SMALL TRANSFORMERS LEARN KEPLER'S EQUATION WITH A NUMBER LINE AND AN ECCENTRICITY REGISTER

Anonymous authors
Paper under double-blind review

ABSTRACT

We train a family of 210 small transformers to solve Kepler's equation $M = E - e \sin E$, then reverse-engineer the algorithm they learn. To our knowledge this is the first circuit-level account of a continuous, transcendental equation. The task is trigonometric, yet the digit embeddings form a **number line**, not a circle. Attention writes e into a single residual-stream direction largely hidden from the readout, the **e-register**. The MLP reads it and writes the $e \sin(nM)$ corrections. The output carries harmonics of M to $n \approx 30$. Scale the register and the corrections scale with it; set it to zero and the model tracks its own circular orbit. The line and the register hold across the 210-model family. We release the models and the analysis code as a testbed: the register's variable is known to be e, so a method that claims to find hidden computation can be scored against it.

1 INTRODUCTION

Can the algorithm a transformer learns for a continuous transcendental equation be read out of its weights? Modular addition in a small transformer has been taken apart down to the algorithm (Nanda et al., 2023). Beyond it the closest work reads representations, not circuits. Integer addition in LLMs runs on a helix, and interventions show the model uses it (Kantamneni & Tegmark, 2025).

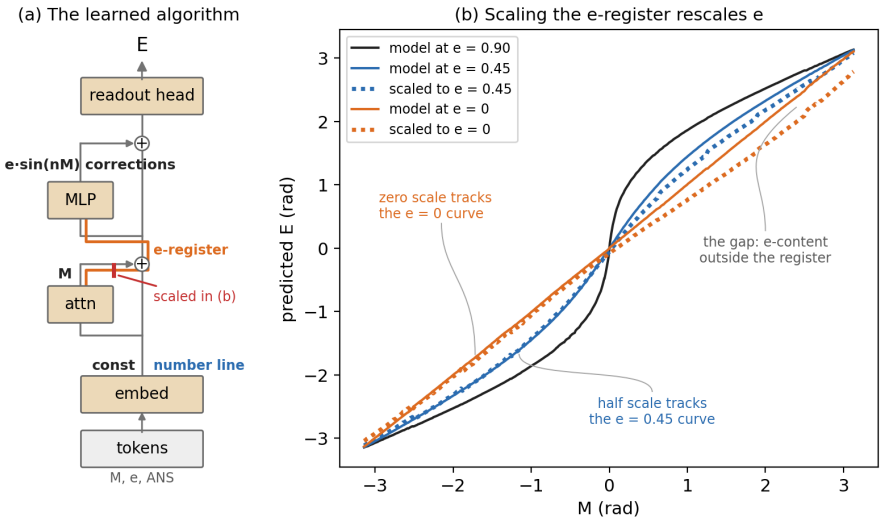

Figure 1: **(a) The learned algorithm.** Three writes sum into the logit through the final LayerNorm (§4.4): embeddings write a **const**, attention **M**, the MLP the $e \sin(nM)$ **corrections**. The **number line** (blue) lives in the digit embeddings, which only attention reads. The **e-register** (orange) carries e to the MLP. Not shown: the pure-M harmonics both components also write. **(b) Scaling the e-register rescales e.** The patch site is the red mark in (a) (§4.5).

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.

Probes on transformers trained on the simple harmonic oscillator find the intermediates of a known numerical method (Kantamneni et al., 2024). Kepler's equation, $M = E - e \sin E$, turns time into position on an elliptical orbit: given mean anomaly $M \in [-\pi, \pi)$ and eccentricity $e \in [0, 1)$, find the eccentric anomaly E (Colwell, 1993). The solution E has no closed form, but Newton's method computes it to machine precision, and the Fourier–Bessel series converges to it for every e below 1.

We train a 1,153-parameter transformer to solve Kepler's equation and call it the *primary*. It is one of a designed family of 210 models: 42 configurations at five seeds each, spanning width, training steps, loss, output map, and architecture (§3.2). We read the algorithm out of the weights, then check it across the family. The primary reaches a median error of 1.5e-3 radians and a max of 8.6e-2, against a classical method's 4e-10 max (Boyd, 2007, degree 15). We do not compete on accuracy. We study representation and mechanism.

- **Representation (§4.1).** The digit embeddings form a **number line**, line-dominated and open, in every model of the family. The modular-addition circle (Nanda et al., 2023) and the integer-addition helix (Kantamneni & Tegmark, 2025) do not appear. Retrained past one rotation, the primary configuration builds no circle either; in the hardest case the line dissolves and error rises $81\times$ (§4.2).
- **Mechanism (§4.3–4.5).** The model's output carries Bessel harmonics (Colwell, 1993) out to $n \approx 30$, the $e \sin(nM)$ corrections to M (§4.6). Attention writes e into the **e-register**, a single residual-stream direction nearly orthogonal to the readout (Fig. 1). The MLP reads it and writes the corrections from a spline basis rather than Bessel terms (Fig. A4).
- **Testbed (§5).** The register is a working specimen of hidden computation. It is causally necessary across the family, and in the primary it is invisible to any read that projects a component onto the readout direction. We release all 240 models (the 210-model family plus the 30 off-family and wrap-study runs) with the training and analysis code.

2 Related Work

Modular-addition lineage. Nanda et al. (2023) reverse-engineered modular addition in transformers and set the template we mirror: Fourier structure in the embeddings, attention-pattern reads, MLP decomposition, ablation as proof. Liu et al. (2022) found the linear form in a network trained on integer addition, and a circle when the addition was modular. At LLM scale the number code for addition is a generalized helix (Kantamneni & Tegmark, 2025), built on Fourier features (Zhou et al., 2024). Tiblias et al. (2026) fit a similar taxonomy to fixed LLMs: cyclic quantities map to circles, ordered magnitudes to log-compressed lines. Engels et al. (2025) find that days and months in LLMs sit on circular features. Zhong et al. (2023) caution that one task can support qualitatively different algorithms across runs and architectures. Nikankin et al. (2025) find LLM arithmetic runs on a bag of heuristics, not a single algorithm. We read the embedding code and the circuit and find one algorithm across the family.

Continuous physics. Kantamneni et al. (2024) train transformers on simple harmonic oscillator trajectories and probe the hidden states, which encode the intermediates of a known numerical method. Interventions through the probes show the model uses them. Our e-register patch in §4.5 is like their steering check, run along a direction obtained by decomposing attention's write into known functions of the input, not by a probe fit on the residual stream. Our model is sized to let us read which component computes each term. Closest in method is Kitouni et al. (2024), who train a network on nuclear observables and find a helix in its embeddings. Their attention is held uniform, so they have no routing circuit to trace as we do in §4.3. A separate line of research treats the model as an oracle and reads planetary laws off its predictions (Lemos et al., 2023; Li et al., 2021; Vafa et al., 2025).

Kepler solvers. Boyd (2007) gives a noniterative classical method with no initial guess: 4e-10 max error at degree 15. Zheng et al. (2022) solved Kepler with a feedforward neural network plus a Halley correction step. Ours is a transformer with no correction step.

Hidden computation. Models compute more than their outputs show: across content-free filler tokens (Pfau et al., 2024), in off-the-shelf frontier models (Brauer et al., 2026), in a workspace of steerable variables apart from a model's automatic processing (Gurnee et al., 2026), and in a residual subspace the unembedding barely sees, where attention-sink signals sit (Cancedda, 2024). Ours is the case where the hidden variable is known (§5). App. A surveys further related work.

Table 1: **The 240 models.** Every configuration is trained at five seeds. The family is all 1-layer with the final LayerNorm on the standard domain; each control varies one axis of the primary at d8 and 800k steps.

family	cluster	what varies	count
✓	primary & scaling grid	d_model $\in \{4, 8, 16, 32, 64, 128\}$ $\times$ steps $\in \{25\text{k}, 50\text{k}, 100\text{k}, 200\text{k}, 400\text{k}, 800\text{k}\}$	36×5
✓	output-map controls	sigmoid; tanh; clamp	3×5
✓	architecture controls	ReLU MLP; 4 heads	2×5
✓	loss control	MAE	1×5
	off-family	2-layer model; no-final-LayerNorm twin (§4.5, §5)	2×5
	wrap study (§4.2)	M range $\in \{\pm 10\pi, \pm 50$ rad$\}$ $\times$ target $\in \{E, E \bmod 2\pi\}$	4×5

3 SETUP

3.1 INPUT / OUTPUT

The model takes M and e as digit sequences and returns E as a single scalar. All three variables are normalized to $[0, 1)$: M and E by $(x + \pi)/(2\pi)$. e needs no map and stops at 0.999 (§3.3). Each input value is written as its first 12 decimal digits. The input sequence is 25 tokens: 12 M-digits, 12 e-digits, ANS token.

The output is read off the residual stream at the ANS position. A linear head maps the final residual stream to a scalar, and an output map turns it into normalized E. The map is the identity on the primary. The controls of Table 1 use sigmoid, tanh, or clamp. We call the pre-output-map scalar the *logit* throughout, by analogy with a classifier's pre-softmax logits. With a linear output the logit is normalized E. Finally, we denormalize by $\times 2\pi - \pi$, so all reported errors are in radians of E. The rationale for digits in and a scalar out is in App. B.

3.2 MODELS: PRIMARY, FAMILY & OFF-FAMILY

We choose a primary for legibility while avoiding d4 degeneracy (App. I): 1 layer, width 8 (MLP width 4× the model width, here and throughout the family), 2 heads, GELU (Hendrycks & Gimpel, 2016), linear output, MSE, seed 0. Params = 1,153, all trainable. Primary numbers are this seed; family claims quote the five-seed spread. Table 1 lists the 210-model family and the 30 off-family and wrap-study runs.

The architecture is a standard transformer (Vaswani et al., 2017) with pre-LN (Xiong et al., 2020), apart from the scalar readout. A final LayerNorm (Ba et al., 2016) precedes the readout head. The remaining details are in App. B.

3.3 TRAINING & EVAL GRID

Each training step draws a fresh batch of 256 (M, e) pairs, $M \sim \text{Uniform}(-\pi, \pi)$ and $e \sim \text{Uniform}(0, 0.999)$, with targets E from a float64 Newton solver converged to machine precision. Every step sees fresh data, so memorization and grokking (Power et al., 2022; Nanda et al., 2023) are out of scope. Every model we analyze is a separate run trained to completion at its horizon, not a mid-training checkpoint. The one exception is the wrap study's incommensurate wrapped cell, whose circle gain is read from saved snapshots (App. D). Training hyperparameters are fixed so cross-model comparisons are like-for-like (optimizer and schedule in App. B). Scaling laws in steps and width are in App. I.

For evaluation we use a fixed 400×200 grid over $M \in [-\pi, \pi)$, $e \in [0, 0.999)$: we split the domain into equal cells and jitter each column's M and each row's e uniformly within its cell, with a fixed seed, so every digit pattern is natural. The domain stops at $e = 0.999$, short of the parabolic boundary. We call $e \geq 0.9$ the *cusp* region, where the solution steepens, and $e < 0.9$ the *bulk* (App. H). The grid-density rationale is in App. B.

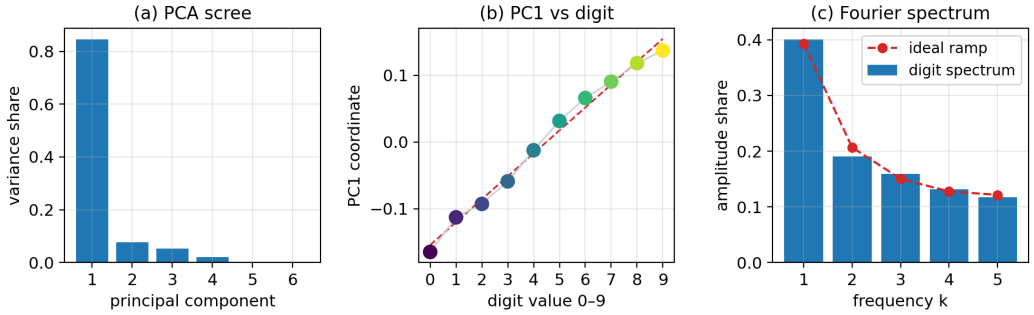

Figure 2: **The primary's digit embeddings form a number line. (a)** PCA scree: PC1 carries an 85% variance share. **(b)** PC1 coordinate vs digit value: ordered and evenly spaced, |Spearman| = 1.00, |Pearson| = 0.995. **(c)** The digit Fourier spectrum sits on the ideal $1/\sin(\pi k/N)$ ramp of a linear code, max deviation 0.02.

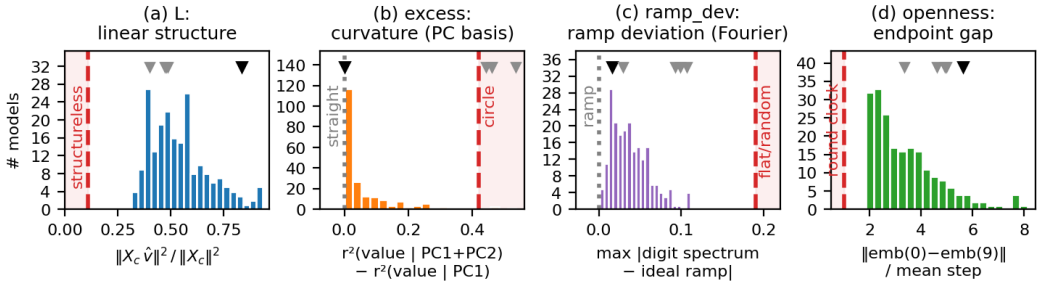

Figure 3: **The number-line battery across all 210 models** (App. C). Gray ▼ the four most curved codes (highest excess; Fig. A1); black ▼ the primary. **(a)** All models sit above the 0.11 structureless expectation on the variance-share statistic. **(b)** Four exceed the 0.42 circle calibration on curvature but stay open arcs. **(c)** All below the 0.19 flat null on ramp deviation; a round clock's 0.61 is far off scale. **(d)** All sit at 1.84 or above, far from a round clock's endpoint gap of ≈ 1.

Mean-ablation, used throughout, replaces a component's output with its per-position average over the eval grid, so the component becomes input-independent (Wang et al., 2023).

4 Results

4.1 Embedding geometry: the number line

The primary arranges its digit tokens on a straight, evenly spaced number line (Fig. 2). PC1 carries the digit values, ordered and evenly spaced, and the Fourier spectrum sits on the ideal ramp of a linear code. A circular code would concentrate at a few sharp frequencies (Nanda et al., 2023), and the primary's spectrum shows none.

All 210 digit codes are line-dominated and open (Fig. 3, Table A1). Three basis-free statistics support it (App. C): *L*, the share of embedding variance that digit value explains; *ramp_dev*, the Fourier spectrum's max deviation from a line's ideal ramp; and *openness*, the 0-to-9 endpoint gap in adjacent-step units. Every model carries linear structure significant against a structureless null (worst p_L = 1.6e-4) and every code remains open. Rotation off PC1 and dilution of its share are common and benign, and four models curve as much as a circle would yet stay open arcs (App. C).

4.2 The wrap study: does wrapping force a circle?

The task is trigonometric but the code is not. The solution is a monotonic ramp plus a Fourier series in sines of nM, each coefficient a function of e (§4.6). Those harmonics could motivate a

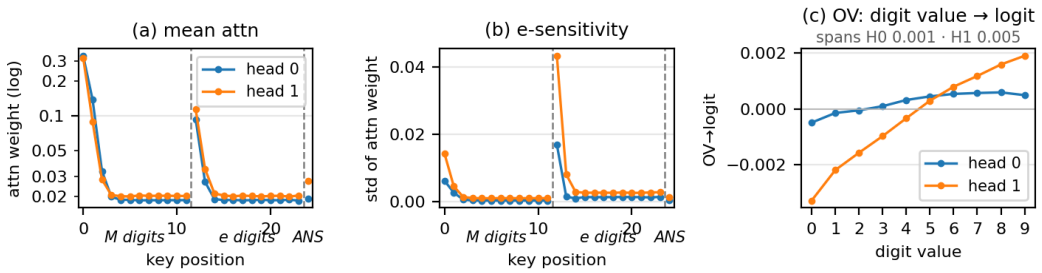

Figure 4: **The reader head: QK routes, OV converts** (primary). **(a)** Per-head mean attention from the readout query concentrates on the leading M and e digits, falling by significance within each field. **(b)** Std over e of each M-averaged attention weight. The leading e digit moves most, 2.6× more in head 1 than head 0. **(c)** Each head's weights-only OV transfer curve from digit value to logit contribution (one curve serves both M and e; App. F). Head 1 rises monotonically with digit value and head 0 is near-flat.

circular code for M, the way the modular-addition lineage builds trig functions from circular codes (Nanda et al., 2023). The theories that explain Fourier features derive them from group or periodic structure (Marchetti et al., 2024; Morwani et al., 2024; Gromov, 2023), and LLMs put quantities that wrap, days and months, on circles (Engels et al., 2025; Tiblias et al., 2026). So we hypothesized wrapping would produce circular embeddings and tested it.

Design. We retrain the primary configuration on extended domains, a 2×2 of M range and target form with five seeds per cell and predictions written before each launch (App. D). One range is digit-aligned at $\pm 10\pi$, where a rotation is a decimal fraction of the input and the phase can be read off the digits. The other is incommensurate at ± 50 radians, 15.9 rotations, where no digit place aligns with the period. One target is the unwrapped continuation E, where the periodic part is about one percent of the output. The other is $E \bmod 2\pi$, reduced to $[-\pi, \pi)$, where the wrap is the entire task. M enters unwrapped in every cell, so the model must reduce it itself. We predicted that the aligned cells would keep the line and that the incommensurate cells would feel pressure toward a circle, weak with the unwrapped target and strongest with the wrapped one, where we expected a circle.

No cell builds a circle, and the incommensurate wrapped cell loses the line. As predicted, both aligned cells keep the line at the primary's accuracy (one seed fails the line floor, App. D). The incommensurate unwrapped cell keeps the line at 2.8× the primary's relative error. Contrary to our prediction, the incommensurate wrapped cell does not build a circle. Its error is 81× the primary configuration's five-seed median, and the models ignore e. Its digit code dissolves, failing every line test in all five seeds, with no circle in its place (Table A2; App. D).

Periodic structure appears at the predicted circle frequencies but never becomes a phase representation. On a range $\pm R$, a digit step at place j advances the phase by $(R/\pi) \cdot 10^{-j-1}$ rotations, so the tokenization predicts the circle's frequencies with zero free parameters. In the incommensurate wrapped cell, the layer-0 OV write (§4.3) carries periodic structure at those frequencies. It sits above the line-code null at every saved snapshot from 50k steps to the end in all five seeds (App. D). The null is the same statistic on the primary's own number line. Whether a circle forms at larger width, or under a schedule that does not decay to zero, is open.

4.3 Input circuit: attention reads the leading digits as quantities

OV converts digit geometry to quantity. We factor the read into its QK and OV circuits (Elhage et al., 2021; Fig. 4; App. F). The OV circuit's transfer curve maps digit value to logit contribution (Fig. 4c). Summed over heads, it is monotone across the family, median |Spearman| 0.99. The 57 exceptions below 0.95 are over-represented at d64 and d128, and most sit on an already deformed line (App. F). The read passes along whatever the line encodes, so each digit reaches the logit as a value on that line.

Table 2: **Component fits against the trig library** (primary specimen; seven of the 11 terms shown). The rows are readout projections in w_{eff} units, without the final LayerNorm's centering and per-input scale, so only the ratios are meaningful (exact split: Table A5). attn0 writes no e term above 7e-4. The output-scale fits in radians of E are in Table A4.

component	R^2	M	$\sin M$	$\cos M$	$\sin 2M$	$e \sin M$	$e \sin 2M$	$M{\cdot}e$
attn0	0.989	0.0044	0.0019	0.0030	0.0004	0.0000	−0.0001	−0.0007
mlp0	0.975	0.0018	0.0041	0.0023	0.0010	**0.0014**	**0.0023**	0.0002

QK routes the leading digits, and reads M one place deeper than e. The QK circuit concentrates the readout position's attention on the leading M and e digits, falling by significance within each field (Fig. 4a). Implicit differentiation says why M gets the extra place. Per unit of normalized input, E moves about 10 times more with M than with e, one decimal place (App. E). The weights carry the gap. In the primary, position geometry and attention mass each resolve M one place deeper than e. Across the family they do in 76% and 58% (Fig. A3). Behavior shows it in 98%, as an accurate solver should.

Dependence instruments. We split the predicted-E surface over the (M, e) grid into the e-averaged M-curve and the residual around it. ***M_dep*** and ***e_dep*** are their stds in radians of E, an exact variance split (App. F). ***e_corr***, the residual's correlation with the true e-correction field, scores whether the residual keeps the true shape. M_dep/e_dep is 9.9 to 10.0 in every model. That is a sanity check on the instruments, since any accurate solver must show that ratio.

In the primary, one head carries the read. Mean-ablation (§3.3) of head 1, the *single reader*, collapses the output nearly flat (M_dep 2.06 → 0.18). What survives of the e-residual is incoherent (e_corr 0.02). Mean-ablating head 0 spares the shape at e_corr 0.94 but raises median error 180× (Table A3). Across the family, head organization is unstable, with 28 single-reader, 141 split across two specialized heads, and 41 distributed (App. F).

4.4 Attention routes, the MLP computes

Component fits. To find which component writes the $e \sin(nM)$ corrections, we fit each component's write against a trig library. At the readout position the residual stream is an exact sum of three writes: the embedding, attn0 from attention, and mlp0 from the MLP. That position always holds the ANS token, so the embedding's write is constant, and everything the MLP sees about the input arrives through attention. We measure attn0 and mlp0 by their *readout projection*, the dot product with $w_{\text{eff}} = w_{\text{head}} \circ \gamma_{\text{ln}}$, the direction the logit reads. The projection keeps the final LayerNorm's learned gain but leaves out its centering and per-input scale (App. F).

The library has 11 terms: 1, M, e, $\sin M$, $\cos M$, $\sin 2M$, $\cos 2M$, $e \sin M$, $e \cos M$, $e \sin 2M$, and $M{\cdot}e$. E is M plus a series of $\sin(nM)$ corrections that grow with e (§4.6), so the signal sits in M, $e \sin M$, and $e \sin 2M$. The constant, e, and the three cosines must fit zero in the output, because E is odd in M: if E solves $M = E - e \sin E$, then $-E$ solves $-M$. The remaining three, $\sin M$, $\sin 2M$, and $M{\cdot}e$, are *absorbers*. The true coefficients on $\sin M$ and $\sin 2M$ are Bessel functions, curved in e. A line in e cannot follow the curve, and its intercept lands on the pure sine. $M{\cdot}e$ appears because the library terms are not orthogonal on the evaluation grid (App. G). The exact solution, fit on the same grid, returns all three too. The model's output matches that fit to within 0.001 in all 11 terms (Table A4). The library stops at $2M$ and at first order in e, so it captures shape but not precision, leaving a residual 18× the model's own error (App. G).

In the primary, the MLP carries most of the $e \sin(nM)$ corrections. All five e terms of the attn0 fit sit at or below 7e-4 (Table 2), while the MLP's write carries $e \sin M$ and $e \sin 2M$. Attention holds 2.4% of the $e \sin(nM)$ signal in the readout projection. Direct logit attribution (Elhage et al., 2021; Nanda & Bloom, 2022), which keeps the final LayerNorm's centering and per-input scale, gives attention 33%. That scale is one number per input, shared by all three writes, and it varies with e and $\cos M$, so every row does too, even the constant embedding's (App. G, Table A5). Direct logit attribution cannot separate a write's own e content from the scale's. The readout query is constant, so each attention score sees one source digit and cannot form a product. The paths that can are the

softmax, the LayerNorm scales, and the MLP. Across the linear-output models, where the logit is normalized E, the MLP carries a median 95% of the $e\sin(nM)$ signal in the readout projection.

Parity binds the logit, not each write. Both writes in Table 2 carry positive $\cos M$, so they cannot cancel each other. After the shared scale, the three rows cancel, the constant embedding's included, and $\cos M$ fits 0.000 in the logit (Table A5).

Ablating the MLP removes the corrections and leaves the M-curve at reduced slope. Mean-ablating the MLP raises median error 400× on the primary and at least 29× across the family (Tables A3 and A7). The e-dependence drops, and what remains is anti-correlated with the true correction field, so the corrections are gone. The M-curve keeps its shape but only 72% of its slope, consistent with attention's 71% share of the M coefficient (Table 2). Direct logit attribution taken per harmonic splits the logit's harmonic energy 53/47 between the two writes, the M ramp included, and sees none of this. It measures energy, not necessity (§5).

Inside the MLP: a spline family and one dominant digit corrector. In the primary, 11 of the 32 MLP neurons are live by contribution (App. G). The top neurons are soft steps and bumps in M that tilt with e, a spline basis over the input (Fig. A4). No live neuron encodes a single Bessel term, and every one already fires at $e \approx 0$. One neuron is different: a sawtooth at the second M digit's period, the digit comb. Mean-ablating this neuron raises median error 12×. It corrects digit corners, and the block pattern of Fig. A5 is the floor left where that correction runs out (App. G).

4.5 THE E-REGISTER

One residual-stream direction carries e from attention to the MLP. The readout projection sees attention's write only along w_{eff}. The reader head's attention on the leading e digit moves with e (§4.3), so the full write should carry e. Regressing the full attention write, summed over heads, on the trig library gives one direction per term. The pure e term's direction is the **e-register**. It carries 94% of the write's variance across e in the primary and sits nearly orthogonal to the readout ($|\cos| = 0.04$). So the readout projection shows little e in attention's write, while the MLP reads it. We test the e-register by patching: scale the write's fitted e term along the register at the readout position by s and rerun the model. At $s = 0$ this is an activation patch (Vig et al., 2020; Meng et al., 2022), and at other scales activation addition (Turner et al., 2023). Full removal zeroes both $e\sin(nM)$ terms ($e\sin M$ +0.767 → −0.05, $e\sin 2M$ +0.441 → −0.00) while the output fit's M coefficient holds, +0.99 → +1.00.

Scaling the register approximately sets e. Scale the e term it carries and the model tracks the eccentricity the register holds. Scale it to half and the model tracks its own $e/2$ curve to 2.5e-2 radians at the median. Scale it to zero and the model tracks its own circular orbit to 5.8e-2 radians at the median (Fig. 1b). Both coefficients fall as the register is scaled down, but the patch removes more of each than the model's own run at se does, and not by a common factor (Fig. 5). The direction is an approximate e register with a measurable calibration error.

Two controls show the direction is specific. Removing the $e\sin M$ term of attention's write raises median error 4.3× but leaves the coefficients unchanged (±0.001). The register is the e input, not the product it feeds. A random direction of matched norm perturbs the model incoherently. What distinguishes the register is what its removal leaves: the model's own $e = 0$ behavior. Distributed alignment search (Geiger et al., 2024), an optimization cross-check, converges on the same direction in all seven models we ran it on when the swapped inputs share M, at cosine 0.92 to 0.99. With fully random sources it finds a direction at cosine 0.59 to 0.89 to the register, read by the same MLP neurons (App. G).

Every model in the family has the register. All 210 models carry a rank-1 e-register. One direction holds 72 to 100% of the write's variance across e at every width, including d128 where 128 directions were available. Scaling approximately sets e. The do-nothing gap is the unpatched model's own distance from its own $e/2$ or $e = 0$ run. The half-scale patch tracks the model's own $e/2$ behavior (median 0.02 radians, max 0.10, do-nothing gap 0.12) and the zero-scale patch its own $e = 0$ behavior (median 0.04, max 0.20, do-nothing gap 0.26). Eight models close less than half their zero-scale gap, all at 200k steps or fewer, and seven their half-scale gap, six of them at 200k or fewer. The register's $|\cos|$ with the readout is 0.16 at the family median (max 0.86, max 0.35 at 800k). The head that writes the register is the head that reads e (App. G). Of the 30 off-family and

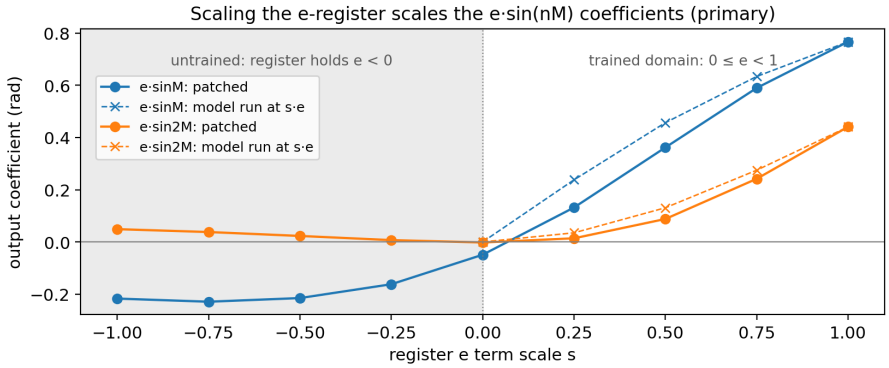

Figure 5: **Scaling the e-register scales the $e\sin(nM)$ coefficients** (primary). The scale s multiplies the write's fitted e term along the register: 1 is unpatched, 0 is full removal, and negative holds an e the model never saw. The dashed curves are the unpatched model run at (M, se), fitted with the same 11-term library in (M, e). At every tested dose $0 \leq s < 1$ the patched coefficients fall under this reference, 0.36 against 0.46 on $e\sin M$ at $s = 0.5$, and full removal overshoots to -0.05.

wrap-study runs, six ignore e and have no register to grade: the five in the wrap study's dissolved cell (§4.2) and one incommensurate unwrapped seed (App. D). The other 24 pass the 72% share floor. The no-final-LayerNorm twin also has the register, with its removal and scaling behavior, but it is not hidden from the readout (§5). Removal is incomplete in 16 models, up to 55% of $e\sin M$ surviving, and scaling still sets e in all 16 (App. G).

The 2-layer model holds the register in two seeds of five. All five pass the share test in layer 0, and two pass removal and scaling (App. G). A high share alone is not a register. We scope the result to one layer; a joint two-layer patch is the first experiment we would run next.

4.6 Classical placement and residual shape

The output carries Bessel harmonics to $n \approx 30$. The primary's harmonics $b_n(e)$ track the true coefficients $(2/n)\,J_n(ne)$ (Colwell, 1993), relative RMS below 0.01 through $n \approx 5$ and 0.18 by $n = 30$ (App. H). Any accurate solver's spectrum matches Bessel by definition, so the finding is the order.

Plotted as bulk worst case against median error, the model's widths track Boyd's degrees, not Fourier–Bessel's orders. Adding Fourier–Bessel orders cuts the median but cuts the worst case far more slowly, log-log slope 0.15 (Fig. 6). Adding width cuts both, slope 0.84. Boyd's degree (Boyd, 2007) does the same at 0.93.

The residual matches no classical truncation. In the bulk the model's signed residual has cosine similarity at most 0.05 with sampled truncations of either family, Fourier–Bessel orders 4–40 and Boyd degrees 3–9 (Fig. A5; App. H).

5 Discussion: A testbed for hidden computation

The e-register is a hidden variable whose identity is known. At scale, a hidden variable must be inferred from the outside, by ablation and steering (Gurnee et al., 2026). Here the variable is e, its consumer is the MLP, and ablation identifies both. The register is causally necessary and hidden from the readout, yet we can read it and set it (Fig. 1b).

Reads that project onto the readout miss the register and misrank the neurons. Direct logit attribution, taken per harmonic, splits the logit's harmonic energy 53/47 between the two writes, the M ramp included, yet the e corrections die only with the MLP (§4.4). The readout projection shows little e in attention's write, and the register that carries e sits at $|\cos|$ 0.04 to the readout, out of its view (§4.5). Attention's share of the $e\sin(nM)$ signal is 2.4% in the readout projection and 33% under direct logit attribution (§4.4). The contribution score $|c| \cdot \sigma$ scores each neuron along the same

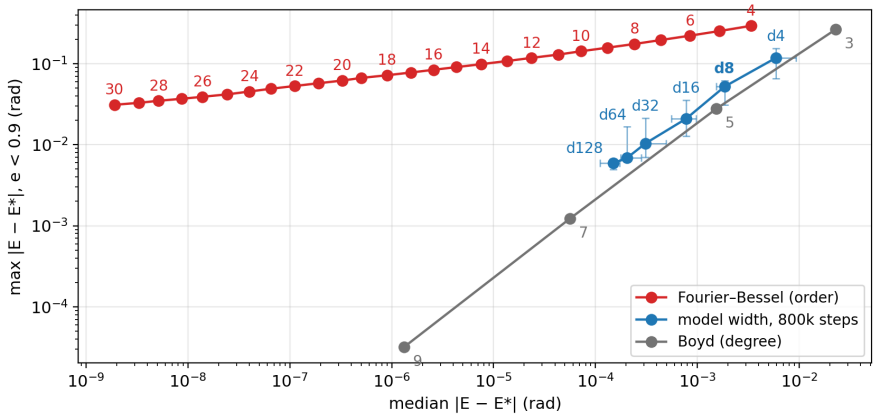

Figure 6: **Plotted as bulk worst case against median error, the model's widths track Boyd's degrees, not Fourier–Bessel's orders.** Bulk ($e < 0.9$) worst-case $|E - E^*|$ against median on the eval grid, log-log; down and left is better. Each point is one rung: Fourier–Bessel orders 4–30, model widths d4–d128 at 800k steps, Boyd odd degrees 3–9. Model points are per-width medians over five seeds; whiskers span seed min to max.

readout and misranks them. On the primary the tenth neuron by contribution, not the comb neuron of §4.4, is the second by damage and raises the median error 100× on removal. Damage and score correlate at 0.77 (App. G).

Removing the final LayerNorm removes both blind spots and leaves the algorithm intact. We trained a twin of the primary with the final LayerNorm removed at five seeds. Median error rises 2.1 to 3.1× against the primary seed, so the normalization does real precision work. The number line and the rank-1 register, with its removal and scaling behavior, reappear in all five seeds, and the spline and comb anatomy and the block residual in the seed we read at neuron grain. What does change is the register's visibility: $|\cos|$ 0.28 to 0.72 across seeds, against the primary's 0.04. The contribution score now predicts ablation damage, correlating at 0.98 to 1.00 against the primary's 0.77. On the primary itself, pinning the final-LayerNorm statistics at their unablated values during the tenth neuron's ablation cuts its damage 14× (App. G). The score misses that per-input scale, the same term the readout projection leaves out (§4.4). LLMs have normalization, so the hidden case is the more realistic one.

The testbed identifies the conditions a method works under. Alignment search finds the register when the swapped inputs share M and falls short of it with random sources, cosine 0.59 to 0.89 (§4.5). The release holds all 240 models. In the 234 that use e the variable to find is e, and the six that ignore it (§4.5) are negative cases. A linear probe for e should find the register, but it would not find that attention writes it, that the MLP reads it, or that setting it sets e. We do not test dictionary-learning reads (Bricken et al., 2023; Cunningham et al., 2024). Whether a sparse autoencoder trained on the stream finds the register is the first question we would put to it.

Limitations. The family is all one layer, trained on one equation, with 1,153 parameters at the primary. The account is at the component level. Replacing attention's write with its 11-term fit keeps every coefficient within 7% and raises median error from 1.5e-3 to 5.3e-2, so the model's precision sits in the 3% of the write the fit discards (App. G). LLMs already carry the integer-addition helix, a linear ramp with periodic components (Kantamneni & Tegmark, 2025). Whether an LLM given Kepler would reuse it, and whether e would take a readable path across layers, we do not know.

Conclusion. The models solve nothing the classical methods do not solve better. What they add is an algorithm that can be read from the weights, and a hidden variable that can be read and set.

AI USE STATEMENT

In this work, we used generative AI tools for research assistance (refining hypotheses, designing experiments, developing the conceptual framing, assisting with mathematical derivations, and interpreting results, in iterative sessions directed by the author), for coding assistance (training, analysis, and infrastructure code), and for writing assistance (drafting and editing text under author direction). We have not used generative AI tools to generate data (training targets come from a float64 Newton solver in authored code), and the remaining required-disclosure tasks (translation, dataset cleaning, qualitative data analysis) are not applicable to this work. We have reviewed all AI-assisted work: all experimental results, quantitative claims, figures, and references were checked by the author, AI-drafted text was reviewed line by line, and AI-suggested citations were verified against the cited papers. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

All 240 model checkpoints, the training code, and the analysis battery are released at `https://anonymous.4open.science/r/kepler-e-register-iclr-D17C`. Every number, figure, and table in the paper regenerates from the committed weights via a single script (`reproduce_analysis.py`). The released checkpoints are the exact analyzed models, so no retraining is required. Retraining could not recover them bit for bit: PyTorch's MPS backend on Apple Silicon is not bit-deterministic, and two runs at the same seed diverge. Retraining from §3 and App. B reproduces the results up to seed-to-seed variation, not the weights.

REFERENCES

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

John P. Boyd. Rootfinding for a transcendental equation without a first guess: Polynomialization of Kepler's equation through Chebyshev polynomial expansion of the sine. *Applied Numerical Mathematics*, 57(1):12–18, 2007. doi: 10.1016/j.apnum.2005.11.010.

Kaley Brauer, Claudio Mayrink Verdun, and Samuel Marks. Reading between the dots: Decoding hidden computation across filler tokens. *arXiv preprint arXiv:2607.03502*, 2026.

Trenton Bricken, Adly Templeton, Joshua Batson, et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. transformer-circuits.pub/2023/monosemantic-features.

Nicola Cancedda. Spectral filters, dark signals, and attention sinks. In *Annual Meeting of the Association for Computational Linguistics*, 2024. arXiv:2402.09221.

Peter Colwell. *Solving Kepler's Equation Over Three Centuries*. Willmann-Bell, 1993.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations*, 2024. arXiv:2309.08600.

Alex Davies, Petar Veličković, Lars Buesing, Sam Blackwell, Daniel Zheng, Nenad Tomašev, Richard Tanburn, Peter Battaglia, Charles Blundell, András Juhász, Marc Lackenby, Geordie Williamson, Demis Hassabis, and Pushmeet Kohli. Advancing mathematics by guiding human intuition with AI. *Nature*, 600:70–74, 2021. doi: 10.1038/s41586-021-04086-x.

Persi Diaconis, Sharad Goel, and Susan Holmes. Horseshoes in multidimensional scaling and local kernel methods. *Annals of Applied Statistics*, 2(3):777–807, 2008. doi: 10.1214/08-AOAS165.

Nelson Elhage, Neel Nanda, Catherine Olsson, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. transformer-circuits.pub/2021/framework.

Joshua Engels, Eric J. Michaud, Isaac Liao, Wes Gurnee, and Max Tegmark. Not all language model features are one-dimensionally linear. In *International Conference on Learning Representations*, 2025. arXiv:2405.14860.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah D. Goodman. Finding alignments between interpretable causal variables and distributed neural representations. In *Causal Learning and Reasoning*, 2024. arXiv:2303.02536.

Siavash Golkar, Mariel Pettee, Michael Eickenberg, Alberto Bietti, Miles Cranmer, Geraud Krawezik, François Lanusse, Michael McCabe, Ruben Ohana, Liam Parker, Bruno Régaldo-Saint Blancard, Tiberiu Tesileanu, Kyunghyun Cho, and Shirley Ho. xVal: A continuous numerical tokenization for scientific language models. *arXiv preprint arXiv:2310.02989*, 2023.

Andrey Gromov. Grokking modular arithmetic. *arXiv preprint arXiv:2301.02679*, 2023.

Wes Gurnee, Nicholas Sofroniew, Adam Pearce, et al. Verbalizable representations form a global workspace in language models. *Transformer Circuits Thread*, 2026. transformer-circuits.pub/2026/workspace.

Baran Hashemi, Roderic Guigo Corominas, and Alessandro Giacchetto. Can transformers do enumerative geometry? In *International Conference on Learning Representations*, 2025. arXiv:2408.14915.

Benjamin Heinzerling and Kentaro Inui. Monotonic representation of numeric attributes in language models. In *Annual Meeting of the Association for Computational Linguistics*, 2024. arXiv:2403.10381.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (GELUs). *arXiv preprint arXiv:1606.08415*, 2016.

Mark O. Hill and Hugh G. Gauch, Jr. Detrended correspondence analysis: an improved ordination technique. *Vegetatio*, 42:47–58, 1980. doi: 10.1007/BF00048870.

Subhash Kantamneni and Max Tegmark. Language models use trigonometry to do addition. *arXiv preprint arXiv:2502.00873*, 2025.

Subhash Kantamneni, Ziming Liu, and Max Tegmark. How do transformers model physics? Investigating the simple harmonic oscillator. *Entropy*, 26(11):997, 2024. doi: 10.3390/e26110997.

Ouail Kitouni, Niklas Nolte, Víctor Samuel Pérez-Díaz, Sokratis Trifinopoulos, and Mike Williams. From neurons to neutrons: A case study in interpretability. In *International Conference on Machine Learning*, 2024. arXiv:2405.17425.

Pablo Lemos, Niall Jeffrey, Miles Cranmer, Shirley Ho, and Peter Battaglia. Rediscovering orbital mechanics with machine learning. *Machine Learning: Science and Technology*, 4:045002, 2023. doi: 10.1088/2632-2153/acfa63.

Amit Arnold Levy and Mor Geva. Language models encode numbers using digit representations in base 10. In *Nations of the Americas Chapter of the Association for Computational Linguistics*, 2025. arXiv:2410.11781.

Zelong Li, Jianchao Ji, and Yongfeng Zhang. From Kepler to Newton: Explainable AI for science. *arXiv preprint arXiv:2111.12210*, 2021.

Ziming Liu, Ouail Kitouni, Niklas Nolte, Eric J. Michaud, Max Tegmark, and Mike Williams. Towards understanding grokking: An effective theory of representation learning. In *Advances in Neural Information Processing Systems*, 2022. arXiv:2205.10343.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. arXiv:1711.05101.

Giovanni Luca Marchetti, Christopher J. Hillar, Danica Kragic, and Sophia Sanborn. Harmonics of learning: Universal Fourier features emerge in invariant networks. In *Conference on Learning Theory*, 2024. arXiv:2312.08550.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, 2022. arXiv:2202.05262.

Gabriela Moisescu-Pareja, Gavin McCracken, Harley Wiltzer, Colin Daniels, Vincent Létourneau, Doina Precup, and Jonathan Love. On the geometry and topology of representations: The manifolds of modular addition. In *International Conference on Learning Representations*, 2026. arXiv:2512.25060.

Depen Morwani, Benjamin L. Edelman, Costin-Andrei Oncescu, Rosie Zhao, and Sham Kakade. Feature emergence via margin maximization: Case studies in algebraic tasks. In *International Conference on Learning Representations*, 2024. arXiv:2311.07568.

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

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. In *International Conference on Learning Representations*, 2023. arXiv:2301.05217.

Yaniv Nikankin, Anja Reusch, Aaron Mueller, and Yonatan Belinkov. Arithmetic without algorithms: Language models solve math with a bag of heuristics. In *International Conference on Learning Representations*, 2025. arXiv:2410.21272.

Jacob Pfau, William Merrill, and Samuel R. Bowman. Let's think dot by dot: Hidden computation in transformer language models. In *Conference on Language Modeling*, 2024. arXiv:2404.15758.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint arXiv:2201.02177*, 2022.

Federico Tiblias, Irina Bigoulaeva, Jingcheng Niu, Simone Balloccu, and Iryna Gurevych. Hypothesis-driven feature manifold analysis in LLMs via supervised multi-dimensional scaling. *Transactions on Machine Learning Research*, 2026. arXiv:2510.01025.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023.

Keyon Vafa, Peter G. Chang, Ashesh Rambachan, and Sendhil Mullainathan. What has a foundation model found? Using inductive bias to probe for world models. In *International Conference on Machine Learning*, 2025. arXiv:2507.06952.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017. arXiv:1706.03762.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, 2020.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *International Conference on Learning Representations*, 2023. arXiv:2211.00593.

Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tie-Yan Liu. On layer normalization in the transformer architecture. In *International Conference on Machine Learning*, 2020. arXiv:2002.04745.

Yifan Zhang, Wei Bi, Kechi Zhang, Dongming Jin, Jie Fu, and Zhi Jin. Weights to code: Extracting interpretable algorithms from the Discrete Transformer. In *International Conference on Machine Learning*, 2026. arXiv:2601.05770.

Maozhang Zheng, Jianjun Luo, and Zhaohui Dang. Machine learning-based solution of Kepler's equation. In *Third International Conference on Computer Science and Communication Technology (ICCSCT 2022), Proc. SPIE 12506*, 2022. doi: 10.1117/12.2661776.

Ziqian Zhong, Ziming Liu, Max Tegmark, and Jacob Andreas. The clock and the pizza: Two stories in mechanistic explanation of neural networks. In *Advances in Neural Information Processing Systems*, 2023. arXiv:2306.17844.

Tianyi Zhou, Deqing Fu, Vatsal Sharan, and Robin Jia. Pre-trained large language models use Fourier features to compute addition. In *Advances in Neural Information Processing Systems*, 2024. arXiv:2406.03445.

Fangwei Zhu, Damai Dai, and Zhifang Sui. Language models encode the value of numbers linearly. In *International Conference on Computational Linguistics*, 2025. arXiv:2401.03735.

A Further related work

Number codes in LLMs. Linear number codes are not a small-model artifact. Numeric values read out along linear or monotonic directions in LLM activations (Zhu et al., 2025; Heinzerling & Inui, 2024), with digit-wise base-10 structure (Levy & Geva, 2025).

Oracles and purpose-built architectures. Davies et al. (2021) belong to the oracle line of §2: attribution on the model's predictions, not its weights, guides conjectures in knot theory. Hashemi et al. (2025) carry probing to research-level mathematics. Zhang et al. (2026) design an architecture, the Discrete Transformer, from which algorithms can be extracted automatically. Our account differs from all three in the same way: the algorithm is read by hand, from the weights of a standard transformer.

B Setup details

Why input digits? We could feed the two floats directly as continuous tokens (each float scaled onto one learned embedding direction, as in xVal; Golkar et al., 2023) and still have a transformer. But with only two tokens and no digit vocabulary, attention has almost nothing to route and the number-line embedding of §4.1 would not exist. Production LLMs also receive numbers as digit or digit-group tokens, so our embedding geometry is comparable (Zhu et al., 2025; Levy & Geva, 2025; Kantamneni & Tegmark, 2025). The 12 digits are also far finer than the model resolves (App. E), so accuracy is limited by the model, not by how the inputs are written.

Why output scalar? Digits out would need autoregressive decoding and a cross-entropy loss, and we did not pursue either. Cross-entropy charges the same for a wrong leading digit as for a wrong trailing one, though the E-error differs by $\sim 10^{11}\times$. A scalar loss weights by the true angular error (MSE on the primary).

Architecture details. The vocabulary has 11 tokens: digits 0–9 and ANS. Learned token and position embeddings are summed at the input. Each layer applies $x = x + \mathrm{attn}(\mathrm{LN}(x))$ then $x = x + \mathrm{mlp}(\mathrm{LN}(x))$. Attention is causally masked. The MLP carries biases, attention does not. There is no weight tying; the readout head is its own d_model$\to$1 map. At initialization, linear and embedding weights draw from $N(0, 0.02^2)$, biases start at zero, and LayerNorm gains at one.

Optimizer. Training uses AdamW (Loshchilov & Hutter, 2019; betas 0.9/0.98), peak LR 1e-3 with a 1000-step linear warmup then cosine decay to zero over the full horizon (800k steps for the primary), weight decay 0.01, and MSE loss in normalized E-space.

Eval grid density and jitter. A naive grid would hit exact corners and round fractions whose digits end in long runs of zeros. Jitter makes every digit pattern natural. The jitter is one offset per column and one per row, not one per point, so a statistic taken down a column or across a row reads at one input. The grid is denser in M because the structure we measure oscillates along M (harmonics out to $n \approx 30$) while the e-dependence is smooth away from the cusp.

Table A1: **The primary and the family's worst digit code on each battery metric.** Each row after the primary is the family model with the worst value in its bolded column. Even the worst codes stay open and ramp-like. Four models exceed the 0.42 circle calibration on excess (the ReLU control's first seed and one primary-configuration seed at its three shortest horizons), all open arcs.

model	L	p_L	|Pearson|	excess	ramp_dev	openness
primary	0.84	2×10^{-22}	0.99	0.00	0.02	5.6
ReLU MLP (s0)	0.40	4×10^{-5}	**0.62**	**0.54**	0.03	3.3
d32 at 200k (s0)	**0.31**	3×10^{-9}	0.81	0.18	0.06	1.9
sigmoid (s3)	0.37	$\mathbf{1.6 \times 10^{-4}}$	0.80	0.14	0.05	1.9
d128 at 400k (s3)	0.40	1×10^{-53}	0.81	0.16	**0.11**	2.7
d32 at 50k (s0)	0.36	5×10^{-12}	0.93	0.02	0.05	**1.8**

C THE NUMBER-LINE BATTERY

A number line is ordered, evenly spaced, straight, and open. Five standard statistics test these properties, one job each. Three are basis-free and test the family claim, line-dominated and open: a regression R^2 (L), a Fourier-spectrum goodness-of-fit (ramp_dev), and an endpoint ratio (openness). Two are PC-anchored and measure deformation: a correlation (Pearson, even spacing and rotation) and an incremental R^2 (excess, curvature). The calibrations come from ideal geometries, not tuning. No single statistic settles the claim. Together they support it. High L with high openness, for example, is also satisfied by a spiral, which ramp_dev rules out.

This battery runs opposite to persistent homology. Persistent homology detects closed structure by counting loops and voids, and recent work uses it to sort modular-addition representations into disc, circle, and torus across hundreds of models (Moisescu-Pareja et al., 2026). It gives no positive evidence of a line. An open arc is contractible, so its first homology is trivial, the same reading as no structure at all.

- **L**, the fraction of embedding variance explained by digit value ($L = \|X_c\hat{v}\|^2/\|X_c\|^2$, with X_c the centered digit embeddings as columns and $\hat{v}$ the unit centered digit-value vector), detects linear structure regardless of which PCs carry it. Its significance, p_L, is the upper tail of the closed-form structureless null $\mathrm{Beta}(d/2, (N-2)d/2)$. Together they rule out a code with no linear value structure. L cannot separate line from circle, since a round clock at $N = 10$ sits at 0.32 and an eccentric ellipse reaches ≈ 0.63; that is the job of openness.
- **ramp_dev**, the max deviation of the normalized digit Fourier spectrum from the ideal $1/\sin(\pi k/N)$ ramp, tests for a smooth magnitude ramp. Ramp = 0.00, flat/random = 0.19, round clock = 0.61. It rules out spirals and multi-frequency codes that excess could miss.
- **openness** = $\|\mathrm{emb}(0) - \mathrm{emb}(9)\|$ / mean adjacent step, the endpoint gap in adjacent-step units. Open $\gg 1$, round clock ≈ 1. It rules out a circle, where 9 wraps back to 0.
- **Pearson** on PC1 tests even spacing, not just order. It rules out a monotonic warp such as a log-like code, but only where PC1 carries the line, which is why the family claim in §4.1 rests on the three basis-free statistics. A drop in Pearson signals rotation or curvature. The four lowest are the curved four of Fig. A1.
- **excess** = $r^2(\mathrm{value} \mid \mathrm{PC1{+}PC2}) - r^2(\mathrm{value} \mid \mathrm{PC1})$ asks whether recovering value needs a second dimension. Straight line 0.00, circle ≈ 0.42, parabola-arc ≈ 0.97. It measures curvature, including a collapsed ellipse, but it too is conditional on PC1 carrying the value.

On the primary, PC1 carries the value axis and almost none of its curvature (quadratic share 0.005, cubic 0.002). PC2 and PC3 together span 98% of the quadratic mode and 79% of the cubic mode. That is the *Guttman ("horseshoe") effect*, which a single 1-D ordering produces whenever the embedding curves along it (Diaconis et al., 2008; Hill & Gauch, 1980). We read the curvature in the PC1-vs-PC2 plot (Fig. A1a) as PC1's shadow, not independent structure. Which of PC2 and PC3 carries which mode is near-degenerate, not stable across the family.

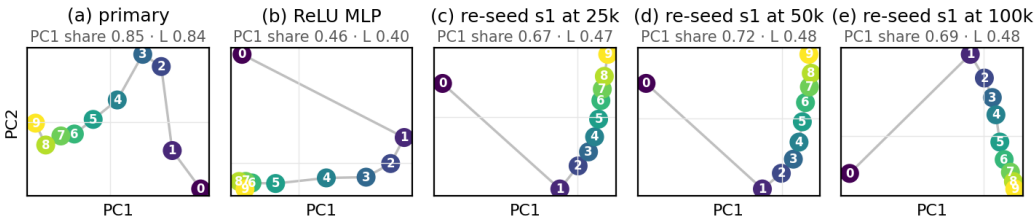

Figure A1: **The four most curved digit codes stay open, ordered arcs** (PC1 vs PC2 of the digit embeddings). The primary and the four models above the 0.42 circle calibration on excess, also the four lowest |Pearson|. The PC1 share drops to 0.46–0.72 while L holds at 0.40–0.48, well above the 0.11 structureless expectation: the value axis bends into PC2 rather than vanishing. The primary's own bend is the Guttman effect described above: PC1 holds a 0.005 quadratic share and the excess is 0.00. ▼ marks the same five models in the histograms of Fig. 3.

Table A2: **Wrap study configurations and results** (five seeds per cell). The aligned cells keep the line at the primary's accuracy; the incommensurate wrapped cell loses both and ignores e. Each cell gives the battery verdict of App. C and the seed-median error as a ratio to the primary's five-seed median, 1.85e-3 rad. The unwrapped column compares at matched relative precision, error over output range, since its output scale grows with the range. The wrapped column shares the primary's output scale.

	target	
M range	E (unwrapped)	$E \bmod 2\pi$ (wrapped)
$\pm 10\pi$ (digit-aligned)	line; 1.2×	line, one seed fails; 1.1×
± 50 rad (incommensurate)	line; 2.8×	**dissolves; 81×**

Rotation and dilution are benign; curvature is the exception. Eighty-one of the 210 fall below |Pearson| = 0.95, bottoming at 0.62, the four curved codes among them. Every model at d16 and wider dilutes below the primary's L = 0.84, running 0.31 to 0.79, the width trend of Table A7. In both cases the claim rests on the three basis-free statistics, clean everywhere. Low L also means most variance sits off the value axis. In the worst case, d32 at 200k, 69% sits off-axis. Curvature is not a width trend. The four curved codes are all d8: one seed of the primary configuration at its three shortest horizons, and the ReLU control. Milder curvature, excess 0.15 to 0.42, appears in 25 models scattered across every width and horizon.

D WRAP STUDY DETAILS

To test whether periodicity encourages a circle, we retrain the primary configuration on a 2×2 of M range and target form, five seeds per cell (§4.2; Table A2).

M-range alignment. The tokenizer writes M as twelve decimal digits of its normalized value $(M + R)/(2R)$ on a range $\pm R$. If the range spans 10^k rotations, one period is a decimal fraction of the input. At $\pm 10\pi$, ten rotations, the first digit is the rotation count and the remaining eleven are the phase. At ± 50 radians every digit place advances the phase by an irrational fraction of a rotation, so the phase is a sum mod 2π across all places.

Target form. The unwrapped target $E = E_0 + 2\pi k$ is the canonical continuation of the solution (Boyd, 2007, Thm. 1(2)). It is a ramp plus a small periodic correction, which is why the periodic part is about one percent of the output at ± 50 rad. The wrapped target reduces E to $[-\pi, \pi)$; we write it $E \bmod 2\pi$. It is a sawtooth in M, discontinuous at the rotation seams. Each target is then normalized to $[0, 1)$ over its own range as in §3.1.

Embedding geometry. The battery of App. C runs on every cell. Both unwrapped cells pass as line-dominated and open, L 0.38–0.77. The aligned wrapped cell passes in four of five seeds. One

seed fails the L floor at 0.24 and shows the largest curvature excess in the 240, 0.89. The next lowest seed passes at 0.32, and both stay open. The incommensurate wrapped cell fails in all five seeds: L 0.07–0.23 with p_L up to 0.79, openness 1.04–1.48, and ramp deviation 0.14–0.20 against the 0.19 flat null. No circle forms either. A circle would pair openness near 1 with a sharp spectral spike, and the spectrum carries no spike with the PCA scrambled. The code is neither line nor circle, and we do not resolve what it is. These models ignore e (e_dep 0.01 to 0.04 against the primary's 0.21, and attention's write along e has norm at most 0.0002 against 0.014), so there is no register to grade and both the share and the scaling check are uninformative. One incommensurate unwrapped seed ignores e too.

Mechanism. Every cell computes the wrap. Error per rotation is flat, max over min 1.0 to 2.4 across interior rotations, and the error field is periodic in M, its phase profile correlating 0.72 to 0.99 with the central rotation's. The cells differ at neuron grain. In the aligned wrapped cell the MLP neurons are phase functions (phase R^2 0.97 to 0.99), because the phase is the shifted digit string. In the incommensurate unwrapped cell they are local structure on the M line, not phase functions (R^2 0.01 to 0.07 against a locality null of 0.06). Circle gain is the extra variance a periodic component at the predicted frequencies explains in the layer-0 OV write beyond a line fit, summed over the two leading M places. In the incommensurate wrapped cell it stays between 0.51 and 1.26 at every saved snapshot from 50k to 800k steps in all five seeds. The same statistic on the primary's number line, the line-code null, is 0.23. Phase R^2 stays between 0.10 and 0.26 over the same snapshots, so the periodic structure of §4.2 never becomes a phase representation.

E POSITION EMBEDDINGS AND READ DEPTH

Position embeddings encode significance, but only for the leading places. The input fields are fixed-width, so each absolute position pins one place of one field (M at positions 0–11, e at 12–23). In each field the leading places get their own well-separated position vectors. Past the model's read depth, the deeper places collapse to one cluster at the noise scale (Fig. A2a,b). On the primary, PCA separates five leading M places from the noise cluster (distances 0.56, 0.52, 0.31, 0.09, 0.011) and four e places (0.55, 0.39, 0.14, 0.012), so M resolves one place deeper. Across the family this gap is one place in 148 of the 194 models with measurable cluster structure. Sixteen models never form the cluster, most in the d4 corner or at short horizons. The collapse is an at-convergence structure: 57 of 60 models at the 800k horizon show it, and short-horizon models show it still forming.

Digit sensitivity, the behavioral read-depth measurement. We vary one input digit over 0–9 with the rest held fixed, take the std of the predicted E over the ten settings, then average over 4096 random (M, e) inputs. Sensitivity falls $\sim$10$\times$ per decimal place (M places: 2.1, 0.18, 0.018, 1.8e-3 radians) until it reaches a $\sim$2e-4 floor (Fig. A2c,d). The slope says nothing about the circuit. A digit at place k moves the true E by $\sim 10^{-k}$ of its range, so any solver accurate at those places shows the same falloff. Where the sensitivity hits its floor is the model's behavioral read depth. That depth sits one place short of where the position geometry collapses, so the weights carry slightly more place structure than the behavior uses. The two reads agree on the pattern: reading stops far short of place 11. §4.3 shows the mechanism.

One rule measures depth at every stage. A place is resolved when its signal clears five times the tail reference, the deepest six places of its field; depth is the length of the leading resolved run. Geometry reads position-vector distances against the tail cluster's noise, attention reads mass in excess of the softmax's common floor, and behavior reads the sensitivity profile against the tail floor. A field whose deep places never form a cluster gets no geometry depth. The gap distributions keep the same mode with the factor set to three or eight. Fig. A3 gives the family distributions; absolute depths track accuracy, so the gap carries the claim. The gap is one place in 148 of 194 models by geometry, 121 of 210 by attention, and 205 of 210 by behavior. Geometry and attention resolve places more coarsely than behavior. Attention mass falls only $\sim$3$\times$ per place (below), and both instruments' misses land mostly at zero and two places, on either side of one.

Position vectors set a per-place gain. Attention alone cannot explain the sensitivity falloff. The pattern separates only the leading M places, and its mass falls $\sim$3$\times$ per place, not 10$\times$ (Fig. 4a). The position table supplies the missing factor, through LayerNorm. The layer's input at place p is LayerNorm(token + position), which couples the two tables: each distinct position vector sets a per-place gain on the digit direction. With LayerNorm removed the gain is flat across places,

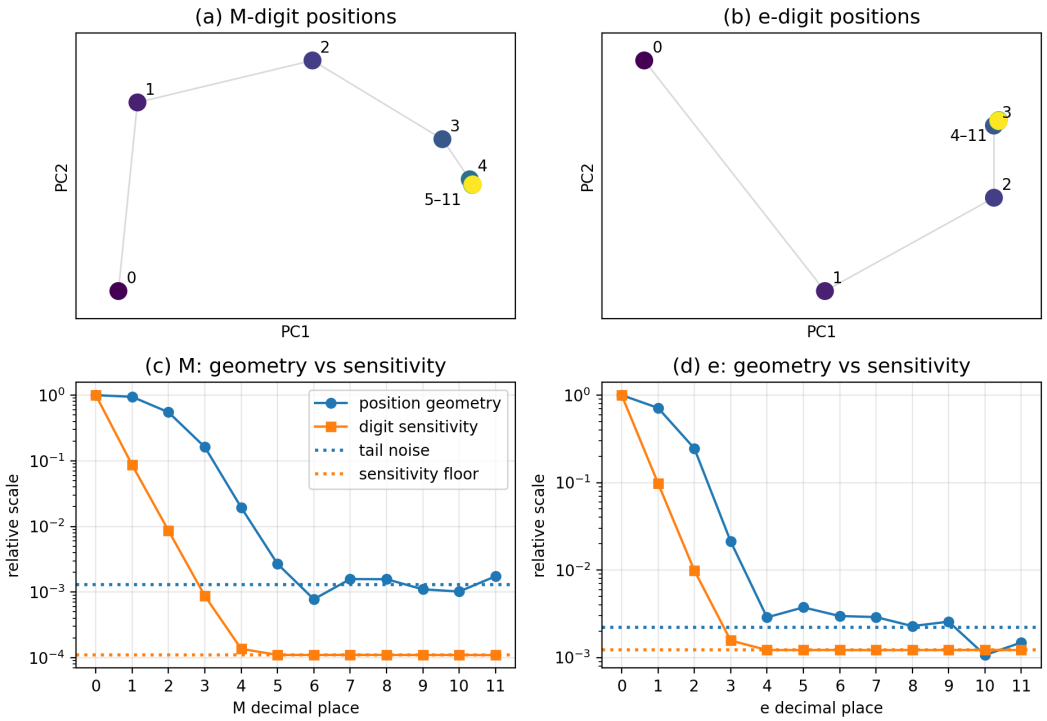

Figure A2: **Position embeddings separate only the leading places** (primary specimen). **(a,b)** PC1 vs PC2 of the M- and e-digit position embeddings. Leading places are distinct and ordered; all deeper places collapse to one cluster at the $\sim$0.001 noise floor. **(c,d)** Per-place distance of the position vector from the collapsed cluster (blue, weights-only) against the behavioral digit sensitivity (orange), each normalized to its leading place, log scale. Dotted lines mark each curve's floor; read depth is where a curve hits its floor. Both stop far short of place 11, and both put M one place deeper than e. The geometry extends one place deeper than the sensitivity.

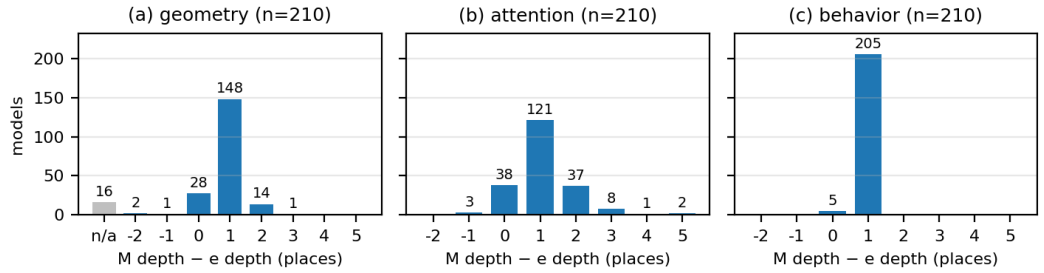

Figure A3: **The $M - e$ read-depth gap across the family, by instrument** (210 models; depth rule in the text). **(a)** Position geometry; gray marks the 16 models whose deep places never form the cluster. **(b)** Attention mass. **(c)** Behavioral digit sensitivity.

by linearity. A first-order account from the weights alone, attention times gain projected onto the output, reproduces the measured sensitivity at every resolved place to within about $2\times$. The floor itself is beyond first order.

The M/e asymmetries are the equation's calculus. Implicit differentiation of $M = E - e\sin E$ gives $\partial E/\partial M = 1/(1 - e\cos E)$ and $\partial E/\partial e = \sin E/(1 - e\cos E)$, so per unit input e moves E less than M by the factor $\sin E$, whose magnitude averages $\sim$0.6. M's range adds a second factor: 2π against e's $\sim$1, $\sim$6.3$\times$. The two factors combine to $2\pi/0.6 \approx 10$: the dependence ratio of §4.3 (M_dep 2.06 vs e_dep 0.209 on the primary) and one decimal place of read depth.

Table A3: **Mean-ablation shows head 1 is the reader on the primary** (§4.3). The −MLP column backs §4.4.

quantity	baseline	−head 0	−head 1	−MLP
M_dep (rad)	2.06	1.90	0.18	1.48
e_dep (rad)	0.21	0.20	0.12	0.09
e_corr	1.00	0.94	0.02	−0.41
M_corr	1.00	0.99	0.82	0.99
median error (rad)	1.5e-3	0.27	1.8	0.61

F READER-HEAD DETAILS

The M_dep/e_dep pair of §4.3 is an exact variance split by the law of total variance: the residual is orthogonal to the e-averaged curve, so the surface variance is M_dep^2 + e_dep^2. The residual holds the e main effect and any $M \times e$ interaction.

How Fig. 4 is computed. The attention pattern separates the leading places, four for M and three for e by the App. E depth rule; deeper places share a common floor (Fig. 4a). The leading M digit's attention moves a third as much as the leading e digit's and in opposition (corr -0.97). When the e digit's weight rises, the softmax takes the mass mostly from the M digit. The OV read is pure weights: the digit's token embedding is passed through $W_V W_O$ and projected onto $w_{\text{eff}} = w_{\text{head}} \circ \gamma_{\text{ln}}$, the readout head weights scaled elementwise by the final LayerNorm gain. The weights-only read is linear, so a position embedding adds only a constant offset: each head has one transfer curve, shared by the two fields. In the live model the per-place weighting comes from the attention mass and the LayerNorm place gain (App. E), not from the OV shape. This weights-only cross-check agrees with the ablations: head 1's transfer curve spans 0.005 of a logit, head 0's 0.001 (Fig. 4c). Spans are in w_{eff} units, uncalibrated by LayerNorm's per-input scale (§4.4), so only their ratio is meaningful.

The family's transfer curves. Across the family the head-summed transfer curve is monotone at median |Spearman| 0.99, and 57 of 210 fall below 0.95. Those 57 skew wide: 26 of the 60 models at d64 and d128 read softly, against 31 of the 150 at d32 and below. A clean line (embedding |Spearman| = 1.00) gives a monotone read (summed-curve |Spearman| ≥ 0.95) in 90 of 98 models; a deformed line does in 63 of 112 (odds ratio 8.8, Fisher exact p = 3e-9).

Head organization. We classify each model from its mean head-ablation sweep. Each dependence, M_dep and e_dep, is assigned to the head whose ablation removes the most of it, provided the removal exceeds 30%. One head owning both is a single reader, different heads is double, and a dependence no head owns makes the model distributed. The output-map and architecture controls match the primary's accuracy (five-seed median error 0.74 to 1.00× the primary configuration's five-seed median) and between them land in all three classes. The primary configuration's own five seeds also cover all three: two single, two distributed, one double.

G REGISTER AUDIT AND FIT DETAILS

The absorbers. The true coefficient on each $\sin(nM)$ is a Bessel function of e (§4.6), curved in e. On $\sin M$ it is $2J_1(e) = e - e^3/8 + \ldots$. The library offers only $e\sin(nM)$ and $\sin(nM)$ for it, a line in e, and the best line through a curve has a nonzero intercept, which lands on the pure sine. The concave $\sin M$ coefficient gives a positive intercept and the convex $\sin 2M$ coefficient a negative one (Table A4). The series has no $M{\cdot}e$ term either, since $E - M$ is periodic in M. The library terms are not orthogonal on the evaluation grid, so the fit returns one for any accurate solver. The exact solution, fit the same way, carries the same intercepts and the same 0.06 on $M{\cdot}e$ (Table A4). The same non-orthogonality makes every coefficient a partial regression coefficient (corr(M, $\sin M$) $\approx$ 0.78), not a stand-alone contribution.

The exact split. Remove the mean from the readout direction, $\tilde{w} = w_{\text{eff}} - \text{mean}(w_{\text{eff}})\,\mathbf{1}$, and divide each write's projection by the per-input scale, $a_c(x) = \tilde{w} \cdot c(x)/\sigma(x)$. The rows and the

Table A4: **Output and truth fits** (primary specimen; the same seven of 11 terms as Table 2, in radians of E). The output row fits the model's prediction itself, the residual stream passed through the actual LayerNorm and denormalization. The truth row is the same fit of the exact solution. They match to within 0.001 in all 11 terms. No single scale maps the component rows of Table 2 onto these rows, since the per-input scale runs from 0.0096 to 0.14 over the grid. The exact split is Table A5.

fit	R^2	M	$\sin M$	$\cos M$	$\sin 2M$	$e \sin M$	$e \sin 2M$	$M{\cdot}e$
output (rad)	0.998	0.988	0.048	0.000	−0.065	**0.767**	**0.441**	0.062
truth (rad)	0.998	0.988	0.047	0.000	−0.065	**0.768**	**0.441**	0.061

Table A5: **Direct logit attribution with the scale kept in** (primary specimen; logit units; the rows and the offsets sum to the logit). R^2 is each row's own 11-term fit. The shared scale puts input dependence into the constant embedding's row and raises attention's share of the fitted $e \sin(nM)$ content from 2.3% to 33%.

row	R^2	M	$\cos M$	e	$e \sin M$	$e \sin 2M$
emb	0.888	0.016	−0.043	**0.044**	−0.002	−0.003
attn0	0.988	0.085	0.069	−0.048	**0.046**	0.015
mlp0	0.987	0.057	−0.026	0.005	0.079	0.058
sum (logit)	0.998	0.157	0.000	0.000	0.122	0.070

offsets reconstruct the logit, $z = b + \sum_c a_c$, to 2e-7 (Table A5). Centering alone changes attention's rms-content share of the $e \sin(nM)$ signal only slightly, from 2.4% to 2.3%. Applying the cached scale to the centered projections raises it to 33%. The scale varies with e and with $\cos M$, so every row acquires e and $\cos M$ content, the constant embedding's included. The fitted pure-e and cosine coefficients of the summed logit round to zero, consistent with the target's odd symmetry after the normalization offset is subtracted. Table 2 is what each write puts on the readout direction. These rows are what reaches the logit once the shared scale is applied. Neither is causal. The ablations of §4.4 are.

The completeness distribution. The patch removes the product terms completely in most models and partially in some. The surviving $e \sin M$ coefficient runs from an overshoot past zero to 55% of baseline. Sixteen of the 210 leave more than 20% in magnitude: six d16 and six d64 MSE runs, one ReLU seed, two d128 runs, and one tanh seed. Where removal completes it often overshoots past zero, since the patch subtracts the whole fitted direction and the fit is linear in e.

The 2-layer split. All five 2-layer seeds fit a high-share direction in layer 0 (share 0.94 to 0.99). In two seeds a layer-0 patch zeroes the products, the surviving $e \sin M$ below 7% of baseline, and the zero-scale patch matches the model's own $e = 0$ behavior to 0.05 to 0.06 radians, inside the family range. However, in the other three, 78% to 99% of $e \sin M$ survives the patch, which is the failure, and both patches land at 0.82 to 1.01 of the do-nothing gap.

Head ownership follows the read. Per head, the share of the e-read (the Fig. 4b e-sensitivity, summed over the e digits) and the share of the e-write (the per-head norm of the write's e term) correlate at 0.92 across the family's 205 two-head models. In the 4-head control the read-share and write-share vectors match at cosine 0.82 to 0.98 in all five seeds. The head that reads e writes it, and distributed readers have distributed writers. The three reversals, read and write on different heads at cosine below 0.5, are d4 seed 3 at 25k, 50k, and 100k steps.

Visibility across the family. The alignment $\cos(\hat{u}_e, w_{\text{eff}})$, with $\hat{u}_e$ the register direction, has a family median of 0.16, with no monotone trend in width or steps. Only the tail moves with training: six models exceed 0.5, all at $\leq$100k steps, the largest 0.86 at d4, and every model trained to 800k stays at or below 0.35. Centering w_{eff}, as the final LayerNorm does, leaves the primary's cosine at 0.04; the two readout directions agree at cosine 0.985.

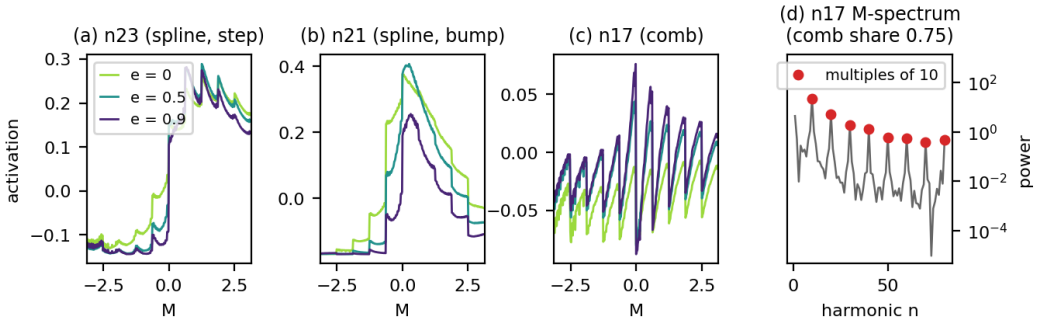

Figure A4: **Inside the MLP: the spline family and the digit comb** (primary specimen; post-GELU activation against M at e = 0, 0.5, and 0.9 on a clean uniform-M grid). **(a,b)** The two largest-contribution smooth neurons, a soft step and a bump in M. Both tilt with e and already fire at $e = 0$. **(c)** The comb corrector n17, a sawtooth at the second M digit's period. A smaller sawtooth shows on the spline neurons too. **(d)** n17's M-spectrum. Power concentrates at harmonic multiples of ten, the digit comb, share 0.75 (§4.4).

Fit sufficiency and the linearity bound. The library's residual against the model's output is 0.027 radians median, 18× the model's own median error against truth. The model carries the high-order tail the 11 terms cannot (§4.6). Replace the attention write with its fitted version and the structure survives whole: e_dep and M_dep unchanged, coefficients within 7% of baseline. Only precision drops (median error 1.5e-3 → 5.3e-2), carried by the 3% of the write the fit discards. The same linearity bounds the zero-scale patch. It removes only linear-in-e content, so deeper e-reads survive and the zeroed model matches the model's own $e = 0$ behavior to 5.8e-2 median, not to full precision. The subtracted content is the pure e term along the register, constant in M by construction, and the residual it leaves concentrates on $M > 0$ (Fig. 1b).

The search cross-check. Distributed alignment search asks what regression cannot: is there a better single direction? We optimize a unit direction in the frozen model against an interchange objective. Swap the write's coordinate along the candidate direction with a source input's coordinate at eccentricity e', and score the output against the model's own behavior at e'. Five random initializations converge to one direction per model. With sources that share the base M, the search lands on the register's channel in all seven models tested: the primary, its four re-seeds, a no-final-LayerNorm twin, and the d128 model. The found direction reads through the same top MLP neurons as $\hat{u}_e$, at cosine 0.92 to 0.99, and scales e better. With fully random sources the task is stricter. The coordinate along $\hat{u}_e$ mixes in M content ($\cos(\hat{u}_e, \hat{u}_M) = 0.25$, with $\hat{u}_M$ the write's M-content direction). On the primary it scores 1.3× worse than the do-nothing floor. Across the seven models it lands between 4× better and 11× worse. The search then finds a compromise that sheds the M overlap, still reading through the same neurons, at cosine 0.59 to 0.89 to $\hat{u}_e$, and width does not close the gap (0.84 at d128). Regression finds the direction along which e is written. The search finds the best single axis to read and write e through, and in a stream whose term directions are not orthogonal the two differ.

Spline and comb detail. A neuron is live by contribution when $|c| \cdot \sigma \geq 1\%$ of the largest, with c the neuron's weight along the readout and σ its activation std. 11 of the primary's 32 qualify. The comb sawtooth has period 0.1 in normalized M, holding 75% of its tuning power at harmonic multiples of ten. Mean-ablating it makes the residual blockier as well as larger: the share of the bulk residual's M-spectrum power at the comb harmonics, multiples of ten, rises from 0.105 to 0.141. The residual's own comb is weaker along e than along M (spectral excess 1.8 against 2.5 at the digit harmonics), consistent with the register carrying e near-linearly into the correction. The blocks mark where the model uses digits nonlinearly.

No live neuron is a single Bessel term. The log-log slope of dominant-harmonic amplitude against e, fit on e in (0.05, 0.40), runs -0.14 to 0.25 across the 11, where a single Bessel term has slope n, at least 1. At the smallest grid e, 0.002, every live neuron still varies over M, with std 0.48 to 1.30 times its std over the whole grid. A single Bessel term would be flat there.

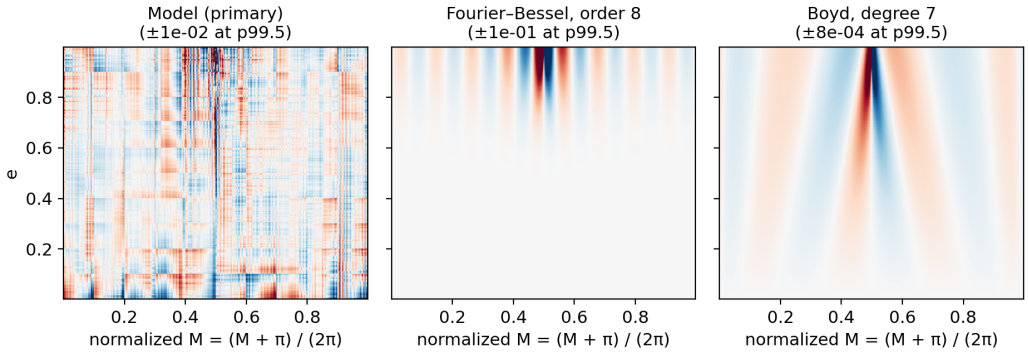

Figure A5: **The model's residual is not a truncated series** (primary specimen). Signed residual over the eval (M, e) grid, each panel self-scaled to its own p99.5 to view the pattern. Fourier–Bessel leaves a cusp-anchored comb, Boyd a smooth fan, and the model a fine block pattern with edges on round values of the normalized M and of e. The block pattern is the digit floor of the correction. Mean-ablating the model's strongest comb-corrector neuron raises error 12× and makes the pattern stronger (§4.4).

The no-final-LayerNorm twin. The number line and the register replicate in all five twin seeds; the spline and comb anatomy and the block residual are read from the first seed. In every seed the attention and MLP fits carry equal and opposite constant, $\cos M$, $\cos 2M$, and e coefficients, matched to 0.0001, so their sum is odd by itself. The neuron-grain comparison is per-neuron mean-ablation damage against the contribution score $|c| \cdot \sigma$ over each model's live neurons, 11 on the primary and 14 to 21 per twin seed. The 14× of §5 comes from the same tool: each live neuron's ablation is rerun with the final LayerNorm's per-input mean and variance at the readout position pinned to their unablated values, so the ablation reaches the head only as content along w_{eff}. For the tenth neuron by contribution the damage falls from 0.163 to 0.012 radians. For eight of the other ten live neurons the pinned damage exceeds the live damage, so the normalization ordinarily absorbs part of an ablation rather than carrying it.

H Classical placement details

The cusp derivation. As $M \to 0$ and $e \to 1$, Kepler's equation degenerates. Set $e = 1$ and expand the sine as its Taylor series, $\sin E = E - E^3/6 + E^5/120 - \ldots$, so $M = E - \sin E = E^3/6 - E^5/120 + \ldots$. For small E, $M \approx E^3/6$ and $E \approx (6M)^{1/3}$. A cube root has infinite slope at zero. The same divergence in general form is $dE/dM = 1/(1 - e\cos E)$, whose denominator vanishes exactly at $e = 1$, $E = 0$. No smooth approximant in M, series or network, can match an infinite slope, so every method concentrates its error at this corner (Fig. A5). That is why we score the bulk and the cusp separately (§3.3).

Deviation profile. The primary's output harmonics deviate from the true coefficients by relative RMS < 0.01 through $n \approx 5$, $\sim$0.02 at $n = 7$, rising to $\sim$0.18 by $n = 30$. Sine energy beyond $n = 30$ is 1.3e-4 and cosine energy 3.8e-5 of the sine energy, matching the odd-in-M parity of §4.4.

Two specific low-order approximations fall short of the model. One test is on d8 and one on the sigmoid control. For d8, the five-seed bulk worst case is 5× below the Fourier–Bessel truncation of matched median (Fig. 6), which rules out a low truncation. On the sigmoid specimen, fitting the 11-term library to the pre-sigmoid logit gives R^2 0.993. The logit has standard deviation 2.1, so the fit's RMSE is 0.17 logit units. The output map has derivative at most $\pi/2$. Through it, the surrogate's median error is 6.4e-2 radians against the model's 2.0e-3, 32× worse. A high R^2 on the logit does not guarantee output accuracy, and this logit-fitted surrogate does not retain the model's precision. At the cusp, where every method concentrates its error, the model's residual correlates at most 0.38 with Fourier–Bessel's and 0.20 with Boyd's.

Table A6: **Error-metric scaling exponents** (2-D power-law fits on the seed-median grid, 30 cells over d8–d128; R^2 0.93–0.96; standard errors ± 0.03 on the steps exponents, ± 0.02 on the params exponents). bulk-max is the worst case over the bulk, $e < 0.9$. The median row is bold; its exponents are the guides in Fig. A6.

metric	steps exp	params exp
median	$\mathbf{-0.43}$	$\mathbf{-0.42}$
bulk-max	-0.41	-0.34
max	-0.33	-0.27

Table A7: **Line share, MLP-ablation factor, and register share by width** (the 6-width $\times$ 6-horizon grid, 30 models per width; median with min–max). The number line's share of embedding variance L falls with width, the MLP-ablation factor (§4.4) grows as params$^{0.54}$ (R^2 0.99, fit on the five 800k cells, d8 to d128), and the register's share of the write's variance across e holds at 0.92 to 0.98.

width	line share L	MLP-ablation factor	register share
d4	0.68 (0.51–0.93)	101 (29–382)	0.98 (0.87–1.00)
d8	0.57 (0.38–0.84)	278 (57–559)	0.92 (0.87–0.97)
d16	0.57 (0.38–0.79)	317 (102–799)	0.94 (0.79–0.99)
d32	0.40 (0.31–0.59)	618 (168–2021)	0.94 (0.73–0.99)
d64	0.47 (0.38–0.55)	743 (174–5157)	0.93 (0.78–0.95)
d128	0.42 (0.33–0.74)	2611 (253–9421)	0.95 (0.85–0.97)

I SCALING DETAILS

Error is a separable power law in steps and params. All three error metrics fit at R^2 0.93 to 0.96 (Fig. A6; Table A6). Adding a steps$\times$params interaction term raises R^2 by 0.006 to 0.03. Floor-term fits return a negative floor for the median and bulk-max, and +4e-4 radians for the max, 4% of its best cell. d4 is excluded as the degenerate low end: its learning curve flattens early (Fig. A6b), and it hosts the worst register-visibility exception and all three head-ownership reversals (App. G).

The worst case improves more slowly than the typical case. The exponent falls from median to bulk-max to max (Table A6). The median and max exponents differ by 2.5 standard errors in steps and 5.6 in params, and the median stays steepest in every per-seed refit at 800k, so the median-to-max gap is not fit noise; the bulk-max rung is softer, swapping with max in two of five seeds. Why the worst case scales slower we do not know.

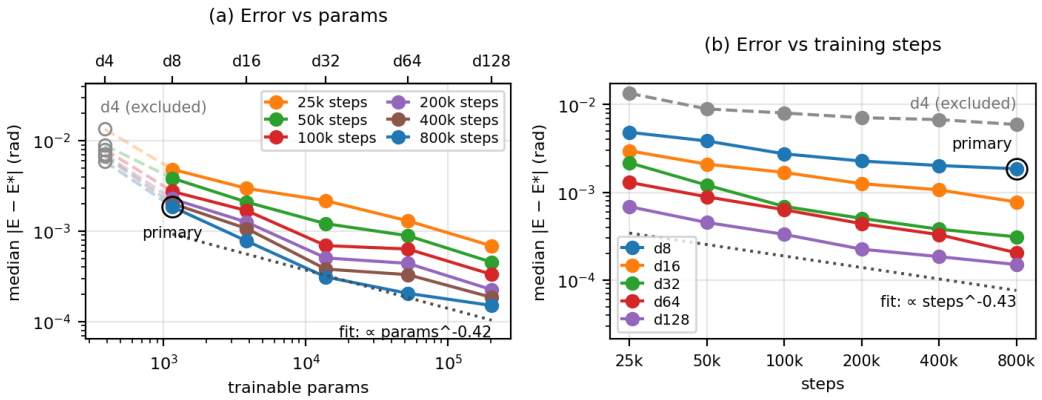

Figure A6: **Median error is a power law in steps and params** (the 6-width $\times$ 6-horizon grid, each point the median over five seeds). The hollow marker is the primary configuration, d8 at 800k. **(a)** median $|E - E^*|$ vs trainable params, one line per training horizon. **(b)** vs steps, one line per width. The d4 points are excluded from the fits. The dotted guides are the median exponents of Table A6, bold there.