THE NUMERAL BASE SETS THE FOURIER FEATURES LANGUAGE MODELS COMPUTE WITH

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding neural networks requires explaining how the geometry of their representations supports the algorithms learned from training data. The Fourier features used by language models to represent numbers are tailored to the base-10 number system, and we ask how these features change when models learn a different numeral system. We pretrain matched language models on text and arithmetic in bases 6, 7, 8, and 10. We find that the features form harmonic families whose fundamental period follows the numeral base, e.g., periods $10, 5, 10/3, 2.5, 2$ in base 10 and $7, 7/2, 7/3$ in base 7. From these experiments, we develop a model in which each Fourier feature contributes a periodic pattern to the scores for possible digits of the sum. For a decimal sum ending in 7, a period-5 feature favors 2 and 7 because its score pattern repeats every five digits; combining it with a period-2 feature that favors odd digits gives 7 as their only shared peak, the digit the Chinese Remainder Theorem assigns to those two residues. Interchanging feature angles between inputs tests which digits each feature favors and how combinations of features determine a prediction. These contributions approximately add near the output in the base-7 and base-8 models and Qwen3.5-9B. In Qwen, patches at earlier layers reveal nonlinear interactions: the combined effect of several features exceeds the sum of their individual effects. We also find that carrying one into the tens position shifts Qwen's tens-place Fourier features toward the phases for adding one to the tens digit, with the parity feature updating last.

1 INTRODUCTION

Language models learn from data that carries traces of how the world is structured. Circles for months and curves for ordered quantities are examples of geometry that emerges inside these models (Engels et al., 2025; Modell et al., 2025; Gurnee et al., 2025; Karkada et al., 2026). Studies of correlations and symmetries in training data help explain how this geometry forms (Saxe et al., 2019; Korchinski et al., 2025; Karkada et al., 2026; Prieto et al., 2026). These representations provide data structures for learned algorithms, so their geometry is part of an account of how models compute (Geiger et al., 2026; Wurgaft et al., 2026).

Prior work has found that language models represent numbers using Fourier features with periods tied to decimal notation (Zhou et al., 2024; Kantamneni & Tegmark, 2025; Zhou et al., 2025; Fu et al., 2026; Feucht et al., 2026). A Fourier feature encodes a number as an angle that repeats with a fixed period. Features with periods such as 2, 5, and 10 supply information about decimal digits. The choice of base 10 is a convention that shapes how numbers appear in text. We ask what these features would look like in a world that used a different numeral system, and how their role in predicting a sum would change.

We pretrain matched language models on text and arithmetic expressed in bases 6, 7, 8, and 10, converting numbers throughout the training corpus. Across these numeral systems, the Fourier features follow harmonic families set by the base. For units digits, the fundamental has a period equal to the base; the other harmonics complete integer numbers of cycles within that period. For example, the units family has periods $10, 5, 10/3, 2.5, 2$ in base 10 and $7, 7/2, 7/3$ in base 7. Moving one digit position to the left multiplies each period by the base.

We develop a model in which each Fourier feature contributes a periodic pattern to the scores for possible answer digits. The base determines how much information each harmonic supplies. For

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.

example, a period-4 feature for base-8 units groups digits four apart, leaving other harmonics to distinguish the members of each pair. To test how models use this information, we perform interchange interventions on feature angles and measure the resulting digit scores. This approach builds on work that tests the role of geometric coordinates through interventions (Engels et al., 2025; Kantamneni & Tegmark, 2025; Gurnee et al., 2025; Sarfati et al., 2026; Wurgaft et al., 2026). Patching features individually and together tests the predicted contribution of each harmonic and how those contributions combine.

Our contributions are:

- We use controlled pretraining across numeral bases to test how numerical representations adapt to the notation in the training data.
- We identify harmonic families set by the numeral base and place value, and explain how the information in each harmonic changes across bases.
- Interchange interventions support approximately additive contributions to digit scores near the output in the base-7 and base-8 models and Qwen3.5-9B-Base.
- Earlier patches in Qwen reveal nonlinear amplification when features are combined. Carrying one shifts the tens features toward the phases for the next tens digit, with the parity feature updating last.

2 Experimental Setup and Methods

Pretraining toy models on data with numbers converted to a different base. We pretrain one eight-layer NanoChat language model for each base $b \in \{6, 7, 8, 10\}$. The training corpus combines ClimbMix text with 9.79 million arithmetic equations. About 82% of these equations add exactly two numbers, which is the task we study in this paper. Every ordered pair of operands from 0 to 2000 is included twice. For each base, we convert digit strings throughout the equations and web text while preserving document order and nonnumeric text. Each digit string is converted separately, including those within decimal fractions and scientific notation.

The runs share an architecture, training recipe, random seed, and digit-level tokenizer. Each model trains for 7,680 steps with a sequence length of 2,048 and a global batch of 262,144 tokens, totaling about two billion tokens. Since numeral length varies with the base, the runs reach different points in the corpus within this budget. Training details appear in Appendix C.1. Each model is a decoder-only transformer trained with the standard next-token objective, with eight blocks, residual width 512, four attention heads and four key/value heads, a vocabulary of 32,768 tokens, and a context length of 2,048.

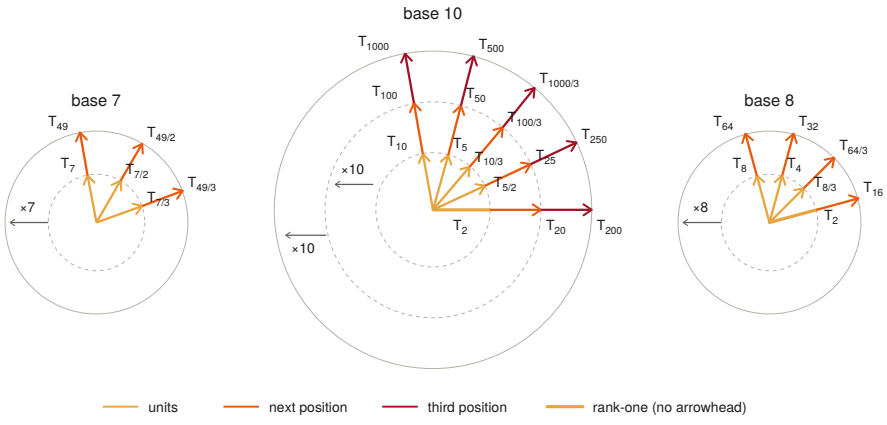


Figure 1: Harmonic families across numeral bases and digit positions. Labels give the feature periods, with the units family on the innermost ring. Moving one ring outward multiplies every period by the base. Period-2 features are one-dimensional and appear without arrowheads.

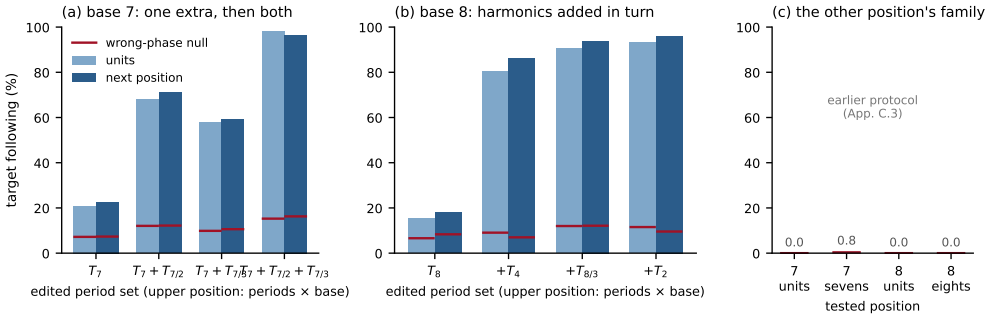


Figure 2: Fourier patching in models trained in bases 7 and 8, after block 7. (a, b) Interchange intervention accuracy as harmonics are combined, at units and at the next digit position. Periods for the next position are multiplied by the base. Red marks show paired controls with incorrect phases and matched patch norms. Each evaluation cohort contains 1,500 initially correct cases. (c) Patching the other position's family in a separate experiment using the earlier uncentered protocol (Appendix C.3).

We also test Qwen3.5-9B-Base (Qwen Team, 2026), with replications in Pythia-6.9B (Biderman et al., 2023) and Llama-3.1-8B (Grattafiori et al., 2024). Unless stated otherwise, interventions act after block 30 in Qwen and after the final block in NanoChat, indexed as block 7. We chose Qwen's intervention layer through a held-out layer sweep.

Fourier features across multi-digit sums. We study one answer digit at a time, providing the correct preceding digits. For example, the decimal prompt `23+14=3` tests prediction of the units digit `7`. Let $n \in \mathbb{Z}_{\geq 0}$ be the scalar sum. A Fourier feature with scalar period $T > 0$ maps n to coordinates on a circle:

$$\phi_T(n) = \left(\cos \frac{2\pi n}{T}, \sin \frac{2\pi n}{T}\right)^\top. \tag{1}$$

The angle on this circle is the feature's *phase*. For $T = 2$, we retain only cosine because sine is zero at integer sums. Thus, $\phi_T(n) \in \mathbb{R}^r$, where the scalar dimension r is 1 for $T = 2$ and 2 otherwise.

Let $p = b^j$ be the scalar place value for base b and position $j \in \mathbb{Z}_{\geq 0}$, counting from the right with $j = 0$ at the units position. For example, $b = 2$ and $j = 3$ give $p = 2^3 = 8$, the place value of the fourth digit from the right in a binary number. We test the *harmonic family* with periods $T = bp/k$ for $k = 1, \ldots, \lfloor b/2 \rfloor$. Here k is the harmonic index and $k = 1$ is the fundamental. In base 10, the units family has periods $10, 5, 10/3, 2.5, 2$. Multiplying each period by ten gives the tens family; another factor of ten gives the hundreds family. Thus, 37 and 47 have the same phases in the units family and different phases in the tens family. With the lower digits fixed, increasing the target digit by one advances harmonic k by k/b of a turn, at every position.

Under this condition, harmonic k identifies the digit modulo $b/\gcd(k, b)$. For decimal units, period 5 assigns the same phase to digits 2 and 7, while period 2 encodes parity. Together, these features identify a digit. Periods 2 and 3 have this property in base 6. In base 8, periods 2 and 4 leave two possible digits, while each harmonic in prime base 7 identifies the digit. We use these cases to test how the information in each feature affects its role in the model's prediction.

Training Fourier probes. Let $D \in \mathbb{N}_{>0}$ be the residual width and $h \in \mathbb{R}^D$ the activation at a chosen layer and the token immediately before the digit being predicted. For each tested period, we fit two affine maps by ridge regression and construct a basis for patching:

$$\phi_T(n) \approx \mathbf{B}h + a, \qquad h \approx \mathbf{W}\phi_T(n) + \mu, \qquad \mathbf{Q} = \mathrm{orth}[(\mathbf{I}_D - \mathbf{P})\mathbf{W}]\,. \tag{2}$$

The decoder has weights $\mathbf{B} \in \mathbb{R}^{r \times D}$ and intercept $a \in \mathbb{R}^r$. The reverse map uses $\mathbf{W} \in \mathbb{R}^{D \times r}$ and $\mu \in \mathbb{R}^D$ to fit activation directions from Fourier coordinates. Each fit minimizes the sum of squared errors plus the squared Frobenius norm of its coefficient matrix, with scalar penalty weight $\alpha = 1$ and unpenalized intercepts. We report decoding performance as the mean held-out R^2 across the predicted coordinates. Period grids and probe splits are specified with each experiment.

The columns of $\mathbf{W}$ give candidate patching directions. Here $\mathbf{P} \in \mathbb{R}^{D\times D}$ is the orthogonal projector onto the protected directions, and $\mathbf{I}_D \in \mathbb{R}^{D\times D}$ is the identity matrix. We use $\mathbf{P} = \mathbf{0}$ when no directions are protected. After removing the protected components, orth returns an orthonormal basis $\mathbf{Q} \in \mathbb{R}^{D\times q}$, where the scalar dimension $q \le r$ is the retained numerical rank. Patching subspaces are fitted on sums disjoint from intervention evaluation; the fitting order and protected directions are specified with each experiment; the numerical-rank threshold is given in Appendix A.2.

Fourier patching. We use *Fourier patching* to perform an interchange intervention on the angle in the plane learned by the probe. We replace the original activation's angle with the counterfactual activation's angle while preserving the original radius. In a one-dimensional subspace, including the period-2 feature, the intervention replaces the sign and preserves the magnitude. The paired inputs match on the answer prefix and lower digits. We also match the incoming carry while changing the digit being tested. For original and counterfactual activations $h_o, h_c \in \mathbb{R}^D$ and subspace center $c \in \mathbb{R}^q$, define $z_o = \mathbf{Q}^\top h_o - c$ and $z_c = \mathbf{Q}^\top h_c - c$, both in $\mathbb{R}^q$. The patched activation $\widetilde{h}_o \in \mathbb{R}^D$ is

$$\widetilde{h}_o = h_o + \mathbf{Q}\left(\|z_o\|_2 \frac{z_c}{\|z_c\|_2} - z_o\right). \tag{3}$$

This preserves the projection onto the orthogonal complement of the edited subspace. Unlike steering, which adds a feature direction with a chosen gain (a gain of ten in Feucht et al. (2026)), Fourier patching replaces only the angle at the activation's own radius, so no magnitude is added and the patch is an interchange intervention rather than an amplification. The centered base-7 and base-8 factorials use $c = \mathbf{Q}^\top \mu$, with μ the intercept of the reverse ridge fit estimated from fitting rows only. The Qwen factorial and layer sweep of Section 4.1 use $c = 0$, and a centered replication of the sweep is reported in Appendix B.5; the Qwen residue and chimera experiments center each plane at its fitting-set mean, and the center used by every other experiment is stated with that experiment in the appendix. A counterfactual whose radius falls below 10^{-8} has no defined direction and invalidates the case; in the centered radix runs an original activation at the center (radius below 10^{-8}) receives a zero edit. Control donors below the threshold are skipped, and a case without an admissible control is excluded from that paired comparison only, not from the all-case accuracy.

For an original sum of 34, we can take the period-5 phase from a counterfactual sum of 32 and the parity phase from a counterfactual sum of 33. These phases jointly specify units digit 7, so we test whether the prediction changes from 4 to 7. In factorial experiments, we evaluate every subset $S \subseteq \{1, \ldots, \lfloor b/2 \rfloor\}$ of harmonics by summing its individual displacements, each computed from the unedited activation. The empty subset $S = \varnothing$ gives the unedited baseline. When subspaces overlap, the combined edit can alter the radius and direction within each subspace. Experiments that preserve a joint radius specify this variant separately.

Evaluating Fourier patching. We evaluate cases where the unedited next-digit prediction is correct. *Interchange intervention accuracy* (IIA) is the fraction of cases in which the model's highest-scoring token across the full vocabulary matches the intended counterfactual digit. We also report the *logit difference score* between the target and original digits. For each case, let $s, t \in \{0, \ldots, b-1\}$ be the scalar original and target digits, and let $\ell_S(d) \in \mathbb{R}$ be the logit of the token for digit $d \in \{0, \ldots, b-1\}$ after patching harmonics in S. The score $m(S)$ and its change from the unedited baseline $\Delta m(S)$ are real scalars:

$$m(S) = \ell_S(t) - \ell_S(s), \qquad \Delta m(S) = m(S) - m(\varnothing). \tag{4}$$

For factorial patches, we compare this change with the additive prediction $\sum_{k\in S} \Delta m(\{k\})$, keeping s and t fixed across subsets within each case. This comparison includes effects too small to change the predicted digit. When the supplied phases identify several possible digits, we examine predictions over that set.

Controls include edits of the same norm that supply the wrong phase, with constructions specified for each experiment. We examine changes across digit logits to see which competing digits each harmonic favors. The definitions above apply to the Qwen and NanoChat experiments, which supply the answer prefix and score single-digit tokens; the Pythia and Llama replications predict a whole one-token answer or a weekday label, and their scoring is given in Appendix D.1 and Appendix D.2.

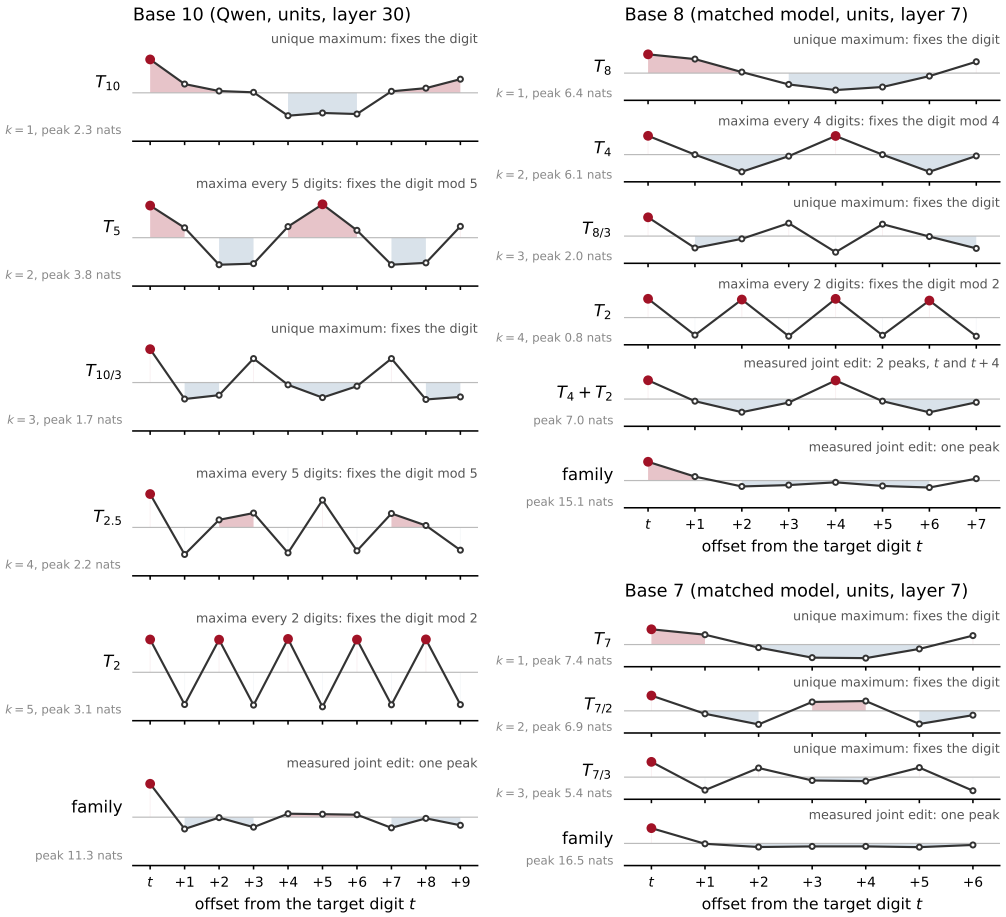


Figure 3: Harmonic contributions to digit scores. Each row shows the change from the unedited digit logits, centered across digits within each case, aligned to the target digit t, and averaged over cases. Curves are divided by their own positive peak, with its size reported in nats; red points mark maxima. The family rows and the base-8 row for the joint patch of periods 4 and 2 show measured joint patches. All panels concern units digits: bases 7 and 8 after block 7 ($n = 1{,}500$), and Qwen after block 30 ($n = 800$).

3 Fourier Features Follow the Numeral Base and Digit Position

We test whether the harmonic families specified by numeral notation control digit predictions in the models trained in bases 7 and 8. For units digits, the predicted periods are $7, 7/2, 7/3$ in base 7 and $8, 4, 8/3, 2$ in base 8. Multiplying each period by the base gives the family for the next digit position. We fit these features and perform Fourier patching at the final block, indexed as block 7, using 1,500 initially correct cases per base and position. Each patch supplies phases from a counterfactual sum with a different target digit. We evaluate every subset of the family, with controls that match the patch norm and supply incorrect phases (Appendix C.3).

At the units position, patching the fundamental alone gives an interchange intervention accuracy (IIA) of 20.9% in base 7 and 15.4% in base 8 (Figure 2a,b). With the second harmonic also patched, IIA reaches 67.9% and 80.6%, respectively. The complete families change the prediction to the target digit in 98.0% of base-7 cases and 93.1% of base-8 cases. Their matched controls reach 15.3% and 11.5%.

The same progression occurs at the next digit position, where the periods scale with the base. In base 7, patching the family with periods $49, 49/2, 49/3$ gives 96.5% IIA. The corresponding base-

8 family, $64, 32, 64/3, 16$, reaches 96.1%. Patching only the fundamental at these positions gives 22.6% and 18.1%, respectively.

Figure 2c reports a separate experiment that patches the family for the other digit position. For example, at base-7 units we patch periods $49, 49/2, 49/3$. These patches yield at most 0.8% IIA across the four conditions. This experiment uses an earlier protocol with uncentered subspaces and patches that preserve the radius in the combined span; its patch norms are between half and two thirds of those for the corresponding native families (Appendix C.3).

In both models the contributions combine approximately additively at the final block: the full-family change in the logit difference score is within 5.2% of the sum of the individual changes at units and at the next position, with per-case vector R^2 of 0.95–0.98 (Appendix C.3).

We next examine how each harmonic changes the scores for possible answer digits. For each patch, we subtract the unedited digit logits, center the changes across digits, and average over cases after aligning the target digit t to offset zero. Figure 3 shows these profiles for units digits in the base-7 and base-8 models, together with Qwen3.5-9B-Base after block 30 on 800 cases. Each curve is scaled by its own positive peak, whose size is reported in nats. Digit offsets are taken modulo the base.

The dominant frequency of each single-harmonic profile matches the patched harmonic. In base 7, every harmonic produces its largest mean increase at the target among the seven candidate digits. The period-4 feature in base 8 favors both t and $t + 4$, and the period-2 feature favors digits with the target's parity. Patching these two features together produces a profile with peaks at t and $t + 4$. The period-8 and period-$8/3$ features each have a unique highest value at the target, as does the measured patch of the complete family.

Qwen shows the corresponding decimal patterns. The period-5 feature favors t and $t + 5$, which share its phase, and period 2 favors digits with the target's parity. Period 2.5 also favors t and $t + 5$, with a different pattern across the other digits. The period-10 and period-$10/3$ profiles each have their highest value at the target. Patching the complete family produces its largest mean increase there. These profiles describe how each patch changes relative digit scores; the final prediction also depends on the scores before the patch.

4 Fourier Features During Addition in Qwen

4.1 Earlier Patches Reveal Nonlinear Interactions

We test how harmonic contributions combine across the layers of Qwen3.5-9B-Base. At each block from 21 to 31, we fit the five units-digit features and patch every nonempty subset on the same 800 initially correct cases. For each subset, we compare the measured change in the logit difference score with the sum of the changes from its individual patches (Appendix B.5).

Patching the complete family gives 7.0% IIA after block 27, rising to 61.8% after block 28 and 92.9% after block 30 (Figure 4a). At block 28, the individual effects sum to a mean score change of 9.55 nats, and the joint patch produces 15.84 nats, about 1.7 times the additive prediction. At this block, positive interactions occur for all ten pairs of harmonics. The excess also grows as further harmonics are patched together (Figure 4c–e).

At block 30, the measured full-family change is 24.08 nats and the sum of individual changes is 24.28 nats. Across all 31 subsets, the additive prediction explains the mean score changes with $R^2 = 0.9992$. Each patch location uses separately fitted subspaces and leaves a different sequence of blocks to process the edited activation. Identifying the operations that produce the earlier amplification remains an open question.

4.2 Residues Compose Across Inputs, and Base 8 Splits

The composition test of Section 2 asks whether two residue phases taken from two different inputs combine into a digit that neither input contains. In Qwen after block 30, taking the period-5 phase from one counterfactual sum and the parity phase from another, chosen so that the digit they jointly specify appears in neither, produces that digit above shuffled-phase controls at units, tens,

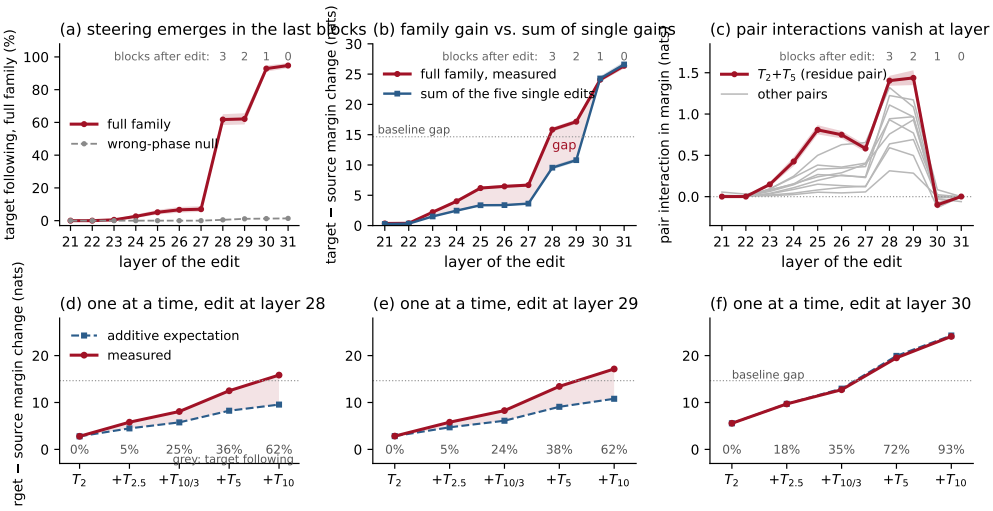


Figure 4: Harmonic interactions across Qwen layers, using the same 800 units-digit cases. (a) Full-family IIA and the incorrect-phase control. (b) Measured full-family score change and the sum of individual changes. (c) Pair interactions, computed as the joint change minus the sum of the two individual changes. (d–f) Harmonics added successively after blocks 28, 29, and 30; grey values give IIA. Dotted lines mark the original digit's mean logit lead over the target. Bands show 95% case-bootstrap intervals.

and hundreds, by 10–16 pp with the fundamental left at its original value and by 32–35 pp with the fundamental removed (Appendix B.3). Residue information supplied from separate inputs can therefore be recombined into a digit present in neither input, the internal counterpart of phase-swap experiments on images (Oppenheim & Lim, 1981; Yıldırım, 2026). Pythia-6.9B gives the same result (Appendix D.1).

In base 8, jointly patching periods 4 and 2 leaves the two-digit ambiguity that Figure 3 predicts. With the period-8 feature left at its original value, the joint patch places two thirds of the outputs whose residues change on the pair $\{t, t+4\}$, split evenly (429 against 430 of 1,283), and outputs at $t+2$ or $t+6$, which would violate the patched period-4 residue, are rare. The patch raises t and $t+4$ by nearly the same amount, so which of the two wins is carried over from the logit advantage present before the patch: at odd shifts from the original digit the nearer candidate is favoured, and at the equidistant shifts 2 and 6 the smaller baseline asymmetry remains. Placing the same patch one block earlier enlarges that advantage by more than half in the same direction (Appendix C.2).

4.3 Carry Shifts the Phases of Tens-Place Features

We compare 900 pairs of additions with and without an incoming carry, matched on the local tens sum before carry, the lower remainder, and the answer prefix. At each layer, we fit and calibrate phase readouts on separate carry-free examples. Carry increases the tens digit by one, so harmonic k should advance by $k/10$ of a turn. For the period-20 feature, which records the tens digit's parity, this correction is half a turn.

By block 30, the mean paired differences in all five planes point along the expected correction, with directional errors of at most 7°. Their components along that correction reach 84–93% of its expected size (Figure 5a,b). The parity feature tracks the local tens sum at blocks 23–24 and acquires the carry correction across blocks 25–27. Its radius briefly shrinks during this transition (Figure 5c).

A second comparison using 1,000 pairs matched on the final sum shows the same delayed parity update (Appendix B.7). These comparisons measure how carry appears in the representations. The computation that produces the carry remains to be traced.

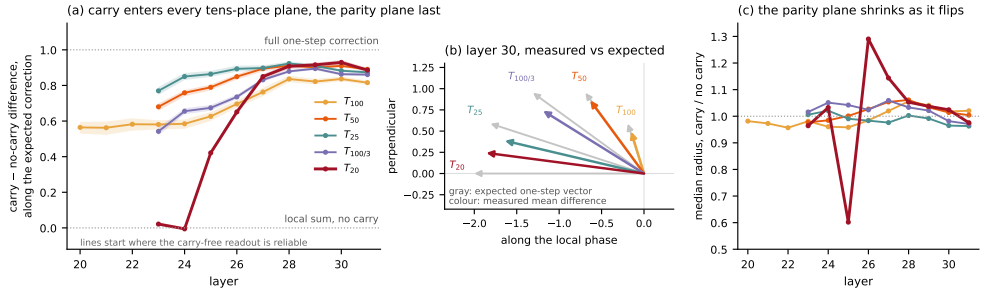


Figure 5: Carry in Qwen's tens-place Fourier features. (a) Mean carry-minus-no-carry difference projected onto the expected correction for a one-digit increase, using pairs matched on the local tens sum. A value of 1 denotes the full expected correction; curves begin where the carry-free phase readout meets the reliability criterion in Appendix B.7. (b) Measured and expected differences at block 30, expressed in each plane's coordinates. (c) Median radius on carry examples divided by the median radius on no-carry examples. Bands show cluster-bootstrap intervals over pair groups.

5 Related Work

Algorithms learned from arithmetic. Studies of models trained on arithmetic have recovered Fourier algorithms for modular addition and examined how architecture and training affect which frequencies contribute (Nanda et al., 2023; Zhong et al., 2023; Ding et al., 2024; Furuta et al., 2024; McCracken et al., 2025). Work on multidigit addition and subtraction traces how circuits combine computations across digit positions (Quirke & Barez, 2024; Quirke et al., 2025).

Numerical reasoning in pretrained language models. Pretrained language models represent numbers and cyclic concepts with periodic features, including periods tied to decimal digits (Engels et al., 2025; Levy & Geva, 2025; Zhou et al., 2024; Kantamneni & Tegmark, 2025; Kadlčík et al., 2025; Fu et al., 2026; Feucht et al., 2026). Fourier features also support precise number embeddings designed for language models (Zhou et al., 2025). Related accounts trace how MLPs contribute to arithmetic predictions (Stolfo et al., 2023; Nikankin et al., 2025; Baeumel et al., 2025), with further work linking carry to continuous geometry (Wen et al., 2026).

Training data and representation geometry. Correlations and symmetries in training data shape the geometry of learned representations (Saxe et al., 2019; Korchinski et al., 2025; Karkada et al., 2026; Prieto et al., 2026). Such geometry can encode relations between concepts and support operations such as counting (Engels et al., 2025; Park et al., 2025b;a; Gurnee et al., 2025; Sarfati et al., 2026; Wurgaft et al., 2026).

Causal analysis of neural network internals. Causal mediation analysis and causal abstraction use interventions on internal states to test their role in a model's behavior (Vig et al., 2020; Finlayson et al., 2021; Meng et al., 2022; Stolfo et al., 2023; Mueller et al., 2026; Geiger et al., 2025b; 2024; Wu et al., 2023; Geiger et al., 2025a). Our phase patches build on interventions that use the geometry of a representation to control model predictions (Engels et al., 2025; Kantamneni & Tegmark, 2025; Gurnee et al., 2025; Sarfati et al., 2026; Wurgaft et al., 2026). Algebraic analysis and equivariance also help recover explicit algorithms from trained networks (Lutz et al., 2025; 2026).

6 Discussion

Across matched models trained in four bases and a pretrained decimal model, the Fourier features that decide a digit form a harmonic family of the place value, and the family is set by the notation. In small networks trained on modular addition the learned frequencies are a choice of the optimizer and need not divide the modulus (McCracken et al., 2025); here they follow the base, and the convex amplification before Qwen's readout does not appear in the eight-layer models, whose last three

blocks are already additive. Mechanisms read off small or single-task networks are worth rechecking in the models they are meant to describe.

Two kinds of linearity should be kept apart. The family itself is an instance of the superposition principle of Klindt et al. (2026) with multidimensional features (Engels et al., 2025): each harmonic is a circular feature, nonlinear within itself, and the residual stream is close to their linear sum, which is what the reverse fit assumes and what lets one plane be patched while the protected coordinates stay put. How the family is read is a separate question. Near the output the harmonics add, each writing a periodic profile at its own frequency, and the prediction follows the peak of the sum; a few blocks earlier in Qwen the same patches interact. Which subsets of the family can reach a digit is then a question of where the maxima of their profiles meet, and how strongly they move the prediction a question of accumulated height: residue phases from two inputs compose because their profiles add without reference to their origin, base 8 splits between the two digits where its period-4 and period-2 maxima meet, and in prime base 7 every profile has a unique maximum, so each harmonic moves the prediction alone and the family only adds height. The rule that complementary residues fix a digit is the Chinese Remainder Theorem; the model implements only the summation, and the theorem describes where the maxima meet.

These results carry a caution for interpretability aimed at monitoring or control. A plane can be decodable and yet not be where the decision is made. In Pythia the exclusive period-10 plane retains nearly all of its fitted norm and decodes the digit, but patching it alone changes no prediction, while patching the residue harmonics changes the digit even though the period-10 plane remains aligned with the original digit through the final layer (Appendix A.2); in Qwen no single plane owns the decision. A probe on one Fourier plane can report the right answer while the sibling harmonics carry the causal work, and suppressing that plane need not suppress the computation. We suggest treating the harmonic family, rather than any one period, as the unit of analysis.

Limitations. Our interventions measure the next digit's prediction rather than tracing how the harmonics are computed from the operands or where carry is resolved, and the primary NanoChat rates are read at the final block, close to the readout. Late-layer additivity and logit profiles are measured in Qwen and in the base-7 and base-8 models; the Qwen layer sweep covers units only, and the harmonics coprime to ten are spectrally less pure than periods 5 and 2. Phase patches use measured counterfactual directions, so intermediate angles and the shape of decision boundaries are not tested. The NanoChat models are small, one seed per base, and trained on a corpus with explicit arithmetic; their cross-position and ceiling arms come from an earlier uncentered run, and base 6 was tested only with the residue design. Interchange intervention accuracy is a coarse endpoint, and several confidence intervals are exploratory and unadjusted for multiple comparisons.

AI USE STATEMENT

Large language models were used most heavily for coding, figure generation, and writing, and we describe our methodology for each. For code, the authors designed the experiments and wrote the core intervention logic; language models were then used as coding agents to implement analysis pipelines, batch runners, figure scripts, and validation scripts around that core. The validation scripts test what the authors specified: intervention geometry and the intervention math on synthetic inputs, targeting of the intended layer and token position, preservation of the same evaluation examples across layers, correct resumption of interrupted runs, and consistency checks that recompute the reported statistics from saved outputs. A run was accepted only when these tests passed. For analysis, the experiments produced a large archive of logged outputs; language models were used to search this archive for patterns and to propose additional analyses of the saved data, which the authors decided whether to run and then reviewed. For writing, language models drafted and polished sections of the manuscript under the authors' direction and were used to check reported numbers against saved outputs. The research questions, experimental design, models, datasets, interpretation, and claims are the authors' own, and the authors take full responsibility for the work and its final content.

Reproducibility statement

The appendix reports the radix-training corpus, model and training configuration, qualification criteria, data splits, and Fourier analysis procedure. The main text specifies the intervention, controls, evaluation sets, and uncertainty estimates used for each causal result.

References

Tanja Baeumel, Daniil Gurgurov, Yusser Al Ghussin, Josef van Genabith, and Simon Ostermann. Modular arithmetic: Language models solve math digit by digit. In *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2025*, pp. 1380–1409, 2025. doi: 10.18653/v1/2025.findings-ijcnlp.86. URL https://aclanthology.org/2025.findings-ijcnlp.86/.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, Usvsn Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar Van Der Wal. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 2397–2430. PMLR, 2023. URL https://proceedings.mlr.press/v202/biderman23a.html.

Xiaoman Delores Ding, Zifan Carl Guo, Eric J. Michaud, Ziming Liu, and Max Tegmark. Survival of the fittest representation: A case study with modular addition. *arXiv preprint arXiv:2405.17420*, 2024.

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

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

Matthew Finlayson, Aaron Mueller, Sebastian Gehrmann, Stuart Shieber, Tal Linzen, and Yonatan Belinkov. Causal analysis of syntactic agreement mechanisms in neural language models. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1828–1843, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.144. URL https://aclanthology.org/2021.acl-long.144.

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.

Hiroki Furuta, Gouki Minegishi, Yusuke Iwasawa, and Yutaka Matsuo. Interpreting grokked transformers in complex modular arithmetic. *arXiv preprint*, 2024.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah D. Goodman. Finding alignments between interpretable causal variables and distributed neural representations. In *Proceedings of the 3rd Conference on Causal Learning and Reasoning (CLeaR)*, pp. 160–187, 2024.

Atticus Geiger, Jacqueline Harding, and Thomas Icard. How causal abstraction underpins computational explanation, 2025a. URL https://arxiv.org/abs/2508.11214.

Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, and Thomas Icard. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, 26(83):1–64, 2025b. URL http://jmlr.org/papers/v26/23-0058.html.

Atticus Geiger, Ekdeep Singh Lubana, Thomas Fel, Jack Merullo, Michael Byun, Owen Lewis, and Thomas McGrath. The world inside neural networks. *Goodfire Research*, May 2026. URL https://www.goodfire.com/research/the-world-inside-neural-networks. https://www.goodfire.com/research/the-world-inside-neural-networks.

Aaron Grattafiori et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

Wes Gurnee, Emmanuel Ameisen, Isaac Kauvar, Julius Tarng, Adam Pearce, Chris Olah, and Joshua Batson. When models manipulate manifolds: The geometry of a counting task. *Transformer Circuits Thread*, 2025. URL https://transformer-circuits.pub/2025/linebreaks/index.html.

Marek Kadlčík, Michal Štefánik, Timothee Mickus, Michal Spiegel, and Josef Kuchař. Pre-trained language models learn remarkably accurate representations of numbers. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025.

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.

David Klindt, Charles O'Neill, Patrik Reizinger, Harald Maurer, and Nina Miolane. A unifying framework from neural superposition to sparse interpretable codes. *Nature Machine Intelligence*, 8:1025–1037, 2026. doi: 10.1038/s42256-026-01259-z.

Daniel J Korchinski, Dhruva Karkada, Yasaman Bahri, and Matthieu Wyart. On the emergence of linear analogies in word embeddings. *arXiv preprint arXiv:2505.18651*, 2025.

Amit Arnold Levy and Mor Geva. Language models encode numbers using digit representations in base 10. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 2: Short Papers, Albuquerque, New Mexico, April 29 - May 4, 2025*, pp. 385–395. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.NAACL-SHORT.33. URL https://doi.org/10.18653/v1/2025.naacl-short.33.

Patrick Lutz, Aditya Gangrade, Hadi Daneshmand, and Venkatesh Saligrama. Linear transformers implicitly discover unified numerical algorithms. *arXiv preprint arXiv:2509.19702*, 2025.

Patrick Lutz, Themistoklis Haris, Arjun Chandra, Aditya Gangrade, and Venkatesh Saligrama. Symmetry reveals layerwise dynamics: How transformers perform in-context classification. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=fDMizWsNoG.

Gavin McCracken, Gabriela Moisescu-Pareja, Vincent Létourneau, Doina Precup, and Jonathan Love. Uncovering a universal abstract algorithm for modular addition in neural networks. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/3edb234091dca2023308398dbf824850-Abstract-Conference.html.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 17359–17372. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/6f1d43d5a82a37e89b0665b33bf3a182-Paper-Conference.pdf.

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

Aaron Mueller, Jannik Brinkmann, Millicent Li, Samuel Marks, Koyena Pal, Nikhil Prakash, Can Rager, Aruna Sankaranarayanan, Arnab Sen Sharma, Jiuding Sun, Eric Todd, David Bau, and Yonatan Belinkov. The quest for the right mediator: Surveying mechanistic interpretability for nlp through the lens of causal mediation analysis. *Computational Linguistics*, pp. 1–48, 02 2026. ISSN 0891-2017. doi: 10.1162/COLI.a.572. URL https://doi.org/10.1162/COLI.a.572.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/pdf?id=9XFSbDPmdW.

Yaniv Nikankin, Anja Reusch, Aaron Mueller, and Yonatan Belinkov. Arithmetic without algorithms: Language models solve math with a bag of heuristics, 2025. URL https://arxiv.org/abs/2410.21272.

Alan V. Oppenheim and Jae S. Lim. The importance of phase in signals. *Proceedings of the IEEE*, 69(5):529–541, 1981. doi: 10.1109/PROC.1981.12022.

Core Francisco Park, Andrew Lee, Ekdeep Singh Lubana, Yongyi Yang, Maya Okawa, Kento Nishi, Martin Wattenberg, and Hidenori Tanaka. ICLR: In-context learning of representations. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 53258–53284, 2025a. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/83fe5a77502e3d4cfab5960aed0ee6c3-Paper-Conference.pdf.

Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. The geometry of categorical and hierarchical concepts in large language models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025b. URL https://openreview.net/forum?id=bVTM2QKYuA.

Lucas Prieto, Edward Stevinson, Melih Barsbey, Tolga Birdal, and Pedro A. M. Mediano. Correlations in the data lead to semantically rich feature geometry under superposition. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=7akSRQS5Xh.

Philip Quirke and Fazl Barez. Understanding addition in transformers. In *The Twelfth International Conference on Learning Representations*, 2024.

Philip Quirke, Clement Neo, and Fazl Barez. Understanding addition and subtraction in transformers. *arXiv preprint arXiv:2402.02619*, 2025.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Raphaël Sarfati, Eric Bigelow, Daniel Wurgaft, Jack Merullo, Atticus Geiger, Owen Lewis, Tom McGrath, and Ekdeep Singh Lubana. The shape of beliefs: Geometry, dynamics, and interventions along representation manifolds of language models' posteriors, 2026. URL https://arxiv.org/abs/2602.02315.

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

Alessandro Stolfo, Yonatan Belinkov, and Mrinmaya Sachan. A mechanistic interpretation of arithmetic reasoning in language models using causal mediation analysis. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7035–7052, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.435. URL https://aclanthology.org/2023.emnlp-main.435/.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 12388–12401. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/92650b2e92217715fe312e6fa7b90d82-Paper.pdf.

Liuyuan Wen, Xun Zhu, Lihao Huang, Wenbin Li, and Yang Gao. The shape of addition: Geometric structures of arithmetic in large language models. *arXiv preprint arXiv:2606.03645*, 2026. URL https://arxiv.org/abs/2606.03645. Accepted at the International Conference on Machine Learning.

Zhengxuan Wu, Atticus Geiger, Thomas Icard, Christopher Potts, and Noah D. Goodman. Interpretability at scale: identifying causal mechanisms in alpaca. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Daniel Wurgaft, Can Rager, Matthew Kowal, Vasudev Shyam, Sheridan Feucht, Usha Bhalla, Tal Haklay, Eric Bigelow, Raphael Sarfati, Thomas McGrath, Owen Lewis, Jack Merullo, Noah Goodman, Thomas Fel, Atticus Geiger, and Ekdeep Singh Lubana. Manifold steering reveals the shared geometry of neural network representation and behavior, 2026. URL https://arxiv.org/abs/2605.05115.

Alper Yıldırım. The importance of phase in neural representations: An internal Oppenheim–Lim test of image classifiers. *arXiv preprint arXiv:2606.17037*, 2026. doi: 10.48550/arXiv.2606.17037. URL https://arxiv.org/abs/2606.17037.

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.

Tianyi Zhou, Deqing Fu, Mahdi Soltanolkotabi, Robin Jia, and Vatsal Sharan. Fone: Precise single-token number embeddings via fourier features. *arXiv preprint arXiv:2502.09741*, 2025.

A HARMONIC FAMILIES AND FITTING

A.1 PHASE SEPARATION AND INFORMATION CONTENT OF HARMONICS

With $n = bpq + pd + r$ as in Section 2, changing only the digit d changes the phase of the harmonic with period bp/k by

$$\Delta\theta_{bp/k} = \frac{2\pi k\, p\, \Delta d}{bp} = \frac{2\pi k\, \Delta d}{b} \pmod{2\pi}, \tag{5}$$

which does not depend on p: the ideal separation between adjacent digit candidates is $2\pi k/b$ at every position, and the lower suffix r adds the common offset $2\pi kr/(bp)$ to all candidates. The phase of harmonic k is $2\pi kd/b$ up to this offset, so two digits d, d' give the same phase exactly when $k(d - d') \equiv 0 \pmod{b}$, that is, when $d \equiv d' \pmod{b/\gcd(k, b)}$. A single harmonic therefore determines d modulo $b/\gcd(k, b)$, a set of harmonics determines it modulo the least common multiple of these moduli, and the joint ideal phase labels uniquely identify the digit exactly when that least common multiple equals b. This is an information criterion and does not by itself guarantee successful steering; the base-7 single-harmonic results are the clearest example. In base 10, $\{T_2, T_5\}$ gives $\mathrm{lcm}(2, 5) = 10$; in base 8, $\{T_2, T_4\}$ gives $\mathrm{lcm}(2, 4) = 4 < 8$, leaving the pair $\{t, t + 4\}$ unresolved; in base 7, every harmonic alone gives modulus 7. Figure 6 illustrates both statements for base 10. Our interventions transplant measured donor directions rather than sweeping continuous angles, so the figure shows ideal phase-label groupings, not measured decision boundaries; the cosine form of the readout (Appendix B.5) makes smooth boundaries plausible but does not establish their location or width.

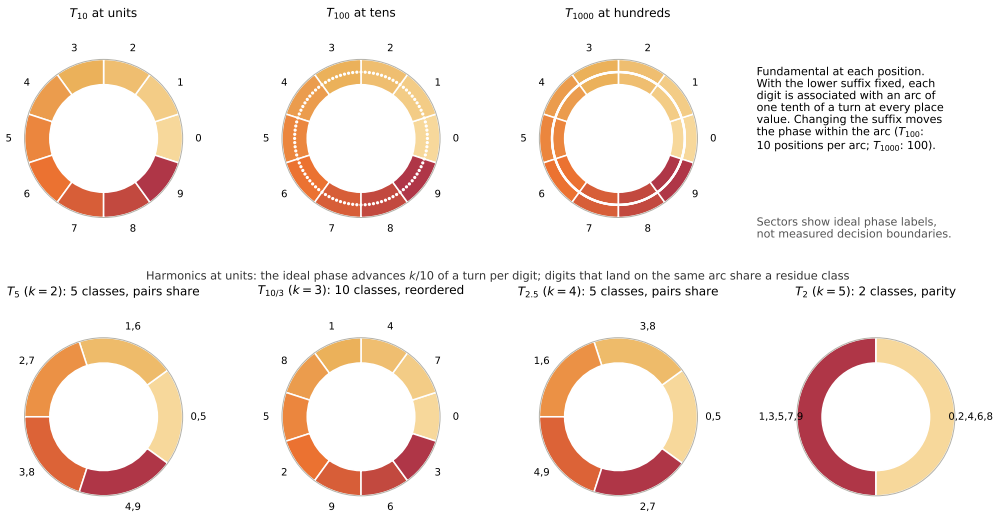


Figure 6: Ideal Fourier phase geometry in base 10. Top: with the lower suffix fixed, the fundamental at units, tens, and hundreds has ten digit-associated arcs of one tenth of a turn regardless of place value, and changing the suffix moves the phase within an arc. Bottom: the units harmonics, whose phase advances $k/10$ of a turn per digit, so digits that land on the same arc share a residue class. Shaded sectors are schematic label groupings, not measured model decision boundaries.

A.2 FITTING ORDERS AND AUDIT OF THE EXCLUSIVE CONSTRUCTION

Numerical rank. After the protected components are removed (the projection is applied twice for numerical stability), orth keeps the left singular vectors of the residualized coefficient matrix whose singular values exceed $\max(10^{-10}, 10^{-6}\,\sigma_{\max})$, where $\sigma_{\max}$ is the largest singular value of the raw, unprojected matrix. This criterion is used by the Qwen factorials and layer sweep and by the centered radix factorials. The Pythia replication instead requests a fixed rank and checks that every retained singular value exceeds 10^{-6} times the largest singular value of the residualized matrix, raising an error if the requested rank is unavailable.

In the residue-first order, used for the decimal residue matrices, the harmonic factorial, and the radix gate matrices, the base-10 units planes are fitted as Q_2, then Q_5 against Q_2, then Q_{10} against $[Q_2, Q_5]$. At tens and hundreds, we first protect the lower-place fundamental (T_{10} or T_{100}), then fit the two residue harmonics, then a rank-one within-period magnitude direction (the scaled sum modulo M), and finally the fundamental (T_{100} or T_{1000}). Additional harmonics ($T_{M/3}$, $T_{M/4}$ for $M = 10, 100, 1000$) are each residualized against this complete span but not against each other. The harmonic factorial uses the same construction and also protects a magnitude direction at units. In the fundamental-first order, used for the two-position experiments in bases 7 and 8, the fundamental T_{bp} is fitted first (after a common lower-place T_b plane at the upper position) and each extra harmonic is residualized against it; no magnitude direction is used. Qwen's exclusive fundamental planes retain 0.69–0.73 of their fitted norm after protection.

The exclusive construction could produce a null effect if protection removed most of the fitted T_{10} plane. We audit this possibility before orthonormalization. Let $Q_{2,5}$ be the protected factor basis, W_{10} the raw ridge coefficient matrix, and $Q_{10}^{\text{raw}} = \text{orth}(W_{10})$. We measure raw principal overlap and the retained fraction as

$$o_{\text{raw}} = \left\| Q_{2,5}^{\top} Q_{10}^{\text{raw}} \right\|_2, \qquad \rho_{10} = \frac{\left\| (I - Q_{2,5} Q_{2,5}^{\top}) W_{10} \right\|_F}{\| W_{10} \|_F}. \tag{6}$$

In Pythia at layer 26, the primary fit gives $o_{\text{raw}} = 0.180$ and $\rho_{10} = 0.989$. After residualization, the maximum principal overlap is 7.4×10^{-16}. A nearest-centroid decoder fitted on the exclusive coordinates reaches 0.439 accuracy on 1,500 held-out sums, compared with 0.1 chance. An independent expanded audit gives $o_{\text{raw}} = 0.148$, $\rho_{10} = 0.994$, and exclusive overlap 1.5×10^{-16}. The null intervention therefore acts on an intact, informative independent plane rather than a small nondecodable remainder.

We also reverse the protection order for a downstream-propagation test. Protecting Q_{10}^{raw} first and residualizing the factor directions against it retains 0.995 of their fitted Frobenius norm at layer 26, while their maximum overlap with the raw clock is 1.8×10^{-16}. These factor edits leave the raw T_{10} coordinate unchanged at the intervention site and remain causally effective (29.9 pp over the keep control on 1,500 pairs). Measured with layer-specific T_{10} bases after blocks 26–31, the edited activations stay aligned with the clean source T_{10} coordinates (cosine 0.97 at layer 31) and gain only a modest relative alignment with the donor (paired change in target-minus-source cosine $+0.049$, 95% interval $[0.039, 0.060]$). Residue edits therefore change the output while the raw T_{10} coordinates remain predominantly source-aligned, and steering the exclusive T_{10} plane alone was insufficient in the tested conditions. Together, these audits establish that the null intervention targets a well-retained independent T_{10} plane.

B Qwen Decimal Experiments

All Qwen experiments use Qwen3.5-9B-Base on six-digit addition with the answer written digit by digit. Unless stated otherwise, interventions act on the residual stream after block 30 at the token immediately before the digit being predicted, subspaces are fitted on 4,000 baseline-correct examples and evaluated on 800 baseline-correct examples from disjoint sums. Intervals are paired and unadjusted for multiplicity; the residue matrices use normal-approximation intervals (mean paired difference ± 1.96 standard errors), the harmonic additions and the factorial use 2,000-draw paired bootstraps, and the carry analysis uses cluster bootstraps over pair groups.

B.1 Fourier Screen

The screen is a descriptive check that the periodic structure assumed by the harmonic family is present in the residual stream before any subspace is fitted. Following Kantamneni & Tegmark (2025), we use four complete grids of sums $1000q + s$ with $q \in \{317, 491, 673, 851\}$ and $s \in \{0, \ldots, 999\}$, keeping an example only when greedy generation produces the full six-digit answer, which leaves 4,000 examples. For each grid we compute power spectra of residual states and of block updates $\Delta h_\ell = h_{\ell+1} - h_\ell$ along two axes. The suffix axis treats the remaining suffix at each generation stage as the transformed variable (length 1,000 before the hundreds digit, 100 before the tens digit, 10 before the units digit). The digit axis reshapes the suffix into hundreds, tens, and units coordinates and transforms along the next digit with the others fixed, so that its bins separate

parity ($j = 5$), mod-5 harmonics ($j = 2, 4$), and full-order harmonics ($j = 1, 3$). Power is averaged over residual dimensions and grids after mean subtraction, and a 64-permutation digit-order null is computed per grid and depth.

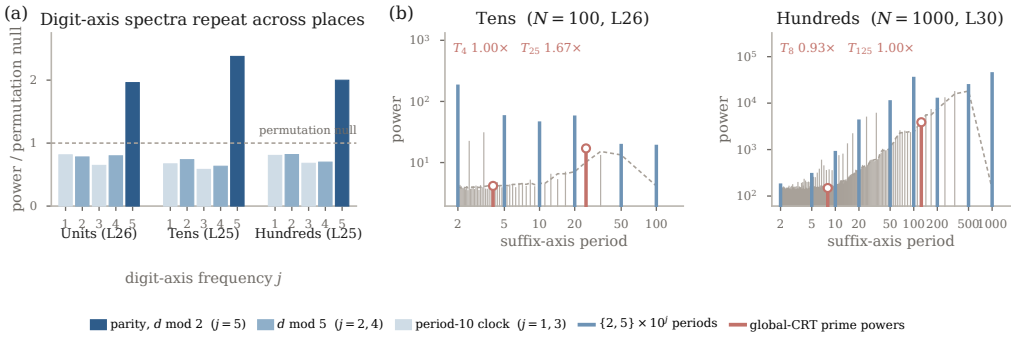


Figure 7: Fourier screen of centered block updates. (a) Along the digit axis the same parity-dominated pattern appears at units, tens, and hundreds. (b) Along the suffix axis, the prime-power coordinates a single global CRT would require (T_4, T_{25} or T_8, T_{125}) form no joint signature.

Figure 7 shows the depths with the largest centered candidate power fraction for each screen (digit axis L26 for units, L25 for tens and hundreds; suffix axis L26 for tens, L30 for hundreds). Parity power is about twice the permutation null at every position, whereas the mod-5 and full-order groups are not enriched, so the spectrum alone would not have predicted the causal importance of T_5 and T_{10}. The suffix-axis panel addresses a different hypothesis, a single global CRT over the whole suffix, and finds no support for it. Spectra describe periodic structure; the interventions in the main text test causal use.

B.2 Held-Out Fourier Probes

The probes reverse the direction of the fit used for interventions, mapping activations to Fourier labels, and measure linear decodability rather than causal use. For every period $T \in \{2, \ldots, 1000\}$ we fit affine ridge probes ($\alpha = 1$) from the residual state after each block to $\sin(2\pi n/T)$ and $\cos(2\pi n/T)$, with n the full sum, at the positions before the hundreds, tens, and units digits. A fixed split (seed 42) holds out 20% of each magnitude grid, giving 3,200 fitting and 800 evaluation examples. We report the mean held-out R^2 of the two targets (cosine only for $T = 2$); negative values are kept in the results and clipped to zero only in the heatmaps.

Two observations matter for the main text. First, decodability and causal use come apart. At the tens position T_2 and T_5 remain decodable, yet adding them to the tens-place pair gives no clear benefit (Appendix B.3). Second, the strongest readability at the hundreds position is selective. Of the 999 periods scanned, only twelve reach held-out $R^2 \geq 0.95$ at any layer, namely $200, 250, 364, 400, 445, 500, 576, 809, 973, 978, 989$, and 1000, all peaking at layer 31; the nominated hundreds family $T_{200}, T_{500}, T_{1000}$ is among them ($R^2 = 0.954, 0.983, 0.991$). The threshold is descriptive, and because fitting and evaluation examples share magnitude grids the R^2 values measure within-grid generalization.

B.3 Residue Matrices and Lower-Place Control

At each Qwen position we transplant one residue harmonic, the other, or both, while the fundamental is kept, deleted, or aligned, and compare each edit with a control that does only the fundamental operation (Figure 9; Table 1). With the fundamental kept at its conflicting source value, single residue edits rarely change the prediction, but the two together do, and aligning the fundamental roughly doubles the joint effect; every joint edit exceeds its norm-matched shuffled control. Each cell is the paired change in target following, in percentage points, of a residue edit combined with a fundamental operation, relative to the same fundamental operation alone, with the change in the

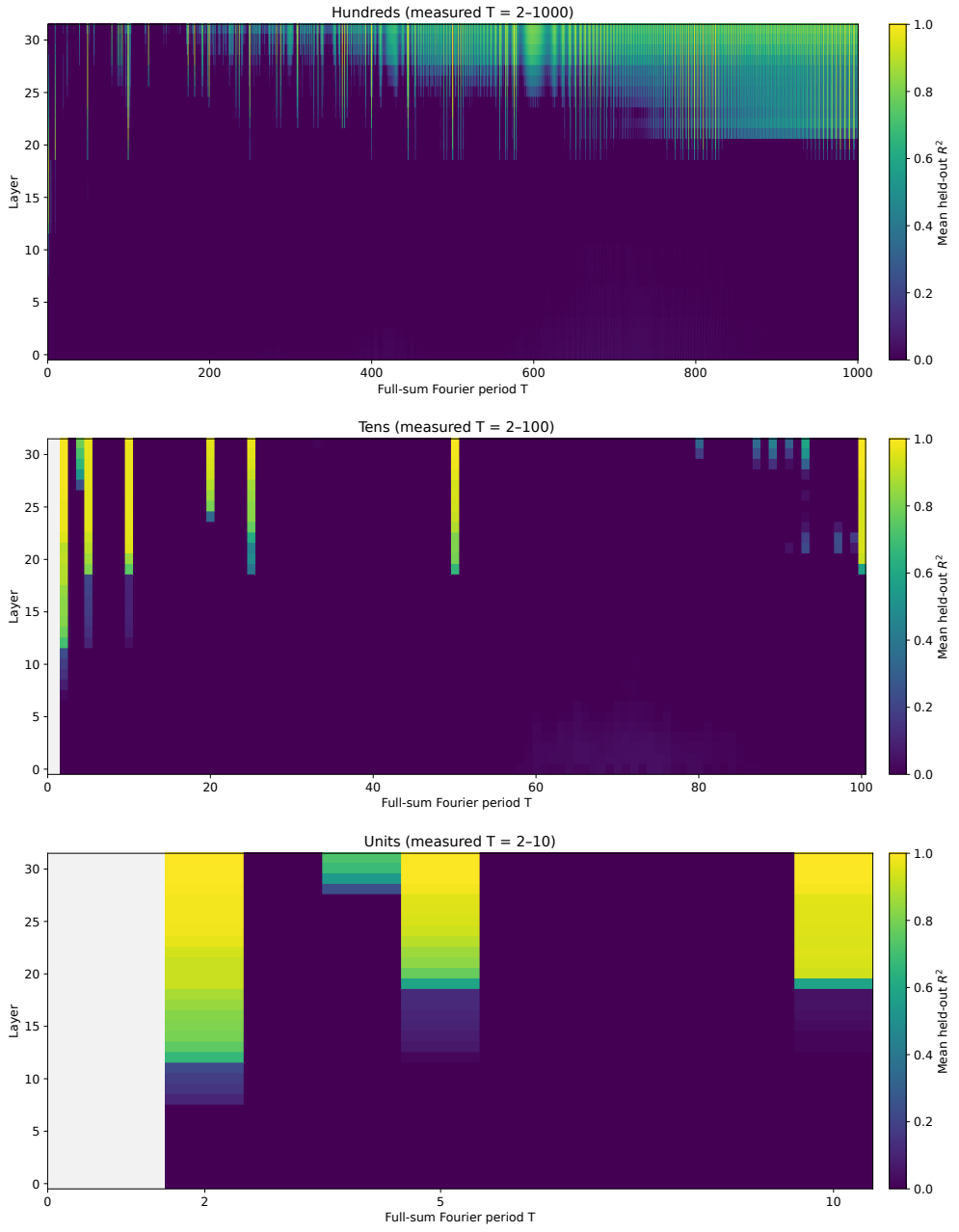


Figure 8: Held-out Fourier-probe R^2 before the hundreds, tens, and units digits (top to bottom), for periods 2–1000, 2–100, and 2–10 across layers 0–31. Periods are not rescaled by position; the colour scale is clipped to $[0, 1]$.

target-minus-source margin in parentheses. The "sum of singles" rows add the two single-edit rows and are not separate interventions; the joint margin changes differ from these sums by at most 0.40 nats. The bottom block compares the joint edit with norm-matched shuffled phases.

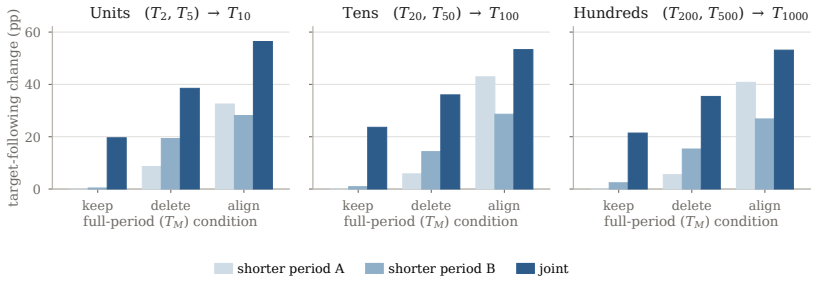


Figure 9: Residue steering in Qwen at layer 30. Bars are gains in target following over a control that only manipulates the fundamental ($n = 800$). With the fundamental kept, single residue edits do little while the joint edit steers at every position.

Table 1: Residue steering at layer 30 by decimal position. Entries are paired target-following changes in pp (margin changes in nats); $n = 800$ per cell.

Position	Periods	Residue edit	Full keep	Full mean-delete	Full align
Units	$(T_2, T_5) \to T_{10}$	joint	$+19.6$ $(+12.44)$	$+38.5$ $(+13.42)$	$+56.4$ $(+12.62)$
Tens	$(T_{20}, T_{50}) \to T_{100}$	joint	$+23.6$ $(+12.08)$	$+36.0$ $(+13.24)$	$+53.3$ $(+12.24)$
Hundreds	$(T_{200}, T_{500}) \to T_{1000}$	joint	$+21.4$ $(+10.94)$	$+35.4$ $(+12.16)$	$+53.1$ $(+11.03)$
Units	$(T_2, T_5) \to T_{10}$	T_2 only	$+0.0$ $(+5.56)$	$+8.6$ $(+5.56)$	$+32.5$ $(+5.37)$
		T_5 only	$+0.4$ $(+7.13)$	$+19.3$ $(+7.53)$	$+28.1$ $(+7.06)$
Tens	$(T_{20}, T_{50}) \to T_{100}$	T_{20} only	$+0.0$ $(+5.13)$	$+5.8$ $(+5.23)$	$+42.9$ $(+4.93)$
		T_{50} only	$+0.9$ $(+7.15)$	$+14.3$ $(+7.61)$	$+28.6$ $(+7.05)$
Hundreds	$(T_{200}, T_{500}) \to T_{1000}$	T_{200} only	$+0.0$ $(+4.55)$	$+5.5$ $(+4.58)$	$+40.8$ $(+4.28)$
		T_{500} only	$+2.4$ $(+6.67)$	$+15.3$ $(+7.18)$	$+26.8$ $(+6.53)$
Units		sum of singles	$+0.4$ $(+12.69)$	$+27.9$ $(+13.09)$	$+60.6$ $(+12.43)$
Tens		sum of singles	$+0.9$ $(+12.28)$	$+20.1$ $(+12.84)$	$+71.5$ $(+11.98)$
Hundreds		sum of singles	$+2.4$ $(+11.22)$	$+20.8$ $(+11.76)$	$+67.6$ $(+10.81)$
Units	joint vs. shuffled		$+19.0$ $(+4.57)$	$+33.0$ $(+4.90)$	$+36.8$ $(+4.71)$
Tens	joint vs. shuffled		$+21.8$ $(+5.09)$	$+29.0$ $(+5.53)$	$+34.0$ $(+5.15)$
Hundreds	joint vs. shuffled		$+20.0$ $(+4.85)$	$+28.6$ $(+5.31)$	$+35.0$ $(+4.84)$

Cross-donor chimeras. The residue and chimera edits at layer 30 operate on plane coordinates centered at the fitting-set mean. Joint steering alone does not show that the two residues are supplied independently. The cross-donor chimeras do (Figure 10a): the T_2-family phase is taken from one donor and the T_5-family phase from another, chosen so that the digit the two residues jointly specify appears in neither donor, and the model produces that digit above norm-matched shuffled phases by 10.0, 16.3, and 13.0 pp at units, tens, and hundreds with the fundamental kept, and by 34.8, 35.3, and 32.3 pp after deleting it ($n = 800$ per condition).

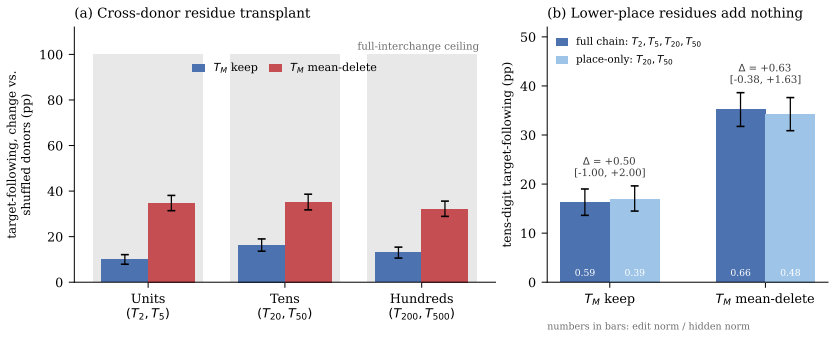


Figure 10: Cross-donor chimeras in Qwen at layer 30 ($n = 800$). (a) Gain over shuffled phases when the two residue phases come from different donors: the model produces a digit present in neither. (b) Lower-place residues give no clear incremental benefit at the tens position; numbers inside bars give edit norm relative to hidden-state norm.

Lower-place control. The lower-place control asks whether the units residues help at the tens position. Adding T_2, T_5 to the tens pair T_{20}, T_{50} changes target following by 0.5 pp $[-1.0, 2.0]$ with the fundamental kept and 0.6 pp $[-0.4, 1.6]$ after mean deletion, despite the larger edit norm (Figure 10b).

B.4 Harmonic Additions

Figure 11 shows the result and the tables give the numbers. Starting from the aligned residue-and-fundamental edit, which is retained in every arm, we add the third harmonic ($T_{M/3}$), the fourth ($T_{M/4}$), or both as a union-radius transplant, for $M = 10, 100, 1000$. Each extra plane is residualized against the complete protected span but not against the other extra, so the unions have ranks 3, 4, and 4 at units, tens, and hundreds. The wrong-phase control adds both harmonics from a donor that matches prefix, suffix, and carry but differs from the source and target digits and has the wrong phase for both harmonics, with the extra edit norm matched per example (rescaling capped at 10). Rates use the examples that admit every control within a position (0, 3, and 2 exclusions).

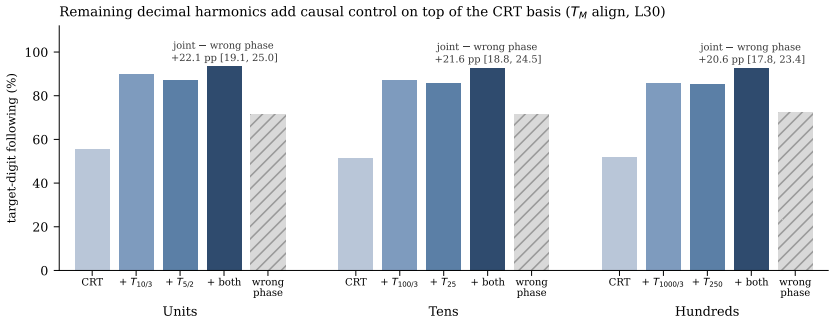


Figure 11: Completing the family in Qwen at layer 30. Adding $T_{10/3}$ and $T_{2.5}$ (and their scaled counterparts) to the aligned residue edit raises target following above 90%; hatched bars are wrong-phase controls that keep the residue-and-fundamental edit and replace only the added harmonics; annotations are paired differences with 95% intervals.

Table 2: Target following (%) after adding harmonics to the aligned residue edit; "Wrong" adds both harmonics with a wrong-phase donor.

Position	(A, B)	n	Residue edit	$+A$	$+B$	$+A, B$	Wrong
Units	$(10/3, 2.5)$	800	55.50	89.88	87.00	**93.50**	71.38
Tens	$(100/3, 25)$	797	51.32	87.08	85.95	**92.85**	71.27
Hundreds	$(1000/3, 250)$	798	51.75	85.84	85.34	**92.73**	72.18

Table 3: Paired gains of the joint addition in pp [95% interval], on the samples of Table 2.

Position	Joint $-$ $(+A)$	Joint $-$ $(+B)$	Joint $-$ Wrong
Units	$+3.63$ $[2.38, 5.00]$	$+6.50$ $[4.88, 8.25]$	$+22.13$ $[19.13, 25.00]$
Tens	$+5.77$ $[4.27, 7.40]$	$+6.90$ $[5.14, 8.78]$	$+21.58$ $[18.82, 24.47]$
Hundreds	$+6.89$ $[5.14, 8.78]$	$+7.39$ $[5.51, 9.27]$	$+20.55$ $[17.79, 23.43]$

Because the joint transplant preserves the union radius rather than each plane's radius, its advantage over the single additions is not a test of interaction; the factorial below is. Among cases where both single additions fail, the joint edit succeeds in 13 of 64, 17 of 72, and 20 of 74 at units, tens, and hundreds, while losing 1, 2, and 4 cases where a single addition succeeds.

Table 4 adds the shorter periods of the lower place instead. With the lower suffix fixed these periods have the same ideal phase for every candidate digit, so their null donors have a wrong digit but not necessarily a wrong phase; the table therefore measures only the incremental steering contribution of these planes, which is negligible.

Table 4: Shorter-period additions at tens and hundreds, paired gains in pp [95% interval].

Position	Period	n	Correct $-$ residue edit	Correct $-$ Null
Tens	$T_{2.5}$	797	$+0.13$ $[-0.25, 0.63]$	$+0.13$ $[0.00, 0.38]$
Tens	$T_{10/3}$	800	-0.13 $[-0.50, 0.25]$	$+0.13$ $[0.00, 0.38]$
Hundreds	T_{25}	799	$+0.25$ $[-0.25, 0.75]$	$+0.00$ $[-0.63, 0.63]$
Hundreds	$T_{100/3}$	797	$+0.50$ $[0.00, 1.13]$	$+0.38$ $[0.00, 0.88]$

B.5 Full Harmonic Factorial

These are the statistics behind Section 4.1. We compute the single-harmonic phase transplant for each of the five family members at layer 30, protecting a magnitude direction as in the residue-first order (Appendix A.2), and apply every nonempty subset of these fixed edits by summing them, giving 31 arms on the same 800 cases per position. Summing fixed edits preserves the protected coordinates but does not re-normalize the joint edit, so the full-family rate differs slightly from the union-radius transplant of Appendix B.4 (92.9% against 93.5% at units). Each arm records the ten digit logits; target following in the main text is the full-vocabulary argmax, while the additive-prediction agreement below is computed among the ten digit logits. Interactions are mean baseline-referenced Möbius contrasts over the 2^5 design, and wrong-phase nulls accompany every arm on the examples that admit one. Figure 12 shows the layer-30 result across positions.

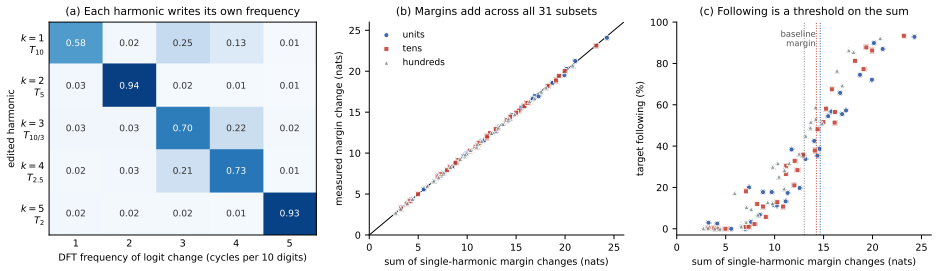


Figure 12: Harmonics add in logit space at layer 30 (Qwen, $n = 800$ per position). (a) Editing one plane changes the digit logits mostly in that plane's own shape (rows: plane edited; columns: shape of the change). (b) The mean margin change of any subset of harmonics is closely predicted by the sum of its members' mean changes. (c) Target following rises with that sum and crosses 50% near the baseline margin (dotted).

Table 5: Additivity of harmonic contributions at layer 30. The additive model predicts a subset's margin change by the sum of its members' single-edit changes.

	Units	Tens	Hundreds
R^2 across the 31 subset means	0.9992	0.9998	0.9997
RMSE across subset means (nats)	0.15	0.07	0.08
R^2 across individual cases	0.976	0.970	0.967
RMSE across cases (nats)	1.11	1.16	1.11
Full family, measured / predicted (nats)	24.08 / 24.28	23.13 / 23.20	20.70 / 20.78
Range of interaction contrasts (nats)	$[-0.21, 0.15]$	$[-0.14, 0.09]$	$[-0.18, 0.09]$
Argmax agreement, all arms (%)	94.1	94.4	94.0
Argmax agreement, full family (%)	97.4	98.0	97.5
Target following, best single (%)	2.9	0.9	1.8
Target following, residue pair (%)	19.8	21.0	19.5
Target following, full family (%)	92.9	93.4	92.3
Spearman, additive prediction vs. following	0.96	0.97	0.97

Table 6: Single-harmonic main effects on the target-minus-source margin (nats) and spectral purity of the per-digit logit change. Purity is the fraction of the centered ten-logit change's power at the harmonic's own frequency, averaged over the 16 backgrounds; the cosine R^2 is that of a target-phase cosine minus a source-phase cosine fitted per harmonic.

Harmonic	k	Main effect (nats)			Purity / cosine R^2		
		units	tens	hundreds	units	tens	hundreds
T_{10} family	1	4.34	4.11	3.88	.56 / .32	.58 / .32	.61 / .34
T_5 family	2	6.99	7.03	6.57	.92 / .84	.95 / .89	.95 / .86
$T_{10/3}$ family	3	3.24	3.24	2.72	.64 / .41	.74 / .50	.72 / .49
$T_{2.5}$ family	4	4.14	3.83	3.20	.70 / .53	.74 / .52	.75 / .49
T_2 family	5	5.57	4.98	4.41	.92 / .76	.93 / .87	.93 / .86

Some interaction contrasts are distinguishable from zero (the units $T_2 \times T_5$ contrast is -0.10 nats with 95% interval $[-0.14, -0.05]$), but all are an order of magnitude below the main effects in Table 6. The residue pair has the largest mean margin gain among pairs, though not always the highest following (at tens, $T_{50} + T_{100}$ reaches 26.5%). The $T_{10/3}$ and $T_{2.5}$ planes, each residualized against the same protected span, share a direction (principal cosine 1.00 at units, 0.68 and 0.71 at tens and hundreds); their pairwise interaction is nevertheless small (-0.03 nats at units). The shared direction is induced by the protected span, not by the model. On the balanced fitting set (400 examples per units digit) the nine nonconstant decimal Fourier labels are orthogonal to floating-point precision, so fitting the extras jointly reproduces the separate ridge fits to roundoff. The raw

$T_{10/3}$ and $T_{2.5}$ coefficient matrices have joint rank four (singular values 26.4, 23.5, 21.8, 21.1), and removing the T_2, T_5, and fitted T_{10} spans leaves rank four (smallest singular value 13.0); only adding the protected magnitude direction reduces the rank to three (smallest singular value 10^{-13}). At units the protected magnitude is a centered scalar function of the digit, which the complete nonconstant Fourier family spans, so protecting it allocates one direction from the same digit-dependent space and the two residual extras lose it between them. The five harmonics are therefore distinct in the raw fits, and the overlap in the final units planes is a cost of the protection order; it may contribute to the mixing between frequencies 3 and 4, which this diagnostic does not test. At tens and hundreds the overlap is partial rather than exact. This is a projection diagnostic on the saved fits, not a rerun with a different protection order.

Layer sweep. The units factorial is repeated with the edit at each layer from 21 to 31, with planes fitted at that layer, the same protection recipe, and the same 800 cases; the layer-30 arrays reproduce the original experiment exactly. Table 8 gives the numbers behind Figure 4. Intervals are paired case bootstraps (2,000 draws, seed 20260922) without multiplicity adjustment; the R^2 column is recomputed over the 31 subset means in each draw without refitting a slope or intercept. At layers 21 and 22 the edits barely move the margin, so agreement there is not evidence of additivity. Digit-only argmax agreement across the 26 multi-harmonic arms is 88.1% at layer 28, 86.9% at 29, 92.9% at 30, and 99.1% at 31 (a different denominator from the 31-arm figure in Table 5). The positive interactions before layer 30 are not specific to the residue pair: at layer 28, $T_5 + T_{10}$ contributes 1.2 nats and $T_2 + T_{2.5}$ 1.3 nats beyond their singles. The unedited baseline margin is -14.7 nats (s.d. 2.3) on these cases at every layer. Predicting each subset's measured gain by a quadratic function of the sum of its members' single-edit gains, fitted through the origin, explains 99.3% of the variance across the 31 subsets at layer 28 and 99.7% at layer 29 (additive model: 51% and 59%); at layer 30 the fitted quadratic coefficient is approximately zero. The Möbius contrasts show the same signature, positive at order two and negative at order three (summing to -2.6 nats at layer 28), as a convex and then saturating function of one variable produces. Pair-specific residuals from the quadratic are within 0.56 nats, negative for the residue pair (-0.55 at layer 28), and largest in the positive direction for $T_{10/3} + T_{2.5}$ at layer 29 ($+0.45$), the pair whose fitted planes share a direction. The one-at-a-time sequence in Figure 4d–f is one of 120 orders; the excess over the additive expectation at the full family is the same whichever order is used.

Centering control. The sweep uses the uncentered fits of the three-position factorial so that its layer-30 arrays coincide with Tables 5 and 6. Repeating it with fit-only centering, as in the radix runs, leaves the picture unchanged (Table 7). The family's gain exceeds the summed singles by 1.69 and 1.65 times at layers 28 and 29 and matches them at layers 30 and 31 (ratios 0.996 and 0.992, the layer-31 difference of -0.22 nats distinguishable from zero), the $T_2 + T_5$ interaction is $+1.42$ nats at layer 28 and positive in 99% of cases, and full-family following moves by at most 8 points (layer 29, 62% to 70%) and by less than a point at layers 28, 30, and 31.

Table 7: Centering control for the units layer sweep. The factorial of Table 8 was repeated with every plane centered at its ridge intercept, estimated from fitting rows only, on the same 800 cases and bases. Following is full-family target following (%); gains are mean margin changes (nats) in the centered run, with a paired case-bootstrap 95% interval on the difference.

Layer	Following, original	Following, centered	Sum of singles	Full family	Difference [95% CI]
21	0.0	0.0	0.50	0.66	+0.16 [+0.12, +0.19]
22	0.0	0.0	0.66	0.80	+0.15 [+0.11, +0.18]
23	0.5	0.8	2.24	3.63	+1.40 [+1.30, +1.49]
24	2.6	2.8	3.03	5.72	+2.69 [+2.56, +2.82]
25	5.1	5.9	3.42	6.76	+3.34 [+3.20, +3.50]
26	6.6	9.0	3.65	7.32	+3.67 [+3.51, +3.84]
27	7.0	11.0	3.83	7.58	+3.75 [+3.58, +3.92]
28	61.8	62.4	9.46	15.96	+6.50 [+6.25, +6.75]
29	62.1	70.4	10.79	17.81	+7.02 [+6.77, +7.27]
30	92.9	93.4	24.66	24.58	-0.09 [-0.25, +0.07]
31	94.8	95.1	27.23	27.00	-0.22 [-0.27, -0.18]

Table 8: Units factorial by layer of the edit. Gains are mean target-minus-source margin changes (nats); the interaction is the baseline-referenced contrast for $T_2 + T_5$ with its 95% interval.

Layer	Blocks after	Following (%)	Sum of singles	Full family	$T_2 \times T_5$	R^2, subset means
21	10	0.0	0.26	0.35	0.00 $[0.00, 0.00]$	0.85
22	9	0.0	0.32	0.38	0.00 $[0.00, 0.00]$	0.95
23	8	0.5	1.51	2.21	0.15 $[0.12, 0.17]$	0.80
24	7	2.6	2.46	4.01	0.43 $[0.37, 0.48]$	0.69
25	6	5.1	3.36	6.19	0.81 $[0.75, 0.87]$	0.56
26	5	6.6	3.37	6.47	0.75 $[0.70, 0.79]$	0.51
27	4	7.0	3.62	6.68	0.58 $[0.54, 0.63]$	0.54
28	3	61.8	9.55	15.84	1.40 $[1.34, 1.46]$	0.51
29	2	62.1	10.79	17.15	1.44 $[1.35, 1.53]$	0.59
30	1	92.9	24.28	24.08	-0.10 $[-0.14, -0.05]$	0.9992
31	0	94.8	26.59	26.38	0.00 $[-0.01, 0.01]$	0.9998

B.6 Mismatched-Period Controls

This control asks whether a rank-two plane fitted to a period unrelated to the decimal family can steer a decimal digit equally well. At each position we replace the T_5-family plane by one fitted to T_7, T_{70}, or T_{700}, delete the fundamental in every arm, match the edit norm per example to the native edit, and evaluate steering toward the specified decimal target digit; null donors carry a different mod-7 residue. We then repeat the control after residualizing the mod-7 plane against the complete decimal span through the edited place (T_2, T_5, T_{10} at units, additionally T_{20}, T_{50}, T_{100} at tens, and $T_{200}, T_{500}, T_{1000}$ at hundreds), again with matched norms. Table 9 shows that most of the mod-7 plane's fitted energy lies inside the decimal span and that its steering effect largely disappears once that overlap is removed. One hundreds example was excluded ($n = 888/888/887$).

Table 9: Mod-7 control planes before and after removing their overlap with the decimal span. "Energy removed" is the fraction of the mod-7 ridge energy inside the decimal span; the last column is the paired decrease in the correct-minus-null exact-target effect.

Position	Period	Energy removed	Held-out R^2 (before/after)	Exact target (before/after)	Effect decrease [95% CI]
Units	T_7	84.36%	.441/.065	12.95/3.83%	5.52 [3.56, 7.47] pp
Tens	T_{70}	82.43%	.437/.074	11.26/3.27%	6.19 [4.35, 8.04] pp
Hundreds	T_{700}	83.61%	.434/.073	11.84/3.61%	5.75 [3.82, 7.68] pp

B.7 Carry Phase Readout

These are the details behind Section 4.3. Phase-readout maps for each period are fitted on 1,000 carry-free examples and calibrated on a separate 1,000, and phases are read as the calibrated $\mathrm{atan2}$ of the plane coordinates, with a separately fitted basis at every layer. Pairs are matched either on the final sum (1,000 pairs) or on the local tens sum before the incoming carry (900 pairs with an unchanged answer prefix), always sharing the lower remainder; pairs at the tens-9-to-0 boundary form a separate stratum and are excluded here. Differences are carry minus no-carry, wrapped circularly, and uncertainty comes from a cluster bootstrap over the 20 pair groups. Raw planes are used throughout; the T_{20} transition is robust to using orthogonalized planes instead, whereas the onset times of some other periods are not.

For the vector comparison in Figure 5, both members of a pair are rotated by their shared ideal local-sum phase, so that this phase lies on the x axis, and scaled by the no-carry median radius of that layer. In this frame the no-carry partner sits ideally at $(1, 0)$ and a one-step carry correction is the fixed vector $(\cos\delta - 1, \sin\delta)$ with $\delta = 180°, 72°, 36°, 108°, 144°$ for $T_{20}, T_{50}, T_{100}, T_{100/3}, T_{25}$. At layer 30 the angular error between measured and expected vectors is $7°$ for T_{20} and $0.4°$ to $4.7°$ for the other planes, and the coefficient along the expected vector is 0.84–0.93. These are paired representation differences, not isolated update vectors, and they do not by themselves show that all planes are corrected by one mechanism. Figure 13 gives the circular-mean phase differences for both pair designs, which agree, and the component perpendicular to the expected vector, which is

small except for a transient excursion of the parity plane at layers 25–27; Table 10 lists the layer-30 values.

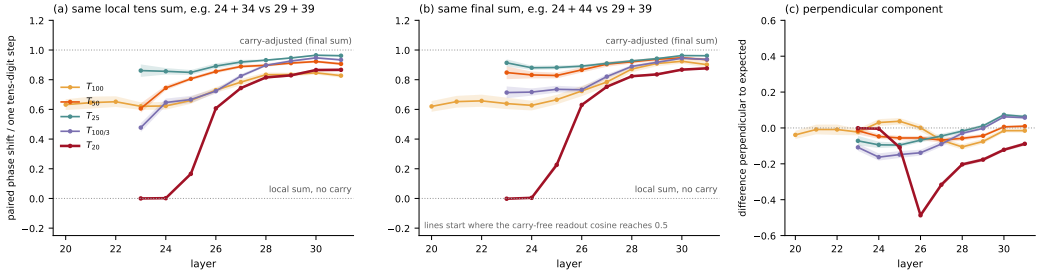


Figure 13: Carry phase differences by layer. Circular-mean paired phase difference $\Delta\theta$ in units of one tens-digit step δ: (a) $\Delta\theta/\delta$ for pairs matched on the local tens sum, (b) $1 + \Delta\theta/\delta$ for pairs matched on the final sum, on a consistent circular branch, so that the carry-adjusted prediction is 1 in both. (c) Component of the mean difference perpendicular to the expected correction. Bands are cluster-bootstrap intervals; lines start where the carry-free readout cosine reaches 0.5.

Table 10: Carry-minus-no-carry circular mean phase differences at layer 30 (degrees; raw planes).

Period	Same result	Same local sum	One-digit increment
T_{20}	-23.8	155.7	180
T_{50}	-3.8	66.4	72
T_{100}	-2.7	30.5	36
$T_{100/3}$	-6.0	102.3	108
T_{25}	-5.4	138.9	144

C Radix Experiments

The four radix models share one architecture, recipe, seed, training step (7,680), and digit-level tokenizer, and differ only in the numeral system of their training corpus. Interventions act on the residual stream at the token before the digit being predicted, with subspaces fitted on 4,000 baseline-correct rows (3,999 at sevens) and evaluated on 1,500 baseline-correct cases from disjoint sums. Primary rates are reported at the final block (L7); the native factorials are swept over L0–L7, and the earlier persistent all-layer edits are stated where used. Intervals are paired and unadjusted for multiplicity; the gate matrices use normal-approximation intervals and the two-position experiments 2,000-draw paired bootstraps.

C.1 Training Corpora and Configuration

The decimal corpus is a frozen ClimbMix corpus with 9,786,141 arithmetic equations interleaved into it (Table 11). Binary addition dominates by design, and five further operation families keep the arithmetic data from reducing to a single prompt template. Equations are shuffled with a deterministic Feistel permutation (seed 42), grouped into documents of 16 equations, and inserted at approximately equal intervals into the first 150 of 170 shards (4,078 documents in shards 0–83, 4,077 in shards 84–149); shards 150–169 contain only web data.

Table 11: Arithmetic equations interleaved into the ClimbMix training corpus. Ordered pairs $a + b$ and $b + a$ are distinct records, and every binary-addition pair receives two exposures.

Family	Construction	Equations	Share
$a + b = c$	$a, b \in \{0, \ldots, 2000\}$, all ordered pairs, two exposures	8,008,002	81.83%
$a + b + c = d$	$a, b, c \in \{0, \ldots, 200\}$, sampled without replacement	1,000,000	10.22%
$a - b = c$	$0 \leq b \leq a \leq 2000$, sampled without replacement	500,000	5.11%
$a \times b = c$	$a \in \{0, \ldots, 2000\}$ and $b \in \{1, \ldots, 20\}$	40,020	0.41%
$c/b = a$	Exact inverses of the multiplication examples	40,020	0.41%
$a \bmod m = r$	$a \in \{0, \ldots, 2000\}$ and $m \in \{2, \ldots, 100\}$	198,099	2.02%
Total		9,786,141	100.00%

The base-6, base-7, and base-8 corpora are derived from the frozen decimal corpus, not from one another. Every maximal ASCII digit string is parsed as a decimal integer and rewritten in the target base in place, in the inserted equations and in numbers occurring naturally in the web text alike; shard order, document order, insertion positions, operators, and all non-numeric formatting are preserved, and digit runs inside decimal or scientific notation are converted independently. Each model is a decoder-only NanoChat transformer with eight blocks, residual width 512, four attention heads, four key/value heads, a vocabulary of 32,768 tokens, a context length of 2,048, and the `SSSL` attention-window pattern, trained with the next-token objective under the same recipe and seed in every base; the saved checkpoint configurations of the four runs agree on these settings. Each model is trained for 7,680 steps at sequence length 2,048 with a global batch of 262,144 tokens, a budget of $7{,}680 \times 262{,}144 \approx 2.0$ billion tokens. Because numeral length changes with the base, the token-level batches are not identical and the runs end at slightly different document positions.

C.2 Gate Matrices and the Base-8 Ambiguity

The residue design of Appendix B.3 was run in bases 6, 8, and 10 at layer 7 with each base's shorter periods (T_2, T_3; T_2, T_4; T_2, T_5) and its fundamental, using the earlier uncentered residue-first fit and clamp-only controls rather than the norm-matched phase nulls of the centered runs (Figure 14; Table 12). Bases 6 and 10, whose two shorter periods peak together at one digit, show the decimal pattern: with the fundamental kept, the joint edit steers far more than either single edit. Base 8, whose shorter periods peak together at two digits, gains least. Base 6 is reported only from this run.

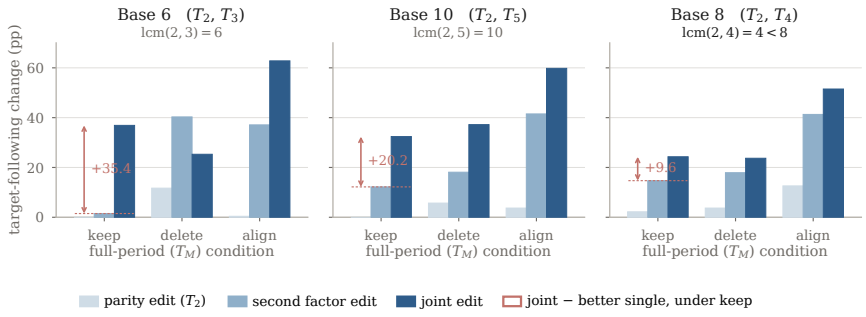


Figure 14: Residue steering in bases 6, 10, and 8 at layer 7, in the layout of Figure 9 ($n = 1{,}500$). Red arrows mark the joint advantage over the better single edit.

Table 12: Layer-7 radix gates. Edit rows show paired changes in pp against their own row-specific clamp-only controls ($n = 1{,}500$); the displayed control rates (absolute target following in percent) refer to the joint-edit row only. K, D, and A denote T_M keep, delete, and align.

Base 6 $\mathrm{lcm}(2,3) = 6$				Base 10 $\mathrm{lcm}(2,5) = 10$				Base 8 $\mathrm{lcm}(2,4) = 4 < 8$			
Edit	K	D	A	Edit	K	D	A	Edit	K	D	A
Joint control (%)	0.0	0.4	12.1	*Joint control (%)*	0.1	0.0	0.9	*Joint control (%)*	0.0	6.1	14.1
T_2	0.0	11.7	0.4	T_2	0.0	5.7	3.7	T_2	2.2	3.7	12.6
T_3	1.5	40.3	37.1	T_5	12.2	18.1	41.5	T_4	14.7	17.9	41.3
$T_2 + T_3$	36.9	25.3	62.8	$T_2 + T_5$	32.4	37.2	59.8	$T_2 + T_4$	24.3	23.7	51.5

The base-8 split. In the centered units run, the joint $T_4 + T_2$ edit with T_8 held at its source value (the T_4+T_2 row of Table 15) places 1,076 of 1,500 layer-7 outputs on the pair $\{t, t+4\}$, 429 at t and 647 at $t+4$. The 217 cases with shift 4, in which $t+4$ is the source digit and neither residue changes, all keep the source prediction; among the remaining 1,283 cases the pair receives 859 outputs (67%), 429 at t and 430 at $t+4$, and 60 outputs (4.7%) fall on $t+2$ or $t+6$, which violate the transplanted T_4 residue (Figure 15a). Within the pair the choice follows the shift as a group tendency: t wins more often at shifts 1, 2, and 7 and $t+4$ at shifts 3, 5, and 6; at the odd shifts this is the candidate nearer the source, while shifts 2 and 6 are equidistant and retain the smaller baseline asymmetry (Figure 15b). The shift-4 predictions all stay at the source, although the edit is not an exact no-op in logit space (mean gains of $+0.02$ and -0.05 nats). The edit itself does not distinguish the two: at every shift it raises t and $t+4$ by nearly the same amount (7.6 and 7.1 nats at shift 1; within about half a nat at every shift), so the logit gap between them after the edit equals the baseline gap, already present before the edit and carried by the coordinates left at their source value (Figure 15c). Placing the same edit at layer 6 enlarges that gap by roughly 60–75% at the odd shifts and by more at the near-zero even ones, in the same direction (Section 4.2). These are intervention-induced outputs on baseline-correct examples, not natural errors. The earlier uncentered run also included a persistent all-layer edit, under which three quarters of outputs fell on the pair; the centered run does not, and we report only single-layer edits.

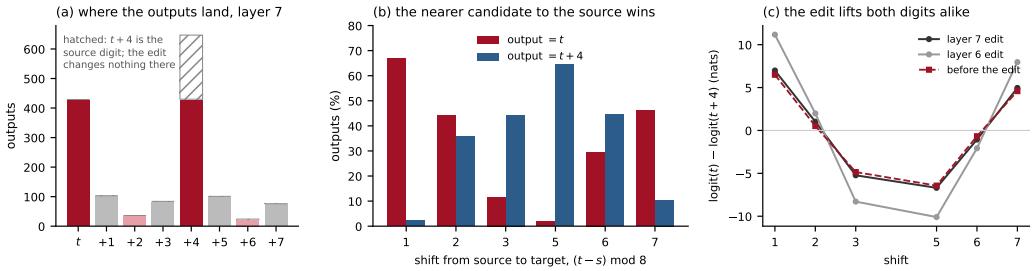


Figure 15: Base-8 outputs under the joint $T_4 + T_2$ edit with T_8 at its source value (units, layer 7, $n = 1{,}500$). (a) Output offsets from the target; red marks the unresolved pair, pink the outputs that violate T_4. (b) Share of outputs at t and at $t+4$ by shift from source to target. (c) Logit gap between t and $t+4$ before and after the edit, by shift; the layer-7 edit leaves the gap unchanged and the layer-6 edit enlarges it.

C.3 Native Harmonic Factorials in Bases 7 and 8

These are the details and complete rates behind Section 3, Figure 2, and the radix rows of Figure 3, for bases 7 and 8 at units and at the next position. The native factorials are centered runs; the cross-position, lower-place, and ceiling arms at the end of this section come from an earlier uncentered run of the same design and are labeled as such.

Data. Sums have the form $n = b^2q + s$ with $s < b^2$ and 12 operand decompositions per sum; fitting uses $q = 8, 10, \ldots, 38$ and evaluation $q = 9, 11, \ldots, 39$, with operands in 0–2000, inside the arithmetic training domain. Evaluation cases are balanced over upper blocks and over ordered

unequal digit pairs (42 in base 7, 56 in base 8), and source and target donors match the answer prefix, lower suffix, and incoming carry. Each position has a fixed cohort of 1,500 cases used at every layer; fitting uses 4,000 rows (3,999 at sevens, where one fitting row is excluded). On a separate unedited diagnostic of 256 examples, full-answer accuracy is 98.4% in base 7 and 99.6% in base 8.

Subspaces. Each plane is centered at the ridge intercept projected onto its basis, estimated from fitting rows only; source and donor both subtract this center, and correct edits preserve the source radius about it. At units the fundamental is fitted first and each extra harmonic independently removes it; at the next position a common lower-place T_b plane is protected, the fundamental removes it, and each extra removes both. Extras are not orthogonalized against one another. T_2 uses a single cosine feature; T_{16} has two Fourier features across the full fitting grid, although a fixed lower suffix leaves it only two antipodal ideal phases; every other component has rank two. No magnitude direction is protected. All four runs pass the projection no-op audit at every layer.

Interventions and controls. Every subset is the sum of the fixed single-harmonic displacements from the same source, as in the Qwen factorial, so the full-family arm is not the union-radius transplant of the earlier run and the two are not directly comparable. Null donors match the prefix, lower suffix, and incoming carry, differ from the source and target digits, and have the wrong phase for every edited period; their displacements are norm-matched per case, and subset nulls are matched again after summing. Parity-changing edits (T_2 at units, T_{16} at eights) use a null that moves outward along a wrong-phase donor direction on the source side of the plane. A case is excluded from a paired comparison only if it admits no null: at layer 7 this removes one base-8 units case from the arms containing T_2 and five eights cases from the arms containing T_{16}; base 7 retains every case. Intervals for paired differences resample the 16 upper blocks (2,000 draws).

Rates. Table 13 summarizes the native families and Tables 14 and 15 list every subset with its paired null. The same arms were run at every layer; large full-family gains over the null emerge from layer 3 at base-7 units, layer 4 at sevens, and layer 5 in base 8, with small but distinguishable gains at some earlier layers, and steering strengthens through the last block in every run.

Table 13: Layer-7 target following (%) under native harmonic transplants in the centered runs ($n = 1{,}500$ correct-arm cases). "+ second harmonic" adds $T_{7/2}$ or T_4, or the radix-scaled counterpart at the next position.

Base	Position	Fundamental	+ second harmonic	Full family	Full-family null
7	Units	20.9	67.9	98.0	15.3
7	Sevens	22.6	71.0	96.5	16.3
8	Units	15.4	80.6	93.1	11.5
8	Eights	18.1	86.0	96.1	9.6

Table 14: Base-7 native harmonic factorial at layer 7 (centered run). Correct and null are target following (%) on the paired cases; the difference is paired, with a 2,000-draw upper-block bootstrap interval.

Position	Periods edited	Paired n	Correct	Null	Difference [95% CI]
Units	T_7	1,500	20.9	7.2	13.7 [11.9, 15.3]
Units	$T_{7/2}$	1,500	19.9	4.7	15.3 [13.4, 17.2]
Units	$T_7 + T_{7/2}$	1,500	67.9	12.1	55.9 [53.6, 58.2]
Units	$T_{7/3}$	1,500	12.7	3.6	9.1 [7.8, 10.8]
Units	$T_7 + T_{7/3}$	1,500	57.7	9.9	47.8 [45.1, 50.6]
Units	$T_{7/2} + T_{7/3}$	1,500	59.3	10.6	48.7 [46.5, 51.0]
Units	$T_7 + T_{7/2} + T_{7/3}$	1,500	98.0	15.3	82.7 [80.5, 84.8]
Sevens	T_{49}	1,500	22.6	7.3	15.3 [13.7, 16.8]
Sevens	$T_{49/2}$	1,500	22.7	5.1	17.6 [15.8, 19.4]
Sevens	$T_{49} + T_{49/2}$	1,500	71.0	12.2	58.8 [57.2, 60.3]
Sevens	$T_{49/3}$	1,500	10.4	1.9	8.5 [7.1, 9.8]
Sevens	$T_{49} + T_{49/3}$	1,500	59.3	10.6	48.7 [46.7, 50.7]
Sevens	$T_{49/2} + T_{49/3}$	1,500	54.5	10.3	44.2 [41.7, 46.4]
Sevens	$T_{49} + T_{49/2} + T_{49/3}$	1,500	96.5	16.3	80.2 [77.7, 82.4]

Table 15: Base-8 native harmonic factorial at layer 7 (centered run), as in Table 14.

Position	Periods edited	Paired n	Correct	Null	Difference [95% CI]
Units	T_8	1,500	15.4	6.6	8.8 [7.1, 10.7]
Units	T_4	1,500	19.7	1.1	18.7 [17.5, 19.8]
Units	$T_8 + T_4$	1,500	80.6	9.1	71.5 [69.7, 73.2]
Units	$T_{8/3}$	1,500	3.7	1.1	2.5 [1.7, 3.6]
Units	$T_8 + T_{8/3}$	1,500	29.3	8.7	20.5 [18.5, 22.6]
Units	$T_4 + T_{8/3}$	1,500	39.8	4.1	35.7 [33.7, 37.8]
Units	$T_8 + T_4 + T_{8/3}$	1,500	90.5	12.0	78.5 [76.9, 80.0]
Units	T_2	1,499	0.8	0.0	0.8 [0.5, 1.1]
Units	$T_8 + T_2$	1,499	27.2	5.6	21.5 [20.2, 23.0]
Units	$T_4 + T_2$	1,499	28.6	1.3	27.3 [25.7, 28.9]
Units	$T_8 + T_4 + T_2$	1,499	86.9	8.9	78.1 [76.1, 79.8]
Units	$T_{8/3} + T_2$	1,499	7.7	0.3	7.4 [6.4, 8.5]
Units	$T_8 + T_{8/3} + T_2$	1,499	43.1	7.8	35.3 [33.8, 36.7]
Units	$T_4 + T_{8/3} + T_2$	1,499	42.9	3.3	39.6 [37.2, 41.9]
Units	$T_8 + T_4 + T_{8/3} + T_2$	1,499	93.1	11.5	81.5 [80.0, 83.0]
Eights	T_{64}	1,500	18.1	8.3	9.8 [8.1, 11.6]
Eights	T_{32}	1,500	18.3	0.9	17.5 [16.3, 18.8]
Eights	$T_{64} + T_{32}$	1,500	86.0	7.0	79.0 [77.6, 80.3]
Eights	$T_{64/3}$	1,500	1.7	0.7	1.0 [0.5, 1.4]
Eights	$T_{64} + T_{64/3}$	1,500	31.4	11.1	20.3 [18.3, 22.4]
Eights	$T_{32} + T_{64/3}$	1,500	41.4	3.5	37.9 [35.6, 40.0]
Eights	$T_{64} + T_{32} + T_{64/3}$	1,500	93.9	12.1	81.7 [80.3, 83.2]
Eights	T_{16}	1,495	0.1	0.0	0.1 [0.0, 0.2]
Eights	$T_{64} + T_{16}$	1,495	31.0	5.2	25.8 [23.9, 27.8]
Eights	$T_{32} + T_{16}$	1,495	30.8	0.6	30.2 [28.4, 31.8]
Eights	$T_{64} + T_{32} + T_{16}$	1,495	93.0	5.7	87.4 [86.4, 88.3]
Eights	$T_{64/3} + T_{16}$	1,495	8.4	0.5	7.9 [6.6, 9.1]
Eights	$T_{64} + T_{64/3} + T_{16}$	1,495	48.2	8.7	39.5 [37.6, 41.4]
Eights	$T_{32} + T_{64/3} + T_{16}$	1,495	44.0	1.8	42.2 [40.4, 44.1]
Eights	$T_{64} + T_{32} + T_{64/3} + T_{16}$	1,495	96.2	9.6	86.6 [85.3, 88.0]

Additivity. Table 16 applies the definitions of Appendix B.5 to the radix factorials at the last three layers. Across the last three layers the family's margin gain differs from the summed singles by at most 9%; at layer 7 by at most 5.2%. The differences are usually negative, with one small positive difference for base-8 eights at layer 7 ($+0.24$ nats), and the positive interactions of the Qwen sweep do not appear in these layers. The centered vector R^2 is 0.95–0.98 at layer 7. Most differences are

distinguishable from zero but small, at most -2.5 nats against family gains of 26–34, with base-7 units the largest at layer 7 (-1.85 nats).

Table 16: Additivity in the radix factorials by layer of the edit. Gains are mean target-minus-source margin changes (nats); the difference carries a paired case-bootstrap 95% interval; R^2 is that of the summed single-edit digit-centered gain vectors against the measured ones over all multi-harmonic subsets, cases, and digits, with no fitted coefficient.

Base	Position	Layer	Full family	Sum of singles	Difference [95% CI]	Ratio	Vector R^2
7	units	5	26.5	28.0	-1.51 [-1.97, -1.04]	0.95	0.91
7	units	6	31.2	33.6	-2.36 [-2.79, -1.94]	0.93	0.93
7	units	7	33.9	35.8	-1.85 [-2.11, -1.60]	0.95	0.95
7	sevens	5	27.9	28.3	-0.39 [-0.68, -0.11]	0.99	0.93
7	sevens	6	31.1	31.4	-0.27 [-0.52, -0.02]	0.99	0.95
7	sevens	7	32.9	33.1	-0.12 [-0.30, 0.06]	1.00	0.95
8	units	5	26.3	28.8	-2.46 [-2.81, -2.11]	0.91	0.88
8	units	6	29.2	31.2	-1.99 [-2.27, -1.69]	0.94	0.91
8	units	7	31.3	31.6	-0.29 [-0.48, -0.09]	0.99	0.97
8	eights	5	29.5	30.3	-0.79 [-1.04, -0.56]	0.97	0.90
8	eights	6	31.2	32.0	-0.85 [-1.05, -0.64]	0.97	0.92
8	eights	7	32.0	31.8	0.24 [0.11, 0.36]	1.01	0.98

Earlier uncentered run: cross-position, lower-place, and ceiling arms. An earlier run of the two-position design, without centering and with union-radius transplants for the extras, supplies the arms absent from the centered factorials. Its native full-family rates differed (92.1%/96.6% in base 7 and 82.2%/93.5% in base 8 at units/next position); centering, the composition rule, and the sevens fitting cohort all changed between the runs, so we do not attribute the difference to one cause and do not mix its native rates with the centered ones. Its cross-position arms transplant the other position's family at the tested token position, fitted there without removing overlap with the native planes; at units they use wrong-phase donors, and at the upper position, where lower periods have suffix-invariant ideal phases, same-phase wrong-digit donors. Table 17 gives the unions; individual cross periods applied one at a time give at most 1 of 1,500 target-following cases in either base, and the cross unions edit three or four planes with half to two thirds of the native edit norm. Adding lower-position periods to the complete upper family changed it little in that run: at sevens, 96.0–96.5% against 96.6%; at eights, 93.3–94.1% against 93.5%, where the all-lower addition gained 0.6 pp $[-0.1, 1.4]$ over the native intervention while beating its own null by 1.9 pp $[1.2, 2.7]$. Two ceilings bound what the family could do there: copying the donor's native-family coordinates without preserving the source radius gave 97.3% and 98.3% at base-7 units and sevens and 96.6% and 98.9% at base-8 units and eights, and copying the donor's entire residual gave 99.9%/97.3% and 100%/100%.

Table 17: Standalone cross-period unions at layer 7 in the earlier uncentered run. Rates are target following (%) on 1,500 cases.

Base	Tested position	Cross periods	Correct	Null
7	Units	$T_{49}, T_{49/2}, T_{49/3}$	0.00	0.00
7	Sevens	$T_7, T_{7/2}, T_{7/3}$	0.80	0.33
8	Units	$T_{64}, T_{32}, T_{64/3}, T_{16}$	0.00	0.00
8	Eights	$T_8, T_4, T_{8/3}, T_2$	0.00	0.00

D Other Models

D.1 Pythia Units-Digit Controls

Pythia-6.9B replicates the units-digit residue matrix at layer 26 and adds three controls that the Qwen experiments do not include (Table 18). Pythia predicts a complete one-token numeric answer without a supplied answer prefix, from the prompt `Output ONLY a number. {a}+{b}=` with

operands in 0–99. We retain initially correct examples whose sums tokenize as one token, and count a patch as following the target units digit when the integer decoded from the full-vocabulary argmax has that residue modulo 10; nonnumeric outputs count as failures. First, the exclusive T_{10} plane retains 98.9% of its fitted norm and decodes the digit, yet steering it alone changes no prediction at any tested layer (Appendix A.2). Second, editing the residue planes at the operand tokens rather than at the answer position steers the digit, with a 35.0 pp advantage over shuffled phases, so the residues enter the computation before the answer is generated. Third, cross-donor chimeras produce the CRT-composed digit with a 24.3 pp advantage, and residue edits that leave T_{10} untouched steer the digit while shifting the downstream alignment of T_{10} only modestly (Appendix A.2). Together these show that the residue planes are used in the computation, support cross-donor residue composition, and are not read off a causally sufficient fundamental. Unlike in Qwen, single residue edits are already effective with the fundamental kept (12.5 and 22.4 pp); whether this reflects the earlier site or the model is not resolved. Intervals in Table 18 are paired normal-approximation intervals.

Table 18: Pythia units-digit interventions at layer 26. Entries are paired target-following changes against matched clamp-only controls in pp, with 95% confidence intervals and $n = 1{,}500$ per cell.

Residue edit	T_{10} keep	T_{10} delete	T_{10} align
T_2	$+12.5\ [10.9, 14.2]$	$+48.2\ [45.7, 50.7]$	$+66.9\ [64.6, 69.3]$
T_5	$+22.4\ [20.3, 24.5]$	$+31.1\ [28.7, 33.4]$	$+45.9\ [43.3, 48.4]$
$T_2 + T_5$	$+45.4\ [42.9, 47.9]$	$+43.4\ [40.9, 45.9]$	$+67.5\ [65.1, 69.8]$

D.2 Cross-Task Transfer in Llama

Feucht et al. (2026) show that Fourier features fitted on addition steer weekday arithmetic in Llama-3.1-8B. We ask whether the internal residue organization transfers as well, applying planes fitted only on addition prompts at layer 18 to weekday questions, with disjoint phase-alignment and evaluation offsets (Table 19). Weekday answers are scored by the largest logit among the seven weekday-label tokens rather than over the full vocabulary, with label log-probabilities normalized over those seven labels. Joint residue steering is stronger than either single plane and increases from `keep` to `align`, reaching 36.5 pp $[24.5, 48.5]$ over shuffled donors and 44.4 pp over the keep-residue control. The residue and T_{10} spans overlap only at numerical precision, and the edit changes the same-layer T_{10} coordinates by at most 0.16%. Addition therefore transfers a structured residue organization rather than a single periodic direction. We treat this endpoint as exploratory because decimal residues only partially specify a weekday answer, and the sample is small.

Table 19: Llama cross-task steering using addition-only subspaces. Entries are paired changes in target-weekday following in percentage points (pp) relative to shuffled donors; $n = 63$ per cell.

Addition residue edit	T_{10} keep	T_{10} delete	T_{10} align
T_2 only	+0.0	+6.3	+11.1
T_5 only	+3.2	+7.9	+19.0
$T_2 + T_5$	+12.7	+19.0	+36.5