AN UNCHANGED ANSWER CAN HIDE AN OPERAND'S CAUSAL ROLE

Anonymous authors
Paper under double-blind review

ABSTRACT

Copying a hidden state between questions can test what information a language model uses. The final answer can conceal a contribution when another part of the computation determines which answer wins. We study this problem in Qwen3-8B on weekday and month arithmetic. We replace states at two positions using independently chosen source questions. Holding one source fixed lets us test whether the other supplies an offset or a complete result. Sources with the same offset favor the corresponding candidate even when their own results differ. This offset contribution persists in cases where every substitution produces the same answer. We trace the same candidate comparison through the receiving computation. The next attention output carries the contribution to the final position. Its effect then depends on the receiving state. Fixing the outputs of the next two feed-forward layers reduces this dependence by roughly two-thirds to three-quarters. These experiments identify an operand contribution hidden by answer recovery and distinguish its transmission from the later processing that determines its effect.

1 INTRODUCTION

To explain a model's calculation, we need to identify the information supplied by its internal states and how later computation uses it. A common test copies a state from one question into another and observes the answer. This intervention can expose the supplied information when it changes the answer. An unchanged answer is harder to interpret because several parts of the computation can support competing outcomes. A contribution may affect their scores without determining which one wins.

Calendar arithmetic makes the competing explanations concrete. Suppose a state is copied from "August + 13 months" into "March + 3 months." The source question's complete result is September. If the copied state instead supplies the offset thirteen, combining it with the recipient's March predicts April. The distinction matters for explaining where a model forms its result. A position that supplies an operand can appear to carry a complete answer if later computation combines it with information from another position.

Crossed starting points and offsets can distinguish these possibilities when their effects determine the output. Calendar studies already use such controls to separate input concepts, offsets, and results (Feucht et al., 2026). Causal abstraction gives a broader framework for relating internal interventions to task variables (Geiger et al., 2021). Our question concerns the contribution that remains hidden when another position controls the preferred answer. Identifying that contribution requires varying the two sources separately and retaining the same candidate comparison throughout the test.

We use two semantic controls. A second source has the same offset as the first but a different result. A third has the same result as the first but a different offset. If the replaced state supplies offset information, the first two sources should have similar effects on candidates formed from the recipient's start. Agreement between the first and third sources would instead support their shared result. Holding the other position's source fixed isolates the contribution under comparison.

In Qwen3-8B, we apply this test to the day/month word following the numerical offset and the final prompt position. Figure 1 shows a recorded month example. The final-position source stays fixed as we change the source at the word "months." August + 13 and April + 13 both increase support for April over November. January + 8 gives less support for April despite sharing August + 13's

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.

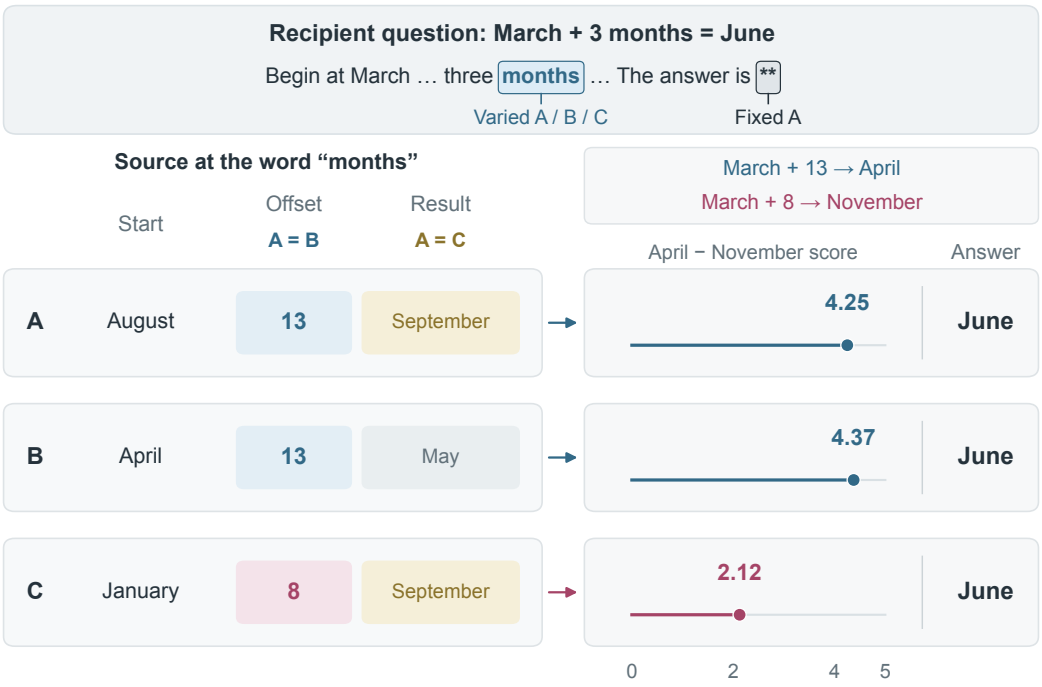


Figure 1: Independent sources expose an offset contribution through the April–November score margin even when every substitution answers June.

complete result. All three runs answer June. The offset comparison reveals a contribution that the final answer omits.

The next question is how the model uses this contribution. Work on communication between positions and interactions between patched components motivates separating transmission from subsequent processing (Mamidanna et al., 2025; Vaidyanathan et al., 2026). We follow the same candidate comparison through attention and feed-forward (MLP) layers. Restoring an attention output tests whether it carries the contribution. Holding that output fixed while changing the receiving state tests how later computation changes its effect.

Our contributions are:

- We combine independently sourced positions with same-offset and same-result controls to identify an operand contribution hidden by an unchanged answer. A fixed candidate comparison distinguishes the supplied information throughout the intervention sequence.
- We trace this contribution through the next attention output and separate its transmission from receiving-state effects. Forward transfer and reverse restoration establish the tested path. Fixing later MLP outputs locates much of the receiving-state dependence in subsequent processing.

2 Related Work

Identifying what a state contributes. Causal abstraction uses interchange interventions to test correspondences between neural states and higher-level variables (Geiger et al., 2021). Distributed alignment search learns these correspondences (Geiger et al., 2024). The pyvene library provides tools for expressing the interventions (Wu et al., 2024). Binding studies distinguish contextual entities and attributes (Feng & Steinhardt, 2024). Mixed answers separate abstract formulations from arithmetic inputs (Cheng et al., 2025). Retrieval work separates a request from the context supplying its answer (Variengien & Winsor, 2025; Geva et al., 2023). Our crossed-answer construction follows

this line of work. The question here is whether a separately sourced position retains an operand contribution when that contribution does not determine the preferred answer.

Arithmetic across layers and positions. Arithmetic studies describe modular algorithms, circular or helical features, and collections of heuristics (Nanda et al., 2023; Zhong et al., 2023; Engels et al., 2025; Kantamneni & Tegmark, 2025; Nikankin et al., 2025). Feucht et al. localize input concepts, offsets, and results in cyclic tasks and relate them to base-10 addition (Feucht et al., 2026). Circuit studies examine operand propagation and result formation (Stolfo et al., 2023; Baeumel et al., 2025). All for One identifies communication between positions that supports later computation at the final token (Mamidanna et al., 2025). Post-routing interventions locate a result-writing boundary (Yan, 2026). We build on crossed-operand controls and these accounts of late computation. Independent sources and a fixed semantic comparison let us follow a contribution that answer recovery conceals.

Processing an incoming contribution. Feed-forward layers have been interpreted as key–value memories and as updates that promote concepts in vocabulary space (Geva et al., 2021; 2022). These accounts motivate testing how later MLPs process the transferred operand information. Our fixed-write experiment measures how this processing depends on the receiving state.

Intervention effects depend on their context. Manual and automated circuit studies use causal interventions to identify task-relevant components and paths (Wang et al., 2023; Conmy et al., 2023). Patching conclusions depend on output metrics, ablations, and corruption choices (Zhang & Nanda, 2024; Heimersheim & Nanda, 2024). Circuit faithfulness and component overlap need not agree (Miller et al., 2024; Hanna et al., 2024). Self-repair can mask local effects (McGrath et al., 2023). Vaidyanathan et al. analyze interactions with the states that bypass a patched component (Vaidyanathan et al., 2026). Causal task performance also differs from invariance across input formats (Opielka et al., 2026). These results establish that causal influence alone does not identify a stable, isolated representation. Our path interventions retain a fixed comparison between candidate answers. This connects the measured interaction to an explicit operand hypothesis.

3 METHOD

3.1 SEPARATE AN OPERAND FROM A COMPLETE RESULT

A source state can affect the recipient by supplying a starting point, an offset, or a complete result. We choose questions for which these possibilities imply different answers. Let $f(s, o)$ be a calendar calculation with starting label s and offset o. The *donor* is the source question whose state is copied into a recipient question. For donor (s_D, o_D) and recipient (s_R, o_R), the candidate answers are

$$\begin{array}{ll} \text{recipient result: } f(s_R, o_R), & \text{donor result: } f(s_D, o_D), \\ \text{donor offset: } f(s_R, o_D), & \text{donor start: } f(s_D, o_R). \end{array}$$

For March + 3 and August + 13, these candidates are June, September, April, and November. Requiring four distinct answers makes the hypotheses behaviorally separable. If a swap makes April win, its effect is consistent with supplying the donor offset to the recipient start.

A state swap copies the donor's hidden state at specified token positions after a complete Transformer block. The remaining computation then runs normally. The *unit position* is the singular or plural day/month word after the numerical offset. The *final position* is the last token of the answer-free prompt. The independent-source experiments replace both positions after layer 22 with states from separately chosen donors. At this layer, position scans show distributed operand effects. By layer 24, final-position swaps predominantly transfer the complete result (Appendix B).

Calendar cycles can make distinct start–offset combinations share an answer. For each donor assignment, we combine the recipient's and two donors' starting labels with all three offsets. This gives nine candidate compositions for checking answer collisions. The primary analysis retains all pairs. A separate subset requires both competing candidates to be unique in this table. Agreement between these analyses checks whether an apparent operand effect depends on an accidental answer match.

3.2 TEST AN OPERAND CONTRIBUTION BENEATH THE WINNING ANSWER

The final answer may follow one replaced position even when another supplies information. An initial competition test gives the unit and final positions states from two donors with the same complete result and different operands. Swapping their assignments tests which source more often controls the answer. The candidate answers combine the recipient start with each donor's offset.

We then hold the final-position donor fixed to examine the unit's contribution. Comparing sources with a shared offset tests whether their effects agree despite different complete results. A source with the same result but a different offset separates those two explanations. We call the first source anchor A. Donor B shares its offset but has a different start and result. Donor C shares A's result but has a different offset.

Independent source assignment makes this comparison possible. If both positions always receive the same donor, changing the unit's source also changes the final-position information. The resulting answer cannot separate their contributions. We evaluate ordered assignments of the sources to the two positions. The unit comparison fixes the final-position donor to anchor A and keeps the recipient unchanged. Any difference between the three runs then follows from changing the unit source and recomputing the subsequent layers.

Figure 1 uses August + 13, April + 13, and January + 8. The unit sources A and B both predict April when their offset is combined with the recipient's March. Source C instead predicts November. We compare these same two candidates under all three substitutions, even if a different answer wins.

Let m_A, m_B, m_C denote the anchor-offset candidate's score minus the other-offset candidate's score. The offset contrast is

$$\chi_U = \frac{m_A + m_B}{2} - m_C. \tag{1}$$

A positive contrast means that the same-offset donors jointly favor their predicted candidate over the different-offset donor. We also compare m_A and m_B separately to check similarity between the same-offset donors. Together, these comparisons test an offset-related contribution under the substitutions. The preferred answer records which token wins in the full vocabulary.

The two controls constrain the interpretation of that score difference. Agreement between A and B links the effect to the shared offset despite different source answers. Separating A from C tests whether sharing the source answer is sufficient to preserve the effect. Keeping the candidate pair fixed is essential here. If each run were scored against its own source answer, the output comparison would change together with the intervention.

3.3 TRACE THE SAME CONTRAST THROUGH THE NEXT COMPUTATION

To locate how the unit affects later computation, we compare runs that differ only in whether its state was replaced. The *final-only run* swaps the final position and leaves the unit unchanged. The *joint run* additionally swaps the unit from A, B, or C. Their final-position states initially match. Their subsequent difference measures what the unit adds.

At layer 23, we transfer the joint run's unit keys, values, or both into the final-only run. We also transfer the *attention write* at the final position. This output is added to the residual stream. The residual stream carries the state between blocks. Reverse restoration puts the corresponding final-only quantity into the joint run. Remaining layers recompute after each replacement. Measuring the same χ_U throughout connects the operand hypothesis to the tested path. Unit-value restoration changes one part of attention. Complete-write restoration replaces the resulting attention output.

The two intervention directions answer complementary questions. Forward transfer asks whether the selected quantity can carry the offset contrast into a run without the unit replacement. Reverse restoration asks how much of that contrast remains when the same quantity is returned to its final-only value. Unit projections are replaced before positional encoding and attention computation. Direct-write transfer instead replaces the head outputs at the final position before the attention output projection. Comparing these operations separates a source of attention information from the complete contribution delivered to the final position.

3.4 SEPARATE TRANSMISSION FROM RECEIVING-STATE EFFECTS

We next separate the write from the state receiving it. We interpolate the layer-22 final-position residual between recipient and donor states. Each strength produces a layer-23 attention write. We then cross every recorded write with every receiving residual. Let r denote residual strength and w the strength used to generate the write. Moving down a column changes the receiving state under a fixed write. Moving across a row changes the incoming write under a fixed receiving state. The diagonal recovers the original matched combinations.

For each unit-donor role, its write remains fixed as r changes. Writes and unit values can differ across A, B, and C. We evaluate the same candidate margins $m_A(r, w)$, $m_B(r, w)$, and $m_C(r, w)$ at every crossing. Their contrast retains the definition in Equation 1. A change in this contrast along a column locates receiving-state dependence after the fixed attention contribution.

We then fix the final-position MLP outputs in layer 23, layer 24, or both to the corresponding role's full-strength endpoint. Each donor role retains its own endpoint output. The clamp holds each role's MLP output constant across receiving states without forcing the three roles to share an output. For each group, we measure the standard deviation of χ_U across five residual strengths at a fixed full-strength write. Comparing this variation before and after fixing MLP outputs tests how much the later processing contributes to the dependence.

4 EXPERIMENTAL SETUP

Questions and models. The main analysis uses Qwen3-8B (Yang et al., 2025) on all starting weekdays or months and offsets from one to fourteen. Prompts state the cycle rule and end with the answer-free prefix "The answer is **". The Original wording is accompanied by Count and Advance prompts. Journey and Tracker wordings test the receiving-state analysis. We use BF16 and the nonthinking chat template. Layer and head indices are zero-based. Full prompts and readout details appear in Appendix A.

Llama-3.1-8B and its instruction-tuned variant (Grattafiori et al., 2024) supply a contrasting result-transfer analysis. We use a separate protocol with different prompts, selected heads, and intervention spaces. Appendix B reports it alongside the position scans and recombination controls.

Pair selection and scores. Natural accuracy includes every question. Swaps require naturally correct donor and recipient answers and matched token lengths. Semantic comparisons require distinct counterfactual predictions. A stricter analysis also requires both crossed questions to be naturally correct. The independent-source analysis uses 24 recipient groups per task–wording condition. Selection uses these natural answers before intervention outcomes. Group summaries describe the fixed task and selection protocol because questions and donors recur across groups.

The preferred answer is decoded from the unrestricted first-token argmax. Candidate scores use log-sum-exp over distinct single-token label encodings with and without a leading space. A positive candidate margin need not make either label the full-vocabulary winner. We report signed contrasts on the preselected groups. The same-answer analysis additionally selects groups whose preferred answer matches under all three unit donors after intervention. Its mean absolute contrast measures the effect hidden by the unchanged answer.

Head selection and random controls. Four layer-23 heads are selected on the original weekday condition and held fixed across wordings and months. Random value directions match the actual replacement norm.

Predicting unmeasured strengths. We also test whether endpoint interventions predict the unit's added effect at other residual strengths. Regression rules are fitted on the Original weekday development group. Each new pair supplies its full-strength final-only and joint interventions for calibration. At each target strength, we first measure the final-only margin. The frozen rules predict the increment from adding the unit replacement before that joint intervention is run. A generic regression uses each role's endpoint effect and the mean endpoint effect. A semantic regression uses the mean effect and the endpoint offset contrast with fixed coefficients for the donor roles. Both have three parameters and receive the same calibration information. Adding the predicted increment to the

measured final-only margin gives the predicted two-candidate margin. Appendix D specifies the features and fitting details.

5 RESULTS

5.1 AN OFFSET CONTRIBUTES EVEN WHEN THE ANSWER DOES NOT REVEAL IT

In the initial competition test, the preferred answer more often follows the final-position donor's offset. This holds under both donor assignments and after excluding algebraic collisions. The unit could help read the final-position information. It could also supply its own offset without winning the competition.

Fixing the final donor reveals a separate offset contribution from the unit. Under Original wording, mean χ_U is 4.289 logits for weekdays and 8.172 for months, with positive contrasts in 22/24 and 19/24 groups. Count and Advance also retain positive means in both tasks. Restricting to unique crossed candidates gives Original-wording means of 6.469 and 7.834. Sources sharing the anchor's offset therefore support its predicted candidate more than sources sharing only its result.

The unchanged answer in Figure 1 also occurs in 21/24 weekday and 19/24 month Journey groups. Their mean absolute offset contrasts are 1.146 and 3.132 logits despite identical preferred answers. Restoring the next attention write reduces these values to 0.104 and 0.293 (Figure 2). Tracker gives the same ordering. The fixed candidate comparison exposes an operand contribution that preferred-answer recovery misses.

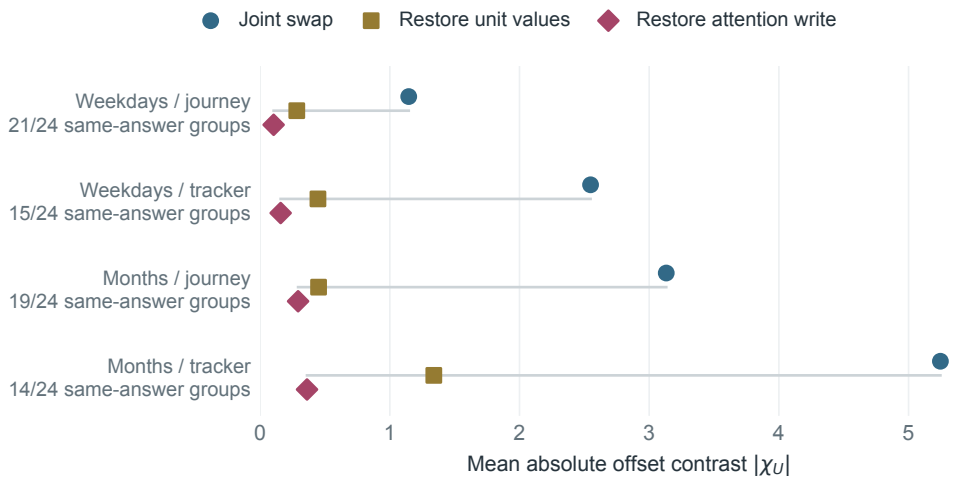


Figure 2: Groups with identical preferred answers retain an offset contrast that decreases when unit values or the next attention write are restored.

5.2 THE CONTRAST REACHES THE FINAL POSITION THROUGH ATTENTION

The final-only and joint runs begin the path experiment with the same final-position state. Under Original weekday wording, transferring the joint run's next-layer unit K/V produces a contrast of 4.112 logits, close to the joint effect of 4.289. Values alone produce 3.299, whereas keys alone produce -0.328. Under Advance wording, joint K/V gives 1.219 against 0.492 for values and 0.125 for keys. The combined effect exceeds their separate sum, showing that keys and values interact in carrying the offset contribution.

Copying the layer-23 attention write retains 4.167 of the 4.289-logit weekday contrast and 7.674 of the 8.172-logit month contrast under Original wording. The four selected heads retain 3.852 and 6.997, compared with 0.096 and 0.205 for norm-matched random values. The heads were selected

only on Original weekday questions. Their effect persists across the other wordings (Figure 3), linking the same counterfactual operand comparison to a specific attention write.

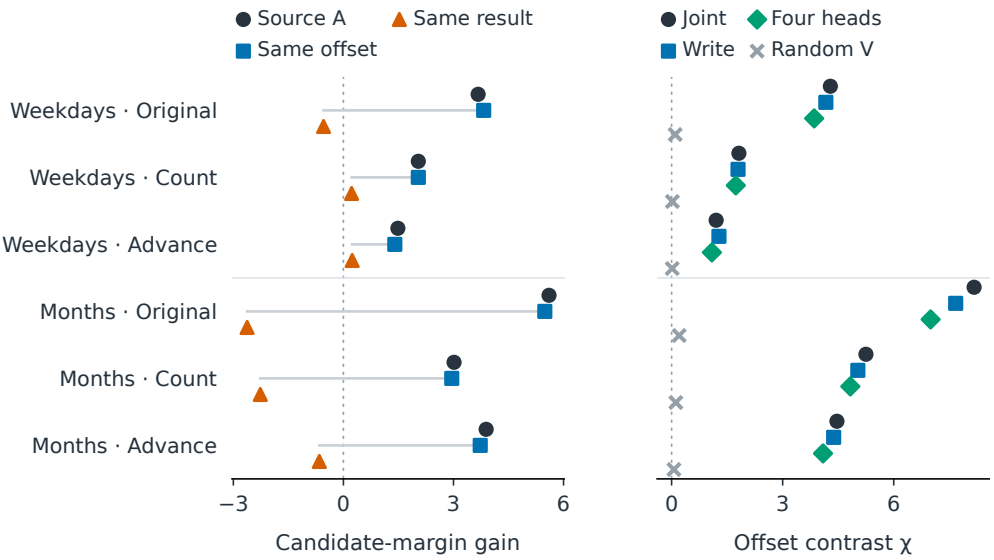


Figure 3: Forward transfer retains the offset contribution through the next attention write across wordings and tasks.

Reverse restoration tests the corresponding path within the joint run (Figure S1). Replacing unit values with final-only values reduces much of the contrast. Mean χ_U across the six task–wording groups ranges from 1.203 to 8.172 before restoration. Restoring the complete attention write reduces this range to 0.000–0.406. Restoring the layer-23 and layer-24 final-position MLP outputs also attenuates it. The offset contribution reaches the final position through attention and depends on subsequent processing.

One Tracker month pair reveals a difference between the two restoration tests. Value restoration changes χ_U from 9.250 to -10.625 logits, whereas whole-write restoration reduces it to 0.813. After value restoration, the same-offset margins return close to the final-only margin. The same-result role retains a different margin, reversing the contrast. Whole-write restoration brings the three margins closer together. The remaining role-specific attention effect matters even after the unit values have been restored. This explains why a value intervention cannot replace the complete-write test. Appendix C reports the individual margins.

5.3 Later processing changes the effect of the incoming offset

The path result leaves open whether the attention write determines the operand's influence by itself. The Journey and Tracker groups retain a positive mean offset contrast across residual interpolation strengths. Its magnitude varies nonmonotonically. Holding the write fixed isolates the contribution of subsequent processing.

At a fixed layer-23 write, changing the receiving residual changes χ_U (Figure 4). Each heatmap cell gives the mean contrast at one measured crossing of write-source and residual strengths. All panels share a color scale. Within each donor role, unit values remain identical across residual strengths. The variation along a column shows that the supplied operand's effect changes even with the incoming attention contribution held fixed.

Fixing the subsequent MLP outputs to their role-specific endpoints reduces this dependence (Figure 5). For Journey, mean within-group standard deviation across residual strengths falls from 0.612 to 0.227 logits in weekdays and from 2.227 to 0.710 in months. Tracker falls from 1.456 to 0.504

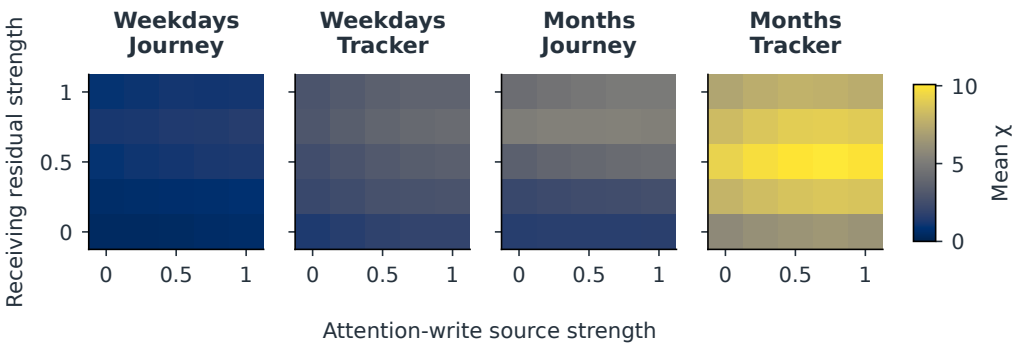


Figure 4: Mean offset contrast χ_U across attention-write and receiving-residual strengths. All panels share a color scale.

and from 2.838 to 0.735. These 63–74% reductions locate much of the receiving-state dependence after the attention write.

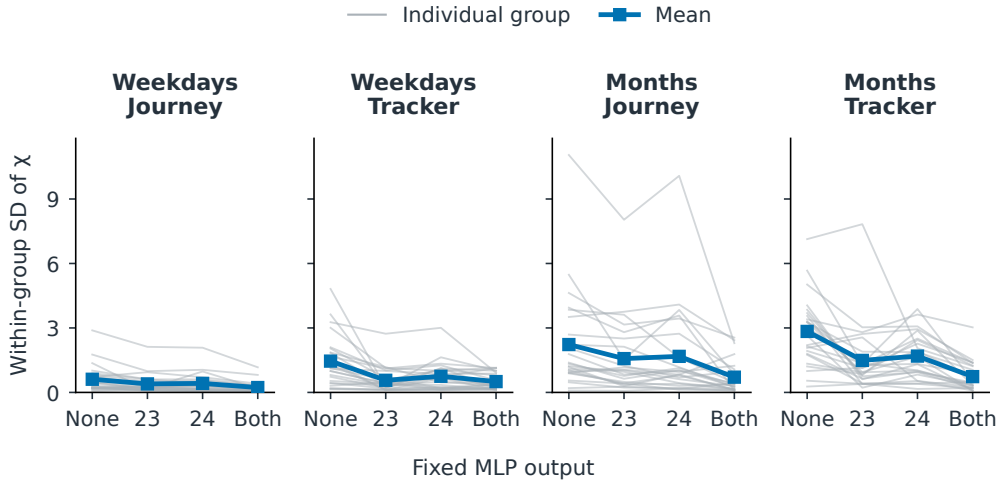


Figure 5: Fixing later MLP outputs reduces variation in the offset contrast across receiving states under fixed attention writes.

The endpoint prediction test asks whether naming the operand also improves prediction of the response curve. Table 1 reports mean absolute error. The final-only baseline predicts no added unit effect. The endpoint-delta baseline adds the full-strength unit effect unchanged at every target strength. With matched calibration information, the generic regression predicts held-out strengths as accurately as the semantic regression. Their similar errors separate prediction accuracy from the additional interpretation supplied by the controls. Same-offset and same-result donors identify which task variable participates in the response. Following their fixed candidate comparison through attention and MLP interventions identifies the tested path.

Where the account applies. The position controls connect the hidden score contribution to cases where the operand determines the answer. Qwen unit states at shallower layers mainly produce the recipient start plus donor offset. Starting-point prefix states transfer the complementary operand. Combining the prefix with the layer-22 final position increases complete donor-result recovery. Adding the unit state and then the offset-token state resolves most remaining cases. By layer 24, the

Table 1: Held-out-strength prediction errors under matched endpoint calibration. Entries are mean ± sample SD of group MAE across 24 recipient groups per row.

Task / wording	Final only	Endpoint delta	Generic	Semantic
Weekdays / Journey	0.794 ± 0.839	0.982 ± 0.987	0.950 ± 0.552	0.950 ± 0.553
Weekdays / Tracker	1.700 ± 1.607	1.949 ± 1.449	1.233 ± 0.881	1.241 ± 0.881
Months / Journey	2.025 ± 1.066	2.758 ± 2.089	1.639 ± 1.115	1.639 ± 1.123
Months / Tracker	5.340 ± 3.466	3.569 ± 1.822	3.301 ± 2.434	3.303 ± 2.431

final position more often transfers the complete result. This progression explains how an operand-bearing state can later acquire a result-transfer effect.

The carrier also changes with context. On the same semantic pairs, early unit-only transfer weakens under Count and Advance wording even when the plural unit form is held constant. Later final-position result transfer persists. In the separate Llama protocol, same-result donors usually preserve the recipient answer, and different-result donors steer it toward their result. Boundary-value interventions locate a result-related path there. The counterfactual candidate test distinguishes these effects across the tested carriers. Appendix B gives the full position scans and recombination controls.

6 Discussion

A state swap changes one part of a calculation that still combines information from several positions. The final answer can obscure the contribution of a position whose information does not dominate that calculation. Our independent-source comparison identifies such a contribution at the unit word. The same-offset controls agree despite different complete results, and the same-result control separates from them. This resolves which operand the unit contributes under the tested replacements.

The path interventions explain why identifying the incoming operand is only part of the account. Forward transfer preserves the counterfactual comparison after attention has combined information from several positions. Fixing that write still leaves different receiving states with different contrasts. The MLP clamps reduce this variation. This links later processing to how strongly the offset affects the candidates. The calculation can transmit the same operand-related contribution and use it differently as the receiving state changes.

This procedure can guide intervention studies in which several components influence the same answer. Competing interpretations of a state first need to imply distinguishable candidates. Independently varying its source then separates its contribution from the other patched states.

Scope. We study naturally solvable calendar questions with matched token lengths. The diagnostic measures offset-related influence under the specified replacements. Calendar cycles make some offsets indistinguishable. Crossed activation states isolate responses that may differ from a natural run. Same-answer subsets are selected after intervention and reuse source questions. They demonstrate information hidden by the answer without estimating its prevalence across tasks.

7 Conclusion

Independent sources and same-offset/same-result controls reveal an operand contribution that an unchanged answer conceals. In Qwen3-8B calendar computations, this comparison identifies offset information supplied through a unit-word position. Following the same comparison through attention and later MLPs separates how the contribution reaches the final position from how the receiving state changes its effect.

AI USE STATEMENT

Generative AI tools assisted manuscript writing and translation.

REPRODUCIBILITY STATEMENT

The appendices specify prompts, token positions, head selection, intervention operators, numerical precision, and answer scoring.

REFERENCES

Tanja Baeumel, Josef van Genabith, and Simon Ostermann. Disentangling mathematical reasoning in llms: A methodological investigation of internal mechanisms. In *Proceedings of The 3rd Workshop on Mathematical Natural Language Processing (MathNLP 2025)*, pp. 203–217. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.mathnlp-main.16.

Ziling Cheng, Meng Cao, Leila Pishdad, Yanshuai Cao, and Jackie Chi Kit Cheung. Can LLMs reason abstractly over math word problems without CoT? disentangling abstract formulation from arithmetic computation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 14306–14333. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.723.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards Automated Circuit Discovery for Mechanistic Interpretability. In *Advances in Neural Information Processing Systems*, volume 36, pp. 16318–16352, 2023. doi: 10.52202/075280-0719.

Joshua Engels, Eric J. Michaud, Isaac Liao, Wes Gurnee, and Max Tegmark. Not All Language Model Features Are One-Dimensionally Linear. In *International Conference on Learning Representations*, pp. 84591–84622, 2025.

Jiahai Feng and Jacob Steinhardt. How do Language Models Bind Entities in Context? In *International Conference on Learning Representations*, pp. 36391–36413, 2024.

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

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pp. 9574–9586, 2021.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah Goodman. Finding alignments between interpretable causal variables and distributed neural representations. In *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pp. 160–187. PMLR, 2024.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer Feed-Forward Layers Are Key-Value Memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5484–5495. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.446.

Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. Transformer Feed-Forward Layers Build Predictions by Promoting Concepts in the Vocabulary Space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 30–45. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.emnlp-main.3.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting Recall of Factual Associations in Auto-Regressive Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12216–12235. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.751.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, et al. The Llama 3 Herd of Models. arXiv:2407.21783, 2024.

Michael Hanna, Sandro Pezzelle, and Yonatan Belinkov. Have Faith in Faithfulness: Going Beyond Circuit Overlap When Finding Model Mechanisms. In *Conference on Language Modeling*, 2024.

Stefan Heimersheim and Neel Nanda. How to use and interpret activation patching. arXiv:2404.15255, 2024.

Subhash Kantamneni and Max Tegmark. Language Models Use Trigonometry to Do Addition. arXiv:2502.00873, 2025.

Siddarth Mamidanna, Daking Rai, Ziyu Yao, and Yilun Zhou. All for one: LLMs solve mental math at the last token with information transferred from other tokens. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 30747–30760. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.1565.

Thomas McGrath, Matthew Rahtz, Janos Kramar, Vladimir Mikulik, and Shane Legg. The Hydra Effect: Emergent Self-repair in Language Model Computations. arXiv:2307.15771, 2023.

Joseph Miller, Bilal Chughtai, and William Saunders. Transformer Circuit Faithfulness Metrics are not Robust. In *Conference on Language Modeling*, 2024.

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

Yaniv Nikankin, Anja Reusch, Aaron Mueller, and Yonatan Belinkov. Arithmetic Without Algorithms: Language Models Solve Math With a Bag of Heuristics. In *International Conference on Learning Representations*, pp. 55939–55965, 2025.

Gustaw Opielka, Hannes Rosenbusch, and Claire Stevenson. Causality $\neq$ invariance: Function and concept vectors in llms. In *International Conference on Learning Representations*, pp. 83393–83417, 2026.

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. doi: 10.18653/v1/2023.emnlp-main.435.

Sankaran Vaidyanathan, David Arbour, Aaron Mueller, Scott Niekum, and David Jensen. The curse of multiple mediators: Hidden interaction effects in activation patching. arXiv:2606.27510, 2026.

Alexandre Variengien and Eric Winsor. Look before you leap: Universal emergent mechanism for retrieval in language models. In *International Conference on Learning Representations*, pp. 69352–69394, 2025.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the Wild: a Circuit for Indirect Object Identification in GPT-2 small. In *International Conference on Learning Representations*, 2023.

Zhengxuan Wu, Atticus Geiger, Aryaman Arora, Jing Huang, Zheng Wang, Noah Goodman, Christopher D. Manning, and Christopher Potts. pyvene: A Library for Understanding and Improving PyTorch Models via Interventions. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: System Demonstrations)*, pp. 158–165. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.naacl-demo.16.

Yao Yan. Post-routing arithmetic in llama-3: Last-token result writing and rotation-structured digit directions. arXiv:2602.19109, 2026.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 Technical Report. arXiv:2505.09388, 2025.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations*, pp. 1651–1678, 2024.

Ziqian Zhong, Ziming Liu, Max Tegmark, and Jacob Andreas. The Clock and the Pizza: Two Stories in Mechanistic Explanation of Neural Networks. In *Advances in Neural Information Processing Systems*, volume 36, pp. 27223–27250, 2023. doi: 10.52202/075280-1186.

A EXPERIMENTAL DETAILS

A.1 TASK CONSTRUCTION AND READOUT

All natural questions are retained for accuracy evaluation. Donor–recipient swaps additionally require correct answers and equal token lengths. Initial same-start pairs compare offsets n, $n+3$, and $n+10$ for $n = 1, \ldots, 4$. The two donor offsets induce the same cyclic answer change. Semantic pairs change both operands. The stricter subset requires correct natural answers to the full crossed square.

Offsets are written as tokenizer-checked English words. Qwen receives the cycle rule and "The answer is **" without an answer. These measurements are separate from Llama's original protocol. All layer and head indices are zero-based. Post-layer states are complete block outputs.

A swap at position set S substitutes source block outputs $h^{\ell}_{S}(x_s)$ for recipient outputs $h^{\ell}_{S}(x_d)$. Identity and full-context swaps reproduce the corresponding full-vocabulary logits. At the final block, only the final-position output affects the next-token prediction.

A.2 PROMPT WORDING AND ANSWER SCORES

Qwen uses BF16, SDPA attention, its nonthinking chat template, and the fixed answer prefix. Each task lists its weekday or month cycle. The following rule accompanies the list:

```
Moving forward one step reaches the next item; after the last item, wrap
↪  to the first. Count the starting item as step zero.
```

The original month question is "If this month is {label}, what month will it be in {word} months? Give only the month." The weekday question substitutes "today," "day of the week," and "weekday." New prompt bodies are:

```
Starting from {label}, count forward {word} {unit}. Name the resulting
↪  {target}.
Take {label} as the starting {target}. Advance by {word} {unit}; what is
↪  the resulting {target}?
Begin at {label}. Move ahead by {word} {unit} along the cycle. Which
↪  {target} do you reach?
A tracker starts on {label} and goes forward {word} {unit}. Report the
↪  final {target}.
```

These correspond to Count, Advance, Journey, and Tracker. Placeholders use the task's starting label, spelled-out offset from one to fourteen, day/month unit, and weekday/month target. New wordings use the singular unit for offset one. Original wording uses the plural form.

Preferred answers are decoded from the unrestricted first-token argmax and stripped of surrounding whitespace. A calendar-label score uses log-sum-exp over the distinct single-token encodings of the label with and without a leading space. Continuous margins use these scores. A calendar candidate need not be the preferred token in the full vocabulary.

A.3 HEAD SELECTION AND VALUE INTERVENTIONS

Llama head selection uses only base-model development pairs with an offset difference of ten and the "It is" prefix. Candidates lie in layers 19–31. The score is the mean increase in the donor-versus-recipient weekday margin after a single-head restoration. The selected heads are L28H18, L30H24, L28H16, and L30H26. They remain fixed across prefixes, instruction formatting, and months.

Head outputs are patched before the attention output projection. A value intervention at layer ℓ writes

$$\Delta h_{\ell} = \sum_{h \in H_{\ell}} W^{O}_{\ell,h} \sum_{t \in S} a^{(d)}_{\ell,h,t} \big(v^{(s)}_{\ell,h,t} - v^{(d)}_{\ell,h,t} \big).$$

Here H_ℓ is the selected head set, S the donor positions, and $a^{(d)}$ the recipient attention weights. Grouped-query attention uses the model's actual key–value mapping. Natural ablation substitutes task-, format-, and length-conditioned mean values without answer labels.

Random heads occupy the same layers and exclude the selected heads. Their projected writes are norm-matched layer by layer. Same-result and random-direction controls match the different-result write norm. Norm matching uses the projected BF16 write, with a maximum relative discrepancy of 0.269%.

A.4 Independent donors and direct reading

The source assignment and collision checks are described in Section 3. For unit donor U, final-position donor L, and recipient start s_d, the competition test compares $f(s_d, o_U)$ with $f(s_d, o_L)$. We report their signed label-score margin and whether the preferred answer matches either candidate.

The fixed layer-23 heads are 30, 0, 25, and 11. Selection uses the Original weekday condition. All direct-write and MLP interventions affect only the final position. Signed contrasts are retained without clipping or normalizing by near-zero joint effects.

Three random unit-value controls match the joint-minus-final-only norm after BF16 conversion. The maximum relative discrepancy is 0.035%.

A.5 Semantic contrast and descriptive statistics

Offset contrasts are summarized by their signed means and medians. Random-value effects first average the three directions within each recipient group. Shared donors and repeated questions are grouped rather than treated as independent samples.

Table 2: Unit-token contrasts under same-offset and same-result donors.

Task	Wording	Mean χ_U	Median	Positive groups	Unique groups	Unique meanχ_U
Weekday	Original	4.289	2.875	22/24	8	6.469
Weekday	Count	1.818	1.281	22/24	9	1.437
Weekday	Advance	1.203	0.969	21/24	9	0.993
Month	Original	8.172	6.125	19/24	20	7.834
Month	Count	5.250	3.938	23/24	15	5.733
Month	Advance	4.466	3.719	24/24	20	4.453

Table 3: Offset contrasts recovered through different reading paths.

Task	Wording	K	V	KV	Direct write	Four heads	Random V
Weekday	Original	-0.328	3.299	4.112	4.167	3.852	0.096
Weekday	Count	0.091	1.471	1.805	1.794	1.734	0.022
Weekday	Advance	0.125	0.492	1.219	1.279	1.089	0.016
Month	Original	-0.198	6.661	7.669	7.674	6.997	0.205
Month	Count	0.878	3.836	5.026	5.031	4.828	0.115
Month	Advance	0.130	2.596	4.443	4.378	4.091	0.063

Table 4: Signed offset contrasts remaining after reverse restoration.

Task	Wording	Restore unit V	Restore write	Restore MLP 23/24
Weekday	Original	-0.234	0.146	0.766
Weekday	Count	0.201	0.021	0.281
Weekday	Advance	0.156	0.000	0.221
Month	Original	0.174	0.406	1.201
Month	Count	0.917	0.073	0.495
Month	Advance	0.253	0.161	0.555

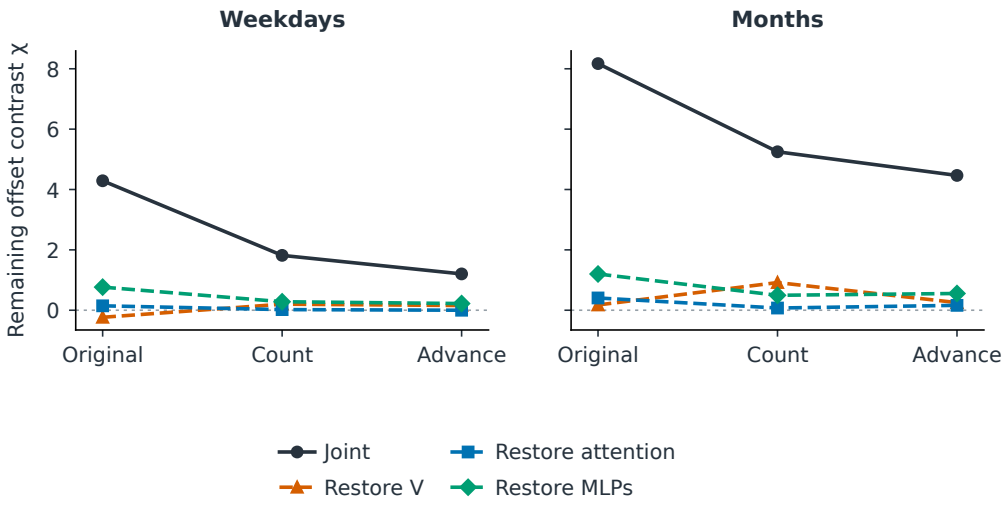


Figure S1: Restoring unit values, complete attention writes, or later MLP outputs attenuates the offset contrast.

B Position and Context Controls

B.1 Unit-Token Swaps Can Control the Crossed Answer

Qwen's main carrier is the unit word after the offset. Ablating its varying state reduces natural accuracy. The assistant boundary lacks the corresponding effect. Crossing both operands shows that a unit swap mainly produces the recipient start plus the donor offset. A donor with the same complete result can still change the answer.

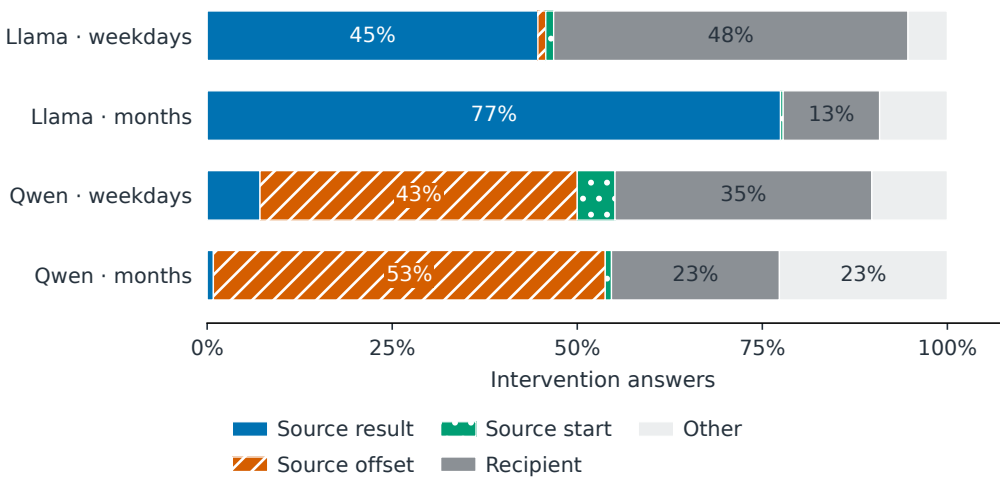


Figure S2: Crossed starts and offsets distinguish result transport from operand transport despite similar donor-answer effects.

The layer-22 final position has a similar offset effect. By layer 24, it mainly transfers the complete result. Requiring both crossed questions to be naturally correct preserves the mixed-answer pattern. Low donor-result recovery can coexist with effective operand transport.

B.2 ANSWER-BOUNDARY VALUES INSTEAD SUPPORT RESULT TRANSPORT

Llama value swaps keep recipient queries, keys, and attention weights fixed. Reverse interventions restore recipient head outputs in the donor run. The fixed head set, mean-ablation operator, and norm-matched controls are defined in Appendix A.

In instruction-tuned Llama, swapping the final position after layer 18 recovers few donor answers. Full-context swapping exactly replays the donor. Restoring the selected later readout heads recovers nearly all donor answers. Reverse restoration disrupts that recovery. The difference depends on subsequent attention reading.

Position-specific value swaps locate the main carrier at the answer-format boundary. This is the colon in plain-text "A:" or the start marker preceding the assistant header. After the relevant reading, boundary influence declines and final-position influence rises. Figure S3 reports the layer dependence.

Natural mean ablation of these boundary values reduces weekday and month accuracy. Matched random heads remain near baseline. Ordinary addition and naturally solvable twelve-hour clock questions do not show the corresponding loss. The effect has a task-specific role under the measured protocol.

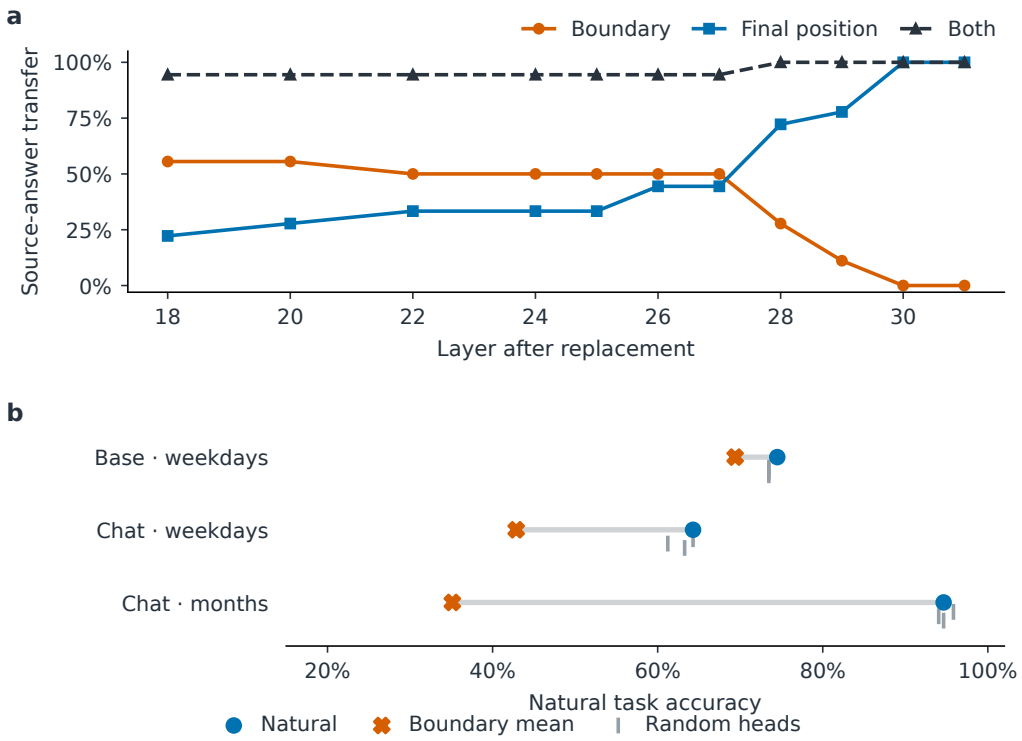


Figure S3: Answer-boundary information is read into the final position and supports natural calendar answers.

Donors with the same result usually preserve the recipient answer even when both operands differ. Different-result donors steer the answer toward their result. Scaling same-result changes to the same write norm does not reproduce this effect. With all four counterfactual answers distinct, Llama month swaps mainly transfer the complete donor result. Restoring later recipient MLP outputs reduces result following.

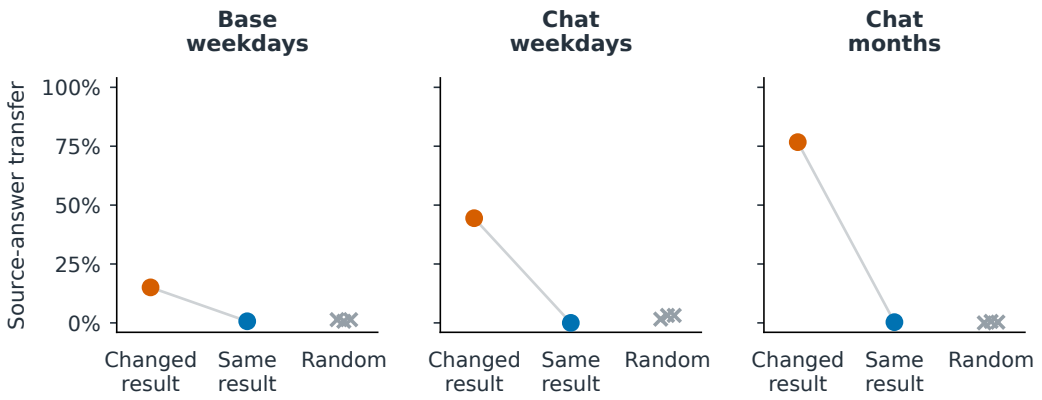


Figure S4: After matching write magnitude, the direction of the boundary-value effect still follows the donor result.

B.3 Operand identity persists beyond one carrier

Positions selected with the original wording stay fixed for "count forward" and "advance" prompts. Each recipient again has anchor, same-offset, and same-result donors. Same-offset invariance requires both swaps to produce the expected mixed answer.

Across the three wordings, the starting-point prefix transfers the donor start with the recipient offset. Joint layer-22 unit and final-position swaps mainly transfer the complementary combination. Layer-24 final-position swaps more often transfer the donor result. Same-offset and same-result donors separate operand identity from result similarity.

Early unit-only transfer is less stable. On identical semantic pairs, both new wordings sharply reduce the mixed-offset effect at layer 18. The same pattern remains when both offsets exceed one. This comparison holds the plural unit form constant. The carrier's ability to control later computation depends on context. Layer scans nevertheless show a common later transition toward complete-result transfer at the final position.

B.4 Complementary positions complete the calculation

The starting-point prefix and layer-22 final position transfer complementary operands. Combining them increases complete donor-result recovery. Some pairs still require more context. Adding the donor's offset-token state resolves most remaining cases. In two original month pairs, the penultimate "is" token supplies an additional conditional contribution.

The penultimate token contributes only when the complementary operand states are also transferred. A correct result at layer 22 can therefore require information outside the starting-point prefix and final position. This conditional contribution differs from the complete-result transfer at layer 24.

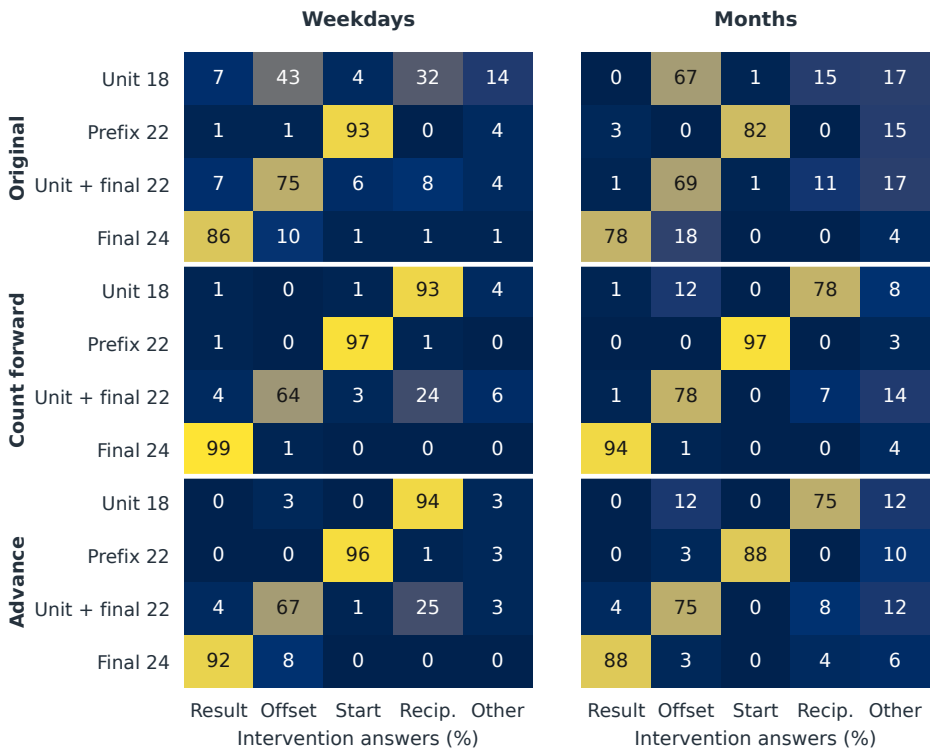


Figure S5: Prefixes and joint states transfer complementary operands across wordings. Early unit-only effects change with wording.

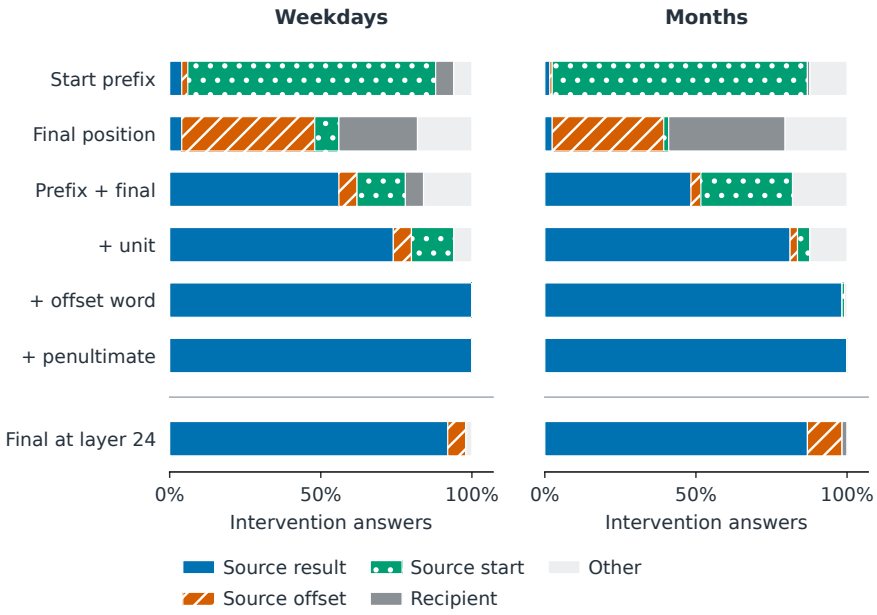


Figure S6: Complementary starting-point and offset states progressively recover the donor result before final-position consolidation.

C INTERACTION AND RESTORATION EXCEPTIONS

C.1 SEMANTIC MEANING OF THE INTERACTION

Let $m_t(r, w)$ be the fixed-candidate margin for unit role t. The coordinate r controls the layer-22 final residual. The coordinate w selects a layer-23 write from the interpolation runs. Their endpoint interaction is

$$I_t = m_t(1,1) - m_t(1,0) - m_t(0,1) + m_t(0,0).$$

This interaction measures the crossed write and residual intervention. Writes depend on donor role, and other positions remain active. The semantic contrast satisfies

$$\chi(r,w) = \frac{m_A(r,w) + m_B(r,w)}{2} - m_C(r,w), \qquad I_\chi = \frac{I_A + I_B}{2} - I_C.$$

Donors A, B share an offset and A, C share a result. Their contrast tests whether the interaction depends on the transferred offset or the source answer (Vaidyanathan et al., 2026).

C.2 IDENTICAL ANSWERS WITH DIFFERENT CANDIDATE PREFERENCES

The same-answer subset is selected after intervention by requiring the full-vocabulary preferred answer to match across A, B, C at full strength. This subset describes effects omitted by the answer metric. The table reports absolute contrasts to avoid cancellation after restoration.

Table 5: Pairs with identical preferred answers retain an offset contrast that shrinks after write restoration.

Task / wording	Same-answer pairs	Joint contrast	Restore V	Restore write
weekdays / journey	21/24	1.146	0.283	0.104
weekdays / tracker	15/24	2.548	0.446	0.158
months / journey	19/24	3.132	0.451	0.293
months / tracker	14/24	5.246	1.339	0.362

Figure 1 uses the same-answer group nearest the pooled median contrast. Its recipient is March plus three months; donors are August plus thirteen, April plus thirteen, and January plus eight. All joint interventions answer June. The April-versus-November contrast falls from 2.188 to 0.375 after write restoration.

A "tracker" month pair behaves differently under value restoration. The recipient is November plus fourteen. Donors A, B, C are February plus eight, April plus eight, and September plus one. All joint swaps answer July. The original contrast is 9.250. Value restoration changes it to -10.625. Whole-write restoration gives 0.813.

After value restoration, margins for A and B are -6.250 and -6.500, close to the final-only value of -6.375. Role C retains 4.250, reversing the contrast despite all three runs answering January. Whole-write restoration brings the margins together. Restoring unit values therefore leaves a role-dependent attention effect that complete-write restoration removes.

C.3 PRECISION AND POSITION-SET SENSITIVITY

Position-set minimality is sensitive to precision near a decision boundary. In the October-plus-ten to March-plus-seven transfer, 17 of 18 individually necessary BF16 positions become individually removable in FP32. Another pair loses its naturally correct donor answer in FP32. The stable comparison is the aggregate operand effect, rather than the smallest successful position set for either example.

Excluding singular-unit changes retains the wording effect. For layer-18 unit-only offset recovery, the original, Count, and Advance wordings yield 6/14, 0/14, and 0/14 in weekdays, and 12/22, 2/22, and 0/22 in months. All comparisons use the same semantic pairs within a task.

D RECEIVING-STATE EXPERIMENTS

D.1 FACTORIAL INTERVENTION AND PREDICTION

Residual and write strengths are $0, 0.25, 0.5, 0.75, 1$. Diagonal combinations pair each write with its original receiving state.

Section 4 describes development and calibration for the two regressions. Each nonendpoint strength has a separately fitted three-parameter rule for the joint-minus-final-only margin. The measured final-only margin at that target strength is added to the predicted effect before scoring. The generic features are an intercept, the role-specific endpoint effect, and the mean endpoint effect. Semantic features are an intercept, the mean endpoint effect, and the endpoint offset contrast multiplied by 1/3 for roles A/B or $-2/3$ for C. Ridge strength is 0.01 and excludes the intercept. Both models use the same calibration information. The outcomes are the two-candidate margin and its sign. Errors are averaged within each recipient group before task–wording aggregation. Table 1 gives their mean and between-group sample standard deviation, with divisor $n - 1$.

Receiving-state variation is the within-group standard deviation of χ_U across five residual strengths at a fixed write, computed with divisor five. The table compares this variation before and after fixing layer-23 and layer-24 MLP outputs.

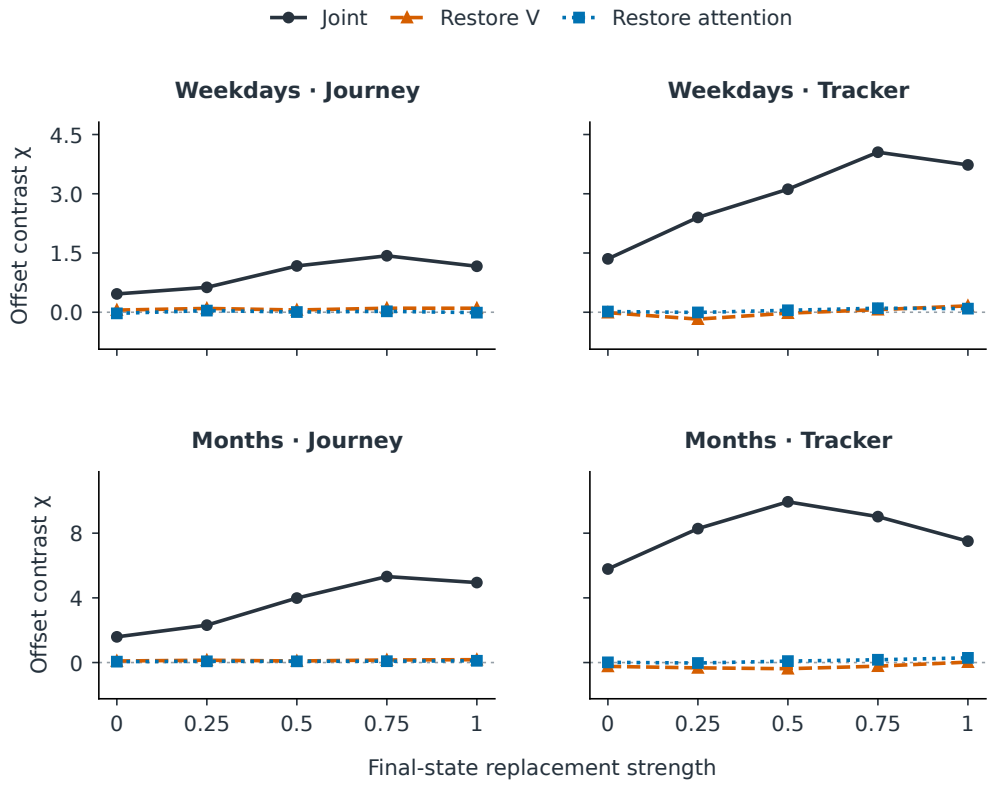


Figure S7: Offset effects vary with the receiving state in new wordings and shrink after write restoration.

Table 6: Receiving-state variation under fixed attention writes and MLP clamps.

Task / wording	No MLP clamp	Clamp 23	Clamp 24	Clamp both
months / journey	2.227	1.573	1.678	0.710
months / tracker	2.838	1.490	1.690	0.735
weekdays / journey	0.612	0.397	0.420	0.227
weekdays / tracker	1.456	0.564	0.750	0.504

E SAME-DONOR PATH COMPARISON

The table compares the donor-offset candidate with the recipient answer. It reports the mean increase in their margin over the final-only run. Unit and final positions use the same donor. Attention and MLP restoration put final-only outputs into the joint run. This is distinct from the independent-donor contrast. Three random four-head sets exclude the selected heads and match their projected BF16 write norm within 0.140%.

Table 7: Same-donor path effects relative to final-position-only swapping.

Task / wording	Joint state	Direct attention	Four heads	Random heads	Restore attention	Restore MLP
Weekday Original	5.823	5.833	5.010	0.014	0.203	1.500
Weekday Count	3.672	3.391	3.250	0.016	0.172	0.641
Weekday Advance	1.922	1.932	2.005	0.075	-0.042	0.214
Month Original	7.536	7.260	6.615	0.087	0.224	1.417
Month Count	7.490	7.214	6.771	0.078	0.359	1.208
Month Advance	7.281	7.130	6.490	-0.219	0.411	1.365