguity as a modal verb and a name distorts its statistics in the same way for every

Model		k=2	k=3	k=4	k=5	k=6
Qwen-2.5-7B	knock out mode k	0.826	0.833	0.826	0.841	0.826
	circle + mode k	0.348	0.485	0.402	0.432	0.121
Mistral-7B-v0.3	knock out mode k	0.682	0.705	0.689	0.689	0.689
	circle + mode k	0.424	0.545	0.492	0.667	0.568
Llama-3.1-8B	knock out mode k	0.614	0.712	0.689	0.682	0.667
	circle + mode k	0.348	0.288	0.205	0.303	0.333
Gemma-2-9b	knock out mode k	0.947	0.947	0.947	0.955	0.939
	circle + mode k	0.508	0.553	0.629	0.705	0.758

Table 5: Hours, per-mode analysis. Baselines are 0.848, 0.689, 0.705, and 0.955 respectively. No single mode is individually necessary, and no single mode added to the circle recovers the full gap.

Condition	Accuracy
Baseline (all components)	1.000
Delete line	0.990
Delete period T=2	0.996
Delete period T=5	1.000
Delete period T=10	1.000
Delete period T=100	0.996
Line + deviation	0.526
Helix (line + periods)	0.480
Deviation only	0.370
Periods only	0.142
Line only	0.014
Deviation + period T=5	0.838
Deviation + period T=10	0.822
Deviation + period T=2	0.718
Deviation + period T=100	0.472
Helix + matched random	0.499
Zero	0.002

Table 6: Two-digit addition (Llama-3.1-8B, n=500, chance 0.01). Structure and deviation are jointly necessary and no single helix block is individually necessary: the costliest single deletion (the line) costs 0.010, while the best isolated pair reaches 0.526. Every row except the baseline and the deletions also keeps the constant.

model. With May excluded, the remaining month correlation is positive but not significant, which we report as a power limit (Table 14).

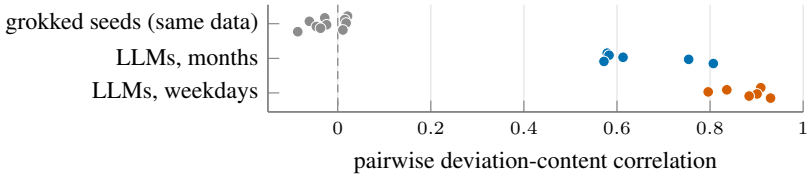

Figure 4: The determination pattern: every pairwise deviation-content correlation. Grokked seeds trained on identical data scatter around zero (dashed line), while language models trained by different organizations converge, more strongly where the token set is smaller. Points are jittered vertically within rows.

A.5 Weekday arithmetic at the first residual site

The weekday results below come from the activation campaign at its first residual site, the state after the first block, rather than from the embedding campaign used elsewhere in this appendix. Deviation

Task	Model pair	r	Null mean	Null p_{95}	p
Months	Qwen–Mistral	0.5785	−0.0046	0.2633	0.002
Months	Qwen–Llama	0.5827	−0.0066	0.2591	0.002
Months	Qwen–Gemma	0.6126	0.0100	0.2764	0.002
Months	Mistral–Llama	0.7542	0.0103	0.2852	0.002
Months	Mistral–Gemma	0.5720	−0.0071	0.2512	0.002
Months	Llama–Gemma	0.8065	−0.0011	0.2828	0.002
Weekdays	Qwen–Mistral	0.9087	0.0224	0.5994	0.004
Weekdays	Qwen–Llama	0.8357	−0.0020	0.6165	0.004
Weekdays	Qwen–Gemma	0.7955	−0.0065	0.5814	0.008
Weekdays	Mistral–Llama	0.9011	−0.0130	0.5634	0.004
Weekdays	Mistral–Gemma	0.8838	0.0161	0.6341	0.002
Weekdays	Llama–Gemma	0.9297	−0.0116	0.5982	0.002

Table 7: Every language-model pairwise deviation-content correlation, with the permutation-null mean, its 95th percentile, and the exact p-value.

sufficiency is full in Qwen-2.5-7B, partial in Mistral-7B-v0.3, and weakest in Llama-3.1-8B, where the circle alone recovers most of baseline.

Model	Baseline	Deviation only	Circle only	Matched random	Zero
Qwen-2.5-7B	0.714	0.714	0.571	0.524	0.119
Mistral-7B-v0.3	0.833	0.738	0.571	0.595	0.143
Llama-3.1-8B	0.905	0.595	0.810	0.810	0.143

Table 8: Weekday arithmetic under component deletion at the activation campaign's first residual site. Chance is $1/7 \approx 0.143$. Matched random is the circle with one strength-matched random deviation.

A.6 Weekday arithmetic at the embeddings

The embedding campaign also ran weekday arithmetic in all four models, and those results are reported here alongside the activation-campaign results of Appendix A.5. The two campaigns used different templates, so no absolute accuracy is compared across them.

Table 9 gives the deletion results. The split varies by model. The deviation alone leads the circle alone in Qwen-2.5-7B and Mistral-7B-v0.3, and the circle alone leads in Llama-3.1-8B and Gemma-2-9b. Mistral-7B-v0.3 is the only model in which the deviation alone reaches baseline. The matched random sits near the circle-only level in three models and above it in Llama-3.1-8B.

Model	Baseline	Deviation only	Circle only	Matched random	Zero
Qwen-2.5-7B	0.667	0.548	0.524	0.400	0.143
Mistral-7B-v0.3	0.714	0.714	0.571	0.562	0.143
Llama-3.1-8B	0.976	0.595	0.714	0.771	0.143
Gemma-2-9b	0.690	0.524	0.548	0.543	0.143

Table 9: Weekday arithmetic under component deletion at the embeddings (n=42, chance $1/7 \approx 0.143$). Matched random is the circle with a strength-matched random deviation at full strength, averaged over five draws.

The rotation conflict was run at four steps for retention fractions 0 and 1 only, and at one and two steps for the full ladder. Table 10 gives the one-step results. At full retention the unrotated day leads the rotated day in all four models, and the true deviation holds the unrotated day at 0.667 to 0.952 against 0.167 to 0.405 for a single strength-matched random draw at the same retention.

At zero retention, where the rotated circle is the only structure present, the rotated day leads the unrotated day in all four models, at 0.524 to 0.714 against 0.167 to 0.429. The weekday circle

is therefore read as geometry when it is all the model has, as the month circle is under the same construction (Appendix B). As the deviation is restored the answers cross to the unrotated day, at a retention fraction that varies by model: by α=0.25 in Qwen-2.5-7B and Mistral-7B-v0.3, by α=0.5 in Llama-3.1-8B, and by α=0.75 in Gemma-2-9b. With a strength-matched random deviation in place of the true one, the rotated day still leads at both retention fractions tested, by 0.167 to 0.643 at α=0.5 and by 0.048 to 0.500 at α=1.

Model	Answers follow	α=0	0.25	0.5	0.75	1
Qwen-2.5-7B	deviation	0.238	0.452	0.619	0.619	0.667
	circle	0.524	0.262	0.095	0.119	0.095
Mistral-7B-v0.3	deviation	0.429	0.619	0.881	0.929	0.929
	circle	0.548	0.357	0.071	0.000	0.000
Llama-3.1-8B	deviation	0.167	0.333	0.524	0.857	0.952
	circle	0.714	0.595	0.405	0.119	0.048
Gemma-2-9b	deviation	0.190	0.214	0.238	0.548	0.690
	circle	0.548	0.500	0.476	0.238	0.143

Table 10: Weekdays at the embeddings: conflict at a one-step rotation of the circle (n=42, chance $1/7 \approx 0.143$). The circle leads at zero retention in every model, and the deviation crosses it as the retained fraction rises: by α=0.25 in Qwen-2.5-7B and Mistral-7B-v0.3, by α=0.5 in Llama-3.1-8B, and by α=0.75 in Gemma-2-9b. With a strength-matched random deviation substituted, in a single draw, the rotated day keeps the lead at both retention fractions tested, reaching 0.476 to 0.762 against 0.119 to 0.405 for the unrotated day at α=0.5, and 0.405 to 0.690 against 0.167 to 0.405 at α=1.

The verdict does not survive a larger displacement of the circle. Table 11 gives the full ladder at a two-step rotation. At full retention the rotated day leads the unrotated day in Mistral-7B-v0.3 (0.405 against 0.143), Llama-3.1-8B (0.476 against 0.071) and Gemma-2-9b (0.286 against 0.024). Qwen-2.5-7B is the only model in which the answers cross to the deviation, by α=0.75, reaching 0.571 against 0.071 at full retention. The separation from the matched random narrows with it. Substituting a strength-matched random deviation at the same rotation, in a single draw at each of α=0.5 and α=1, leaves the unrotated day at 0.000 to 0.119, so the true deviation separates from the random one by 0.452 in Qwen-2.5-7B at α=1 and by at most 0.119 in the other three models at either fraction. At a four-step rotation, run at retention fractions 0 and 1 only, the rotated day leads in all four models at both fractions, at 0.524 to 0.738 against 0.000 to 0.024 at zero retention and at 0.452 to 0.595 against 0.000 to 0.048 at full retention.

Model	Answers follow	α=0	0.25	0.5	0.75	1
Qwen-2.5-7B	deviation	0.119	0.119	0.167	0.405	0.571
	circle	0.524	0.500	0.429	0.143	0.071
Mistral-7B-v0.3	deviation	0.024	0.024	0.024	0.048	0.143
	circle	0.548	0.619	0.548	0.548	0.405
Llama-3.1-8B	deviation	0.000	0.000	0.000	0.024	0.071
	circle	0.738	0.786	0.786	0.548	0.476
Gemma-2-9b	deviation	0.024	0.000	0.000	0.024	0.024
	circle	0.548	0.548	0.524	0.429	0.286

Table 11: Weekdays at the embeddings: conflict at a two-step rotation of the circle (n=42, chance $1/7 \approx 0.143$). The circle leads at zero retention in every model, as at one step, and the deviation crosses it only in Qwen-2.5-7B, by α=0.75. Compare Table 10, where the deviation crosses in all four models at a one-step rotation.

A.7 Where the pattern weakens

One account of where the pattern weakens is the number of components the deviation has to work with. Weekdays fit it in part, since seven concepts leave the deviation only two planes. Deviation sufficiency at the first residual site runs from full in Qwen-2.5-7B, through partial in Mistral-7B-v0.3, to lowest in Llama-3.1-8B (§3.1), where the structure alone also stays ahead of the deviation

alone at every layer of the window. The weekday circle is the hardest structure in the paper to establish in the first place, certified only marginally in Llama-3.1-8B and Mistral-7B-v0.3 and not certified at all in Qwen-2.5-7B (Appendix D).

The conflict condition cuts the other way at one step. At a one-step rotation it resolves for the deviation in all four models at the embeddings (Table 10) and in all three patched models through the residual stream (§3.3), so the few-planes account covers the weekday deletion results and not the one-step weekday conflict. At larger displacements the conflict result turns back toward the account. At a two-step rotation the weekday conflict resolves for the deviation in Qwen-2.5-7B alone, and at four steps in none of the four models (Table 11), while the same two-step rotation resolves for the deviation in all four models in months, at 0.682 to 0.970 against 0.000 to 0.174 at full retention and crossing by α=0.5, and in all four models in hours, at 0.599 to 0.765 against 0.091 to 0.227 and crossing by α=0.75. Twelve concepts leave the deviation five components and seven leave it two, and we offer as a proposal, not a finding, that this is what the two-step contrast tracks. We do not adjudicate why the one-step weekday conflict behaves differently.

Hours do not fit the account at all, since twelve hours admit as many frequencies as twelve months, yet the deviation alone reaches only 0.49 to 0.72 against baselines of 0.69 to 0.95 in the four models, and we do not adjudicate the cause (§5).

B THE BARE STRUCTURE THROUGH ITS FULL ORBIT

Rotating the bare structure through its full orbit, accuracy at the rotation-predicted month holds at the structure-only level at every whole-step angle, while accuracy at the unrotated labels decays away from $0°$ (Fig. 5). Whole-step angles, multiples of $30°$, map month labels onto month labels, so their predicted-month accuracy doubles as a validation that rotation is an exact relabeling. Half-step angles place the circle between two months.

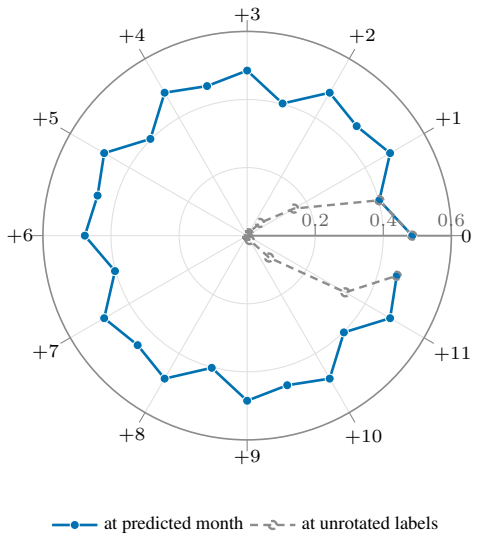

Figure 5: The bare structure rotated through its full orbit (Llama-3.1-8B, months). Accuracy at the rotation-predicted month (solid) holds at the structure-only level at every whole-step angle while accuracy at the unrotated labels (dashed) decays: the circle is read as geometry, and the geometry alone falls well short of baseline.

C EMBEDDING–UNEMBEDDING BLOCK ALIGNMENT

Table 12 reports, for each structured block, the fraction of embedding strength and unembedding strength it carries and the maximal principal cosine between the corresponding subspaces. In the three untied models every cosine is at most 0.046: the structured subspaces of embedding and

unembedding are nearly orthogonal, so embedding-side interventions are not mirrored in the unembedding. Gemma-2-9b ties embedding and unembedding, giving identical fractions and cosine 1, and is untied by cloning before any edit (§2.2).

Domain	Model	Block	Frac. W	Frac. U	Principal cos.
Hours	Qwen-2.5-7B	const	0.342	0.391	0.017
Hours	Qwen-2.5-7B	k=1	0.170	0.203	0.044
Hours	Mistral-7B-v0.3	const	0.400	0.400	0.014
Hours	Mistral-7B-v0.3	k=1	0.167	0.202	0.046
Hours	Llama-3.1-8B	const	0.450	0.461	0.008
Hours	Llama-3.1-8B	k=1	0.157	0.185	0.026
Hours	Gemma-2-9b	const	0.601	0.601	1.000
Hours	Gemma-2-9b	k=1	0.150	0.150	1.000
Letters	Qwen-2.5-7B	const	0.273	0.383	0.032
Letters	Qwen-2.5-7B	ramp	0.038	0.032	0.004
Letters	Llama-3.1-8B	const	0.355	0.477	0.000
Letters	Llama-3.1-8B	ramp	0.035	0.028	0.009
Letters	Gemma-2-9b	const	0.612	0.612	1.000
Letters	Gemma-2-9b	ramp	0.021	0.021	1.000
Numbers	Llama-3.1-8B	const	0.247	0.323	0.030
Numbers	Llama-3.1-8B	ramp	0.046	0.071	0.001
Numbers	Llama-3.1-8B	T=100	0.054	0.067	0.038
Numbers	Llama-3.1-8B	T=10	0.028	0.024	0.028
Numbers	Llama-3.1-8B	T=5	0.029	0.024	0.020
Numbers	Llama-3.1-8B	T=2	0.015	0.015	0.020

Table 12: Structured-block strength fractions in the embedding (W) and unembedding (U) and the maximal principal cosine between the corresponding subspaces.

D THE CALIBRATED SIGNIFICANCE TEST

Separately from the interventions, every structure-existence claim in the paper is checked by a significance test constructed and validated here. The obvious readout for a claimed structure, the fraction of the total strength lying in that structure's subspace, is uncalibrated. Applied to null data generated at the GPT-2 embedding spectrum, which contains no structure at all, it certifies structure in 93.5% of trials. Our test reduces that rate to 4.0% at the same spectrum and supplies a threshold below which a structure of a given size cannot be distinguished from chance.

The natural statistic for a claimed structure is the fraction of the embeddings' total strength that lies in the structure's subspace. This statistic is unreliable on its own, because a structureless matrix with the singular-value spectrum of a trained model concentrates strength in low-dimensional subspaces by chance. We measure the size of this failure by the calibration protocol below: draws of structureless matrices constructed to match the token-embedding spectrum of GPT-2, each scored by each readout, with a false certification counted whenever structure is certified in a wrong mode ($k \geq 2$). We run 200 trials with B=400 rotations per test. Our test conditions on the spectrum. The null distribution is generated by applying random orthogonal transformations on the index side of the matrix, which preserves the spectrum exactly while destroying any alignment with the hypothesized structure, and the observed strength fraction is compared against this null. The construction yields a closed-form *detection floor*: the smallest strength share at which a structure of a given dimension can be distinguished from chance at a given significance level, computable before any experiment is run. We validate the test in the finite-group setting, where the embedded structure is known exactly, obtaining agreement at the level of floating-point rounding error on the symmetric groups S_4 through S_6.

Applied to the paper's models, the test certifies the helix periods in Llama-3.1-8B, each at $p \leq 0.001$, the smallest value its 1000 rotation draws can return, and the month circle in all four models at $p \leq 0.0005$, the floor for 2000 draws. The weekday circle is weaker in every model. It is certified in Gemma-2-9b at $p = 0.0015$, in Llama-3.1-8B at $p = 0.029$ and in Mistral-7B-v0.3 at $p = 0.039$, and the last two do not survive a Bonferroni correction across the four weekday tests. In Qwen-2.5-

7B the test declines to certify it at all ($p = 0.179$), a negative result we report as the test withholding certification.

Table 13 gives the calibration rates. Thresholded strength readouts certify structure on structureless inputs over 90% of the time. The spectrum-preserving rotation null alone reduces this to 20.0%, and only the Bonferroni-corrected spectrum-preserving test is calibrated, at 4.0%.

The existence tests reported in §3.4 apply the spectrum-preserving null to one pre-specified block at a time. For the cyclic domains the statistic is the frequency-one plane's share of the strength left after removing the constant, compared against 2000 random orthogonal transformations of the index side. For the helix it is each period block's share after removing the constant and the ramp, compared against 1000. The p-value adds one to the count of null draws at least as extreme and to the number of draws, so its smallest attainable values are $1/2001$ and $1/1001$. All four helix periods reach that floor in Llama-3.1-8B and remain significant after a Bonferroni correction across the four ($p \leq 0.004$).

Readout	False certification rate
Thresholded strength, uniform baseline	93.5%
Thresholded strength, isotropic Beta	91.5%
Spectrum-preserving null, uncorrected	20.0%
Spectrum-preserving null + Bonferroni	4.0%

Table 13: False certification on structureless inputs constructed at the GPT-2 embedding spectrum. Only the spectrum-preserving rotation null with Bonferroni correction is calibrated.

E Deviation content: corpus and exploratory analyses

Per the Limitations section, the analyses in this section were designed after the causal results and are labeled exploratory. The corpus analysis uses a public encyclopedia sample as a proxy for four proprietary training corpora, and the month-level match rests heavily on the May-adjacent deficits, a single interpretable component.

E.1 Corpus correlations

Table 14 gives the deviation-content versus corpus co-occurrence correlations per model, with permutation nulls. On weekdays every model is individually significant at $p \leq 0.018$ and the pooled correlation is 0.4054. On months the pooled correlation is 0.2946, and every model is individually significant at $p < 0.05$, Mistral-7B-v0.3 narrowly at $p = 0.048$. With May excluded the correlations remain positive but none is significant, which we report as a power limit. Excluding both May and March recovers part of the pooled correlation without recovering significance.

E.2 Semantic regressions (exploratory)

Table 15 regresses deviation content on categorical predictors. Months show no reliable categorical structure in any model. On weekdays the day-type coefficient is positive in all four models and reaches significance in Gemma-2-9b, consistent with the workweek pattern in the corpus analysis.

E.3 Ellipticities

Table 16 gives the per-mode ellipticities of the embedding rows for all four models, offered descriptively. The ellipticity of a plane is the ratio of the smaller to the larger singular value of the rows' component in that plane (Appendix H), so it is 1 for a circle. The deviation's planes ($k \geq 2$) are close to circular in every model, with ellipticities averaging 0.928 against 0.988 for the grokked network's key frequencies ($k \in \{2, 5, 8, 9, 15, 19, 22\}$: 0.9817, 0.9883, 0.9925, 0.9852, 0.9800, 0.9958, 0.9911).

Task	Variant	Model	r	Null p_{95}	p
Months	all (n=66)	Qwen-2.5-7B	0.2331	0.1376	0.006
Months	all	Mistral-7B-v0.3	0.1567	0.1535	0.0479
Months	all	Llama-3.1-8B	0.2538	0.1450	0.004
Months	all	Gemma-2-9b	0.2840	0.1778	0.002
Months	all	Pooled	0.2946	–	–
Months	excl. May (n=55)	Qwen-2.5-7B	0.1579	0.1604	0.0579
Months	excl. May	Mistral-7B-v0.3	0.1046	0.1830	0.1816
Months	excl. May	Llama-3.1-8B	0.1303	0.1959	0.1397
Months	excl. May	Gemma-2-9b	0.0876	0.2102	0.2555
Months	excl. May	Pooled	0.1348	–	–
Months	excl. May, March (n=45)	Qwen-2.5-7B	0.2146	0.2197	0.0559
Months	excl. May, March	Mistral-7B-v0.3	0.0063	0.2354	0.5070
Months	excl. May, March	Llama-3.1-8B	0.1492	0.2304	0.1776
Months	excl. May, March	Gemma-2-9b	0.1967	0.2226	0.0798
Months	excl. May, March	Pooled	0.2242	–	–
Weekdays	all (n=21)	Qwen-2.5-7B	0.3568	0.2370	0.006
Weekdays	all	Mistral-7B-v0.3	0.3209	0.2549	0.018
Weekdays	all	Llama-3.1-8B	0.4053	0.2670	0.002
Weekdays	all	Gemma-2-9b	0.3961	0.2579	0.002
Weekdays	all	Pooled	0.4054	–	–

Table 14: Deviation-content versus corpus co-occurrence correlations per model, with the null 95th percentile and exact p-values.

	Months			Weekdays		
Model	R^2	p	β (season/quarter/length)	R^2	p	β (day type)
Qwen-2.5-7B	0.0504	0.4331	0.159/−0.098/0.006	0.0272	0.0918	0.165
Mistral-7B-v0.3	0.0178	0.8144	0.094/−0.004/0.088	0.0371	0.0619	0.193
Llama-3.1-8B	0.0068	0.9501	−0.035/−0.061/0.067	0.0200	0.1756	0.142
Gemma-2-9b	0.0166	0.8982	0.031/−0.106/−0.028	0.0427	0.0259	0.207

Table 15: Exploratory semantic regressions of deviation content on categorical predictors, with permutation p-values.

F Correspondence to the structures of prior work

The structures of equation 1 are intended as the objects prior causal analyses intervene on, and the correspondence is proved or measured per domain below. Two scope notes apply. The line is our own idealization, constructed so the ordinal hypothesis receives the same test the cyclic structures receive, and curved manifolds not contained in a subspace of functions on the index set, such as the counting manifolds of Gurnee et al. (2026), are outside the protocol.

Basis fairness. Nor can the pattern be an artifact of a decomposition biased toward cyclic structure. No Fourier vector enters the construction of the line's structure term. Its projector in equation 1 is determined by the constant and the ramp alone, and the orthogonal complement that defines the deviation is basis-free. The Fourier basis enters the letter domain only through the matched-random control, which groups the deviation by frequency to match its strength. The deletion verdicts therefore track the model's use of each structure, not the basis in which the deviation is described.

Blockwise Gram–Schmidt. Taking the structure blocks $B_1, \ldots, B_m$ in a fixed order, block j is replaced by the orthonormalized columns of $(I - Q_{<j}Q_{<j}^\top)B_j$, where $Q_{<j}$ collects the orthonormal bases of the preceding blocks. This leaves the total subspace the blocks span unchanged while making the blocks mutually perpendicular.

	Months, mode k					Weekdays, mode k		
Model	1	2	3	4	5	1	2	3
Qwen-2.5-7B	0.9597	0.9439	0.9062	0.9406	0.9467	0.9074	0.9726	0.9917
Mistral-7B-v0.3	0.8605	0.9609	0.9365	0.9703	0.9460	0.8950	0.9616	0.9662
Llama-3.1-8B	0.8679	0.9554	0.9388	0.9741	0.9064	0.8300	0.9240	0.8912
Gemma-2-9b	0.9080	0.9639	0.8413	0.9128	0.8171	0.7994	0.8405	0.8653

Table 16: Per-mode ellipticities of the operand embedding rows, all four models. Mode k=1 is the circle's plane, part of the structure, and is shown for reference. The months mode k=6 is a single alternating vector and has no ellipticity.

Canonicity of the circle subspace, and ellipses. A set of n concepts placed at equal angular steps about a common center, in whatever plane, at whatever phase and radius, lies exactly in the span of the constant vector and the plane at frequency one, so every idealized circle over the ordered concepts determines the same subspace, and the intervention target of Engels et al. (2025), the pair $[\cos(2\pi\alpha/n), \sin(2\pi\alpha/n)]$, is this plane written out. The plane also expresses every centered ellipse traversed at equal angular steps. We follow prior work in calling it the circle, and report the measured ellipticities of the planes descriptively in Appendix E.3.

Alignment of the fitted circle plane. The relation between the analytic plane and the realized intervention subspace of Engels et al. (2025), a probe fitted by least squares inside a five-dimensional principal-component subspace, is empirical, and we measure it at the embedding level by applying their construction to the operand rows of our four models. This correspondence is measured at the embeddings, not at their original activation sites. Across months and weekdays in all four models, the largest principal angle between the fitted plane and the analytic plane lies between $9.7°$ and $20.5°$, against a mean of $71.6°$ to $72.2°$ for random planes drawn in the same principal-component subspace, and no random plane aligns as closely as the probe (Table 17). The strength of the operand rows projected onto the fitted plane is predominantly at frequency one, with the share in the deviation frequencies at most 0.093 (Qwen-2.5-7B, months). An intervention written through the fitted plane therefore writes predominantly, though not exclusively, onto the idealized structure.

Model	Domain	Principal angles (°)		Null mean	Null p_5	Share k=1	Share $k\geq$2
Qwen-2.5-7B	months	13.1	17.4	72.2	47.0	0.898	0.093
Qwen-2.5-7B	weekdays	10.4	20.5	72.1	47.0	0.916	0.073
Mistral-7B-v0.3	months	7.5	13.9	72.2	47.9	0.907	0.048
Mistral-7B-v0.3	weekdays	3.7	9.7	71.7	46.1	0.974	0.005
Llama-3.1-8B	months	7.8	12.6	71.8	46.9	0.858	0.042
Llama-3.1-8B	weekdays	6.9	12.5	71.9	47.5	0.949	0.004
Gemma-2-9b	months	5.0	10.2	71.8	46.5	0.968	0.021
Gemma-2-9b	weekdays	8.1	14.6	71.6	47.5	0.990	0.000

Table 17: Alignment of the fitted circle probe of Engels et al. (2025) with the analytic frequency-one plane at the embeddings. The null columns give the mean and 5th percentile of the largest principal angle for 2000 random planes in the same five-dimensional principal subspace. The share columns give the fraction of the strength of the rows projected onto the fitted plane that lies at frequency one and at the deviation frequencies. The remainder is the constant.

Identity of the helix reconstruction. Kantamneni & Tegmark (2025) fit their helical form by projecting the operand-position states onto their first one hundred principal components, regressing the projected coordinates onto the basis functions by ordinary least squares, and mapping the fitted coordinates back through the components with the mean re-added. Three facts make this pipeline equal to the structure term of equation 1. Least-squares prediction from a set of basis functions is the orthogonal projection onto their span, whatever the conditioning of the basis. Subtracting the mean is a shift the basis absorbs, because the basis contains the constant. And retaining one hundred principal components loses nothing, because the operand-position states take at most one hundred distinct values, one per operand, so their span has dimension at most one hundred. The reconstruction their causal claims rest on is therefore the same matrix as our structure term. We

confirmed the identity numerically on synthetic states, with a maximum absolute difference of 5×10^{-15} between a reimplementation of their released fitting code and our projection. The identity holds wherever the operand alone determines the state, as it does for the first-operand fits that carry their headline claims and ours. States at the second operand vary with both operands, and there the principal-component step can truncate.

G TASKS, TEMPLATES, CHECKS, AND CAMPAIGN BASELINES

Each domain's prompts ask for the concept displaced from the operand by a stated offset: which month follows a given month by a given number of months, which letter follows a given letter in the alphabet, and so on. The tokenization check is verified programmatically against the tokenized prompts. Table 18 lists every template. The embedding campaign evaluates each model on all of a domain's templates and runs surgery on the best one. The activation campaign does the same over its operand-first templates (§2.4). The two sets differ, both sets of baselines are reported below, and no cross-campaign comparison of absolute accuracies is made anywhere in the paper.

Campaign	Domain	Template

Embedding	months	`{item} plus {o} months is`
		`{o} months after {item} is`
		`The month {o} months after {item} is`
		`Question: What month comes {o} months after {item}? Answer:`
		`If it is {item}, then in {o} months it will be`

	weekdays	as months, with `day`/`days`, and `If today is {item}, then in {o} days it will be`

	hours	`{oword} hours after {item} o'clock is`
		`If it is {item} o'clock, in {oword} hours it will be`
		`Two hours after one o'clock is three o'clock. {oword} hours after {item} o'clock is`
		`Question: What time is {oword} hours after {item} o'clock? Answer:`

	letters	`The letter 2 letters after A is C. The letter {o} letters after {item} is`
		`In the alphabet, 3 letters after B is E, and {o} letters after {item} is`
		`The letter {o} letters after {item} is`
		`Question: What letter comes {o} letters after {item}? Answer:`

	addition	`The following is a correct addition problem.\n{a}+{b}=`
		`{a}+{b}=`
		`Question: What is {a}+{b}? Answer:`

Activation	months	`We know January plus 2 months is March. {item} plus {o} months is` `{item} plus {o} months is`
	weekdays	`We know Monday plus 2 days is Wednesday. {item} plus {o} days is` `{item} plus {o} days is`
	letters	`We know B plus 2 letters is D, and E plus 3 letters is H. {item} plus {o} letters is` `{item} plus {o} letters is`
	addition	`{a}+{b}=`

Table 18: Prompt templates by campaign and domain, one per row. Hours use number words for operands and offsets, and letter offsets run from 1 to 10.

Why rotations and translations. We restrict the transformation conditions of §2.2 to rotations and translations, rather than arbitrary affine maps, because only they carry a determinate answer hypothesis, a rotation by one month predicting the displaced month, whereas a general affine map asserts content that defines no answer key. Unstructured content changes are covered by the strength-matched random substitution of §2.3.

Scoring. Rotation and translation conditions are scored against every answer hypothesis: for each shift s, we record accuracy against the labels displaced by s, restricting to prompts whose displaced label remains in range for non-cyclic domains. The two possible outcomes separate cleanly. A computation that uses the structure as geometry produces accuracy concentrated at the displaced labels. A computation anchored by the remaining components produces accuracy concentrated at $s = 0$. The month and weekday campaign also records a 95% percentile bootstrap interval over prompts. For the numerical domain the answer set is the hundred single-token integers from zero to ninety-nine, and the addition prompts are drawn from operand pairs whose sum is at most ninety-nine, 500 pairs in the embedding campaign and 300 in the activation campaign.

The capability threshold. The capability threshold is a design choice set in the campaign code rather than a derived quantity, and two properties matter more than its exact value. It is uniform across domains, set above the highest chance level among them ($1/7$ for weekdays), so that no domain admits a model at a laxer standard than another. A per-domain threshold would make the set

of qualifying models harder to compare across domains. The threshold is also a floor for admission, not a description of the models we study. Admitted baselines range from 0.46 to 1.00, and every comparison in the paper is within-model, against that model's own unedited accuracy on the same prompts, so the threshold enters no inference. Its only role is to exclude models too weak for an intervention to show an effect in either direction.

The letter template sweep. The letter templates were revised after a first sweep in which Qwen-2.5-7B, Mistral-7B-v0.3, and Llama-3.1-8B all fell below the capability threshold, with best accuracies of 0.205, 0.122, and 0.259. On the revised templates Qwen-2.5-7B and Llama-3.1-8B reached 0.766 and 0.459 and entered surgery, Gemma-2-9b reached 0.571, and Mistral-7B-v0.3 did not enter surgery. Mistral's accuracy on the revised templates was not logged, so the first-sweep figure is the only one available for it.

Preservation checks by campaign. In the month and weekday campaign the preservation check covers the circle: its full vector wherever the condition leaves the circle in place, its constant part and its plane's strength under rotation, and its constant part alone when the circle's plane is deleted. In the hour, letter, and addition campaign it covers every block the condition leaves untouched, and the deviation wherever it is left untouched, except in the numerical domain's three block-retention conditions. The grokked network's conditions carry no preservation check. The bound applies to the constructed rows. Writing them into a language model rounds every entry to the nearest value representable in the model's half-precision format, changing it by up to 2^{-8} of its magnitude in bfloat16 or 2^{-11} in float16, and the check does not cover this rounding.

Matched random for the ordinal and numerical domains. For the ordinal and numerical domains, whose deviation is not a set of whole planes, the per-frequency components of the strength-matched random deviation (§2.3) are taken after orthogonalizing each Fourier plane, in order of increasing frequency, against the structure and the lower frequencies.

Campaign baselines. Table 19 reports the activation campaign's baselines. Where they differ from the embedding campaign's, most visibly in Llama-3.1-8B months (0.500 to 0.515 against 0.970), the two campaigns ran different template sets. Predictions for every condition are archived for reanalysis.

Model	Domain	Baseline	Causal window	Drift
Qwen-2.5-7B	months	0.902	layers 0–21	0.000
Qwen-2.5-7B	letters	0.576–0.595	layers 0–21	0.020
Qwen-2.5-7B	weekdays	0.714	layers 0–21	0.000
Mistral-7B-v0.3	months	0.894	layers 0–18	0.000
Mistral-7B-v0.3	weekdays	0.833	layers 0–18	0.000
Llama-3.1-8B	months	0.500–0.515	layers 0–15	0.015
Llama-3.1-8B	weekdays	0.905–0.929	layers 0–15	0.024
Llama-3.1-8B	numbers	0.963–0.967	layers 0–15	0.003

Table 19: Activation-campaign baselines, causal windows, and observed do-nothing drift (maximum baseline spread across layers, residual stream) per model and domain. Drift never exceeds 0.024 against the 0.04 tolerance. Gemma-2-9b was not run at depth.

H THE DEVIATION THROUGH DEPTH: CONSTRUCTIONS AND FULL RESULTS

Templates, layers, and conditions. Letter arithmetic was also attempted in Mistral-7B-v0.3 and Llama-3.1-8B, and neither run produced results. Each domain uses the better of two operand-first templates, one behind a constant few-shot prefix and one without, except addition, which uses the single template `{a}+{b}=` (Table 18). The tested layers are the first three, every third layer after them, the middle layer, and the last. At each tested layer the conditions are the do-nothing patch, the zero patch, the structure alone, the deviation alone, the structure with one strength-matched random deviation at full strength, and, for cyclic domains, a one-step rotation of the circle with the

full deviation retained. These conditions are built by the same construction functions as the embedding conditions and inherit their start-up self-tests, but per-condition preservation is not asserted for activation states.

Alongside the interventions, we track the operand states descriptively through depth: the deviation's share of the total strength of the states, where the total includes the constant block shared by all operands, the mean ellipticity of the states across all two-dimensional Fourier planes, and the correlation of the deviation's content with the content of the embedding-level deviation, each defined below.

Layer indexing. Tracking layer 0 is the input embedding, and tracking layer $\ell + 1$ is the residual stream after block ℓ, which is intervention layer ℓ.

Deviation share. The deviation's share at a layer is $\|D\|^2/\|M\|^2$, where M holds the operand-position states and D their deviation. The total includes the constant block shared by all operands.

Content correlation. For the operand-position states at a layer, let D be their deviation and $G = DD^\top$ its Gram matrix. From each of the $n(n-1)/2$ entries above the diagonal we subtract the mean of all entries whose concept pair lies at the same cyclic index distance $\min(|i-j|,\, n-|i-j|)$, which removes the part of the similarity pattern that depends only on that distance. The content correlation at the layer is the Pearson correlation between these entries and the same entries computed from the deviation of the input embeddings, so it equals 1 at tracking layer 0 by construction. The same cyclic distance is used for the ordinal and numerical domains.

Ellipticity. For each two-dimensional Fourier plane B_k, let $\sigma_1 \geq \sigma_2$ be the singular values of the $2 \times d$ matrix $B_k^\top M$, where M holds the full operand-position states. The ratio σ_2/σ_1 equals 1 when the concepts trace a circle in that plane and 0 when they collapse onto a segment. The tracked ellipticity is the mean of this ratio over all two-dimensional planes, including the frequency-one plane.

How to read the depth claims. One clarification governs how these depth claims should be read. The decomposition is applied to each layer's own states, so the deviation at layer ℓ is defined at that site, and its content is rewritten along the way. In Qwen-2.5-7B months its content correlation with the embedding-level deviation falls from 1.00 at the embeddings to 0.34 after block 11, the middle of the window, while its causal advantage persists, and its share of the total strength of the states falls from 0.39 at the embeddings to between 0.10 and 0.19 inside the window, where the total includes the constant block shared by all operands (Appendix H). The persistent claim therefore concerns the functional partition between the structure subspace and its complement at each causally addressable site, not a single fixed content carried through the network. The reading on which any complement of a small subspace would do is ruled out by the matched random, which carries the same strength as the layer's own deviation and never restores the do-nothing baseline.

Summary. Table 20 summarizes the depth campaign per model and domain: the causal window, the range of the deviation-only and structure-only conditions across the window, the maximum of the structure with a strength-matched random deviation, and the tracking quantities. Full per-layer values are in the archived surgery and track files. Beyond the window the conditions converge, coinciding exactly in Qwen-2.5-7B and by the last tested layer in every model. Zeroing the MLP output at the operand position collapses accuracy at no tested layer of any model and domain except the first block of Llama-3.1-8B months (0.091 against 0.515). The largest other drop is 0.213, at the first block of Llama-3.1-8B addition.

I THE GROKKED NETWORK

The grokked network adds two integers modulo $p = 59$. Each operand is embedded in $\mathbb{R}^{128}$, the two embeddings are concatenated, and a hidden layer of 512 rectified units without biases feeds a linear readout over the 59 answers. It is trained full-batch with AdamW (learning rate 10^{-3}, weight decay 1.0, $\beta = (0.9, 0.98)$) on a random 30% of all p^2 operand pairs, for up to 40,000 steps, stopping once training and held-out accuracy have both stayed at 1.0 across 20 consecutive checks taken every 100

Model	Domain	Window	Dev. only	Struct. only	Rand. max	Corr. $0 \to$mid	Dev. share
Qwen-2.5-7B	months	0–21	0.83–0.91	0.32–0.39	0.47	$1.00 \to 0.34$	$0.39 \to 0.10$
Qwen-2.5-7B	letters	0–21	0.55–0.59	0.05–0.14	0.14	$1.00 \to 0.61$	$0.69 \to 0.18$
Qwen-2.5-7B	weekdays	0–21	0.67–0.76	0.45–0.62	0.69	$1.00 \to 0.28$	$0.30 \to 0.06$
Mistral-7B-v0.3	months	0–18	0.70–0.87	0.49–0.68	0.67	$1.00 \to 0.09$	$0.39 \to 0.14$
Mistral-7B-v0.3	weekdays	0–18	0.71–0.83	0.57–0.69	0.71	$1.00 \to 0.08$	$0.25 \to 0.08$
Llama-3.1-8B	months	0–15	0.42–0.52	0.22–0.33	0.29	$1.00 \to 0.65$	$0.40 \to 0.06$
Llama-3.1-8B	weekdays	0–15	0.57–0.64	0.71–0.83	0.88	$1.00 \to 0.21$	$0.20 \to 0.08$
Llama-3.1-8B	numbers	0–15	0.27–0.43	0.56–0.67	0.68	$1.00 \to 0.53$	$0.58 \to 0.08$

Table 20: Depth campaign summary (residual stream, in-window layers). Llama-3.1-8B weekdays retain at depth the structure-leaning profile they show at the first residual site (§3.1). The deviation alone leads the structure alone at every window layer except in Llama-3.1-8B weekdays and addition, where the structure leads. The structure with a strength-matched random deviation never reaches the do-nothing baseline. Corr.: content correlation of the layer deviation with the embedding deviation, from the embeddings to the state after block $\lceil W/2 \rceil$, where W is the last window layer. Dev. share: the deviation's fraction of total strength, from the embeddings to its minimum inside the window.

steps. Surgery is scored on the held-out pairs. The structure is the constant together with the plane that carries the largest share of the embedding strength, the key frequencies are the planes whose share exceeds twice the uniform share $2/p$, and the matched random deviation spans every plane outside the structure. The conditions are those of §2.2 without rotations, plus knockout and addition of each key frequency and retention of the top m frequencies for m from 1 to 6.