ROUTING-GUIDED SINGLE-EXPERT INTERVENTIONS MEASURED AGAINST A MODEL'S OWN EXECUTION VARIABILITY

Anonymous authors
Paper under double-blind review

ABSTRACT

How large is the effect of changing one expert at one token in a sparse mixture-of-experts model, and how precisely can it be measured? We answer both questions by comparing the model with itself: five mathematically equivalent BF16 executions of the same tokens establish how much each measured quantity varies with execution, and an end-to-end FP64 reference measures the intervention at higher arithmetic precision. We apply this approach to Qwen3-30B-A3B solving linear systems. A linear detector on executed gate weights achieves 94% mean recall of annotated computation events at a 5% review budget per held-out trajectory. Its ranking is highly stable as expert membership changes across executions. A fixed rule on the same weights nominates the expert to remove. A frozen reader trained to predict the target digit measures the changed state; in same-state comparisons, its average position-score drop grows with the removed expert's natural weight. Attenuating the most frequently nominated expert improves local digit evidence while reducing the reader's weight on that position. These opposing contributions retain their average directions across executions. At the model's answer, median high-precision effect magnitudes are 0.02–0.04 nat against a median natural execution range of 0.50 nat. Single BF16 pairs report several-fold larger median magnitudes and agree with the reference in sign only slightly more often than the chance baseline. Raising only the precision of the final normalization and output head preserves most of that spread, which is therefore already present before these operations. Small gaps between competing router or output scores identify the choices most likely to change. Routing guides where to look and which expert to test, interventions reveal what changes, and execution comparisons make numerical precision part of experimental design.

1 INTRODUCTION

A sparse mixture-of-experts model records which experts it uses at every token and how much weight each receives (Shazeer et al., 2017; Fedus et al., 2022; Jiang et al., 2024). Can this record guide the study of computation inside a long reasoning trace? Once it points to an expert, how do we measure the effect of changing that expert at a single token? Routing finds a local intervention; equivalent executions establish its measurement scale.

Consider the model writing $-497y = 0 \Rightarrow y = 0$ in trajectory R039. The first 0 appears in an elimination result that fixes y; the second states the value. Our annotation distinguishes progress in computing a target from the mere appearance of its digit. A detector using gate weights ranks the first 0 eighth among 2,175 positions in this trajectory. Across held-out systems, reviewing the highest-scoring 5% of positions retrieves 94% of annotated computation events. Routing provides a useful index before individual experts have semantic descriptions.

The index also nominates experts for intervention. A fixed selection rule chooses a naturally active component, then a same-state comparison measures its removal against the other seven active components. A frozen numerical reader, trained to predict the target digit, measures the changed state at Layer 22. Removing the nominee lowers its position score more, on average, than removing its peers. Natural gate weight helps explain this effect: it determines how much mixture weight is

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.

redistributed when the component is attenuated. Attenuating expert 29 improves the event position's local digit evidence, yet lowers the reader's probability of the correct digit. The position adds its content advantage over history in proportion to the weight the reader assigns it; the intervention improves the advantage and reduces the share.

Measuring these effects raised a second problem. In our first replay of an intervention, the natural run itself selected a different Top-8 set at the intervention token. We therefore compared the model with itself. Changing prefix chunking, batching or attention implementation preserves the mathematical calculation while changing its floating-point execution (Shanmugavelu et al., 2024; Yuan et al., 2025; He et al., 2025). Five BF16 executions give each measured quantity an execution range, while a numerically tested FP64 calculation provides a high-precision paired reference. This makes precision part of the intervention design.

The comparisons reveal distinct forms of stability. Routing membership changes in about one decision in six, while the gate-weighted index remains highly stable. The reader's content gain and allocation loss preserve their family-level directions across executions. At the answer, high-precision effects are much smaller than the execution range: the median effect-to-range ratio is 0.07. A single BF16 pair often reports a larger effect and a different direction. Pairing shares the input, while the intervention changes the subsequent computation; the measured difference includes the difference between the branches' rounding errors. Raising the precision of the final normalization and output head leaves most of this spread in place.

We contribute a routing-guided path from locating computation to testing natural components, an exact account of their opposing content and allocation effects inside a frozen reader, and an execution-based comparison of intervention measurements with a high-precision reference. Together, they show how to study a selected component and measure its effect accurately. Natural margins provide a complementary guide to where a small perturbation is likely to change an expert selection or a generated token.

2 What we change, read and compare

We change one expert mixture at one token, read the resulting state and answer, and compare each response with its own execution variability. Table 1 gives the route through the analyses.

Model, task and intervention. Frozen Qwen3-30B-A3B (Yang et al., 2025) has 48 layers and selects eight of 128 experts per token. It solves 240 full-rank three-variable integer systems with coefficients and solutions in $[-9, 9]$, generating up to 4,096 tokens. An intervention changes the current token's Layer-16 expert mixture with the input and past cache fixed. We call quantities of the unmodified run *natural*. Attenuation removes a fraction of one expert's natural weight and redistributes it proportionally among the other seven; later layers and tokens follow the model's own routing. Layers are one-indexed, and state measurements follow the residual addition.

Computation events. Two human annotators independently score reader-guided candidates for solution progress S and visible correct-digit exposure N, with model assistance. Both scores agree exactly in 95% of newly rated records; discussion resolves the rest. Previously carried-forward labels are excluded from this agreement denominator. Computation events satisfy $S \geq 5$ and $S > N$, deduplicated by system and token. These ordered categories distinguish a digit's computational role from its visibility; Appendix B gives their definitions.

A frozen reader. The reader combines evidence across positions: each position offers ten digit scores, and the reader combines these scores using a weight for each position. For target variable r and normalized Layer-22 states $\widetilde{h}_j$,

$$q_{rj} = w_{q,r}^{\top} \widetilde{h}_j, \quad v_{rj} = V_r \widetilde{h}_j, \quad \alpha_{rj} = \frac{e^{q_{rj}}}{\sum_{k=0}^{t} e^{q_{rk}}}, \quad z_r = \sum_{j=0}^{t} \alpha_{rj} v_{rj}. \tag{1}$$

The learned vector $w_{q,r}$ gives the *position score* q, which determines the *allocation* α, the reader's pooling weight on that position, distinct from the router's gate weights. The learned matrix V_r gives local digit content $v \in \mathbb{R}^{10}$; $\mathrm{softmax}(z_r)$ predicts the solution's absolute digit. Training

Table 1: **A map of the study.** Routing supplies locations and candidates; reader and answer measurements examine the resulting interventions. Each execution comparison uses the quantity named in its row.

Object	Question	Comparison	Data
Routing	Where is computation, and does its ranking persist?	Gate weights versus membership; rankings across executions	234 positives / 120 systems; stability on 43 trajectories
Nomination	How does nominee removal compare with peer removal?	Nominee versus seven peers at the same state	121 events / 44 families
Reader	How do content and allocation change the reader's prediction?	Four combinations of content and allocation; paired execution range	101 shared events; 99 re-executed
Answer	How large is the effect, and how accurately is it measured?	FP64 reference; natural and paired BF16 ranges	99 events / 43 families; 73-event primary set

uses five whole-system folds and two initializations. Each measured system is held out, and reader parameters stay fixed during intervention (Appendix A). The reader and the model's output head, which projects the final normalized state to vocabulary logits, provide distinct measurements of the intervened computation.

Comparing executions. Five executions, E_1–E_5, keep model weights and input tokens fixed while varying prefix chunking, batch duplication and attention implementation. All five retain the model's native float32 calculations in normalization, routing softmax and rotary embeddings. The maximum minus minimum of a quantity is its *execution range*. An end-to-end FP64 reference passes output-logit agreement tests at about 10^{-13} on the numerical test inputs (Appendix I).

We use two comparisons. The *natural-range comparison* asks whether an absolute FP64 answer effect exceeds the natural answer range, with a lower bound given by BF16 spacing at the natural FP64 log-odds. The *paired-range comparison* asks whether the absolute mean of five BF16 intervention effects exceeds their own range; it is applied to both reader and answer. These compare magnitude with execution variation and the stability of the paired response, respectively. Appendix I gives the exact rules and execution settings.

Samples and averaging. A family contains events from one equation system. Table 1 distinguishes the samples. The 99-event execution study excludes two precision pilots from the 101 shared events. Its *primary analysis set* contains 73 events in 35 families: intervention precedes the first correct assignment, and every natural pre-answer top-two output margin is at least 0.05. Nonlinear reader quantities are computed per initialization before averaging. Mean effects average events within families and then weight families equally. Intervals are per-quantity 95% family-bootstrap intervals; the appendix gives analysis-specific resampling details.

3 Routing locates computation and nominates sensitive components

Gate weights locate computation events at small review budgets, and the index's ranking is highly stable under re-execution. Comparing expert removals at the same state then links natural weight to the reader's response. These weights are both the record used to find computation and the coefficients changed by the intervention.

3.1 A weighted index finds computation

Reviewing the top 5% of positions of each held-out trajectory, ranked by a linear score on Layers 1–22 gate weights, retrieves 94.23% of annotated events (Figure 1a). For the executed gate vector

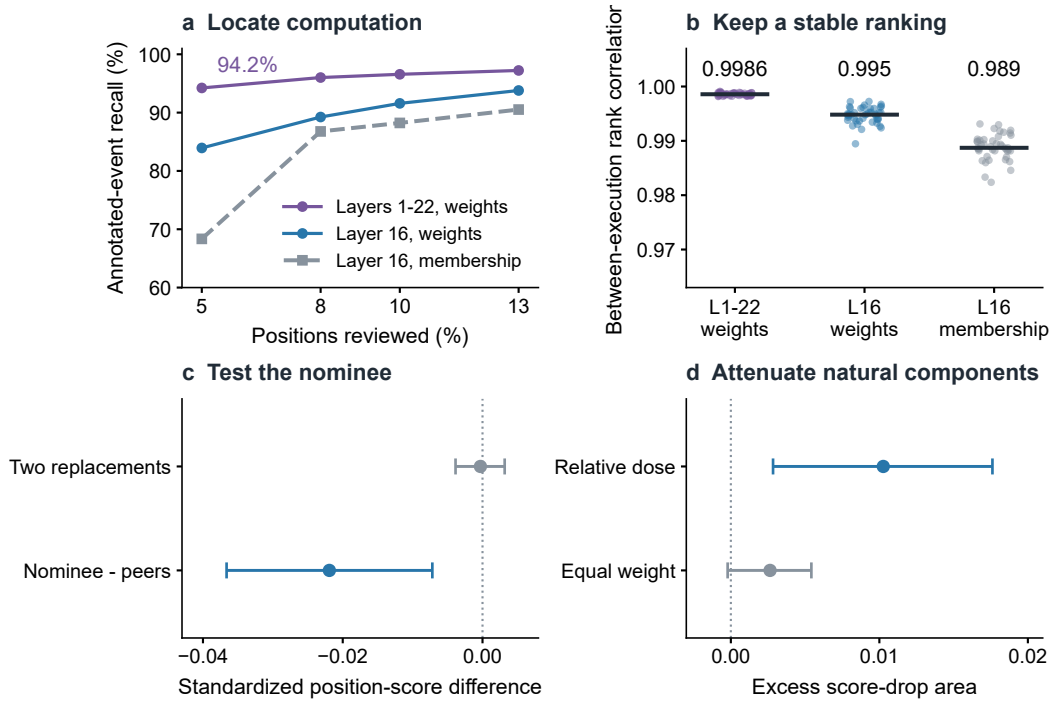

Figure 1: **Gate weights locate computation and form a highly stable index; same-state comparisons measure the effect of removing the nominated expert.** (a) Recall on held-out systems; membership-only assigns $1/8$ to each selected expert. (b) Rankings remain highly stable across BF16 executions; points are trajectories and bars mark medians. (c) Removing the nominee changes standardized position score Q more than removing its peers. (d) Relative attenuation preserves a positive excess score-drop area; equal-weight removal gives the smaller contrast shown. Panels (c,d) have distinct quantities and scales, with 95% family intervals.

$a_{\ell j} \in \mathbb{R}^{128}$ at layer ℓ and position j, with zeros outside the selected eight, the score is

$$s_j = b_s + \sum_{\ell \in \mathcal{L}} w_\ell^\top (a_{\ell j} - \mu_\ell). \tag{2}$$

Here w_ℓ are fitted coefficients, μ_ℓ training means, b_s a bias, and $\mathcal{L}$ one layer or Layers 1–22. Training combines annotated positives with unlabeled positions; tuning and evaluation hold out whole systems (Appendix C). A budget of $b\%$ reviews the top $b\%$ of each trajectory.

At the same 5% budget, Layer 16 alone retrieves 83.94%, compared with 68.35% when each selected expert receives uniform weight $1/8$ in the detector. How much an expert is used therefore adds information to which expert runs. Controls for token identity and digit exposure preserve routing differences (Appendix B).

The Layer-16 signature is simple: the four largest positive coefficients belong to experts 107, 124, 29 and 28 in every fold. They carry 53.6% of the summed gate weight (gate mass) at events versus 22.5% elsewhere. Their summed mass alone retrieves 75.3% of events at the 5% budget (Table 4).

3.2 The weighted ranking survives re-execution

Re-execution changes membership more than the gate-weighted index. Across the execution study, 17.5% of position–layer decisions change Top-8 membership while gate mass moves by 2.4%. In the 20-event replacement analysis, 88% of replaced experts rank seventh or eighth. The index weights each identity by its use: moving weight a from expert e to f, with the other weights held fixed, changes the score by $a(w_{\ell f} - w_{\ell e})$.

Across 43 trajectories, median rank correlation between BF16 executions is 0.9986 for Layers 1–22. The same-layer comparison is 0.995 for Layer-16 gate weights versus 0.989 for membership alone, with the weighted ranking more stable in every trajectory (Figure 1b). At a 13% budget, weighted Layer-16 recall is 94.3% under FP64 and each BF16 execution. Summing expert use over a trajectory gives a median between-execution correlation of 0.99996 (Appendix I.6).

R039 illustrates the distinction. Expert 29 ranks eighth by router score in the FP64 natural run, with weight 0.060 and an eighth-minus-ninth router-logit gap of 0.00214. Two BF16 executions retain it and three replace it. The selection rule below nominates expert 124, whose FP64 natural weight is 0.268.

3.3 Removal sensitivity grows with natural use

At the same state, removing an expert with greater natural weight produces a larger average drop in the reader's position score. Removing the nominated expert lowers the score more, on average, than removing its peers. We measure Q, the position score standardized within its held-out reader and averaged over initializations. Among one-for-one swaps with a valid matched counterpart, the rule selects the *score-lowering swap*, which decreases the Layer-16 index most. Its outgoing expert is the nominee, chosen before measuring downstream responses. When all eight outgoing experts are compared under two fixed replacements, removing the nominee lowers Q by 0.02192 [0.00718, 0.03662] more than removing its seven peers (Figure 1c).

Two controls clarify this comparison. The *matched swap* removes the same expert through a different replacement, with similar gate-vector displacement and a different index change. The mean Q difference between the two swaps is -0.00030 [-0.00386, 0.00317]. Attenuation scaled to natural weight gives the nominee an excess *score-drop area* of 0.01026 [0.00284, 0.01761]; this area integrates the decrease in Q over five doses and subtracts the peers' mean. Removing equal absolute weight gives a smaller contrast of 0.00264 [-0.00022, 0.00542] (Figure 1d; Appendix D).

Natural weight helps explain the response. We fit the change in Q against removed weight across the eight natural experts within each event. The family-weighted mean slope is -0.116 [-0.174, -0.057]. At the mixture output, full attenuation changes the state by the removed weight times the difference between the remaining experts' weighted mean output and the removed expert's output. Weight sets how much is reassigned; output differences set what changes. Relative to its closest-weight peer, the nominee's additional decrease is 0.017 [-0.005, 0.038]. The heaviest natural expert alone shows 0.027 [0.014, 0.041] against its seven peers (Appendix D.5). How much weight the router gives an expert is part of what its removal changes. Experts 29, 107 and 124 receive 61, 36 and 22 nominations; the next section follows them on 101 shared events.

4 Better local evidence can receive a smaller share of the readout

The reader already has evidence from the past. A new position contributes according to what it adds and how much it is heard, its allocation. On average, attenuating expert 29 improves the local digit evidence while reducing the position's share; the reader's probability of the correct digit falls.

4.1 Content improves as the position score falls

Let $m_X = v_X(d^*) - v_X(c_{\mathrm{N}})$ be the correct digit's local margin over the strongest wrong digit in the natural state, held fixed across natural and attenuated states $X \in \{\mathrm{N}, \mathrm{C}\}$. Full attenuation of expert 29 gives a raw position-score decrease $D_q = q_{\mathrm{N}} - q_{\mathrm{C}} = 0.03558$ [0.01239, 0.06148] and a margin gain $G_m = m_{\mathrm{C}} - m_{\mathrm{N}} = 0.02158$ [0.01148, 0.03188]. Both are positive in 39/101 initialization-averaged events, and in 36/101 under both initializations.

4.2 Content advantage times allocation

Hold the past states fixed. Let b be their pooled digit logits, $B = b(d^*) - b(c_{\mathrm{N}})$ their margin, and P_X the complete pooled margin. With the current position's allocation α_X,

$$z_X = (1 - \alpha_X)b + \alpha_X v_X, \qquad P_X - B = \alpha_X(m_X - B). \tag{3}$$

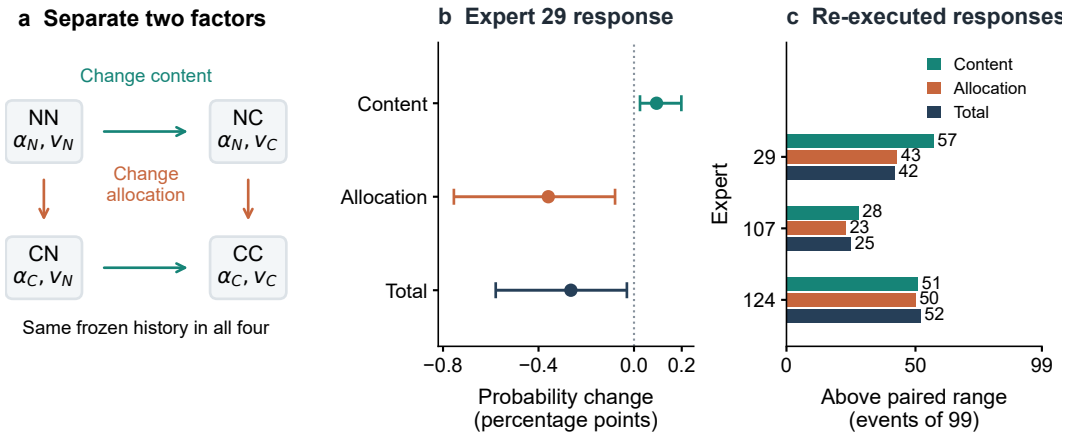

Figure 2: **A content gain and a larger allocation loss explain the reader's decrease.** (a) Four combinations of allocation and content, using the same frozen history. N denotes natural and C attenuated; the first letter chooses allocation and the second content. Each arrow changes one factor. (b) Contributions to the same correct-digit probability under full attenuation of expert 29, with 95% family intervals. (c) Counts exceeding the paired execution range, on 99 re-executed events. Each term is evaluated separately; these counts include either sign.

The current position adds its *content advantage over history*, $m_X - B$, multiplied by its allocation. Under attenuation of expert 29, mean allocation falls from 15.77% to 15.29%. The natural content advantage averages 2.178 [1.797, 2.553] in digit-logit units and is positive in 90/101 events. Reducing the share of this useful position can outweigh the improvement in its local content.

4.3 Two contributions to the same probability

We measure that balance on the complete correct-digit probability. For allocation from state X and content from Y, set $F_{XY} = \mathrm{softmax}((1-\alpha_X)b + \alpha_X v_Y)_{d^*}$. The four combinations share the same history (Figure 2a). Averaging the two orders of changing factors splits their interaction equally (Lundberg & Lee, 2017):

$$\begin{aligned}\phi_{\text{content}} &= \tfrac{1}{2}[(F_{\text{NC}} - F_{\text{NN}}) + (F_{\text{CC}} - F_{\text{CN}})],\\ \phi_{\text{allocation}} &= \tfrac{1}{2}[(F_{\text{CN}} - F_{\text{NN}}) + (F_{\text{CC}} - F_{\text{NC}})].\end{aligned} \tag{4}$$

Their sum is exactly the total probability change. Expert 29's content contribution is $+0.0945$ percentage points [0.0251, 0.1984], its allocation contribution is -0.3590 [-0.7546, -0.0794], and the total is -0.2645 [-0.5793, -0.0298]. The allocation loss explains the reversal on this single endpoint (Figure 2b). Expert 124's response is concentrated in allocation: in R039, full attenuation of expert 124 lowers the reader's correct-digit probability by 0.72 percentage points, 0.65 of them through allocation. Expert 107's contributions are small, with intervals that include zero (Appendix F). Dose and reader-distribution checks preserve expert 29's opposing contributions (Appendices F and G).

4.4 The balance persists across executions

Recomputing the current state under the five executions while holding the original history fixed preserves expert 29's family-level content gain ($+0.093$ percentage points), allocation loss (-0.350) and total decrease (-0.256); all three intervals lie on the same side of zero as before (Appendix I). The natural correct-digit probabilities at these prefixes, median 0.157, set the scale for the percentage-point changes (Appendix I).

Event-level contributions and their execution ranges have similar scales: the ratios of the absolute value of the five-execution mean to the paired range have medians of 1.20, 0.67 and 0.70 for content, allocation and total. The paired-range comparison is met in 57, 43 and 42 events (Figure 2c). Direction is counted separately: content remains positive in all five executions in 53 events, allocation

remains negative in 45, and 22 retain both. The next section measures the same interventions at the model's answer.

5 MEASURING THE ANSWER EFFECT AGAINST EXECUTION VARIABILITY

For the next correct assignment's log-odds, the median effect-to-range ratio is 0.07. Single low-precision pairs report several-fold larger typical magnitudes, with sign agreement slightly above the chance baseline. Higher precision in the final normalization and output head preserves most of the range; the main differences are already present before these operations.

5.1 THE EFFECT AND ITS EXECUTION SCALE

The FP64 natural run supplies a greedy continuation. We measure the joint log-odds $y = \log(P/(1-P))$ of the signed value at the next correct assignment of the target variable, where P multiplies the full-vocabulary conditional probabilities of its answer tokens. This includes the sign for negative answers. The effect is $\Delta = y_{\mathrm{C}} - y_{\mathrm{N}}$; the fixed parser, answer spans and free continuations are described in Appendix I.

Return to R039: five tokens after the event the model completes $y = 0$. Attenuating expert 124 changes that answer's log-odds by -0.052 nat in FP64, while the five natural executions span 0.655 nat. Across the 99 events, median absolute effects are 0.031, 0.019 and 0.040 nat for experts 29, 107 and 124, against a median natural range of 0.50 nat (Figure 3a). The pooled effect-to-range ratio has median 0.07, and 90% of the 297 expert–event pairs fall below a quarter of their range.

Two pairs, both for expert 124, exceed the natural range with its spacing floor. The paired-range comparison is met by 1, 1 and 2 events for the three experts. Across 4,059 strictly pre-answer steps, effects exceed the step-specific natural-range scale at 1.4%, 0.9% and 1.5% of steps.

Direction. Before formal validation we fixed a predicted negative mean for expert 29, matching its reader total, and two selectivity contrasts. On the 73-event primary analysis set the mean is $+0.023$ nat [-0.002, 0.056]. The contrasts $|\Delta_{29}| - |\Delta_{107}|$ and $|\Delta_{29}| - |\Delta_{124}|$ are $+0.012$ [-0.015, 0.039] and -0.027 [-0.061, 0.007]. A follow-up contrast gives expert 124 a larger amplitude than expert 107 by $+0.039$ nat [0.014, 0.068].

Reader and answer each have a direction and a size, and a follow-up analysis measures how the two endpoints correspond. Six diagnostic reader directions carry about 1.1% of the squared norm of the Layer-22 state change. On all 99 events, the signed reader–answer correlation pooled over the three experts is -0.15 [-0.27, -0.02], and expert 29's absolute reader and answer effects correlate at 0.28 [0.04, 0.48] (Appendix I).

5.2 WHAT A SINGLE LOW-PRECISION PAIR REPORTS

A single BF16 pair subtracts the natural outcome from the attenuated outcome within one execution. Its median absolute effect is 0.128, 0.143 and 0.135 nat for the three experts, respectively: about four, eight and three times the corresponding FP64 medians. In R039 the five paired values for expert 124 range from -0.29 to $+0.15$ nat around a reference of -0.052 (Figure 3c).

The discrepancy affects both magnitude and direction. At a 0.05 nat threshold, BF16 alone crosses in 46%, 53% and 46% of the 495 pairs per expert. Both measurements cross in 28%, 23% and 30%; within those subsets their signs disagree in 38%, 54% and 41%. The reference and single-pair rank correlations are $+0.12$, -0.02 and $+0.13$ (Figure 3b). Across five reference-magnitude bins below 0.5 nat, sign agreement remains between 52% and 57% (Appendix I.8). The chance baseline pairs measured and reference signs independently, preserving each sign's observed frequency. These descriptions pool events and executions; formal estimates give families equal weight.

The measurement identity explains the role of pairing:

$$\Delta_v = \Delta^{\mathrm{FP64}} + (\epsilon_{\mathrm{C},v} - \epsilon_{\mathrm{N},v}), \tag{5}$$

where $\epsilon_{X,v} = y_{X,v} - y_X^{\mathrm{FP64}}$ is the branch's deviation from the reference. Pairing shares inputs and settings; subsequent branch computations produce their own rounding errors. Across executions, the

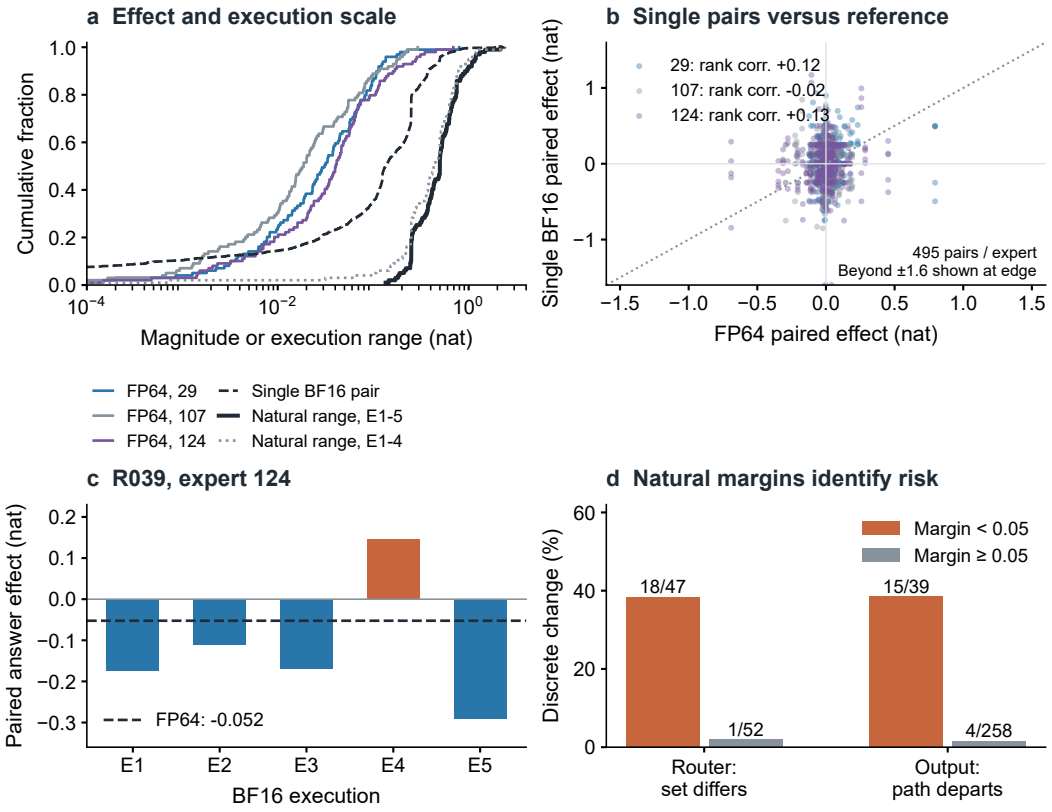

Figure 3: **Small answer effects require matching measurement precision.** (a) High-precision effects, single BF16 paired effects and natural execution ranges, each with its own cumulative distribution. (b) BF16 pairs versus their FP64 references; each expert contributes 99 events under five executions. (c) Changing execution reverses one paired direction in the R039 example. (d) Small natural margins identify higher rates of membership changes and path departures. Output departures use the full saved window; counts give the denominators.

paired effect has a median range of 0.48 nat, close to the natural outcome's 0.50 nat. High-precision paired arithmetic resolves the small reference response, and repeated BF16 executions show how its low-precision measurement changes.

Where the spread arises. Answer-token logits lie near 47, where the BF16 output head works in steps of about 0.25 nat. We keep each forward pass fixed and raise only the precision of the final normalization and output head. The whole-step structure of paired logit differences disappears, and the natural answer range keeps most of its size: the per-event ratio of the range with both operations in FP32 to the original range has a family-weighted median of 95.8% [87.0%, 99.9%]. Single-pair magnitudes fall from 0.13–0.14 to 0.11–0.12 nat, and pooled sign agreement with the reference moves from 53.4% to 53.9% (Appendix I.9). The main differences between executions are therefore already present before the final normalization and output head. Raising their precision measures those differences more finely; the high-precision reference raises the precision of the computation that produces the states.

5.3 Natural margins identify near-tied choices

A small margin puts a discrete choice close to changing. The router margin is the eighth-minus-ninth router-logit gap; the output margin is the top-two vocabulary-logit gap. At natural router margins below 0.05, the executions disagree on the Layer-16 Top-8 at 18 of 47 sites, compared with 1 of 52 at larger margins. Where a natural pre-answer output margin falls below 0.05, attenuation produces a path departure in 15 of 39 expert–event pairs, compared with 4 of 258 elsewhere (Figure 3d).

These departures use the full saved window: 17 begin strictly before the answer, one within its span and one afterward. The 19 free continuations reach the next correct target assignment within the generation window. Across 3.09 million routing decisions, membership disagreement falls from about 42% at margins below 0.003 to 1.7% at margins between 0.1 and 0.3 (Table 17). Raising the output-head precision changes the top-token choice in 488 of 84,260 records along the reference paths, all at exact ties of the BF16 output head. At router and output head alike, natural margins provide a risk scale available before intervention.

6 Related work

Routing and specialization. Studies describe specialization (Muennighoff et al., 2024; Herbst et al., 2026), domain alignment (Lo et al., 2025), semantic routing (Olson et al., 2025), coalitions (Wang et al., 2026b), and geometry (Wang et al., 2026a; Ahrac et al., 2026). Our index locates computation before assigning expert meanings, and re-execution separates the stable weighted index from token-level membership.

Expert importance, probes and patching. Pruning (Lu et al., 2024), steering (Fayyaz et al., 2026), and importance analyses (Su et al., 2026; Bai et al., 2025; Engmann et al., 2026) connect experts to behavior; our comparison holds the state fixed and measures the nominee against the other naturally active components. Probes and interventions distinguish decodability from use (Hewitt & Liang, 2019; Belinkov, 2022; Elazar et al., 2021; Geiger et al., 2021; Meng et al., 2022); metrics (Zhang & Nanda, 2024) and compensation (McGrath et al., 2023; Rushing & Nanda, 2024) shape patching results. Our decomposition explains opposing local and pooled responses through two contributions to the same probability, and the execution comparisons measure reader and answer responses on their own scales.

Numerical precision. Reduction order (Shanmugavelu et al., 2024), inference settings (Yuan et al., 2025; He et al., 2025), arithmetic precision (Feng et al., 2025), and patching precision floors (Baligar, 2026) affect computation and measurement; Gu (2026) relates margins to quantization-induced routing changes. We turn execution changes into the benchmark for each endpoint and show how a single BF16 pair changes both threshold crossings and directions for one expert at one token.

Arithmetic and reasoning localization. Causal mediation locates arithmetic processing (Stolfo et al., 2023), and sentence-level interventions identify reasoning steps (Bogdan et al., 2025). We locate token-level computation using executed gate weights, with annotations that distinguish progress toward a target value from the appearance of its digit.

7 Discussion and conclusion

Routing supplies a stable research entry point, intervention reveals what a selected component changes, and execution comparison establishes how precisely that change is measured. In Qwen3-30B-A3B, weighted routing locates computation despite changes in individual selections. The reader exposes a content–allocation trade-off: better local evidence receives less weight from the reader. At the answer, a small high-precision response can produce substantially different single-pair BF16 measurements.

Gate weight determines the mixture share reassigned by attenuation; the reader's pooling rule determines how a changed state contributes; arithmetic determines the measured response. Pairing fixes the inputs, and precision controls how accurately their difference is evaluated. Raising the precision of the final normalization and output head removes the coarse BF16 rounding at these steps and leaves most of the execution range in place: reading a state more finely and computing it more accurately are different operations, and the FP64 reference performs the second. Here we intervene at Layer 16 in a 128-expert, top-eight model. Comparing other routing designs would measure how the effect-to-range balance varies with expert count, selection size and natural margins. Numerical precision belongs in the design of a mechanistic experiment alongside the intervention and the quantity being measured.

REPRODUCIBILITY STATEMENT

We recomputed nine analysis chains from saved inputs, rebuilt the four figures, and reran model checks on three events with a single A100. The checks cover result tables, family-level uncertainty, numerical precision, caches and intervention controls. Appendix J specifies the inputs, tests and results, with executable entry points for each stage.

AI USE STATEMENT

AI assistants supported semantic annotation, evidence organization, manuscript polishing and the model experiments reported here. Their use included operation of compute resources, retrieval and export of results, and assistance with analysis code. Two human annotators independently scored the semantic labels with model assistance. The authors specified and reviewed the study design and remain responsible for the claims, references and final manuscript.

REFERENCES

Sagi Ahrac, Noya Hochwald, and Mor Geva. Routers learn the geometry of their experts: Geometric coupling in sparse mixture-of-experts. *arXiv preprint arXiv:2605.12476*, 2026. URL https://arxiv.org/abs/2605.12476.

Jun Bai, Minghao Tong, Yang Liu, Zixia Jia, and Zilong Zheng. Understanding and leveraging the expert specialization of context faithfulness in mixture-of-experts LLMs. In *Proceedings of EMNLP*, pp. 21927–21942. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.emnlp-main.1114/.

Yash Baligar. Float32-induced distortion in activation patching: Precision floors and displacement-dependent endpoint-curvature error. *Computation*, 14(9):212, 2026. URL https://doi.org/10.3390/computation14090212.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022. URL https://aclanthology.org/2022.cl-1.7/.

Paul C. Bogdan, Uzay Macar, Neel Nanda, and Arthur Conmy. Thought anchors: Which LLM reasoning steps matter? *arXiv preprint arXiv:2506.19143*, 2025. URL https://arxiv.org/abs/2506.19143.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021. URL https://aclanthology.org/2021.tacl-1.10/.

Leonard Engmann, Christian Medeiros Adriano, and Holger Giese. From observation to intervention: A causal audit of expert importance in mixture-of-experts models. *arXiv preprint arXiv:2606.10703*, 2026. URL https://arxiv.org/abs/2606.10703.

Mohsen Fayyaz, Ali Modarressi, Hanieh Deilamsalehy, Franck Dernoncourt, Ryan Rossi, Trung Bui, Hinrich Schütze, and Nanyun Peng. Steering MoE LLMs via expert (de)activation. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2509.09660.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022. URL https://www.jmlr.org/papers/v23/21-0998.html.

Guhao Feng, Kai Yang, Yuntian Gu, Xinyue Ai, Shengjie Luo, Jiacheng Sun, Di He, Zhenguo Li, and Liwei Wang. How numerical precision affects arithmetical reasoning capabilities of LLMs. In *Findings of the Association for Computational Linguistics: ACL 2025*. Association for Computational Linguistics, 2025. URL https://arxiv.org/abs/2410.13857.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/4f5c422f4d49a5a807eda27434231040-Abstract.html.

Parvel Gu. Detecting a route flip is easier than knowing whether to fix it: Causal route-mediated damage in quantized mixture-of-experts. *arXiv preprint arXiv:2608.11212*, 2026. URL https://arxiv.org/abs/2608.11212.

Horace He et al. Defeating nondeterminism in LLM inference. Thinking Machines Lab blog, 2025. URL https://thinkingmachines.ai/blog/defeating-nondeterminism-in-llm-inference/.

Jeremy Herbst, Stefan Wermter, and Jae Hee Lee. The expert strikes back: Interpreting mixture-of-experts language models at expert level. *arXiv preprint arXiv:2604.02178*, 2026. URL https://arxiv.org/abs/2604.02178.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of EMNLP-IJCNLP*. Association for Computational Linguistics, 2019. URL https://aclanthology.org/D19-1275/.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024. URL https://arxiv.org/abs/2401.04088.

Ka Man Lo, Zeyu Huang, Zihan Qiu, Zili Wang, and Jie Fu. A closer look into mixture-of-experts in large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 4427–4447. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.findings-naacl.251/.

Xudong Lu, Qi Liu, Yuhui Xu, Aojun Zhou, Siyuan Huang, Bo Zhang, Junchi Yan, and Hongsheng Li. Not all experts are equal: Efficient expert pruning and skipping for mixture-of-experts large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6159–6172. Association for Computational Linguistics, 2024. URL https://aclanthology.org/2024.acl-long.334/.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://arxiv.org/abs/1705.07874.

Thomas McGrath, Matthew Rahtz, János Kramár, Vladimir Mikulik, and Shane Legg. The hydra effect: Emergent self-repair in language model computations. *arXiv preprint arXiv:2307.15771*, 2023. URL https://arxiv.org/abs/2307.15771.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/6f1d43d5a82a37e89b0665b33bf3a182-Abstract-Conference.html.

Niklas Muennighoff, Luca Soldaini, Dirk Groeneveld, et al. OLMoE: Open mixture-of-experts language models. *arXiv preprint arXiv:2409.02060*, 2024. URL https://arxiv.org/abs/2409.02060.

Matthew Lyle Olson, Neale Ratzlaff, Musashi Hinck, Man Luo, Sungduk Yu, Chendi Xue, and Vasudev Lal. Probing semantic routing in large mixture-of-expert models. *arXiv preprint arXiv:2502.10928*, 2025. URL https://arxiv.org/abs/2502.10928.

Cody Rushing and Neel Nanda. Explorations of self-repair in language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 42836–42855, 2024. URL https://proceedings.mlr.press/v235/rushing24a.html.

Sanjif Shanmugavelu, Mathieu Taillefumier, Christopher Culver, Oscar Hernandez, Mark Coletti, and Ada Sedova. Impacts of floating-point non-associativity on reproducibility for HPC and deep learning applications. *arXiv preprint arXiv:2408.05148*, 2024. URL https://arxiv.org/abs/2408.05148.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017. URL https://arxiv.org/abs/1701.06538.

Alessandro Stolfo, Yonatan Belinkov, and Mrinmaya Sachan. A mechanistic interpretation of arithmetic reasoning in language models using causal mediation analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7035–7052. Association for Computational Linguistics, 2023. URL https://aclanthology.org/2023.emnlp-main.435/.

Zunhai Su, Qingyuan Li, Hao Zhang, Weihao Ye, Qibo Xue, Yulei Qian, Yuchen Xie, Ngai Wong, and Kehong Yuan. Unveiling super experts in mixture-of-experts large language models. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2507.23279.

Xi Wang, Soufiane Hayou, and Eric Nalisnick. The myth of expert specialization in MoEs: Why routing reflects geometry, not necessarily domain expertise. *arXiv preprint arXiv:2604.09780*, 2026a. URL https://arxiv.org/abs/2604.09780.

Yan Wang, Yitao Xu, Nanhan Shen, Jinyan Su, Jimin Huang, and Zining Zhu. The illusion of specialization: Unveiling the domain-invariant "standing committee" in mixture-of-experts models. *arXiv preprint arXiv:2601.03425*, 2026b. URL https://arxiv.org/abs/2601.03425.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Jiayi Yuan, Hao Li, Xinheng Ding, Wenya Xie, Yu-Jhe Li, Wentian Zhao, Kun Wan, Jing Shi, Xia Hu, and Zirui Liu. Understanding and mitigating numerical sources of nondeterminism in LLM inference. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/f80094a824ba5912d4a2de169c404a40-Abstract-Conference.html.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2309.16042.

A Dataset, state capture, and reader training

A.1 Linear systems and model inference

The frozen model is `Qwen/Qwen3-30B-A3B`, evaluated in BF16. Its revision is `ad44e777bcd18fa416d9da3bd8f70d33ebb85d39`. Each system has three equations and three unknowns. Nine coefficients and three solution components are drawn from $\{-9, \ldots, 9\}$ with probability $1/10$ at zero and $1/20$ at each nonzero integer; the right-hand side is their exact matrix–vector product. Two initially singular systems, R157 and R193, had all nine coefficients and all three sampled solution components redrawn before the formal run, until the determinant was nonzero. The other 238 systems were unchanged. All 240 formal systems are full rank and uniquely solvable. The request presents the full equations; uniqueness does not rely on restricting the answer to a finite digit set.

Generation uses temperature 0.6, top-$p = 0.95$, top-$k = 20$, and a 4,096-token limit. Of the 240 trajectories, 230 terminate with EOS and ten reach the length limit. There are 516,357 saved states included in the analysis: 516,117 post-non-EOS-token states plus one initial boundary state per trajectory. This count is not a count of generated tokens. State $j = 0$ is the prompt's last-token boundary state. A generated token with zero-based index k corresponds to saved state $j = k + 1$, after that token is processed. The state after EOS is excluded. The reader prefix $0{:}t$ includes the boundary and included generated-token states, not every prompt-token state.

A.2 Reader architecture and training

For each layer, outer fold, and initialization, the reader has two bias-free parameter tensors: $W_q \in \mathbb{R}^{12\times 2048}$ and $V \in \mathbb{R}^{12\times 10\times 2048}$. The twelve heads correspond to nine matrix coefficients and three solution variables. Heads have separate parameter rows and are jointly optimized using their mean numerical cross-entropy. For head r, the pooled logits are those in Equation 1; the label is the absolute digit value. The reader does not predict signs or arbitrary multi-token numerical values.

Input normalization subtracts one scalar mean and divides by one scalar population standard deviation computed over all elements of the training systems' legal states, independently for each layer and fold. It is not per-coordinate whitening. Each of five outer folds trains on 192 systems and excludes 48. Two initializations, 2026082311 and 2026082312, are trained independently. For each trajectory, training samples 20 prefix cut points: 0, the final included state J, and one cut point in each of 18 internal time strata. Training lasts 20 epochs; the sampling schedule and epoch count are fixed before inspecting manual stage labels or held-out performance.

The diagnostic reader learns its own attention-pooling rule, jointly optimizing local content scores and position weights. State-dependent weights make the complete reader nonlinear and allow it to combine evidence across positions. We measure the response of this learned rule after freezing its parameters. Semantic-stage characterization uses Layer 22; corresponding readers at Layers 17–22 support the separate depth analyses.

A.3 Which standardization each analysis uses

The three normalization conventions apply to different quantities:

1. **Reader inputs.** Training-fold input normalization defines $\widetilde{h}$.
2. **Intervention scores.** The standardized endpoint Q for the three intervention comparisons uses frozen out-of-fold output means and population standard deviations indexed by target, fold, and initialization. The standardized target scores are then averaged over the two readers.
3. **Cross-layer contrasts.** Later cross-layer standardized contrasts use natural-state sample standard deviations.

The functional analysis in Section 4 uses *raw* q, local logit margins, and actual reader probabilities, not Q or cross-layer standardized scores. No standardized q is exponentiated to construct allocation.

Table 2: **Notation guide.** Reader allocation is distinct from expert gate weight. Italic S, N denote annotation scores; upright N, C denote natural and attenuated states.

Quantity	Meaning and scale
S, N; s	Ordinal progress (0–10) and correct-digit exposure (0–7) annotations; routing detector score (Equation 2)
a; e^*; τ	Expert gate weight; nominee; nominee-minus-others contrast in Q
q, Q; α	Raw position score and its standardized version; reader allocation
v, b, z; m, B, P	Current, historical and pooled digit logits; corresponding fixed-digit margins
d^*; c_{N}	Correct digit; strongest wrong digit in the natural local scores
D_q; G_m	Position-score decrease $q_{\mathrm{N}} - q_{\mathrm{C}}$; local content gain $m_{\mathrm{C}} - m_{\mathrm{N}}$
$\phi_{\mathrm{content}}, \phi_{\mathrm{allocation}}$; F_{XY}	Contributions to one endpoint, in its units; reader outcome with allocation from X and content from Y
$\Delta^{\mathrm{FP64}}, \Delta_v$; r	Paired answer effect under FP64 or BF16 execution E_v (nat); natural execution range floored at BF16 spacing

B Annotation, cohorts, and matched routing comparisons

B.1 Operational annotation

The annotations separate progress toward the target value (S) from visible sources of its correct digit (N). Both are ordered categories. Ratings use the center token and its left context.

S	Progress in the calculation already written
0	Restatement, simplification or rearrangement of the original equations.
1–2	Computation has begun; intermediate results or other-variable information precede a locally sufficient relation for the target.
3–4	New calculations constrain the target; category 4 is closer to a complete target-specific constraint.
5	The required local information is present; a short substitution or calculation remains before a univariate relation.
6–8	A univariate relation fixes the target; increasingly direct arithmetic remains before the correct digit functions as the computed result.
10	The correct digit functions as the target's computed result, including while the final assignment is being written.

Category 9 is unused. These scores describe the written calculation at the position; the original full system uniquely determines the answer from the start.

N	Visible source of the target's correct digit
0	Correct-digit source absent from the visible context.
1	A weak background source of the digit.
3	A nearby active numerical object contains the digit; the object lies outside the target's nearest left or right numerical neighbors in the same active equation.
5	In the same active equation, the nearest numerical object to the target variable on its left or right contains the digit.
7	The center token itself is the target's correct digit.

For example, with correct digit 8, both 18 and 28 in $18z = 28$ supply category-5 exposure around z; a center token 8 supplies category 7. Exposure records where the digit is visible, while S records its role in the calculation. Thus an intermediate digit can receive $N = 7$ together with $S < 10$.

Candidate selection uses numerical-reader information; manual semantic definitions and candidate sampling must therefore not be conflated. The detector's frozen combined-120 annotation table is filtered by $S \geq 5$ and $S > N$, then deduplicated on system and central generated-token index. Its preparation code does not add a two-reader-correctness condition at that step. The separate 1,000-position semantic annotation set is not substituted for this input list. This is a reader-guided annotation sample rather than a random sample of every position in a reasoning trajectory.

Table 3: Data used by each analysis. A family denotes the underlying equation system; multiple target positions can belong to one family.

Analysis	Coverage	Quantity compared
Natural generation / readers	240 systems	Prefixes included in reader training and numerical targets
Routing detector	120 systems	234 labeled positives; 263,932 unlabeled positions; recall over 80 positive-bearing systems
Matched routing semantics	526 candidates	Positions matched on token type and digit exposure
Local causal edits	121 events, 44 families	Fixed computation events; family-balanced effects
Three experts on the same events	101 events, 43 families	All three experts present in natural Top-8; same events for all comparisons

Two human annotators independently scored solution progress and correct-digit exposure, with model assistance. The two 60-system annotation packets displayed the target variable, its true digit, the marked center token and its index, and up to 100 tokens to the left and 12 to the right. Instructions based the ratings on the center and left-hand text; the right-hand tokens served as reading orientation. Annotators could request wider left context. Diagnostic fields—position scores, digit margins, entropy, reader predictions and probabilities—and router/expert information were kept outside the annotation sheet. Rows were ordered by system, target and token index. Exact matches to previously frozen annotations carried their scores forward; fresh rows were issued with blank score fields. Both scores matched exactly between annotators in 95% of newly rated records; previously frozen labels carried forward were excluded from this denominator. The annotators resolved remaining differences by discussion. The workbook records the agreed scores, and the combined table supplies the event labels used here.

B.2 Matched semantic evidence

The 526-candidate routing cohort contains 121 computation events, 338 low-stage positions ($S \leq 4$) passing the same reader-correctness and before-assignment filters, and 67 other events. The 121 computation events comprise 53 positions selected by the reader's entropy-based certainty boundary and 68 other computation positions; their keys match the causal cohort. These counts are not the detector's 234-positive list.

The computation-versus-low-stage comparison restricted to token identities and central-digit exposure ($N = 7$ versus $N < 7$) represented in both groups uses 396 positions (119 computation, 277 low-stage) from 53 runs. The phase comparison $S7/8$ versus $S10$ uses 82 positions from 35 runs. The analyses examine actual Top-8 membership and full-router percentile-rank representations, with run-balanced weighting, clustered inference, and 10,000 wild-bootstrap draws. The 22-layer maximum-statistic correction retains effects for both comparisons; the smallest corrected probability is approximately 10^{-4}. These findings concern structured differences in routing, not classification of every computation type. The certainty-boundary comparison additionally matches solution phase: 92 positions (47 at the boundary, 45 elsewhere) from 39 runs. The matched comparison gives a minimum layer-adjusted probability of 0.0641 and does not establish a routing difference specific to the certainty boundary. The corresponding values are 0.0259 for $N = 7$ and 0.1425 for $N < 7$, showing sensitivity to direct digit exposure.

C Routing detector and budget evaluation

Let $\mathcal{P}$ be annotated positives and $\mathcal{U}$ unlabeled positions in the training fold. The ridge detector minimizes

$$\frac{1}{|\mathcal{P}|}\sum_{j\in\mathcal{P}}\log(1+e^{-s_j})+\rho\left[\max\left(0,\frac{1}{|\mathcal{U}_s|}\sum_{j\in\mathcal{U}_s}\sigma(s_j)-\pi\right)\right]^2+\frac{\lambda}{2}\sum_{\ell}\|w_\ell\|_2^2, \tag{6}$$

Table 4: Held-out budget recall (%). Only the 13% columns have displayed intervals; the four-expert row and the 8% and 10% membership-only values come from the same evaluation rule (Appendix I.6). Multi-layer features use separate expert coordinates at each layer.

Representation	5%	8%	10%	13%	13% 95% interval
Layer 16 executed weights	83.94	89.24	91.58	93.80	[90.16, 96.91]
Layers 1–22 ridge	94.23	96.01	96.57	97.23	[94.96, 99.08]
Layer 16 uniform membership	68.35	86.77	88.23	90.52	[85.89, 94.60]
Layer 16 four-expert mass sum	75.32	84.64	90.59	92.33	–

where $\mathcal{U}_s$ is a random subsample of unlabeled training positions. The penalty enforces an upper bound $\pi = 0.13$ on their mean predicted value; this is a fitting constraint. Each layer's coefficient vector is constrained to sum to zero. Features are centered using that training fold's unlabeled mean. Sampling is without replacement, up to ten unlabeled positions per positive. Three grouped inner folds select regularization using budget recall, with stronger ridge regularization preferred in ties. For the single-layer detector, $\lambda \in \{10^{-3}, 10^{-2}\}$ and $\rho \in \{10^3, 10^4\}$. Final fits average ten such subsamples. If the mean sigmoid score over all unlabeled training positions exceeds 0.135, the fit is repeated with ten times the penalty; a remaining excess is corrected by lowering the intercept using those training positions. The ensemble is checked against the same bound. The sigmoid score is a ranking score; calibration as a probability of computation is a separate question.

For a trajectory of n positions and budget b, evaluation takes the top $\lceil bn \rceil$ held-out scores. Ties use the fixed row ordering. Recall is calculated within each positive-bearing system and averaged equally across 80 systems; the 40 systems without labeled positives still contribute to background evaluation. Five outer folds hold out complete systems. All 22 single layers are compared out of fold, yielding Layer 16 as the reported best single layer. The 95% recall operating threshold is calibrated inside training folds and then evaluated on outer-fold data; it does not guarantee exactly 95% recall in the outer fold.

At the inner-calibrated 95% recall threshold, the outer-fold mean shares of unlabeled positions above threshold are 32.19% for the best single layer and 12.83% for the joint ridge detector. This operating-point statistic differs from within-trajectory budget recall. This measures selection from the unlabeled positions; a false-positive rate would require complete negative labels.

The fitted sparse-group alternative retains all 22 layers. Its result does not identify a much smaller sufficient layer set; the main multi-layer curve is the joint ridge model.

D Intervention design and original endpoints

D.1 Nomination and matching

The selection rule enumerates one-for-one edits of the natural Top-8. Eligible incoming experts rank within the top 32 router scores, including ties, and lie outside the natural Top-8. Each candidate's selected weights are recomputed under the model's frozen gate semantics. Among edits with a negative detector-score change and a valid match, the selection rule selects the largest score decrease, breaking ties by gate displacement and fixed expert ordering. Its match uses the same outgoing expert and a different incoming expert, with a higher detector-score change and a difference of at most 0.05 between the two edits' total-variation distances from natural. Among valid matches, it first minimizes the absolute score change, then displacement mismatch and incoming-logit mismatch. This fixes both the nominated outgoing identity and two replacement identities before outcomes are measured.

The matched swap is an active comparison condition: it removes the same natural expert as the score-lowering swap. The *matched-swap comparison* measures the score-lowering-minus-matched difference in standardized Q. The *outgoing-expert comparison* holds the two replacement identities fixed while crossing them with all eight natural outgoing experts. Its 2,057 conditions are $121 \times (1+8\times 2)$. If g is the average effect over all outgoing–incoming combinations, the nominee's effect relative to natural obeys $\overline{Q}_{e^*} - Q_{\mathrm{N}} = g + (7/8)\tau$. Hence the generic mean, nominee effect, and specificity contrast are distinct quantities.

Table 5: Primary intervention results. The matched-swap and outgoing-expert comparisons use differences in standardized target score; attenuation comparisons use excess score-drop areas; each contrast keeps its own scale.

Contrast	Estimate	95% interval	What the interval says
Score-lowering minus matched swap	-0.00030	$[-0.00386, 0.00317]$	point estimate near zero; compare the nominee contrast in the next row
Nominee minus other seven	-0.02192	$[-0.03662, -0.00718]$	interval below zero
Relative-dose score-drop area	0.01026	[0.00284, 0.01761]	interval above zero
Equal-weight score-drop area	0.00264	$[-0.00022, 0.00542]$	interval includes zero

D.2 Attenuation and equal-weight removal

For dose d, natural weights a_i and their total $A = \sum_i a_i$, the intervention is

$$a'_e = (1-d)a_e, \qquad a'_i = a_i + \frac{da_e}{A - a_e} a_i \quad (i \neq e). \tag{7}$$

Equation 7 preserves total natural mixture mass A, using the original weights without assuming exact unit mass in finite precision. In the ideal linear expert mixture, removed mass $\delta = da_e$ produces the output displacement

$$\Delta o = \delta \left(\frac{\sum_{i \neq e} a_i E_i(u)}{A - a_e} - E_e(u) \right), \tag{8}$$

where the expert input u is fixed. This is a contrast against the other naturally recruited expert outputs, not an isolated expert output. Actual model arithmetic remains BF16. Equal-weight removal conditions instead remove $\delta = \gamma \min_i a_i$ from each candidate expert at a given γ, so the comparison equalizes removed weight rather than relative dose. The complete attenuation experiment has 8,289 unique conditions. The natural-plus-relative series alone has 4,961 unique conditions (121×41); the full plan also includes the equal-weight controls.

The corresponding reported two-sided probabilities are 0.8737, 0.00520, 0.00890, and 0.07369. These results keep their original family-based procedures, and the later functional analyses build on them. A follow-up asked whether increasing attenuation shifts routing change toward Layers 18–19. For the 22 events in 11 families where expert 124 was nominated, define $R(d) = (\mathrm{TV}_{19} - \mathrm{TV}_{18}) / \sum_{\ell=17}^{22} \mathrm{TV}_\ell$, with zero assigned when the denominator is zero, and subtract the other seven experts' mean to obtain $R_{\mathrm{spec}}(d)$. Here TV_ℓ is the total-variation distance between natural and attenuated gate vectors. The change $R_{\mathrm{spec}}(0.50) - R_{\mathrm{spec}}(0.25)$ is -0.061 $[-0.166, 0.040]$; the mean over doses 0.50, 0.75 and 1.00 minus the mean over 0.10 and 0.25 is -0.092 $[-0.157, -0.032]$. The first interval spans zero; the high-minus-low interval lies entirely below zero, indicating a decrease in R_{spec} opposite to the predicted increase.

D.3 Why these three experts

Nomination counts are 61 for expert 29, 36 for 107, and 22 for 124. Natural Top-8 appearance counts across the 121 events are instead 110, 107, and 119. The three-identity set was discovered from the earlier nomination distribution; it is a follow-up set assembled from nomination results. Their shared support consists of the fixed 101 events where all three naturally appear. Nomination and natural membership determine this support before the content–allocation analysis.

D.4 How the cache is built and the current token changed

The causal execution builds the prefix cache through the token before the event, copies it for each condition, and runs the current token with a Layer 16 mixture modification. Causal attention preserves past Layer 17–22 states; downstream routing follows each condition's own computation. For pooling reconstruction, the available states are the saved production-natural history and the current states from attenuation replay. These define the shared-history reader evaluated here. Holding that history fixed gives an exact comparison of how the current state changes the reader's prediction.

Table 6: Expert 29 local effects and joint-response counts. Intervals are the existing pointwise 95% family-bootstrap intervals. Joint-family coverage means at least one joint event, not all events in that family.

Dose	D_q [95% interval]	G_m [95% interval]	Joint	Both readers	Families
0.10	.00630 [.00130, .01129]	.00147 [−.00158, .00468]	30/101	21/101	23/43
0.25	.00801 [.00105, .01507]	.00488 [.00085, .00911]	34/101	28/101	27/43
0.50	.01818 [.00650, .03084]	.00986 [.00366, .01626]	38/101	32/101	29/43
0.75	.02152 [.00337, .04122]	.01561 [.00704, .02437]	37/101	34/101	28/43
1.00	.03558 [.01239, .06148]	.02158 [.01148, .03188]	39/101	36/101	28/43

D.5 Natural weight and same-state sensitivity

The follow-up analysis uses all 121 events and the eight natural outgoing experts, averaging each expert's Q over the same two incoming replacements. For each event, an intercept and slope relate $Q_e - Q_{\mathrm{N}}$ to the outgoing expert's natural weight. Event-wise slopes are averaged within families and then equally across families: -0.1160 [-0.1739, -0.0573]. The intervals below use 50,000 family resamples and apply to individual follow-up comparisons.

The nominee is the heaviest expert in 57/121 events. On the other 64 events (35 families), nominee-minus-heaviest Q is $+0.00163$ [-0.01669, 0.02109]. Comparing each nominee with the peer of closest natural weight gives -0.01726 [-0.03801, 0.00524]; the mean absolute weight gap is 0.0260. This is nearest-weight matching, with no fixed maximum-gap restriction. Subtracting each event's fitted linear weight response gives a nominee-minus-peers residual of -0.00942 [-0.02103, 0.00209]. Selecting the heaviest expert alone gives a same-state contrast of -0.02694 [-0.04098, -0.01416] against its other seven peers. These estimates place the nomination result alongside the role of natural mixture weight.

E Local joint responses across all doses

For every event, expert, and dose, continuous effects are computed separately under each of the two readers, then averaged to one event-level pair (D_q, G_m). The main joint indicator is $\mathbf{1}\{D_q > 0, G_m > 0\}$; the strict stability indicator requires both inequalities in each initialization. Neither uses an added magnitude threshold. Main means and proportions average events within each family and then weight 43 families equally. Raw event counts are descriptive, not 101 independent-system observations.

The raw full-dose joint fraction is $39/101 = 38.6\%$, and the strict two-reader fraction is $36/101 = 35.6\%$. The strict events span 27 families. With families weighted equally, the corresponding proportions are 42.59% [31.00, 54.71]% and 39.80% [28.48, 51.78]%. Family coverage, raw event fraction, and family-balanced fraction answer different questions. These counts establish that the opposing responses co-occur within individual intervention states. Both local effects are computed directly from the current state, independently of the historical vector used in the subsequent pooled decomposition.

F Splitting the correct-digit probability change into content and allocation

F.1 Fixed past positions, fixed competitor, four combinations

For each event and initialization, retain all legal past positions $j < t$. Define

$$L_{\mathrm{prev}} = \mathrm{logsumexp}_{j<t}\, q_j, \quad \beta_j = e^{q_j - L_{\mathrm{prev}}}, \quad b = \sum_{j<t} \beta_j v_j. \tag{9}$$

b is a weighted average of logits, not probabilities; there is no requirement that its ten entries sum to one. Each current state contributes $z = (1-\alpha)b + \alpha v$. This equation follows exactly from separating the current term in the reader's prefix softmax. The four mixed evaluations hold b and

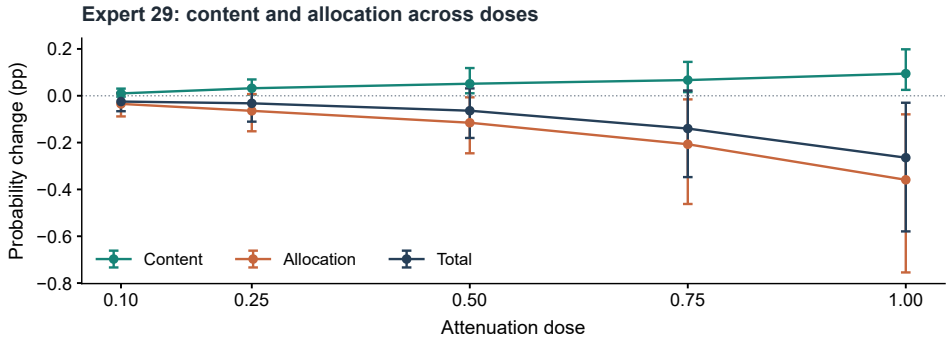

Figure 4: **Expert 29's content and allocation contributions across doses.** Values are probability changes in percentage points on the same 101 events and 43 families. Error bars are per-quantity 95% family intervals; the main text uses the full-dose comparison.

L_{prev} fixed. They change only the current allocation scalar or current ten-class vector inside the diagnostic reader.

For the fixed endpoint, the competitor is selected from natural *current* logits. Eleven of the 101 events have different selected competitors between the two readers; each reader retains its own natural competitor throughout. Dynamic endpoints instead select the highest wrong *pooled* logit independently at each corner. The complete correct-digit probability depends on all ten pooled logits. The fixed contrast measures separation from one wrong digit; the complete probability combines all ten. Positive content contribution to that complete probability is therefore a further measured result.

F.2 Exactness, interaction, and uncertainty

For the complete probability, the four evaluations are

$$F_{XY} = \text{softmax}((1-\alpha_X)b + \alpha_X v_Y)_{d^*}\,, \qquad X, Y \in \{\text{N}, \text{C}\}. \tag{10}$$

More generally, write $F_{XY} = F(z_{XY})$ for any scalar function F. Summing the two expressions in Equation 4 cancels F_{NC} and F_{CN} and leaves $F_{\text{CC}} - F_{\text{NN}}$. No linearity or differentiability assumption is needed. Equivalently, with interaction $I = F_{\text{CC}} - F_{\text{CN}} - F_{\text{NC}} + F_{\text{NN}}$, $\phi_{\text{content}} = F_{\text{NC}} - F_{\text{NN}} + I/2$ and $\phi_{\text{allocation}} = F_{\text{CN}} - F_{\text{NN}} + I/2$. For the fixed margin this reduces to

$$\phi_{\text{content}} = \tfrac{1}{2}(\alpha_{\text{C}} + \alpha_{\text{N}})(m_{\text{C}} - m_{\text{N}}), \tag{11}$$

$$\phi_{\text{allocation}} = (\alpha_{\text{C}} - \alpha_{\text{N}})\left[\tfrac{1}{2}(m_{\text{C}} + m_{\text{N}}) - B\right]. \tag{12}$$

The allocation sign depends on both the current-versus-history contrast and the weight change. The decomposition assigns the paired intervention's reader response to two factors inside the frozen diagnostic rule.

The saved four-corner calculation covers 1,414 event–reader records: expert 29 at five doses and experts 107/124 at full dose. The largest recorded validation discrepancy is 7.11×10^{-15}; four-corner sums close at the recorded arithmetic precision. The fixed-endpoint decomposition matches the previous direct fixed-margin decomposition to less than 10^{-15}. Natural and counterfactual probability and dynamic endpoints reproduce the previous event-level endpoints.

All nonlinear quantities, corner choices, and contributions are computed per reader first. Scalar results are averaged over two readers, then over events within each family, then equally over 43 families. Intervals use 50,000 percentile family-bootstrap draws with fixed event effects; the analysis reuses seed 20260912. These post-hoc intervals apply to individual quantities. Expert 29 at full dose is the primary comparison for this follow-up; the complete five-dose trajectory and full-dose controls are retained regardless of interval sign.

For expert 29's fixed margin, the content interval is [0.00253, 0.00825], allocation $[-0.02404, -0.00205]$, and total $[-0.01704, 0.00202]$. For expert 124's dynamic margin, the content interval is $[-0.00454, 0.00329]$, allocation $[-0.02796, -0.00550]$, and total

Table 7: Expert 29: complete correct-digit probability decomposition in *percentage points*. All entries are existing means and pointwise 95% family intervals, converted by multiplying probability units by 100.

Dose	Content contribution	Allocation contribution	Total
0.10	.0100 [−.0039, .0305]	−.0348 [−.0883, .0078]	−.0248 [−.0661, .0122]
0.25	.0320 [.0048, .0696]	−.0644 [−.1517, .0070]	−.0324 [−.1106, .0340]
0.50	.0513 [.0106, .1183]	−.1152 [−.2457, −.0067]	−.0639 [−.1803, .0314]
0.75	.0670 [.0155, .1446]	−.2071 [−.4622, −.0154]	−.1401 [−.3474, .0224]
1.00	.0945 [.0251, .1984]	−.3590 [−.7546, −.0794]	−.2645 [−.5793, −.0298]

Table 8: Expert 29: dynamic highest-wrong pooled-margin decomposition in original digit-logit units. The wrong-digit competitor is reselected independently at every corner. Intervals are pointwise 95% family intervals.

Dose	Content contribution	Allocation contribution	Total
0.10	.00065 [−.00005, .00162]	−.00201 [−.00438, .00022]	−.00136 [−.00355, .00082]
0.25	.00174 [.00034, .00352]	−.00342 [−.00834, .00060]	−.00168 [−.00667, .00228]
0.50	.00269 [.00052, .00577]	−.00739 [−.01535, −.00102]	−.00470 [−.01257, .00132]
0.75	.00336 [.00035, .00713]	−.01010 [−.02409, .00080]	−.00674 [−.02004, .00356]
1.00	.00480 [.00104, .00939]	−.01767 [−.03527, −.00434]	−.01287 [−.02918, −.00034]

$[-0.03015, -0.00433]$. All expert 107 dynamic-margin intervals cross zero. The table describes each expert's contribution profile; paired expert-difference estimates appear in Appendix I.

In probability percentage points, expert 124's content contribution is -0.0036 $[-0.0655, 0.0571]$, allocation -0.2287 $[-0.4366, -0.0468]$, and total -0.2323 $[-0.4742, -0.0246]$. Expert 107 has content $+0.0085$ $[-0.0109, 0.0308]$, allocation $+0.0231$ $[-0.0566, 0.0986]$, and total $+0.0316$ $[-0.0528, 0.1178]$. All three nominated experts use identical endpoint definitions and the same event set.

G Reader responses across distribution-shift checks

We compare each natural and attenuated state with the reader's training-fold reference states using four quantities:

- The normalized residual norm measures input magnitude.
- Nearest-neighbor distance measures distance to the closest reference state.
- The ten-neighbor measure averages normalized distances to the ten closest reference states.
- Shrinkage Mahalanobis distance measures distance after accounting for the reference covariance.

Reference data, folds, and metrics are fixed. The comparison distinguishes distance already present in a natural computation state from the additional shift caused by intervention.

The audit ranks families by their mean signed counterfactual-minus-natural standardized distance and removes the largest 5 or 9 of 43 families under each metric. Across these eight specified sensitivity summaries, probability and dynamic-margin content intervals remain positive and allocation intervals remain negative. All net point estimates remain negative, but some net intervals cross zero. For example, removing nine families by the ten-neighbor measure gives probability contributions of $+0.0487$ [0.0147, 0.0912], -0.2590 $[-0.5235, -0.0655]$, and net -0.2103 $[-0.4663, -0.0295]$ percentage points. The opposing contributions persist under these fixed, training-reference-based sensitivity checks.

The measurement objects are explicit: local q and fixed margins use current states; integrated reader effects use shared frozen-production history; and Section 5 evaluates the model's vocabulary output at the next correct assignment. The last quantity measures the log-odds of that assignment. Final-answer accuracy is a separate behavioral endpoint.

Table 9: Full-dose endpoint decomposition. Probability is in percentage points; both margins are in original digit-logit units. Axes and scales are not interchangeable.

Expert	Endpoint	Content contribution	Allocation contribution	Total
29	Probability	.0945	−.3590	−.2645
124	Probability	−.0036	−.2287	−.2323
107	Probability	.0085	.0231	.0316
29	Dynamic margin	.00480	−.01767	−.01287
124	Dynamic margin	−.00033	−.01641	−.01674
107	Dynamic margin	.00030	.00141	.00172
29	Fixed margin	.00513	−.01145	−.00631

Table 10: Additional next-layer prediction from orthogonal direction. Values are $(\mathrm{MSE}_{M1} - \mathrm{MSE}_{M2})/\mathbb{E}\|Z\|^2$, not percentage reduction relative to M1 error. Intervals condition on the existing fixed out-of-fold predictions.

Transition	Increment	Pointwise 95% interval	Five-transition 95% interval
17 → 18	.08745	[.04143, .14616]	[.01748, .15742]
18 → 19	.10872	[.02147, .19386]	[−.00549, .22292]
19 → 20	.01349	[−.03822, .05625]	[−.04819, .07517]
20 → 21	.05561	[.01780, .09345]	[.00563, .10559]
21 → 22	.05079	[.01969, .08169]	[.01030, .09129]

H Do changes outside the reader directions predict the next layer?

The hidden-state response contains structure beyond the reader coordinates. A separate analysis tests whether perturbations outside six diagnostic reader directions predict later task changes. Let $Y_\ell = h_\ell^{\mathrm{C}} - h_\ell^{\mathrm{N}}$ and let $\mathcal{T}_\ell$ span the three variable-specific q directions and three fixed-margin directions. The source input is $u_\ell = (I - P_{\mathcal{T}_\ell})Y_\ell$, and the next-layer target is

$$Z_{\ell\to\ell+1} = P_{\ell+1}^{\mathrm{target}}(Y_{\ell+1} - Y_\ell). \tag{13}$$

The main spaces use averaged raw-coordinate gradients across the two readers, not two independent replication studies. The six task directions span a rank-six space in every event, and the next-layer target plane has rank two. The source predictor is the response outside the rank-six space.

The nested models control current task responses, intervention conditions, and the geometric projection of the current response through the next-layer reader (M0); add orthogonal norm (M1); then add normalized orthogonal direction (M2). Predictor and probe isolation share the original five whole-family outer folds. Each outer fit defines training and testing features with that fold's frozen reader; regularization and fallback to a simpler model use training-family inner folds only. Separate models are fit for x, y, z. The 121 events, eight natural experts, and five doses yield 4,840 conditions per transition, but not 4,840 independent samples.

Errors first average experts and doses within an event, then events within a family, then 44 families equally; the reported score is the ratio of these aggregated quantities. Family resampling keeps fitted predictions fixed. Simultaneous intervals use the maximum standardized absolute deviation across five transitions. Direction adds information for several adjacent transitions beyond norm and geometric reader rotation, but lowest-dose point estimates are negative for all five. Direct $19 \to 21$ and $19 \to 22$ follow-ups do not establish stable overall delayed gains. These findings concern predictive information, not a causally isolated orthogonal channel.

Four related analyses compare prediction within and across expert directions:

- Same-expert dose prediction meets its rule at 6/6 layers: the 50%–100% response-direction cosine and the full-dose error advantage of a same-expert two-point linear fit over prediction from the other seven experts both have 95% interval lower bounds above zero.

- Predicting one expert's response as a negative weighted combination of the other seven meets the joint positive-cosine and improvement-over-zero rule at 0/24 layer–dose points.
- A quadratic fit to the 25% and 50% responses meets the four-part prediction rule at 0/12 held-out layer–dose points: improvement over the same-expert linear fit, improvement over the designated first-order combination, and improvement over zero for both predicted responses. The designated first-order term uses the quadratic fit's linear coefficient.
- A shared higher-order component meets its rule at 0/6 layers. The rule requires lower bounds above 0.5 for cross-expert coefficient alignment, shared energy and held-out combination-direction cosine, together with improvement over zero at both held-out doses.

These comparisons distinguish scaling a measured expert direction from predicting another expert's response. Their counts retain the original fixed comparison rules.

I Numerical variation across executions: methods and results

I.1 Executions, endpoints and strata

Table 11: **Each analysis with its cohort, execution and benchmark.** All executions use the same token sequences; only the numerical execution path changes.

Analysis	Cohort	Execution	Endpoint	Comparison
Detector recall (Sec. 3)	120 systems, 234 positives	original saved BF16 run	within-trajectory budget recall	membership-only detector
Nomination (Sec. 3.3)	121 events, 44 families	BF16 re-run	standardized Q at Layer 22	the other seven natural experts
Content–allocation (Sec. 4)	101 events, 43 families	BF16 re-run	reader probability, frozen history	same-endpoint decomposition
Execution study (Secs. 3–5)	99 events, 43 families	E_1–E_5 BF16 and FP64	index ranking, reader terms, answer log-odds	Natural range with spacing floor; paired range

The five executions keep the model, weights and input tokens fixed:

Execution	Prefix processing	Batch	Attention implementation
E_1	Full prefix	1	Explicit attention (eager)
E_2	Split near one third	1	Explicit attention
E_3	Split near two thirds	1	Explicit attention
E_4	Full prefix, duplicated	2	Explicit attention; first row used
E_5	Full prefix	1	Mathematical SDPA

All five use BF16 with native float32 normalization, routing softmax and rotary embeddings. The FP64 reference uses explicit attention. Results for E_1–E_4 show the comparison within that attention implementation; E_1–E_5 also include a change of implementation.

Answer endpoint. The answer position is the next explicit correct assignment of the target variable on the natural greedy continuation, found by a fixed parser that was checked on texts it should and should not recognize before the formal run; if the intervention token itself completes an assignment, the following one is used. The endpoint is $\ell_a - \mathrm{logsumexp}_{j \neq a} \ell_j$ over the full vocabulary, computed from logits. For the 55 two-token negative answers it is $\log P - \log(1 - P)$ with P the joint probability of the signed sequence. The attenuated branch is evaluated by teacher forcing on the natural FP64 greedy path; where its greedy choice first departs from that path, a free greedy branch is run to the answer or to 512 tokens and analysed separately. Attenuation follows Equation 7 at dose one; natural is dose zero through the same code path, which is bit-identical to the unmodified path.

Events and strata. The validation set is 99 events in 43 families (the 101 shared-support events minus two pilots used to settle numerical precision). In 79 events (36 families) the intervention

precedes the first correct assignment; in 20 it follows it or coincides with it. An event is *near-tied* if the natural path has a greedy step before the answer with a top-two logit margin below 0.05: 13 events overall, 6 within the 79. The primary analysis set is the 73 clear-margin events among the 79 (35 families). The event in which expert 29 has zero natural weight in FP64 is retained; weight-normalized quantities use the 98 events with positive weight. Single-pair direction comparisons use the nonzero reference effects defined in Appendix I.8.

Comparisons fixed before the formal run. After the two pilots, the written analysis plan fixed five questions: magnitude (threshold 0.05 nat on answer $|\Delta|$); direction (expert 29 mean below zero); selectivity ($|\Delta_{29}| - |\Delta_{107}|$ and $|\Delta_{29}| - |\Delta_{124}|$); discrete departures; and the natural-range and paired-range comparisons. It also fixed the grouping rules, including natural pre-answer margins below 0.05 (0.125 for sensitivity), equal family weighting and 50,000 family-bootstrap draws (seed 20260912). Applying these rules yielded the 73-event primary analysis set. Inference is reported through estimates and the 2.5th and 97.5th percentiles of family-bootstrap draws. The 124–107 contrast, margin-group comparisons, index-stability analysis, reader execution audit, cross-endpoint associations, the four-expert signature and the membership-change curve are post-hoc follow-ups; their intervals describe each quantity individually. We report both the full set and the primary analysis set so that each pattern can be seen on both event sets.

Answer comparison definitions. For the joint signed-answer log-odds y, natural-range comparison uses

$$r = \max\{\mathrm{range}_v\, y_{N,v}, \mathrm{ULP}_{\mathrm{BF16}}(y_{N,\mathrm{FP64}})\},$$

where $\mathrm{ULP}_{\mathrm{BF16}}(x) = \max(2^{-133}, 2^{\lfloor \log_2 |x| \rfloor - 7})$ for $x \neq 0$, and 2^{-133} at zero. The comparison is strictly $|\Delta_{\mathrm{FP64}}| > r$. BF16 spacing sets a reference scale; the log-odds calculation itself uses FP64. Paired-range comparison compares the absolute mean of the five within-execution paired effects with their range. Both rules measure exceedance of the execution benchmark; inferential uncertainty comes separately from family-bootstrap intervals. For each answer-token probability p_i on the shared reference path, the joint endpoint is $S - \log[-\,\mathrm{expm1}(S)]$, where $S = \sum_i \log p_i$: the sequence calculation first combines token log probabilities and then converts the joint probability to log-odds.

I.2 Numerical validity of the FP64 reference

All four places where the reference implementation casts to float32 (normalization, routing softmax, attention softmax, rotary frequencies) were kept in float64. On the numerical test inputs, processing a prefix in one pass or in two chunks changes the 151,936 output logits by at most 2.7×10^{-14}; incremental cache decoding and full recomputation differ by at most 1.14×10^{-13}; storing weights in BF16 and upcasting per layer is bit-identical to FP64 storage. These calculations use the saved BF16 weights. The smallest recorded eighth–ninth router-logit gap is 5.1×10^{-8} over the prefixes and 1.0×10^{-7} over 846,912 continuation decisions. These tests measure output-logit agreement; the router gaps separately record the distance to each selection boundary. A pilot has a gap of 5.8×10^{-6} at one Layer 31 decision and a reported FP32–FP64 paired-effect discrepancy of 1.1×10^{-3}; this is evidence of a paired-effect discrepancy near a routing boundary. Direct localization of the precision-dependent change would require paired router records. FP64 and BF16 natural log-odds have a correlation of 0.9996.

I.3 Answer-level results

Natural range and spacing floor. With five executions the natural range exceeds the BF16 spacing in all 297 pairs, so the floor of natural-range comparison is inactive; with four executions it is active in 9 pairs. The answer-token logits themselves have magnitudes near 47, where the BF16 step is 0.25 nat; a floor equal to the largest BF16 step among the answer tokens' natural FP64 logits is active in 36 pairs under five executions and 60 under four, and the comparison is then met in 2 and 4 pairs (Table 13).

Direction. On the primary analysis set 41, 35 and 34 of 73 events shift upward for experts 29, 107 and 124 (family-balanced share downward 0.49, 0.52 and 0.48; the zero-weight event stays put), and on all 99 events 49, 50 and 50 shift upward.

Table 12: **Answer effects and precision.** Signed means and 95% family-bootstrap intervals use the primary analysis set (73 events); medians and counts use all 99 events. The natural-range comparison includes the BF16-spacing floor; Table 13 gives the alternatives. Single-pair BF16 summaries pool five executions per event and describe 495 paired values per expert.

	Expert 29	Expert 107	Expert 124
Median $\lvert\Delta^{\mathrm{FP64}}\rvert$ (nat)	.031	.019	.040
Median natural range (nat)		.50 (E_1–E_4: .43)	
Signed mean (nat)	$+.023$ [$-.002$, .056]	$-.007$ [$-.028$, .013]	$-.008$ [$-.045$, .028]
Events above natural / paired range (of 99)	0 / 1	0 / 1	2 / 2
BF16 pair: median $\lvert\Delta\rvert$ (nat)	.128	.143	.135
BF16 pair: rank correlation with FP64	$+.12$	$-.02$	$+.13$
BF16 pair: sign agreement	54%	51%	56%

Table 13: Expert–event pairs (of 297) exceeding an execution benchmark, under four rules and two sets of executions.

Rule	E_1–E_5	E_1–E_4
Natural range, with spacing floor (main)	2	6
Natural range, without spacing floor	2	12
Paired range	4	6
Natural range, floor at the largest answer-token logit step	2	4

Steps before the answer. On the primary analysis set the family-balanced mean of the natural range is 0.618 nat [0.524, 0.741]; the maximum $\lvert\Delta\rvert$ before the answer has means of 0.214, 0.231 and 0.231 nat for experts 29, 107 and 124 and exceeds the corresponding natural maximum in 1, 2 and 1 of 73 events. Along the 4,059 pre-answer steps of the reference paths, natural-range comparison is met at 57, 35 and 62 steps (1.40%, 0.86% and 1.53%), while effects exceed the natural range alone at 96, 84 and 101 steps; at the first step after the intervention the median $\lvert\Delta\rvert$ is 0.033, 0.017 and 0.045 nat against a median per-step range of 0.25 nat.

Downstream routing changes. An attenuation changes a median of 17, 9 and 20 downstream position–layer routing decisions up to and including completion of the reference answer for experts 29, 107 and 124; this count follows the length of the reference path (rank correlation 0.64 to 0.69) rather than the size of the answer effect (0.06 to 0.17).

Effect magnitudes. Of the 297 expert–event pairs, 296 have a nonzero FP64 effect; the one zero is the event with zero natural weight. The ratio $\lvert\Delta^{\mathrm{FP64}}\rvert/r$ has a median of 0.07; 62% of pairs lie below 0.1, 90% below 0.25 and 295 of 297 below 1.

Single-pair threshold crossings. Table 14 separates the four outcomes at 0.05 nat and gives sign disagreement within the both-crossing subset. At 0.20 nat, BF16 alone crosses in 39%, 40% and 35% of pairs, and FP64 alone in 0%, 1% and 4%, respectively. For the threshold-crossing shares reported in Section 5.2, equal family weighting changes the pooled estimates by at most 3.6 percentage points. The answer-level paired-range ratio $\lvert\,\mathrm{mean}_v\,\Delta_v\rvert/\,\mathrm{range}_v\,\Delta_v$ has a median of 0.13, 0.15 and 0.17, and the paired effect spans a median 0.477 nat across executions against 0.500 for the natural log-odds, a per-event ratio with median 1.00.

Expert comparisons. The two fixed selectivity contrasts on the primary analysis set are: $\lvert\Delta_{29}\rvert - \lvert\Delta_{107}\rvert = +0.012$ [-0.015, 0.039]; $\lvert\Delta_{29}\rvert - \lvert\Delta_{124}\rvert = -0.027$ [-0.061, 0.007] (on all 99 events $+0.015$ [-0.004, 0.035] and -0.015 [-0.043, 0.011]); per unit natural weight, $29 - 124 = -0.20$ [-0.44, 0.05]. Post hoc: $\lvert\Delta_{124}\rvert - \lvert\Delta_{107}\rvert = +0.030$ [0.008, 0.055] on 99 events and $+0.039$ [0.014, 0.068] on the primary analysis set; $\lvert\Delta_{124}\rvert - \lvert\Delta_{29}\rvert = +0.015$ [-0.011, 0.043] on 99 events.

Table 14: **Threshold crossings in a single BF16 pair and its FP64 reference.** Crossing means an absolute answer effect above 0.05 nat. The first four rows give counts out of 495 pairs per expert (99 events, five executions), with percentages in parentheses. The final row uses only pairs where both effects cross.

Threshold outcome	Expert 29	Expert 107	Expert 124
BF16 alone above	229 (46.3%)	260 (52.5%)	227 (45.9%)
FP64 alone above	39 (7.9%)	26 (5.3%)	47 (9.5%)
Both above	136 (27.5%)	114 (23.0%)	148 (29.9%)
Both at or below	91 (18.4%)	95 (19.2%)	73 (14.7%)
Opposite signs among both above	52/136 (38.2%)	61/114 (53.5%)	60/148 (40.5%)

Free continuations. Departures from the reference path: 8, 5 and 6 pairs for experts 29, 107 and 124; 17 strictly before the first answer token, one within the answer span and one after it; 15 of the 19 occur in near-tied events (threshold 0.125: 17 of 60 pairs against 2 of 237); every free branch reaches the next correct assignment within the window. The parser returns only assignments equal to the truth, so the recorded correctness flag of a free branch follows from its stop reason; a separate scan with the same parser rules and the value test removed finds, in each of the 19 free continuations, that the first completed assignment of the target variable is the correct value, with no wrong-valued assignment before it (fraction-form intermediates such as $z = 540/108$ evaluate to the truth), and the natural paths of the same events behave identically. The free continuations reach their first correct assignment 6 to 328 tokens past the intervention (branch lengths 8 to 330 tokens), and the record ends at the next correct assignment or at 512 tokens.

Changes along the path. Pooled event medians of $|\Delta|$ for expert 29 by step: 0.033 (step 1), 0.024 (2–3), 0.025 (4–10), 0.017 (11–20), 0.010 (21–50), 0.003 (beyond 50); these are aggregate medians across events.

Averaging under an independent-noise assumption. In the BF16 arm the paired effect has a median range across executions of 0.48 nat and a median absolute mean of 0.07 nat. Under an additive-noise model, in which a single run measures the reference effect plus independent, unbiased noise, the single-run Pearson correlation with FP64 ($r_1 = 0.131$ for expert 29, 0.110 for expert 124) predicts a k-run correlation of $r_1\sqrt{k}/\sqrt{1+(k-1)r_1^2}$: 0.284 for five runs (observed 0.256) for expert 29 and 0.241 (observed 0.232) for expert 124, and a correlation of 0.9 at about 244 and 346 runs. These extrapolations use the stated independent, unbiased-noise model; the correlations here are Pearson correlations, whereas Figure 3b reports rank correlations.

I.4 Reader-level results

At the 99 event prefixes, natural E_1 correct-digit probabilities, computed per initialization and then averaged within event, have median 0.157 and mean 0.254; 16 of 99 exceed 0.5 and 65 of 99 exceed the uniform ten-digit reference of 0.10. These are pooled probabilities with the original frozen history, as used for the execution comparisons.

The history vector b and L_{prev} are the original frozen values; only the current-token state is recomputed, under each execution, from a full re-execution of the prefix. The reader version of natural-range comparison compares the absolute originally saved contribution with its natural execution range. The total uses the range of five natural endpoints; each component uses the range of the four natural–natural contributions described below, with no added zero or spacing floor. Component benchmarks for natural-range comparison take the E_1 natural state as "N" and each other natural state as "C" in Equation 10, so that opposite-signed component noise cannot cancel inside the total. Under paired-range comparison in probability units, all three terms exceed together in 23, 12 and 21 events for experts 29, 107 and 124, and a total exceeds while neither component does in 3, 3 and 0 events; under natural-range comparison the corresponding counts are 18, 7 and 30, and 1, 1 and 1. Paired-range comparison implies a common sign across the five executions. With initializations averaged, expert 29's content is positive in all five executions in 53 events and its allocation negative in all five in 45, with 22 events holding both; 40, 35 and 32 of the paired-range events for content,

Table 15: Events (of 99) meeting each comparison in the reader; content / allocation / total. Log-odds counts differ from probability counts by at most one event.

Expert	Natural: saved effect vs. natural range	Paired: absolute mean vs. range
29	50 / 46 / 43	57 / 43 / 42
107	25 / 31 / 25	28 / 23 / 25
124	59 / 54 / 49	51 / 50 / 52

allocation and total carry the direction of the family-balanced mean. The remaining 17, 8 and 10 paired-range events keep the opposite sign across all five executions. The event-level ratio of the absolute mean contribution to its paired range has a median of 1.20, 0.67 and 0.70 for expert 29, 1.02, 1.03 and 1.07 for expert 124 and 0.49, 0.56 and 0.54 for expert 107. On the 73-event primary analysis set the signs of expert 29's means are unchanged; the content interval excludes zero in log-odds units and marginally includes it in probability units. On that set expert 107 has a small positive allocation contribution in log-odds units ($+0.005$ [0.0001, 0.011]) whose probability-unit interval spans zero. In natural-state standard-deviation units at Layer 22, full attenuation of expert 29 lowers the position score by 0.017 [0.006, 0.029] and raises the fixed correct-digit margin by 0.015 [0.008, 0.022]. Taking both changes as attenuated minus natural, the fixed-margin change exceeds the position-score change by $+0.031$ [0.019, 0.045] for expert 29 and by $+0.023$ [0.011, 0.036] for expert 124 on the 101 shared events; replacing the entire natural Top-8 by the next eight experts or by eight random experts gives fixed-margin-minus-position-score changes of $+0.054$ and $+0.057$ while moving the position score by only 0.058 and 0.071 natural standard deviations. Re-executing E_1 reproduces the originally saved natural reader values to within 0.022 in probability, with 0 of 198 records bit-identical; the original pipeline processed the prefix in chunks of 128 tokens, replayed the recorded Layer 16 route, and used batched rather than row-wise matrix products. The resulting drift, passed through the four-corner formula, has a median of one half of the corresponding execution range for all three terms; its correlations with the natural-range comparison indicators are 0.06/0.25/0.09 for expert 29, 0.30/0.33/0.11 for expert 107 and 0.05/0.19/0.19 for expert 124 (content/allocation/total).

I.5 Routing membership, index stability and margins

Routing arrays (Top-8 identities and executed weights at every prefix position and layer) were saved by the FP64 run and by each BF16 execution. Membership change is the fraction of position–layer decisions whose Top-8 sets differ between E_1 and another execution, averaged over E_2–E_5; FP64 against E_1 gives 18.1% (family-balanced mean; event median 18.2%). The single-swap and rank statistics of Section 3 come from the first 20 events, comparing E_1 with E_2. Per-event fractions of decisions with an FP64 gap below 0.05 range from 50.7% to 55.2%; 22% of decisions have a gap below $1/64$. Executions disagree on the natural Layer 16 Top-8 at 11 of the 14 intervention sites whose gap is below $1/64$. Experts 29, 107 and 124 are in the natural Top-8 under all five executions at 97, 98 and 99 of the 99 sites, and their gate weights at the site have a median relative range of 1.4%, 2.1% and 1.9% across executions. For index stability, each family contributes its longest prefix; generated token g sits at prefix position g plus the prompt length, an alignment confirmed for all 43 trajectories by comparison with shifted alignments. Each trajectory is scored with the detector of the outer fold that held it out. The membership-only Layer 16 detector (uniform weight $1/8$ on the selected experts, Table 4) was scored on the same executions: rank correlation between executions 0.989 (minimum 0.982) against 0.995 (minimum 0.990) for the weighted detector, top-13% set retention 0.970 against 0.977, and recall of the 75 positives 0.931 under the FP64 execution and 0.940 under each of the five BF16 executions, against 0.943 under all six for the weighted detector; the weighted detector is the more stable in each of the 43 trajectories. The association between per-event membership-change rate and the natural answer range is 0.33, while the number of changed decisions is negatively associated with it (-0.35) because longer prefixes have smaller answer ranges; the answer range and membership changes capture different aspects of execution variability.

Table 16: Index stability and endpoint comparisons. Figure 1b reports index correlations between BF16 executions (Layers 1–22 median 0.9986); the first row here instead compares each execution with production scores.

Level	Quantity	Value
Index, Layers 1–22	rank correlation with production scores; recall at 13%	0.9985; 0.965 in all six executions
Index, Layer 16	rank correlation between executions; recall at 13%	0.995; 0.943 in all six executions
Index, Layer 16, membership only	rank correlation between executions; recall at 13%	0.989; 0.931 (FP64), 0.940 (BF16)
Reader, paired-range comparison (events of 99)	content / allocation / total	29: 57/43/42; 107: 28/23/25; 124: 51/50/52
Answer, natural / paired comparisons (events of 99)	natural range with spacing floor / paired range	29: 0/1; 107: 0/1; 124: 2/2

I.6 Routing signature, flip rate against margin, and usage stability

The four-expert signature of Section 3 uses the Layer-16 coefficients of the five single-layer fold models: in every fold the four largest positive coefficients belong to experts 107, 124, 29 and 28, and the five largest negative coefficients to experts 118, 108, 83, 52 and 22. Mean gate weights at the 234 annotated positives and at the 263,932 unlabeled positions give the mass shares of Section 3; the sum over experts of the positive differences between the two mean gate vectors is 0.451 at Layer 16 and ranges from 0.37 to 0.59 over Layers 1–22, with the two largest experts of each layer carrying 25–56% of the shift. The simplified score sums the Layer-16 gate mass on the four largest positive coefficients of the fold model that held the run out; the expert set is read off the fitted detector and involves no further fitting. Recall follows the rule of Table 4 exactly (the top $\lceil bn \rceil$ positions, at least one, ties broken by row order), and the same code reproduces the fitted detector's 83.94/89.24/91.58/93.80 and the membership-only detector's 68.35/86.77/88.23/90.52 (the 8% and 10% membership-only values in Table 4 come from this evaluation); a floor rule gives the same four values for the four-expert score. Sums over the eight largest positive coefficients give 82.55/88.71/89.79/92.27, and the four largest positive minus the four largest negative give 74.54/88.04/89.75/92.02 (82.41 and 74.33 at 5% under a floor rule). At the 121 intervention events the signature holds a median 0.575 of the gate mass (interquartile range 0.535–0.612), and experts 29, 28, 107 and 124 appear in the natural Top-8 at 110, 115, 107 and 119 events. The nominee of the score-directed edit coincides with the natural expert of largest coefficient-weighted mass under the single-layer fold model in 101 of 121 events and is among the two largest in 117, against 57 events for the expert with the largest natural gate weight.

Table 17: Top-8 membership change as a function of the FP64 eighth-minus-ninth router-logit margin, over the 3,091,920 unique position–layer decisions of the 43 longest prefixes.

FP64 margin	Decisions	E_1 vs. E_2	FP64 vs. E_1
below 0.001	49,742	42.4%	59.7%
0.001–0.003	95,998	42.6%	57.0%
0.003–0.01	313,187	40.4%	48.0%
0.01–0.03	699,763	29.3%	28.3%
0.03–0.1	1,210,219	10.6%	9.0%
0.1–0.3	648,802	1.7%	1.4%
0.3–1	73,974	0.1%	0.0%
1 and above	235	0.4%	0.4%

The median margin is 0.045 and 53.5% of decisions lie below 0.05. A Gaussian model of the perturbed margin, $P(\text{change} \mid g) = \Phi(-g/s)$, fits with $s = 0.048$ for E_1 against E_2 (0.034 per execution) and $s = 0.046$ for FP64 against E_1; it over-predicts below 0.001 (49.6% against 42.4%) and between 0.001 and 0.003 (48.3% against 42.6%) and under-predicts between 0.1 and 0.3 (0.4% against 1.7%), so the binned rates are the reported object. Per-layer rates for E_1 against E_2 run from 1.8% (Layer 1) through 11.9% (Layer 8), 15.9% (Layer 16), 19.1% (Layer 22) and 23.0% (Layer 32)

to 24.0% (Layer 45) and 15.1% (Layer 48); Layers 1–22 average 12.9% and Layers 23–48 average 21.0%. The usage-stability statistic sums each expert's gate weight over the longest prefix of each family, per layer; the 128-dimensional usage vectors of E_1 and E_2 have a Pearson correlation of median 0.99996 and minimum 0.99960 over the 2,064 layer–trajectory pairs. These statistics use the 43 longest prefixes; the 17.5% of Section 3 averages 99 events against E_2–E_5, and the same 43 prefixes give 17.3% for E_1 against E_2.

I.7 Reader–answer association and response geometry

Associations use average ranks and a weighted correlation in which every family has total weight one; whole families are resampled. The reference for sign agreement preserves each endpoint's negative-effect frequency and pairs their signs independently. Its value is $p_r p_a + (1 - p_r)(1 - p_a)$, where p_r and p_a are the weighted frequencies of negative reader and answer effects. Signed rank correlations of the total term are -0.22 [-0.45, 0.02], $+0.01$ [-0.23, 0.26] and -0.21 [-0.42, 0.03] for experts 29, 107 and 124. On the primary analysis set the size association for expert 29 is 0.22 [-0.06, 0.48] and the pooled signed correlation is -0.09 [-0.22, 0.05]; the size associations for experts 107 and 124 on all events are 0.04 and 0.00. Nomination frequency against effect size: the nominee-minus-others contrast of Section 3.3, split by nominee identity (disjoint event sets), is -0.094 [-0.138, -0.054] where expert 124 is the nominee (22 events), -0.021 [-0.042, -0.001] for expert 29 (61) and -0.004 [-0.019, 0.010] for expert 107 (36); on the 101 shared events the Layer 22 target position-score decrease of expert 124 exceeds expert 107's by 0.053 [0.024, 0.088] and expert 29's exceeds it by 0.038 [0.013, 0.066], while expert 124 minus expert 29 is 0.015 [-0.027, 0.057]; at the answer, $|\Delta_{124}| - |\Delta_{29}| = +0.015$ [-0.011, 0.043] on all events and $+0.027$ [-0.007, 0.061] on the primary analysis set. The response-energy shares come from the task-subspace analysis of the saved Layer 22 responses (121 events, all natural experts; 101 shared events for the three named experts): the six diagnostic reader gradients span a rank-six subspace in every event; its squared projection norm is 1.10% [0.93, 1.29] of the full response's squared norm, against 0.29% for random rank-six spaces; the current target's two-dimensional plane holds 0.41% of that same full-response energy.

I.8 Single-pair direction agreement across effect sizes

Table 18 groups expert–event pairs by their absolute FP64 answer effect. Each contributes five BF16 executions. The zero-reference event for expert 29 remains in magnitude and threshold analyses and is set aside here because it has no positive or negative reference direction. This leaves 296 expert–event pairs and 1,480 execution pairs. Agreement requires both effects to be strictly positive or both strictly negative; a BF16 zero therefore counts as disagreement with a nonzero reference.

Table 18: **Direction agreement across reference-effect magnitudes.** Counts pool experts and events; five executions belong to each expert–event pair. Bins include their lower boundary and exclude the upper boundary.

$\lvert\Delta^{\mathrm{FP64}}\rvert$ (nat)	Expert–event pairs	Execution pairs	Same direction
$(0, 0.02)$	116	580	304/580 (52.4%)
$[0.02, 0.05)$	78	390	202/390 (51.8%)
$[0.05, 0.10)$	57	285	155/285 (54.4%)
$[0.10, 0.20)$	31	155	89/155 (57.4%)
$[0.20, 0.50)$	12	60	34/60 (56.7%)
$[0.50, \infty)$	2	10	6/10 (60.0%)

Overall agreement is 790/1,480 (53.4%). Independently pairing the two sign distributions while preserving their marginal frequencies gives a descriptive reference of 49.9%, using the same nonzero-FP64 subset and treating a BF16 zero as a third outcome. This pooled reference is separate from the family-weighted reader–answer comparison above. Averaging the five BF16 executions gives direction agreement of 58/116, 41/78, 33/57 and 18/31 in the first four bins, and 9/14 when the last two bins are combined.

Execution consistency and the high-precision reference. All four answer pairs exceeding their own paired execution range have an FP64 effect below one tenth of their natural execution range (4 of 185 such pairs; 0 of 112 larger-ratio pairs). The paired-range comparison describes the mean BF16 response relative to its spread, while the natural-range comparison uses the high-precision effect. Across events, correlations between the five paired-error vectors range from 0.03 to 0.31. The five executions are treated as specified numerical configurations, and the empirical range is their sensitivity comparison.

I.9 THREE READOUTS OF ONE FORWARD PASS

In the BF16 executions the final normalization computes in float32 and returns a BF16 state, and the vocabulary projection runs in BF16, so saved token logits lie on the BF16 grid (4,411 of 4,411 steps); the grid step at magnitude x is $2^{\lfloor \log_2 |x| \rfloor - 7}$, 0.25 nat for answer-token logits near 47. To separate rounding in the final normalization and output head from variation in the state, we reran the 99-event BF16 study (four conditions, five executions, teacher forcing on the FP64 reference path up to answer completion, no free branches) and read every forward pass through three output configurations from the same hidden state. Configuration A is the original normalization and BF16 head. Configuration B keeps that normalized BF16 state and raises only the output head to float32. Configuration C starts from the residual before the final normalization and keeps normalization and head in float32. Routing, caches and reference tokens are identical across configurations. Five acceptance checks on three pre-check events preceded the run: configuration A reproduced the saved BF16 metrics exactly; the float32 head applied to FP64 states matched the FP64 head within 10^{-4}; configurations B and C lie off the BF16 grid; repeated E_1 passes agree exactly; and zero dose equals the unmodified path. On the rerun common reference path, configuration A matched every saved BF16 token logit and log-odds value for value.

Table 19: **Three output configurations of the same forward passes.** Per-event ratios are family-weighted medians (each event weighted by the inverse of its family's event count) with 50,000 family resamples. Pooled sign agreement uses the 1,480 execution pairs with a nonzero reference effect; family-weighted values in parentheses. The natural-range row floors each configuration at the largest answer-token step: the BF16 step for A and the FP32 step for B and C.

Quantity	A: BF16 head	B: FP32 head	C: FP32 norm, head
Natural answer range, family-weighted median (nat)	0.502	0.508	0.503
Per-event range ratio to A, family-weighted median	1	0.963 [0.876, 0.993]	0.958 [0.870, 0.999]
Single-pair median $\|\Delta\|$, experts 29/107/124 (nat)	.128/.143/.135	.112/.119/.112	.111/.117/.112
Sign agreement with FP64, pooled (family-weighted)	53.4% (55.6%)	52.8% (53.4%)	53.9% (54.4%)
Rank correlation with FP64, 29/107/124	$+.12/-.02/+.13$	$+.14/+.01/+.08$	$+.14/+.02/+.08$
Paired-range comparison met, 29/107/124 (of 99)	1/1/2	0/1/1	0/1/1
Natural-range comparison met, five executions (of 297)	2	4	4
Pre-answer per-step natural range, median (nat)	.426	.358	.359
First answer-token paired logit differences equal to whole BF16 steps	1,485/1,485	8/1,485	8/1,485
Top-token choice differs from A (of 84,260 records)	–	488	483

Under configuration A, the paired difference of the first answer token's logit (the sign token for negative answers) is a whole number of BF16 steps in every pair: zero steps in 40% of pairs, one in 42% and two or more in 18%, while the FP64 effect on the joint answer log-odds is nonzero in 296 of 297 expert–event pairs; the two statements concern different quantities. The per-event ratio of configuration C's range to configuration A's exceeds 1 in 42 of 99 events and lies below 1 in 57. Rounding to a grid pulls some readings together and pushes others apart, so higher precision

widens some ranges and narrows others; the 95.8% is the family-weighted median of these ratios and compares ranges. The eight whole-step differences that remain under configurations B and C are all zero: the attenuated expert is absent from that execution's Top-8, so attenuation is the identity. Configuration A has an exact top-two tie in 1,073 of 84,260 records; configuration B splits all of them, changing the top-token choice in 488 and keeping it in 585. Rounding erases differences between tokens, and only part of the erased difference changes the final choice. Family-weighted sign agreement exceeds the marginal reference by 5.56 points [2.32, 8.81] under configuration A and by 4.48 points [1.63, 7.42] under configuration C; in the four larger reference-magnitude bins, configuration C's agreement is 53.7% [47.9, 59.8], 53.6% [44.6, 62.3], 64.0% [48.0, 80.0] and 60.0% (two families), and its excess over each bin's own marginal reference has an interval that includes zero. Correlations between the five paired-error vectors range from 0.03 to 0.31.

J REPRODUCTION CHECKS AND CODE ORGANIZATION

Results recomputed from saved inputs. Nine checks passed: (1) rebuilding five event and step tables from 2,376 condition files; (2) answer effects and family-bootstrap intervals; (3) applying the frozen reader to 495 saved states to produce 3,960 readouts; (4) reader contributions, intervals and execution comparisons; (5) independently reconstructing three four-corner endpoints from 1,414 inputs; (6) detector recall and the four-expert score; (7) same-state nomination contrasts and dose-integrated effects; (8) threshold crossings, signs, reader baselines and step-level comparisons; and (9) family-balanced reader–answer associations. Comparisons match identities, row counts, missing values and text, with absolute numerical tolerance 10^{-12}. Bootstrap checks use the original draw counts: 50,000 for the answer, reader and cross-endpoint analyses, and 10,000 for nomination.

Fresh model runs. We reran both pilots and the validation event with zero natural weight for expert 29. Each natural or attenuated condition follows the shared reference path for four steps, using FP64 and all five BF16 executions. On one A100-SXM4-80GB, 364 FP64 assertions and 300 BF16 assertions passed; these count numerical and identity checks within the three events. The runtime used Transformers 4.51.3, PyTorch 2.13.0 and CUDA 13.0.

In FP64, full-prefix and split-prefix executions differed by at most 4.62×10^{-14} over the complete 151,936-entry output vector. Incremental-cache and full-prefix recomputation differed by at most 1.14×10^{-13}. Natural repeats, zero-dose versus unmodified runs, and attenuation of the zero-weight expert gave identical output vectors. Against saved references, the recorded token logit, greedy token and top-two margin matched exactly; the largest difference in the other-token log-sum or token log-odds was 7.11×10^{-15}. The BF16 checks confirmed within-execution repeatability, zero-dose identity and unchanged Layers 1–16 routing sets on the paired common path. On the rerun common reference path up to answer completion, the later three-configuration rerun of the 99-event BF16 study (Appendix I.9) matched every saved BF16 token logit and log-odds value for value under the same software versions.

How to run the checks. In `reproduction/`, the README lists inputs and dependencies. `run_reproduction.py` checks results from saved model outputs and fitted readers or detectors; `gpu_canary.py` reruns the selected forward passes from model weights and fixed prefixes; `build_paper.py` rebuilds the manuscript and four figures. These starting points distinguish statistical recomputation from fresh model runs.