COMPUTED AHEAD, COPIED BLIND: UNCHECKED HANDOFFS IN CHAIN-OF-THOUGHT ARITHMETIC

Anonymous authors
Paper under double-blind review

ABSTRACT

When one position in a chain of thought hands a computed result to the next, does the receiver check it against the other input it can see? We study the final combination of two products in two models' own solutions to 749 linear systems, using sublayer readouts, attention cuts, and value substitutions. (1) Erroneous delivery passes unchecked on the copying route. The answer's first digit is computed at the equals position before it is written. A space receiver copies it through late attention, writing a donor's digit with probability near one with the correct operands in view. A minus receiver computes from the operands and shows no donor following. After operand damage, correct content at the equals position supports some answers, while most remain wrong. (2) The receiving state controls route selection and content influence. Changing the carrier changes input dependence. Cutting late equals reads makes copying fail; also cutting early equals reads restores operand support. Recovery reuses natural computation's operand-reading heads in both models. (3) Enabling operand computation and protecting its output are separate. With all operand reads at the equals position cut, cutting its remaining delivery restores further correct answers. On the computing route, delivered content shifts the output distribution while answers remain largely stable; after operand damage, it changes answers, mostly without donor following. Faithful transfer and effective verification are therefore different capabilities.

1 INTRODUCTION

Chain-of-thought reasoning turns intermediate answers into inputs for later steps (Wei et al., 2022). Before a number enters that text, it can already pass between positions inside the model. When the model writes `47 + 24 = 71`, processing the space after = predicts `7`. This position could compute the digit from the operands, copy a result from the equals sign, or compare the two. When the answer is correct, all three can produce the same text. Changing delivered content tests which input controls that first written digit, linking reasoning faithfulness (Turpin et al., 2023) to a specific generation decision.

We study the final combination of two products in a determinant, repeated throughout two models' own solutions to 749 linear systems using Cramer's rule. The inputs needed for correction are in the context, the answer is known, and errors can be matched to specific digits. Readouts locate numerical information; cuts test dependence on particular reads; substitutions follow how injected content changes the answer.

The central contrast is sharp: substituting values from another problem (the *donor*) for the late delivery values at the receiving problem's equals position makes a space receiver write the donor's digit in all 566 positive 14B units. Receivers processing a minus sign show zero donor following across 553 negative units, and the 30B model shows the same contrast. The space copies an answer; the minus derives it from the operands. Operand-value substitutions establish what derivation means: the output can follow the receiver's own expression with an injected digit—a *mixed answer* distinct from both the original and donor answers. After operand damage, late equals reads still support some answers, yet most remain wrong (§6).

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.

The receiving state shapes what delivered content can do. Changing the carrier changes how the receiver uses delivered content; with late equals reads cut, cutting early equals reads makes the answer depend on the operands. Early equals reads change later operand access; the same operand-reading heads support natural and recovered computation in both models. Under equals-side damage, cutting the remaining late delivery restores additional correct answers. Enabling operand support and preserving its effect on the output are separate.

Effective checking concerns whether the other available input helps correct an erroneous answer before the digit is written. We test this by replacing delivered content while preserving operand access, and by damaging operand reads while preserving the equals position's computation and delivery.

Three questions organize the evidence: how readable content is used (§4), how the route is selected (§5), and how delivered content acts on the computing route (§6).

2 SETTING: POSITIONS, INPUTS, AND INTERVENTIONS

Following one handoff. We study Qwen3-14B and Qwen3-30B-A3B (Yang et al., 2025), using each model's own frozen solutions to the same 749 two-variable linear systems. In `47 + 24 = 71`, processing the equals sign predicts a space; processing that space predicts `7`. For negative results, the intervening token is a minus sign. We call the position processing the space or minus the *receiver*. The space or minus itself is the *carrier*: the equals position's output token and the receiver's input token (Figure 1a).

A *unit* is one determinant-combination event with a two-digit result. We track the last operand digit o, equals sign e, carrier c, and first answer digit d_1; teacher forcing supplies the recorded text at each step. Correctness uses the recorded expression's arithmetic value. The main endpoint is full-vocabulary top-1 first-digit correctness at c; digit-restricted choices are labeled. Additional endpoints predict the carrier and, given the recorded first digit, the second digit.

Fixed pools and defined subsets. The main series uses fixed pools of 1,236 units in 14B and 600 in 30B. Main tables report units whose two products are both nonzero: 566 positive / 553 negative units in 14B and 235 / 246 in 30B. Two separate filters distinguish output destinations. The *non-overlap* subset distinguishes answers from operand digits: the correct first digit differs from both operand first digits. The *strict* subset separates original, donor, and mixed first digits; donor and mixed digits also differ from both operand first digits, while the original may match an operand. Table 4 lists subsets and head cohorts by protocol; §6 identifies the earlier 14B pool separately.

Readouts locate candidate content. Linear answer-digit probes cover the embedding and every attention and feed-forward sublayer at o, e, and c, separately by sign with five folds grouped by problem. The 30B probes train on a 1,688-unit superset and report held-out predictions on the fixed pool. Attention exchanges information across positions; feed-forward layers transform each position in place. Readout changes locate candidate sites for the causal tests below.

Cuts measure conditional dependence. We cut a read by setting its attention logits to $-\infty$ at selected query positions and layers. Operand cuts target the *semantic operand set*: all explicit writings of the two products and their attached signs anywhere in the preceding text; cutting only the local expression leaves earlier writings available (Appendix B). A *route* is a pattern of input dependence: which reads change the output when cut under a stated condition. In the tested natural copying state, late equals cuts change first-digit top-1 outputs, while operand cuts preserve them. Natural computation combines operand dependence with arithmetic transformation of substituted input. Combining cuts and substitutions tests how these patterns change with content and receiver state.

Substitutions distinguish input use from donor following. We replace a position's cached value vectors at selected layers with a donor's, keeping keys fixed. Equals-value substitutions test donor following. Operand-value substitutions test arithmetic input use: the *mixed answer* is the receiver's expression evaluated with the injected digit (Figure 2a). Strict-subset selection distinguishes its

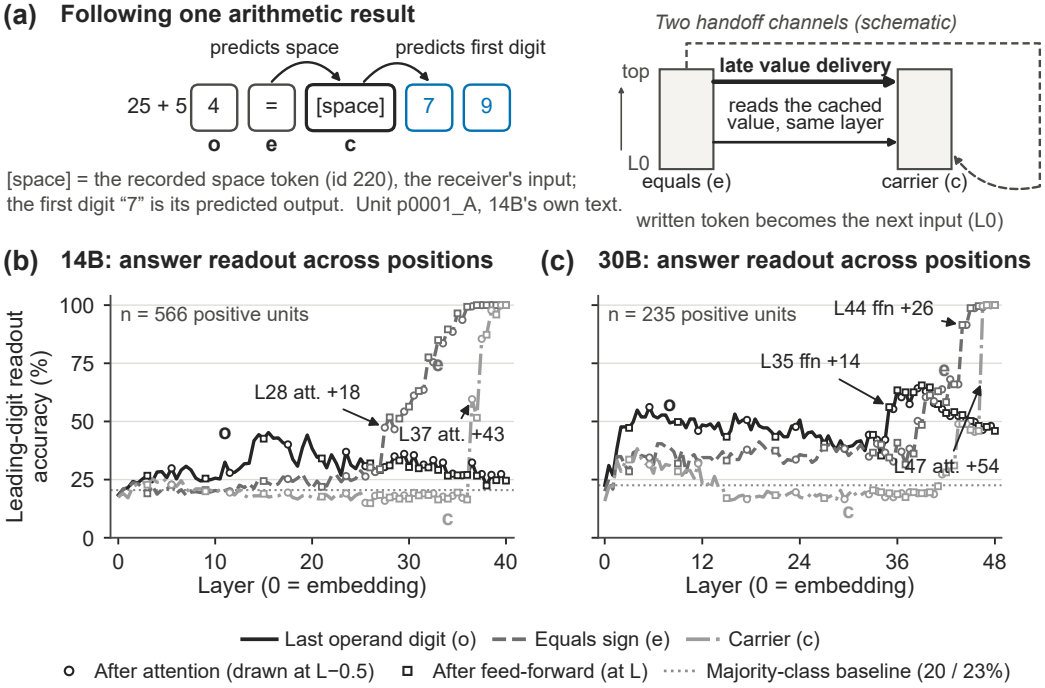


Figure 1: The first answer digit is readable before it is written. (a) One recorded unit of Qwen3-14B's own solution text: after the last operand digit (o), the equals sign (e) predicts the recorded space token, and that carrier (c) predicts the first answer digit. Right: two handoff channels—same-layer cached-value reads (solid) and the carrier entering the receiver at layer 0 (dashed). (b, c) Held-out leading-digit readouts after attention (circles, drawn at $L-0.5$) and feed-forward (squares) sublayers; positive units with two nonzero terms. Arrows mark the annotated curves' largest single-point rises; the dotted line is the same-subset majority-class baseline. Readouts locate content; subsequent interventions establish its use.

leading digit from the original and donor answers. Output destinations track content use; restricted exposure locates the path; intervened readouts follow content along it.

Controls and comparisons. Cuts use size-matched random-position controls; substitutions use null substitutions. Each contrast uses its own run's reference. Confidence intervals are 95% intervals clustered by problem. Results for the two donors appear in donor-1, donor-2 order; ranges across donors give the smaller and larger result. Donor probabilities are unit means. We count gained and lost donor matches separately, relative to each unit's pre-substitution output (Appendix G).

3 The first digit is available before it is written

Where does the answer's first digit become available? We follow the final combination of two products across the neighboring positions in Figure 1a.

The three-position readouts place first-digit information before the digit is emitted (Figure 1): it rises at the last operand digit, then at the equals position, and only in the final layers at the space carrier (largest rises after layers 35, 44, and 47 in 30B; 28/32 and 37–38 in 14B). In 14B, a natural minus carrier's first-digit information rises after layer-28 attention and strengthens through subsequent feed-forward sublayers, most after layer 32; in 30B, it rises most after layer-44 feed-forward processing. Where operands are read and where the first digit becomes linearly readable are different measurements; the arithmetic itself is identified by the interventions that follow.

These profiles suggest receiving the first digit through attention or computing it locally. The next experiments test which readable content controls the output.

Table 1: Natural input dependence, measured by first-digit correctness. Each intervention row cuts one specified read. Baseline pairs list positive, then negative units. LATE denotes layers 37–40 (14B) and 47–48 (30B); KEY denotes layers 28–31 and 41–44. Size-matched random controls reproduce their natural baselines.

Reader	Read cut	14B	30B
baseline	—	566/566; 553/553	235/235; 245/246
space	equals, LATE	**129**/566	**69**/235
space	products, all layers	566/566	235/235
minus	products, KEY	**235**/553	**111**/246
minus	products, all layers	166/553	56/246
minus	equals, LATE	553/553	245/246
equals	products, KEY	**145**/566	**97**/235

4 TWO ROUTES TO THE SAME WRITTEN ANSWER

A readable answer could be copied, used in computation, or replaced. Cuts establish input dependence; substitutions track the content that controls the output.

4.1 NATURAL RECEIVERS DEPEND ON DIFFERENT INPUTS

The carriers have contrasting dependence patterns (Table 1). For the natural space, cutting late equals reads reduces correctness, while cutting all product writings preserves the top-1 outputs. For the natural minus, operand cuts reduce correctness and cutting late equals reads preserves it. The equals position itself depends on operand writings.

4.2 SUBSTITUTIONS SEPARATE DONOR FOLLOWING FROM INPUT USE

To establish copying, we change the answer available at the equals position while keeping the receiver's own operands in view. Substituting the late delivery values makes the 14B space write the donor digit in 566/566 units, at recorded probability rounding to one; the 30B space follows in 235 and 232 of 235 with layers 47–48 substituted (210 and 214 with layer 48 alone; both layers read the equals position and the later read decides, Appendix C). Natural minus receivers show zero donor following in either model (Figure 2)—a direct contrast in donor following under the same content intervention.

To establish computation, we substitute values at the last operand digit. The mixed answer is the largest output category in every strict subset (Figure 2c; Table 11 gives all destinations). The output also moves in the direction predicted by the arithmetic. Here changing the final operand digit makes the mixed leading digit one above or below the original. In a post hoc comparison with the opposite neighbor, the arithmetic-consistent direction wins by 25–29 percentage points across all eight model–sign–donor groups, and wins separately in both directions (Appendix C.2). This ties the response to the injected digit's role in the operation.

Equals-only exposure preserves the positive first-digit effect; exposure starting after the equals preserves the negative effect. Thus the equals computes the digit copied by the space, and the minus computes its own. Given the true first digit, the second digit's positive mixed response is strongest under full exposure; its negative response remains strong under later exposure. Table 9 separates both digits and all exposures, including the later exposure covering the carrier and the next prediction step. Figure 3 summarizes, layer by layer, where Sections 3–4.2 place each position's reads, readable answer, and delivery in the two models.

5 HOW THE ROUTE IS SELECTED

Can the copying receiver use its own operands when delivery fails? We test three ways to change input dependence: the carrier, the receiver's early equals reads, and the equals position's processing (Figure 4).

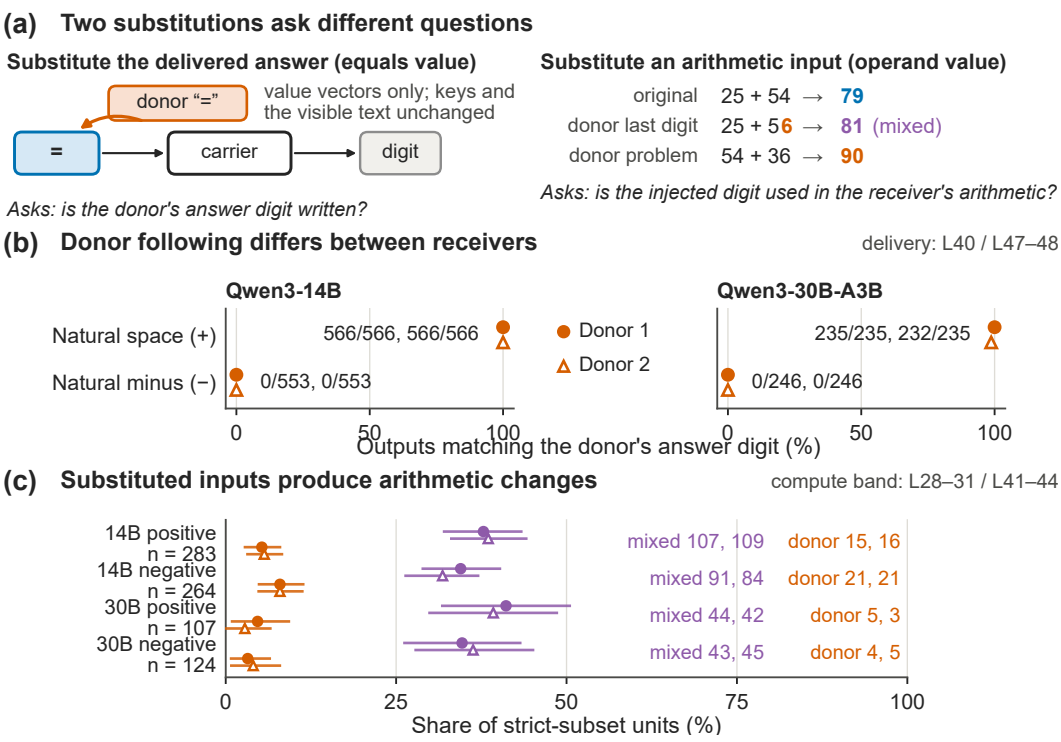


Figure 2: Substitutions separate donor following from arithmetic input use. The endpoint is the full-vocabulary top-1 first digit. (a) The equals-value substitution changes the delivered answer; the operand-value substitution defines a mixed expression. Keys and visible text stay fixed. (b) The delivery-layer substitution makes natural spaces write the donor digit at mean probability near one (1.000 rounded in 14B; 0.999 and 0.988 in 30B); natural minuses show zero donor following. Counts are exact, with donors in fixed order. (c) The compute-band operand-value substitution on the strict subset: original, donor, and mixed leading digits are distinct. Lines are pointwise 95% problem-clustered intervals. Equals-only exposure preserves the positive first-digit effect; exposure after the equals step preserves the negative effect (Appendix C).

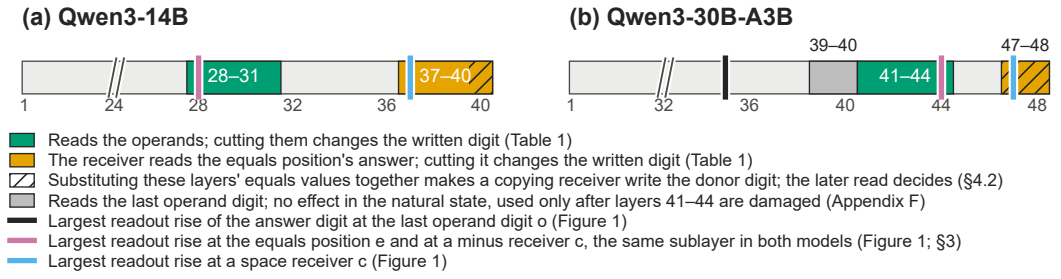


Figure 3: Layer map of the handoff.

5.1 The carrier changes input dependence

We keep the equals cache fixed and replace a natural minus with a space. The donor becomes the highest-probability digit in two thirds of 14B units and a third of 30B units, and is emitted in about a quarter (Figure 4a). Whenever these criteria differ, a re-written minus outranks the donor digit. One carrier changes use of delivered content, exposing two competitions: which digit to write, and whether to write a digit yet.

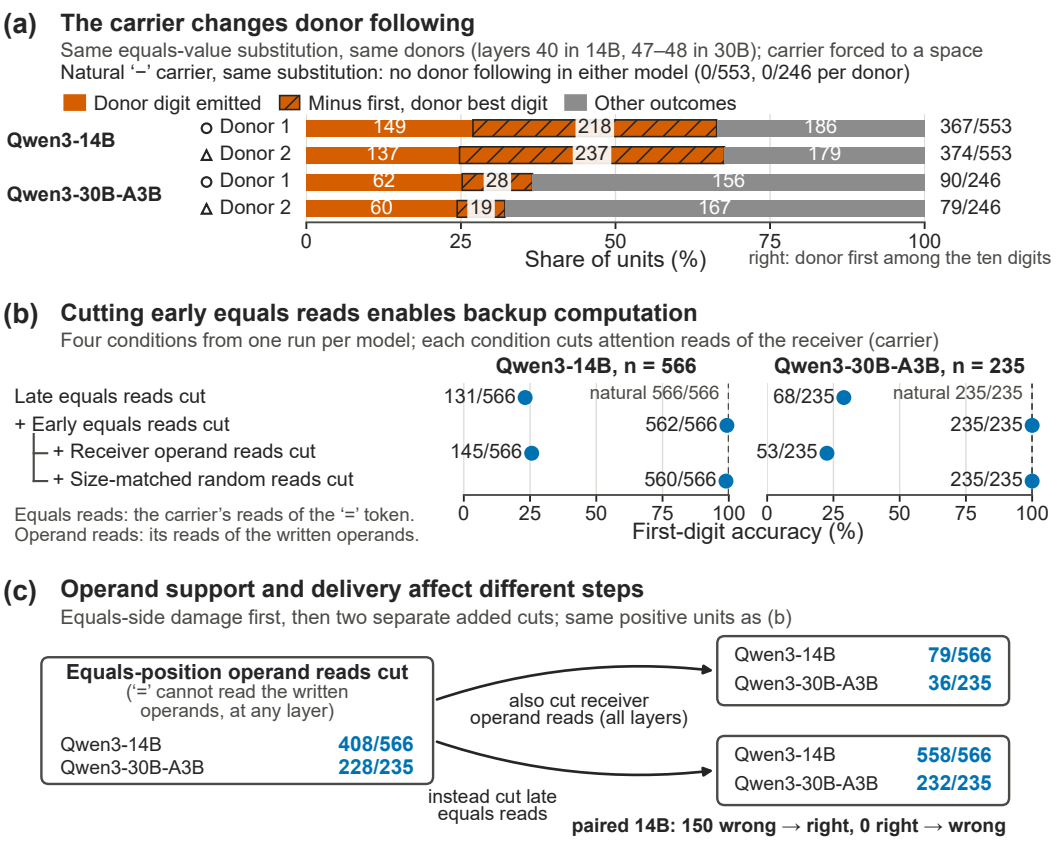


Figure 4: Enabling operand computation and protecting its answer from the remaining delivery are separate questions. **(a)** The same equals-value substitution and donors as Figure 2, negative units: a natural minus never emits the donor digit, while a forced space makes it the top-1 token (solid orange) or the top digit behind a re-written leading minus (hatched); categories are exclusive and sum to n; right totals give the ten-digit criterion. **(b)** Carrier-side read cuts, positive units, one run per model; late equals reads are layers 37–40 / 47–48, early equals reads 1–12 / 1–7, and operand reads cover the semantic operand set at all layers. **(c)** The two added cuts depart from the same equals-damaged start, not a chain; cutting the remaining late delivery raises 14B accuracy by $+26.5$ pp (95% problem-clustered CI $[+23.1, +30.0]$; paired 150 wrong→right, 0 right→wrong; 30B 4 and 0). Where the true first digit differs from both operand first digits, the receiver-operand cut leaves 0/386 and 0/151 correct, down from 252/386 and 144/151. Arrows mark added interventions between conditions, not information flow.

Masks confirm dependence on late equals reads in both models. The forced space retains measurable operand support in 30B, while the 14B operand cut preserves digit accuracy (Appendix B). An earlier reverse test, forcing a minus onto positive 30B units, also retains dependence on equals reads (Appendix B.1). The carrier's effect depends on the receiving state.

5.2 Early equals reads select the route

The token change shifts computation toward use of delivered content; we now test the other direction by preserving the natural space and changing its equals reads. With late equals reads cut, additionally cutting early equals reads raises correctness from 131 to 562 of 566 in 14B (23% to 99%). In 30B, it rises from 68 to 235 of 235 (29% to 100%). Each comparison uses the recovery run's own reference; Table 1 uses the separate natural-dependence run (Appendix G).

Cutting the receiver's operand reads undoes most of the recovery, while size-matched random cuts preserve it (Figure 4b). Recovered answer readouts also weaken under operand cuts. Both measurements locate the recovered computation at the receiver.

Table 2: Shared operand-reading heads, different dependence across states. Full-vocabulary first-digit correctness before → after the selected twenty-head cut. These are fixed-pool subsets of the head experiments; input sets, recovery cuts, and full-cohort results are in Appendix H.

Receiver state	14B	30B
Natural space	$48/48 \rightarrow 48/48$	$44/44 \rightarrow 44/44$
Natural minus	$50/50 \rightarrow 24/50$	$53/53 \rightarrow 25/53$
Recovered space	$46/48 \rightarrow 20/48$	$44/44 \rightarrow 19/44$

Position controls locate the effect at the equals token: matched cuts of a random or adjacent operand position leave performance near delivery failure. Prefix scans narrow the early equals reads to specific layers (Appendix D). This *backup* is the receiver's own operand computation, made effective by changing how it reads the equals position.

5.3 Recovery reuses the heads of natural computation

Recovery uses the same operand-reading heads as natural computation. To identify these heads, we rank layer–head pairs by natural writes from numerical inputs, then cut the top twenty heads' reads in three states (Table 2). The cuts hurt both natural and recovered computation while leaving natural copying unchanged. Random twenty-head cuts preserve each baseline.

What changes when these components start supporting recovery? The companion 30B recordings measure their attention with operand reads undamaged. Across 74 positive units, the selected heads' mean attention on the operands rises from 0.015 with late equals reads cut to 0.514 with early equals reads also cut; the natural-minus reference is 0.606. Both positive conditions have the same late cut. Changing early equals reads increases later operand access; head cuts show that these reads support the recovered answer.

A 14B decomposition separates removing equals writes from scaling up the remaining writes. With late equals reads cut, early removal alone raises correctness from 14 to 41 of 60; the standard cut reaches 57, while scaling up alone stays at 14. Removing early equals writes releases substantial operand computation, and the rescaling strengthens it (Appendix I).

5.4 The equals position's state selects the route; late delivery still competes

We next change the equals position's own computation while initially keeping the receiver's reads available. Mask the equals position's operand reads at all layers, and the space retains 408/566 correct answers in 14B and 228/235 in 30B. In the non-overlap subset, adding a cut of the receiver's operand reads leaves no correct answers (Figure 4c). The receiver's own operands now support its output.

Cutting only the equals position's main operand-reading band differs: it leaves 144/566 and 94/235 correct in this run. Adding a cut of the receiver's operand reads changes these counts only to 140 and 92. Damage confined to the equals position's numerical computation leaves the receiver reliant on the damaged delivery; broader damage changes the input dependence and permits substantial recovery.

An earlier 30B experiment isolates shallow equals-position processing: with deep equals-position operand reads already cut, adding a shallow cut restores 25 to 65 of 75 answers. All eight middle-layer configurations show a gain (Appendix D.1). Both equals-position processing and receiver reads therefore affect operand-supported recovery.

In this all-layer equals-position damage background, cutting the remaining late delivery raises 14B correctness from 408 to 558 of 566 (72% to 99%): 150 wrong-to-right changes and zero reverse changes. The 30B gain is four (Figure 4c). One intervention makes operand computation support the answer; the other cuts a competing delivery that limits its success.

6 How delivered content acts on the computing route

Sections 4 and 5 establish that the minus receiver computes from the operands and follows no donor. What does the equals position's delivered content do on this route? We compare the undamaged route

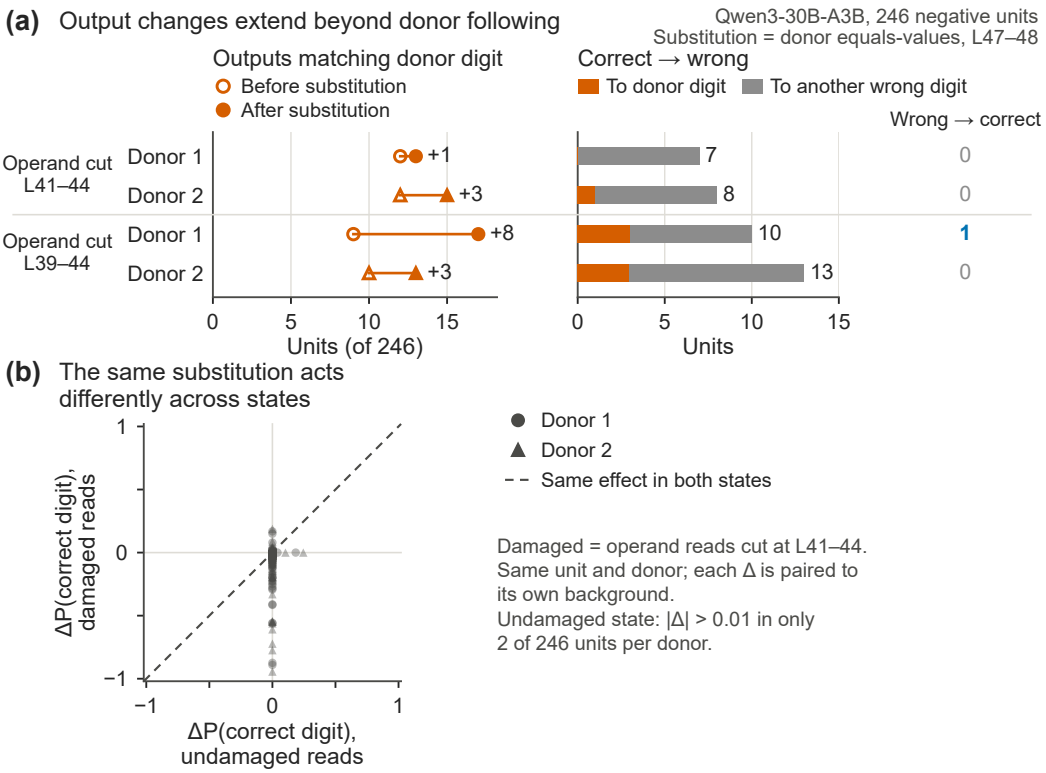


Figure 5: Content influence and direct donor following are different measurements (30B; the substitution replaces the equals position's cached values at layers 47–48 with a donor's; 246 negative units). (a) Pre-existing donor matches stay in the baseline; new matches are counted separately from correctness losses. The 38 correct-to-wrong transitions are condition–unit events (units repeat across conditions): 7 end at the donor digit and 31 at other wrong digits. Under an all-layer cut, one donor leaves correctness at 56/246 while moving two units each way. (b) One point per unit and donor: the same substitution's change in the correct digit's full-vocabulary probability, paired to the undamaged and to the damaged background; the dashed line marks equal effects in both states.

with the route after operand damage, distinguishing accuracy, changes in the output distribution, and donor following (Figure 5).

6.1 Answers stay largely stable as delivered content shifts the distribution

With operand reads undamaged, cutting late equals reads preserves accuracy in the non-overlap subset: 374/374 in 14B and 169/170 in 30B (Table 14). The cut still reduces the correct digit's relative advantage over digits matching neither it nor either operand's first digit in 89% of 14B units and 98% of 30B units (Appendix E). Value substitutions preserve all 553 14B answers. In 30B, both donors correct the same baseline error, raising correctness from 245/246 to 246/246. Donor following remains zero in both models (§4.2). For each donor, only two of the 246 30B units change the correct digit's probability by more than 0.01 (Figure 5b). In the earlier 14B pool of 257 negative units, the layer 37–40 substitution changes the mean of $\log_{10}[p(\text{donor})/p(\text{correct})]$ by 0.65 (Appendix E). Answer stability and sensitivity to delivered content can therefore coexist.

6.2 Damaged computing route: delivered content changes answers, mostly without donor following

After operand damage, keeping the late equals reads yields 142/374 correct first digits in 14B and 69/170 in 30B. Cutting these reads reduces correctness to 82/374 and 62/170 (Table 14). Late equals reads therefore support some answers, while most outputs remain wrong with delivery available.

Substitutions test whether this support is donor following. With the 30B minus receiver's main operand-reading band (layers 41–44) damaged, substituting the equals position's late delivery reduces correctness from 113/246 to 106/246 and 105/246 for the two donors. The paired losses are 2.8 and 3.3 percentage points (pointwise problem-clustered 95% intervals $[0.8, 5.4]$ and $[1.2, 5.7]$). In the cutting run, the same late equals-read cut costs 4.1 points (111 to 101 of 246), a similar size (Appendix E).

Across four damage–donor conditions, 38 correct-to-wrong events end at the donor digit seven times, an operand's first digit fourteen times, and another digit seventeen times. After accounting for baseline matches, donor following gains only 1–8/246 (Figure 5a; Appendix E). Probabilities show the same contrast between the two states: in the 14B pool, operand damage raises the layer 37–40 substitution's mean effect on $\log_{10}[p(\text{donor})/p(\text{correct})]$ from 0.65 to 3.59, an increase of 2.94 (Appendix E), and Figure 5b shows the 30B per-unit shifts growing after damage. Delivered content disturbs calculations much more often than it supplies their answer: on the damaged route, its influence is mostly not donor following.

7 Related work

Arithmetic mediation studies trace attention-based input transfer and feed-forward result processing (Stolfo et al., 2023); other work identifies arithmetic heuristics (Nikankin et al., 2024). Future-token planning (Wu et al., 2024) and depth-dependent intermediate use (Yang et al., 2024; Biran et al., 2024) motivate distinguishing readable content from causal use. The validation gap places checking before deeper arithmetic for supplied solutions (Bertolazzi et al., 2025); we measure content use within the model's own generation.

Mechanistic analyses of chain-of-thought find parallel answer paths (Dutta et al., 2024), and intermediate tokens can act as program variables (Zhu et al., 2025). Our mixed-answer and isolated-exposure interventions distinguish arithmetic input use from donor following. This distinction complements the choices of intervention and outcome metric studied by Zhang & Nanda (2024).

Backup heads (Wang et al., 2023), the Hydra effect (McGrath et al., 2023), and self-repair (Rushing & Nanda, 2024) establish compensation after damage. We connect early equals reads to later operand access and the reuse of natural computation's heads.

8 Discussion and conclusion

From ability to correction. The experiments separate computing an answer, making it support the output, and protecting its influence against erroneous delivery. Recovery enables operand support; cutting the remaining delivery can strengthen its effect. Reliability requires both an effective calculation and an output that preserves its answer.

Identifying use requires tracking content. Different uses of content predict different outputs: arithmetic-consistent shifts identify input transformation, paired destinations distinguish donor following from other errors, and exposure controls locate each digit's response. Head cuts identify the supporting heads. Together, these tests connect readable content to the operation and path that use it.

Dependence belongs to a state and its inputs. Early equals reads shape later operand access; deeper reads deliver the answer. Equals-position processing also changes operand support. Selective shallow-cache replacement can test which equals-position features connect these effects. Combining cuts and substitutions shows that receiver state changes content influence. These tests measure correction at the written digit; internal comparison remains a separate question.

Why this setting supports a broader method. Cramer's rule supplies known inputs and answers within generated solutions. Semantic operand cuts cover earlier product writings. Method transfer requires identifiable intermediates, distinguishable original, donor, and mixed outcomes, and precisely scored decisions. A different tokenizer requires locating the relevant positions and endpoint anew. Textual errors followed by correct later use (Appendix A.1) distinguish this handoff from subsequent revision. At the point of writing, faithful transfer and effective verification are different capabilities.

Reproducibility statement

Section 2 defines the prediction endpoints and interventions. Appendix A describes the fixed pools, subset selection, and semantic operand annotations. Appendix C gives the substitution exposures and digit-level comparisons; Appendix E reports paired output transitions. Execution-matched references and statistical procedures appear in Appendix G, with head selection and attention-cut decomposition in Appendices H–I. Upon acceptance, we will publicly release the intervention and analysis code and the data needed to reproduce the reported results.

AI use statement

Generative AI tools assisted with experimental planning and plotting code, initial interpretation of recorded results, and manuscript polishing. The authors remain responsible for the paper's claims, experimental interpretation, and final text.

References

Leonardo Bertolazzi, Philipp Mondorf, Barbara Plank, and Raffaella Bernardi. The validation gap: A mechanistic analysis of how language models compute arithmetic but fail to validate it. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 29387–29424. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.1495. URL https://aclanthology.org/2025.emnlp-main.1495/.

Eden Biran, Daniela Gottesman, Sohee Yang, Mor Geva, and Amir Globerson. Hopping too late: Exploring the limitations of large language models on multi-hop queries. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024. URL https://arxiv.org/abs/2406.12775.

Subhabrata Dutta, Joykirat Singh, Soumen Chakrabarti, and Tanmoy Chakraborty. How to think step-by-step: A mechanistic understanding of chain-of-thought reasoning. arXiv:2402.18312, 2024. URL https://arxiv.org/abs/2402.18312.

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. URL https://arxiv.org/abs/2307.15771.

Yaniv Nikankin, Anja Reusch, Aaron Mueller, and Yonatan Belinkov. Arithmetic without algorithms: Language models solve math with a bag of heuristics. arXiv:2410.21272, 2024. URL https://arxiv.org/abs/2410.21272.

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.

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. URL https://aclanthology.org/2023.emnlp-main.435/.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2305.04388.

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. URL https://arxiv.org/abs/2211.00593.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://arxiv.org/abs/2201.11903.

Wilson Wu, John X. Morris, and Lionel Levine. Do language models plan ahead for future tokens? In *First Conference on Language Modeling*, 2024. URL https://arxiv.org/abs/2404.00859.

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

Sohee Yang, Elena Gribovskaya, Nora Kassner, Mor Geva, and Sebastian Riedel. Do large language models latently perform multi-hop reasoning? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10210–10229. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.550. URL https://aclanthology.org/2024.acl-long.550/.

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.

Fangwei Zhu, Peiyi Wang, and Zhifang Sui. Chain-of-thought tokens are computer program variables. arXiv:2505.04955, 2025. URL https://arxiv.org/abs/2505.04955.

A POOLS AND SEMANTIC OPERAND SETS

The main series uses each model's own frozen solutions to 749 problems. Units are the final combinations of two products in determinant calculations, selected for two-digit results. The 14B pool contains 1,236 units (619 positive and 617 negative); the 30B pool contains 600 (298 and 302). Main comparisons use the two-nonzero-term subsets and the further restrictions defined in §2. The content-and-damage experiments combine substitutions of the equals position's values with cuts to the receiver's operand reads. The earlier 14B experiment uses 257 negative units from a separate 515-unit pool.

The frozen prompts request Cramer's rule alone, followed by answers in the form `x=..., y=...`, with integers or reduced fractions. For example, problem 1 gives $-5x + 9y = 3$ and $-6x - 5y = 6$. The 14B generation reuses the frozen prompt token sequences, with greedy decoding and a 4,096-token cap. All intervention endpoints replay the recorded prefix and score the mathematical value of its local expression. This fixed-prefix design holds the textual inputs constant while the internal reads change.

The 30B pool is drawn from 1,688 position-aligned two-digit events. Its construction retains eligible events from the earlier annotated pool and fills the remaining places by strata defined by sign, determinant role, and zero-term status. The resulting 600 events bound the intervention workload while preserving these groups. Five-fold probe evaluation groups events by problem; every reported probe prediction uses the held-out fold for that problem.

A semantic operand set contains the tokens writing either product before the carrier, including attached signs and earlier occurrences outside the local expression. Across the fixed pools, these sets contain 12,179 tokens for 14B and 5,447 for 30B. A review of 40 units per model found 10 annotation errors among 547 inspected strings in 14B and zero among 447 in 30B; these are pre-correction string-level counts. The identified errors and matching cases in the pool were corrected before the annotations were frozen for the experiments. Table 3 maps the functional band names to the protocols used in the main comparisons.

Table 3: Layer bands in the fixed-pool series. Layer numbers are one-based. A cut also specifies the query step and the historical positions cut; the same band at another step is a different intervention. WIDE is used in recovery-state cuts (Appendix D); the 30B content-and-damage experiment adds layers 39–44 (Appendix E).

Read or intervention	14B	30B
KEY operand band	28–31	41–44
WIDE operand band	22–31	39–48
30B content-and-damage operand band	—	39–44
LATE equals reads	37–40	47–48
EARLY equals reads, main recovery	1–12	1–7
Late delivery substitution	40	47–48
ALL	1–40	1–48

A.1 CHECKING ARITHMETIC ANOMALIES IN THE EXISTING TEXT

We screened all 749 stored continuations per model with the rules used to extract arithmetic candidates when constructing the pools, retaining candidates before correctness filtering, deduplication, and determinant-role selection. Each flagged occurrence was read in its full expression and subsequent context. This screen covers the parser's recognized forms: a space or minus carrier, a nonzero result of up to three digits, and the allowed operand/operator pattern. We distinguish the actual carrier, the value implied by the written operands, and the determinant value implied by the original problem.

The original 14B counter of 23 arithmetic-or-sign exclusions is a count of parser candidates. All 23 refer to correct compound expressions whose multiplication, division, or outer sign was omitted by the local parse. For example, the suffix of `7*(8/7) - 2 = 8 - 2 = 6` is read locally as `7 - 2 = 8`. Applying the same candidate rules to the 30B text yields 34 flags. Reading the full context classifies 25 as correct expressions obscured by compound syntax or formatting, four as inconsistent

Table 4: Reading guide to cohorts and protocols. Counts within a cell are positive / negative events. Each result uses the subset and intervention named in its row.

Comparison	14B	30B	Inputs and location
Main nonzero-term events	566 / 553	235 / 246	Semantic operand cuts; bands in Table 3; Figures 2–5.
Non-overlap answers	386 / 374	151 / 170	Correct first digit differs from both operand first digits; Table 14.
Strict mixed answers	283 / 264	107 / 124	Distinct original, donor, mixed targets; compute-band operand-value substitution; Table 9.
Neighbor comparison	258 / 236	96 / 108	Strict events with a valid opposite neighbor; Table 10.
Main-pool head overlap	48 / 50	44 / 53	Matched event and query positions; component protocols below; Table 2.
Full head cohorts	60 / 59	74 / 65	14B numerical history, layers 28–31; 30B $C_2 \cup C_3$, layers 39–44; Appendix H.
Earlier state interaction	257 negative	—	Separate 515-unit pool; cut 22–31, main substitution 37–40; Table 16.
Earlier carrier/equals-position tests	—	75 positive	Annotated product positions; reverse carrier in Appendix B.1, equals-position processing in Appendix D.1.

operand/operator writings with the correct determinant result, three as wrong or malformed result writings, and two as generation ending inside the result.

Table 5: 30B text anomalies and later use, from the flagged candidates. The displayed carrier is the actual space or minus. The local value uses the printed operands; the determinant value uses the original problem. These are observations of text and continuation.

Problem / carrier	Written fragment	Interpretation and later text
61 / space	$32 - 30 = 29$	Local and determinant value 2; immediately corrected with "32 - 30 is 2."
129 / space	$16 + 72 = 8$	Value 88; the later expression for x uses $88/19$.
77 / minus	$-5 - 24 = -2$ followed by "constants"	Value -29; later expressions for both x and y use denominator -29.
43 / minus	$-42 + 7 = -15$	Printed operands give -35; the original products give -15, which is used later.
47 / minus	$-25 - 81 = -70$	Printed operands give -106; the original products give -70, which is used later.
79 / minus	$0 + 18 = -18$	Printed operation gives 18; the determinant is -18, used in $x = -18/9$.
105 / minus	$-407 - 49 = -89$	The first product is immediately corrected to -40; the determinant remains -89.
442 / space	$21 + 5 = 5$ at end of record	Value 26; generation ends at its token limit.
644 / space	$-12 + 45 = 3$ at end of record	Value 33; generation ends at its token limit.

These cases distinguish a local textual inconsistency from its later use. The two complete wrong/malformed result writings with two-digit true values, problems 77 and 129, retain the correct leading digit and later use the correct full value. Problem 61 shows an explicit correction for a one-digit true value. The controlled interventions study a complementary endpoint: which internal content determines the next digit while the preceding text is held fixed. The candidate screen and these cases describe the existing continuations; natural-error prevalence requires a denominator covering all expression forms.

B ADDITIONAL CUT AND CARRIER COMPARISONS

The natural-dependence protocol contains 37 conditions per model, spanning equals- and carrier-step cuts, individual layers, layer bands, combined cuts, and random controls. Single-layer scans concentrate the equals position's operand reading in two adjacent 14B layers (a layer-29 cut leaves 523/566 correct; layer 30, 489) and one 30B layer (layer 44, 191/235). The operand set matters for interpreting the cuts: the same product can appear both in the local expression and earlier in the text. When cutting the equals position's main operand-reading band, targeting only the local expression leaves 216/235 positive 30B units correct; targeting every explicit writing reduces this to 97/235, with both text-form groups losing accuracy. Under a key-band equals cut, the 14B text-form groups retain 130/492 correct answers when earlier product writings exist and 15/74 otherwise. The corresponding 30B counts are 75/194 and 22/41. These groups are defined using the semantic annotation. The semantic set closes this alternative textual access path.

For negative units forced to receive a space, digit-restricted correctness at baseline is 415/553 in 14B and 157/246 in 30B. Cutting late equals reads reduces these to 104 and 87; cutting equals-step key-band operand reads gives 118 and 90. Cutting the carrier's own operand reads at all layers gives 415 and 125. The latter contrast records a model difference within the same type of carrier intervention.

Two output criteria separate a preferred digit from the token actually written. Table 6 gives both for delivery substitutions. In every case, the difference between the two counts consists entirely of units whose highest-probability token is a minus sign and whose highest-probability digit is the donor's.

Table 6: Donor following under a forced space. The two criteria use the same unit and donor pairing.

Model	Donor	Highest digit	Highest token	Minus first, donor highest digit
14B ($n = 553$)	1	367	149	218
	2	374	137	237
30B ($n = 246$)	1	90	62	28
	2	79	60	19

B.1 A FORCED MINUS RETAINS DEPENDENCE ON EQUALS READS FOR POSITIVE INPUTS

An earlier 30B protocol tests the reverse carrier change on 75 positive events: it replaces the natural space with the token − (id 481), retaining the original prefix and position. Table 7 scores the immediate full-vocabulary top-1 prediction against the original result's first magnitude digit. The forced minus retains substantial dependence on equals-region reads; cutting operand reads across all layers also changes accuracy. Its input dependence therefore reflects the carrier together with its context.

This earlier cohort includes results of varying lengths. Operand cuts use its annotated product-token set (C_2). Equals-region cuts target every token between the last operand digit and the carrier: the equals token in a two-token gap, and also intervening punctuation in longer gaps. Cuts act only at the carrier query; the preceding cache is computed on the original text. These definitions distinguish this protocol from the main series' two-digit pools and semantic operand sets.

Table 7: Reverse carrier test in the earlier 30B cohort ($n = 75$). Every row forces a minus onto a positive event. Counts concern the next digit, scored over the full vocabulary.

Carrier reads cut	Layers	Correct first digits
None	—	40/75
Product positions	39–48	41/75
Equals region	39–48	18/75
Product positions	1–48	27/75

B.2 Errors retain a structured numerical form

When late equals reads are cut from a natural space receiver, its output often becomes an operand's leading digit: 354/566 units in 14B and 157/235 in 30B. Cutting all explicit product writings preserves much of this bias (350 and 157), although its composition changes. These errors identify a recurring numerical fallback that the semantic-set control separates from direct access to those writings; a wrong digit's identity and its response to targeted change jointly distinguish error pathways.

C Value substitutions and exposure controls

Each receiver has one donor pair for equals-value substitutions and a separate pair for operand-value substitutions. Donors come from a different problem, have the same result sign as the receiver, and have a two-digit answer. Their answer's first digit differs from the receiver's correct first digit and both operand first digits. For each receiver and substitution type, a seeded generator first samples two distinct eligible answer digits, then samples a donor unit within each digit group. Both models use this selection rule.

For operand-value substitutions with a defined mixed answer, the donor's last operand digit must differ from the receiver's. Replacing that digit must give a nonzero mixed result of the same sign, and the donor's answer digit must also differ from the mixed first digit. The strict subset additionally requires a two-digit mixed result whose first digit differs from the original answer and both operand first digits. Both donors are selected from this same strict candidate set, giving a common receiver denominator. The original answer's first digit may match an operand, so retaining it alone does not identify the computation's location. This subset is separate from the non-overlap subset.

When fewer than two donor-answer digit groups satisfy the strict rule, selection first relaxes the mixed result's digit count and its separation from operand first digits, then relaxes its separation from the original answer. These relaxed cases stay outside the strict subset.

Substitutions keep keys fixed and replace the selected value vectors. A null substitution uses the original values.

A full-exposure operand-value substitution is available at three steps: the equals position, the carrier predicting the first digit, and the first-digit position predicting the second. Equals-only exposure restores the recipient's operand values after the equals step, while retaining the resulting equals cache. Exposure after the equals step inserts the donor operand values after that step and retains them through both subsequent steps. Table 8 concerns the first digit; Table 9 separates the exposures for both digits. At the first-digit endpoint, these controls isolate arithmetic use at the equals position for positive units and at the carrier for negative units.

Table 8: First-digit content interventions. Operand-value substitutions use the strict subset, whose original, donor, and mixed leading digits are distinct. Paired entries give the two donors. Table 9 gives the separate second-digit endpoint and full exposure matrix.

Substitution	Outcome	14B	30B
equals, delivery layers	space writes donor	566/566	235, 232/235
equals, delivery layers	minus writes donor	0/553	0/246
operand, compute band	mixed, positive	107, 109/283	44, 42/107
operand, compute band	mixed, negative	91, 84/264	43, 45/124
operand, compute band	donor, positive	15, 16/283	5, 3/107
operand, compute band	donor, negative	21, 21/264	4, 5/124

C.1 Exposure and digit are separate measurement axes

Second-digit measurements supply the recorded true first digit before predicting the next token. Thus each cell in Table 9 measures one digit's response under a fixed prefix. The injected digit changes the mixed value's final digit in every strict unit. These counts measure that target at the full-vocabulary top-1 endpoint; they are separate from free generation of a complete mixed answer.

Table 9: Mixed-answer responses by exposure and digit. The compute band is layers 28–31 in 14B and 41–44 in 30B. Each pair lists donors 1 and 2. Full exposure covers equals, carrier, and second-digit prediction. After-equals exposure covers the latter two; the operand values remain substituted at both.

Model / sign	Exposure	n	First digit	Second digit
14B positive	full	283	107, 109	144, 144
	equals only	283	107, 109	7, 13
	after equals	283	0, 0	1, 2
14B negative	full	264	91, 84	137, 114
	equals only	264	0, 0	0, 0
	after equals	264	99, 90	127, 109
30B positive	full	107	44, 42	66, 62
	equals only	107	43, 41	1, 1
	after equals	107	0, 0	4, 7
30B negative	full	124	43, 45	81, 87
	equals only	124	0, 0	0, 0
	after equals	124	44, 45	83, 87

Positive first-digit responses survive equals-only exposure, whereas the main positive second-digit response appears under full exposure. Negative responses at both digits remain substantial when the equals step processes the original operand values. For the second digit, the after-equals condition combines direct access to the substituted operand with any effect carried through the carrier cache. The measured path is therefore the post-equals sequence; the first-digit result provides the finer localization to the carrier itself.

C.2 The arithmetic predicts which neighboring digit wins

Replacing the last operand digit changes its value by at most nine. In all four strict groups and both donors, the mixed value retains the original sign and two-digit magnitude, and its leading digit differs from the original by exactly one. We therefore compare the mixed-direction neighbor with the equally distant neighbor in the opposite direction. If the original leading digit is t and the mixed digit is $t + \delta$, the comparison digit is $t - \delta$, where $\delta \in \{-1, +1\}$. Requiring both candidates to lie in $1, \ldots, 9$ leaves 258/236 positive/negative 14B units and 96/108 in 30B. This is a post hoc restriction of the strict subsets, with the same exposure and endpoint as Figure 2c.

Table 10: Direction of the first-digit response under the full-exposure operand-value substitution. Advantage is mixed-direction minus opposite-direction frequency, in percentage points, with pointwise 95% problem-clustered intervals. Donors and directional strata share units.

Model / sign	Donor	n	Mixed	Opposite	Advantage [95% CI]
14B positive	1	258	93	27	25.6 [17.8, 33.3]
	2	258	99	28	27.5 [19.8, 35.0]
14B negative	1	236	82	22	25.4 [17.6, 33.3]
	2	236	77	15	26.3 [18.9, 33.3]
30B positive	1	96	39	11	29.2 [15.6, 42.1]
	2	96	38	14	25.0 [11.5, 38.5]
30B negative	1	108	39	11	25.9 [14.2, 37.6]
	2	108	38	7	28.7 [17.1, 39.8]

The mixed direction wins in all eight comparisons. Splitting each by δ gives sixteen descriptive strata, all with the same ordering. The response follows the operation's predicted direction both when the mixed leading digit increases and when it decreases. These shared-unit comparisons add directional evidence to the distinction between mixed and donor answers.

C.3 ALL OUTPUT DESTINATIONS UNDER THE FULL-EXPOSURE SUBSTITUTION

Table 11 classifies every strict-subset first-digit output under the full-exposure compute-band operand-value substitution. All top-1 outputs are digits. The mixed answer is the largest predefined category in all eight model–sign–donor groups. The remaining outputs retain the original answer or match the donor, an operand's first digit, or another digit. Table 10 separately groups outputs by their direction relative to the original answer.

Table 11: Complete full-vocabulary top-1 destinations under the full-exposure operand-value substitution at layers 28–31 (14B) or 41–44 (30B). Entries give donor 1, donor 2. Categories are assigned in column order after n and are mutually exclusive. An original answer matching an operand first digit is counted as original: this occurs in 23 and 35 of the 78 and 82 retained originals in positive 14B units.

Group	n	Original	Mixed	Donor	Operand first	Other digit	Non-digit
14B positive	283	78, 82	107, 109	15, 16	20, 17	63, 59	0, 0
14B negative	264	52, 60	91, 84	21, 21	27, 33	73, 66	0, 0
30B positive	107	39, 31	44, 42	5, 3	9, 4	10, 27	0, 0
30B negative	124	33, 38	43, 45	4, 5	11, 7	33, 29	0, 0

Later reads and processing can replace readable content. Readouts and substitutions separate when content becomes available from when it controls the answer. A 30B layer-47 equals-value substitution makes the donor digit fully readable at the carrier after attention, yet the layer-48 read restores the original digit and only 3/235 outputs follow the donor; substituting layer 48 instead produces 210/235. Two further trajectories show local processing replacing arrived content. A 14B layer-38 equals-value substitution becomes readable, but the layer-39 feed-forward sublayer restores original-digit readability to 0.95 and the output has zero donor following. 30B operand-value substitutions at layers 39–40 are readable, yet the layer-44 feed-forward computation—the site the mixed-answer interventions identify—strengthens the original answer, and 94–98/107 strict positive units retain it.

D RECOVERY: PREFIXES AND EQUALS-SIDE ATTRIBUTION

With late equals reads already cut, increasing the early cut prefix produces a structured recovery pattern. In 14B, prefix endpoints 4, 8, 9, 10, 12, and 16 yield 130, 147, 432, 505, 562, and 551 correct answers out of 566. In 30B, endpoints 3, 5, 10, and 14 yield 68, 235, 235, and 233 out of 235. The non-contiguous 14B set 1–4 together with 9–12 gives 492/566. Cutting the recovered receiver's WIDE operand band reduces correctness to 193/566 and 58/235, complementing the KEY and ALL cuts in Table 12.

Table 12: Receiver-side recovery. Entries are correct first digits / non-overlap correct first digits. EARLY is layers 1–12 in 14B and 1–7 in 30B; LATE and KEY follow Table 1. Each contrast uses references from the same run.

Reads cut at the carrier	14B	30B
equals, LATE	131/566; 26/386	68/235; 6/151
equals, EARLY	566/566; 386/386	235/235; 151/151
equals, EARLY+LATE	**562**/566; 382/386	**235**/235; 151/151
plus products, KEY	213/566; 71/386	90/235; 45/151
plus products, all layers	145/566; 30/386	53/235; 7/151
plus random size-matched set	560/566; 380/386	235/235; 151/151
random position early + equals LATE	127/566	67/235
operand position early + equals LATE	131/566	72/235

Equals-side all-layer operand cuts leave 408/566 and 228/235 correct first digits. Adding carrier-step KEY operand cuts reduces these to 116 and 73; adding ALL cuts reduces them to 79 and 36. Cutting late delivery instead raises correctness to 558 and 232. These comparisons assign the recovery to the carrier's operand access and separately measure the continuing effect of late delivery.

The equals cut also affects its prediction of the carrier token: carrier correctness is 441/566 and 181/235. First-digit comparisons supply the recorded true carrier, so they hold the receiver's input token fixed while changing its equals-position state.

D.1 Shallow equals-position processing changes recovery under deep damage

The earlier 30B equals-side experiment uses the 75 positive events and annotated product positions described in Appendix B.1. Here cuts act at the equals query, and the recorded carrier is supplied before the first-digit prediction. Correctness is 75/75 both at baseline and when only equals-position operand reads at layers 1–7 are cut. Cutting layers 39–48 instead leaves 25/75; also removing layers 1–7 raises this to 65/75. The effect of shallow equals-position processing thus depends on the state of deeper processing.

Table 13 holds deep damage fixed and varies the three intermediate bands. Adding shallow damage improves accuracy in every configuration. These interventions connect the equals position's shallow processing to the downstream answer. A shallow equals cut also changes its later states; selective replacement of the shallow cached values would test their role as the intermediate link to the receiver's early reads.

Table 13: Equals-side shallow/deep interaction in the earlier 30B cohort ($n = 75$). Equals-position operand reads at layers 39–48 are cut in every row. Entries are full-vocabulary correct first digits with shallow equals reads undamaged or additionally cut. Each row is paired on the same events.

Additional middle bands cut	Layers 1–7 undamaged	Layers 1–7 cut
None	25	65
8–17	28	71
18–27	29	71
8–27	48	71
28–38	25	55
8–17 and 28–38	27	61
18–38	31	57
8–38	34	65

E Paired influence and donor following

Stable answers with a distributional response. For natural minus receivers in the non-overlap subset, we measure the correct digit's advantage over digits matching neither it nor an operand's first digit. For unit i, let t_i be the correct digit and N_i this set of neutral digits. Define

$$A_i = \log_{10} p_i(t_i) - \frac{1}{|N_i|} \sum_{d \in N_i} \log_{10} p_i(d). \tag{1}$$

Probabilities are normalized over the full vocabulary. All stored probabilities used here are positive, so the logarithms are evaluated directly. We compare the natural reference with the carrier's late equals reads cut (layers 37–40 in 14B; 47–48 in 30B), with its operand reads undamaged.

The mean change in A_i is -0.453 in 14B and -0.402 in 30B. Pointwise 95% intervals are $[-0.499, -0.409]$ and $[-0.433, -0.372]$, using 10,000 problem-clustered bootstrap draws over 305 and 159 problems, respectively, and retaining each sampled problem's events. The advantage falls in 331/374 units (88.5%) and 167/170 (98.2%). This relative measure records an equals-read effect on the distribution while first-digit accuracy remains unchanged. It complements the donor-following test, which asks whether the output follows the donor's answer.

Table 14 gives the full two-by-two of receiving state and backup removal summarized in §6.

The 30B content-and-damage experiment uses the fixed pool and substitutes equals values at layers 47–48. Table 15 reports negative-unit transitions against the corresponding damaged reference, with the designated donor digit also evaluated before the substitution. New and lost donor matches distinguish intervention-induced following from pre-existing digit coincidences.

Table 14: Conditional support from the other input. Each entry is non-overlap correctness before and after the additional cut. Main-route damage is the matched KEY operand cut for computing and the LATE equals cut for copying. Counts use the same-run references.

Receiving state	Additional cut	14B	30B
copying, undamaged	operands, all layers	386/386 → 386/386	151/151 → 151/151
computing, undamaged	equals, LATE	374/374 → 374/374	169/170 → 169/170
copying, damaged	operands, all layers	25/386 → 22/386	7/151 → 4/151
computing, damaged	equals, LATE	142/374 → 82/374	69/170 → 62/170

Table 15: Paired 30B transitions ($n = 246$ in every row). C and W denote correct and wrong first digits. *To donor* is the subset of C→W transitions ending at the donor digit. Counts across rows include repeated units.

Operand cut	Donor	Matches before	After	New	Lost	C→W	To donor	W→C
41–44	1	12	13	1	0	7	0	0
41–44	2	12	15	3	0	8	1	0
39–44	1	9	17	8	0	10	3	1
39–44	2	10	13	3	0	13	3	0
1–48	1	2	3	1	0	2	0	2
1–48	2	3	4	1	0	2	0	0

The key-band reference has 113 correct outputs; the layers 39–44 reference has 71, and the all-layer reference has 56. The natural reference has 245/246 correct, increasing to 246/246 for each donor, with zero donor following. For each donor, only two units change the correct digit's probability by more than 0.01. This distinguishes a rare substantial response from a small response shared across the whole population.

Under the all-layer cut, donor 1 preserves 56/246 correct despite two changes in each direction.

The four partial-damage conditions in Table 15 produce 38 correct-to-wrong condition–unit events across 21 distinct units. Of the 31 ending away from the donor, 14 match either recipient operand's leading digit and 17 are other digits; these 31 events involve 18 distinct units. The per-row decomposition (donor / operand / other) is $0/4/3$, $1/3/4$, $3/3/4$, and $3/4/6$. This separates a recognizable numerical fallback from other changes in the computed digit.

Under the key-band cut, the two substitutions change accuracy by -2.8 pp $[-5.4, -0.8]$ and -3.3 pp $[-5.7, -1.2]$. The corresponding changes under the layers 39–44 cut are -3.7 pp $[-6.3, -1.2]$ and -5.3 pp $[-8.2, -2.8]$. These are pointwise problem-clustered intervals. In the separate cutting run, cutting late equals reads after key-band damage changes correctness from 111 to 101 of 246, compared with the substitution run's own reference of 113. The two interventions give complementary evidence of equals-position influence under matched functional damage.

In the positive subset, donor following is 235/235 for donor 1 with and without operand cuts. Donor 2 changes from 232 to 233 of 235 at each cut depth.

Table 16 fixes the 14B substitution while varying operand access. Each effect uses its own reference; their difference is the interaction. Donor matches rise from 1 to 10 and 2 to 15 under the layer 37–40 substitution, or 1 to 20 and 2 to 30 for layers 22–40.

Table 16: 14B state interaction on 257 negative units from the earlier pool. Entries average substitution-induced changes in $\log_{10}[p(\text{donor})/p(\text{correct})]$ over both donors within each unit, then over units. The damaged state cuts carrier operand reads at layers 22–31.

Equals-value substitution	Undamaged reads	Damaged reads	Interaction
37–40	0.65	3.59	2.94
22–40	0.79	4.73	3.94
1–40	0.78	4.76	3.98

For the layer 37–40 substitution, the operand cut at layers 22–31 and the value substitution act in different attention operations. The measured interaction therefore extends beyond direct rescaling within the same softmax. An earlier cut can change the later query and its attention to the equals position, as the undamaged-trajectory head recordings demonstrate for recovery. The separate removal/scaling-up experiment in Appendix I identifies those components of the recovery intervention.

F THE 30B OPERAND-POSITION READ AT LAYERS 39–40

A 25-condition run cuts the last operand position at layers 39–40. Natural positive correctness stays at 235/235 under equals-step cuts; negative correctness changes from 245/246 to 244/246 under carrier-step cuts (non-overlap: 169/170 in both).

Its contribution increases after the later operand-reading band is damaged. With layers 41–44 cut, additionally cutting this one position at layers 39–40 changes positive correctness from 94/235 to 84/235. In the non-overlap subset, the change is 55/151 to 45/151, or -6.6 percentage points $[-12.0, -1.4]$. For the negative carrier, the corresponding counts are 110/246 to 71/246, and 68/170 to 43/170, or -14.7 points $[-21.2, -8.1]$. These subset-specific intervals accompany their own denominators.

Under damage, equals-step substitutions yield donor following in 10 units per donor and 16, 19 mixed answers among 107 strict positive units. Carrier-step substitutions yield donor following in 10, 14 units and 19, 15 mixed answers among 124 strict negative units. These are raw output counts; Appendix E's paired counts concern the separate 30B content-and-damage experiment.

Intervened readouts connect the earlier signal to the operand position: cutting it reduces equals-position ones-digit readability immediately after layer-40 attention from 0.79 to 0.20, while cutting the remaining operand positions leaves readability at 0.81. The later mixed-answer interventions establish how the receiver transforms operand content into its answer.

G EXECUTION CONTROLS AND STATISTICAL COMPARISONS

The series uses bfloat16 execution, frozen prefixes, checked query-step cuts, and held-out probes. Null substitutions reproduce their in-run references. Paired effects use each job's own baseline (damaged 30B: 111/246 in the cutting run, 113/246 in the 30B content-and-damage experiment). The jobs use different condition batch sizes. In the 30B content-and-damage experiment, a five-unit test compared single-condition with fifteen-condition execution. It preserved 222 of 225 top tokens; the maximum absolute difference in any full-vocabulary probability was 0.094. Condition-order checks within the matched path were identical. Accordingly, each reported transition and effect uses its own execution-matched reference. Cross-job comparisons describe direction and scale.

Intervals use 3,000–10,000 bootstrap draws clustered by problem, retaining all sampled problems' events. They quantify problem variation conditional on the stated protocol. Table 17 distinguishes exploratory locating, fixed-family tests, and content-tracking counts. The layer scans inform later interventions; their pointwise intervals describe the selected comparisons. The fixed head-group families address the designated comparisons within the component experiments.

H COMPONENTS SHARED BY NATURAL AND RECOVERED COMPUTATION

Selection and protocols. Each selected component is a layer–head pair. Before inspecting the component interventions, heads are ranked by mean natural attention-write magnitude: operand writes at equals and natural-minus queries identify the operand group, and equals writes at natural-space queries identify the delivery group. The top twenty operand pairs are fixed, alongside twenty random pairs outside that set and its full complement. Head-specific cuts cut only the selected heads' reads of the targeted positions.

The 14B experiment targets numerical history (written numbers before the equals sign) in layers 28–31, with early and late equals cuts at 1–14 and 33–40 for recovery. Its full cohort contains 60 positive and 59 negative units. The 30B experiment targets the annotated $C_2 \cup C_3$ numerical inputs (the products and prior result text) in layers 39–44, with equals cuts at 1–7 and 39–48; it contains 74

Table 17: Evidence and statistical treatment. Selection, sampling, and content tracking answer different questions.

Evidence	Treatment
Numerical readouts	Five folds by problem; out-of-fold predictions.
Fixed-pool cuts and recovery	Exploratory localization and matched contrasts; pointwise problem-clustered intervals.
Value substitutions	Two designated donors; paired output destinations, baseline matches, and isolated exposure; mixed-answer intervals are pointwise.
Neighbor-direction comparison	Post hoc strict-subset restriction; eight pointwise problem-clustered intervals, with sixteen shared-unit direction strata reported descriptively.
Full-cohort head tests	Fixed families of 75 comparisons in 14B and 14 in 30B; problem-level sign flips with Holm adjustment.
Fixed-pool head subsets	Descriptive restrictions of the recorded component experiments; full-cohort tests reported separately.
Content-and-damage experiments	Paired transitions and pointwise intervals.

positive and 65 negative units. These component protocols establish shared reading functions using their specified position sets. The main fixed-pool cuts use the semantic operand sets in Appendix A.

Matching unit identities and equals/carrier token positions to the main pools yields 108 overlapping 14B units and 112 overlapping 30B units. Applying the main nonzero-term filter leaves 48 positive / 50 negative units in 14B and 44 / 53 in 30B. Table 2 reports these subsets. Selected-head losses are exclusively correct-to-wrong transitions, and random-twenty cuts preserve all baseline top tokens. Table 18 retains the original cohorts used for the fixed-family inference.

Table 18: Full component cohorts. Entries are correct first digits. Selected and random groups contain twenty layer–head pairs; the remainder contains 140 pairs in 14B and 172 in 30B.

Model	State	Reference	Selected	Random	Remainder	All
14B	natural space ($n = 60$)	60	60	60	60	60
	natural minus ($n = 59$)	59	33	59	59	32
	recovered space ($n = 60$)	57	30	57	56	25
30B	natural space ($n = 74$)	74	74	74	74	74
	natural minus ($n = 65$)	65	35	65	65	20
	recovered space ($n = 74$)	71	46	71	62	24

The selected heads carry a shared contribution, with additional support from other heads. In 30B, selected and remainder cuts separately lose 30 and zero natural-minus answers, while the joint cut loses 45. In recovery, those losses are 25, nine, and 47. The full-band comparison thus exposes support that can survive either single cut.

Reading before operand intervention. The 30B capture records attention with all operand positions available throughout. Across the same 74 positive units, the selected heads' average mass on $C_2 \cup C_3$ is 0.0150 under the late-only cut and 0.5143 under the early-plus-late cut; the 65 natural minuses give 0.6060. The paired increase is 0.4993, with problem-clustered 95% interval $[0.4660, 0.5298]$. Every positive unit increases, and 71 increase by more than 0.2. Most of the added attention reaches the products: C_2 rises from 0.0140 to 0.5111. Both positive conditions cut the same late equals reads while keeping operand reads intact, so their difference measures the effect of the added early equals cut. The 14B diagnostics measure layer 28 before the first operand cut; later diagnostic layers follow earlier operand cuts. The complete-band undamaged-trajectory attention evidence comes from 30B.

Delivery heads. In 30B, cutting the top five equals-reading pairs at layers 45–48 leaves 33/74 natural-space answers correct; cutting the top twenty or all 128 pairs leaves 28/74, with identical failed-unit sets. Random-twenty and remaining-108 cuts preserve 74/74. In 14B, removing the top five late equals writes while preserving other write magnitudes leaves 12/60 correct, compared with

10/60 under the all-head removal. Natural-minus correctness stays at 59/59. Thus localized delivery and shared operand-reading components have different roles across receiving states.

Table 19: Selected operand groups, grouped by layer; layers are one-based and heads zero-based. The five leading delivery heads are L39H2/H4 and L40H10/H12/H13 in 14B, and L47H1/H4 and L48H25/H26/H28 in 30B.

Model	Layer	Operand head indices
14B	28	20, 21, 22
	29	20, 21, 22, 24
	30	10, 11, 12, 14, 21, 23, 30, 32, 33
	31	10, 11, 12, 14
30B	39	27, 31
	40	27, 29
	42	18, 20
	43	10, 11, 12, 13, 14, 19
	44	3, 5, 13, 16, 17, 20, 21, 23

Matched execution and inference. Head cuts use matched references that recompute the same heads while retaining their writes. The 30B references preserve all original top tokens across 21 references and 139 units, while some probabilities vary with numerical execution; correctness comparisons use the matched references. The 14B family combines 72 natural-state comparisons and three recovery head cuts. Selected-head losses in natural and recovered computation have Holm-adjusted Monte Carlo $p < 0.02$. The fixed 30B fourteen-comparison family covers recovery, operand dependence, head specificity, delivery, and undamaged-trajectory attention; selected-head losses, the state interaction, and the attention increase have adjusted $p < 0.0003$. Tests condition on the selected bands and natural-write rankings. The fixed-pool restrictions in Table 2 describe the same component effects on the main events.

I Separating write removal from attention rescaling

The 14B component cohort also separates two effects of attention cuts. At each targeted head, write its output as $M + U$, where M is the contribution from the targeted positions and U from the rest; let u be the original attention mass on the rest. Removal alone gives U, scaling up alone gives $M + U/u$, and the standard cut gives U/u. Each operation is applied to the current trajectory. Table 20 fixes a standard late-equals cut at layers 33–40 and varies the early operation at layers 1–14.

Table 20: Early-write decomposition on the same 60 positive 14B units. The added operand cut targets numerical history at layers 28–31 with the standard cut. Each entry is full-vocabulary first-digit correctness.

Early operation	Operands available	Added operand cut
Original early writes	14/60	14/60
Scale up remaining writes	14/60	14/60
Remove equals writes	41/60	23/60
Remove and rescale (standard cut)	57/60	25/60

Removal alone produces substantial operand-supported recovery; the rescaling adds sixteen correct answers with zero reverse changes. Holding both the late and early interventions to removal alone gives 10/60 before the early removal and 42/60 afterward. Across seventeen pure scaling-up conditions, all 2,023 condition–unit records retain their original top token. These comparisons identify removal and rescaling as interacting parts of the standard cut. The decomposition concerns recovery; the content-and-damage experiments in Appendix E use their own standard operand cuts.