Sparse Attention-Head Patching for Cross-Notation Numerical Comparison: Donor-Position Transfer

Anonymous authors
Paper under double-blind review

Abstract

Prompt-final attention-head patches can restore mixed-notation answers by transferring a donor-selected option position. In stored Llama and Qwen confirmations, aligned donors improve held-out pair consistency from 0/16 to 14/16 and 1/16 to 15/16; summary-level swapped-donor counts are 61/64 and 30/32 presentations, respectively. A relabeled-option intervention shows that, in this Llama setup, target choices follow the donor-selected option position even when donor and target use different labels. This is an output-level transfer result in one model and prompt protocol, not evidence for a portable latent router. A separately preregistered, outcome-informed Llama follow-up finds that knockout of the frozen six-head set reduces signed margins beyond all 30 matched nulls and that operand-span K- and V-vector donor patches shift margins toward a counterfactual answer. The aggregate knockout loss is 0.880 (95% pair-bootstrap CI $[0.805, 0.962]$), but is concentrated in digits/digits prompts (2.390 vs. 0.127 and 0.123 in reciprocal mixed-notation conditions). A later held-out screen across scientific, comma, zero-padded, and word formats met its 80% intact-accuracy floor in only 2/8 format-template cells and stopped before causal tests. A same-cohort chat-format sensitivity run failed donor competence (43/64), so transfer under the original double-BOS rendering remains untested. The evidence supports cross-label donor-position transfer in the standalone prompt setup and a bounded operand-linked effect in one Llama task, not a broad sparse numerical circuit. Unsupported older results are omitted.

1 Introduction

Magnitude is a property of a number, not of its spelling: a model that can compare 2722 with 2847 should also compare 2722 with 2.847e3. Open-weight language models violate this. They compare numbers in a single familiar notation near ceiling, yet fail when the two operands are written differently (Yuchi et al., 2026).

Activation patching (Meng et al., 2022; Wang et al., 2023) offers a causal test of where that failure is repaired. Copying a few prompt-final head outputs from a competent digits-only run into a failing mixed-notation run does restore the answer, in the two models for which we have stored confirmation files. The same patch from an orientation-swapped donor, whose numbers are unchanged and whose correct letter is reversed, makes the model emit the donor's letter. The intervention therefore imports a donor-selected answer position. It does not show that the target's operands were reread, and the same-alphabet swap does not separate a copied letter from a copied comparison result.

Contributions.

1. Aligned patching is sufficient for held-out pair consistency in two stored confirmations (§4). Frozen sets of 6 and 14 heads take Llama and Qwen from 0/16 and 1/16 consistent pairs to 14/16 and 15/16. Qwen recovers 54.1% of its band effect, below a preregistered 75% bar. The random-set file on disk is a 30-draw Llama comparison, reported as such.

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.

2. The swapped donor is followed (§5). Stored summaries report donor-letter output on 61/64 Llama and 30/32 Qwen presentations. These are descriptive presentation-level rates; the current tree has no per-trial logs for these two contrasts, so we report no presentation-level confidence intervals. A stored Qwen summary also records donor tracking of 30/32 on failing numbers, 19/32 on competent numbers, 16/32 on word length, and 23/32 on alphabetical order. Those four rates are reported by domain and are not given a single label.
3. Cross-label donor-position transfer in Llama (§6). Presenting clean donors with labels (1)/(2) and targets with (A)/(B) shows that target choice follows the donor-selected option position: aligned patching restores 16/16 four-way pairs (64/64 trials), swapped patching steers choice to the donor-selected position on 61/64 presentations, and literal donor tokens remain suppressed ($\Delta z \approx -10$). In the raw trial records, 13/16 pairs have donor-slot matches on all four presentations and the remaining three pairs have three of four. The preregistration defines pairs as the inferential unit but sets the donor-slot threshold over presentations; we therefore report these as descriptive counts and do not treat the 64 presentations as independent. This is an output-level transfer pattern in this prompt setup, not evidence for a portable latent router.
4. A bounded native and operand-linked effect, without format generalization (§8). In a new Llama cohort, frozen-set knockout exceeds 30 matched nulls on an equal-weight margin estimand, and counterfactual operand-span K/V patches shift the answer margin; the knockout effect is much smaller in both cross-notation conditions. A separate four-format held-out screen fails its competence gate in six of eight cells and therefore yields no causal breadth result. These data do not establish a general numerical algorithm, cross-format circuit, or cross-model mechanism.

2 Related Work

Number formats in language models. Neural models are sensitive to the surface form of numbers (Spithourakis & Riedel, 2018; Wallace et al., 2019), and arithmetic benchmarks expose large gaps (Patel et al., 2021; Cobbe et al., 2021). Yuchi et al. (2026) documented mixed-notation comparison failures, including in Llama-3.1-8B-Instruct, together with successful probing. We use that behavioral setting and do not restate matrix counts whose trial file is absent from this tree. Number representations are partly linear (Zhu et al., 2025) and partly not (Engels et al., 2024); recent work reports broadly shared sinusoidal number representations across model families and arithmetic-error reductions from enhancing them (Stefánik et al., 2026). We do not claim universal numeric geometry or general arithmetic improvement. Naganna et al. (2026) repair cross-format arithmetic through shared MLP neurons; we study comparison and attention-head outputs.

Recent behavioral work also tests exact-quantity-preserving rewrites, including scientific notation, in numerical word problems (Atta-Duncan, 2026). This broad invariance benchmark does not localize the causal answer-transfer effect studied here, which is limited to pairwise magnitude choices. In arithmetic in-context learning, counterfactual patching shows that strict consistency in demonstration formatting can strongly affect performance (Polyakov et al., 2025). Causal patching has also shown that abstract formulations transfer and compose across math word problems (Cheng et al., 2025). Those studies use demonstrations or word-problem formulations; they do not test donor-selected answer-position transfer in zero-shot pairwise comparison.

Cross-format circuit reuse has also been studied directly: de Varda et al. (2026) find that numeric-arithmetic circuit engagement predicts performance on verbal arithmetic across 13 models, including Llama-3.1-8B. We therefore do not claim cross-format circuit reuse as a general contribution. Their attribution-based circuit-overlap and item-prediction analysis does not test mixed-notation comparison or transplant a natural donor-selected option position across answer alphabets; our claim is limited to that task-specific intervention and does not identify a reusable numerical circuit.

Comparison circuits. Hanna et al. (2023) traced a greater-than circuit in GPT-2 small. Stolfo et al. (2023) used causal mediation to localize components in arithmetic reasoning.

Bhardia et al. (2026) found integer comparison localized in Qwen but distributed in Llama-3-8B. These studies do not test our mixed-notation prompt family; our results are likewise bounded and do not establish a general comparison algorithm.

Interpreting activation patching. Patching evaluates whether components are sufficient to restore an output (Meng et al., 2022; Wang et al., 2023; Conmy et al., 2023). Chowdhury et al. (2026) use the same prompt-final head-output primitive on a different numerical task in Llama-3.1-8B. The choice of metric and corruption changes which components appear causal (Zhang & Nanda, 2024), and denoising establishes sufficiency rather than necessity (Heimersheim & Nanda, 2024). That a final-position component can carry an answer is not new: name-mover heads are the standard example (Wang et al., 2023). The measurement added here is the paired contrast, on stored files, between an aligned donor and an orientation-swapped donor. Function vectors show that attention heads can transport compact task signals (Todd et al., 2024).

Cross-format head-output interventions are also established: Opielka et al. (2026) compare causal function vectors and format-stable concept vectors, and report cross-format activation patching between open-ended and multiple-choice prompts. This rules out novelty based on cross-format head patching alone. Their experiments do not test numerical comparison under notation mismatch or whether an orientation-swapped donor transfers its selected option position across answer alphabets; that paired counterfactual is the narrower distinction examined here.

Correct-label heads and counterfactual routing. Lieberum et al. (2023) study correct-letter heads in multiple-choice answering, including label substitutions and randomized labels; their analysis already cautions that an enumeration-position account is incomplete on broader label distributions. Thavarasa & Sukumar (2026) use four counterfactual donors to separate predicate truth from answer mapping and test transfer of learned routing subspaces across label pairs. Thus neither label manipulation nor counterfactual donor analysis is new here. Our narrower empirical distinction is a prompt-final activation patch in mixed-notation magnitude comparison: numerical values are held fixed while donor operand order changes the selected option position, and a relabeled donor uses (1)/(2) while the target uses (A)/(B). Unlike a four-donor factorial, this experiment does not independently identify truth and mapping factors, and unlike a transfer test across new item-label pairs, it does not establish a reusable routing subspace. The experiment establishes an output-level donor-position effect across the tested label alphabets, while leaving the comparison computation unidentified.

Symbol binding and answer-symbol mechanisms in MCQA. Multiple-choice symbol-content binding is an established problem: Xue et al. (2024) study selection bias as a failure to associate option contents with their symbols and mitigate it through supervised fine-tuning. More directly relevant to mechanistic novelty, Wiegreffe et al. (2025) use activation patching to localize causal answer-symbol prediction to multi-head attention, while Yang et al. (2025) vary option-symbol sets and identify and intervene on heads contributing to symbol bias. Wong et al. (2026) distinguish winner selection in content space from later binding to an output symbol, supported by symbol/content permutations and causal interventions. These papers occupy symbol-content binding, option-symbol sensitivity, and causal head-level analysis of MCQA. They do not test the specific mixed-notation numerical-comparison setting or the natural cross-alphabet donor-head transplant studied here; the distinction is the task/intervention combination, not discovery of a general answer router or causal answer-symbol heads.

3 Setup

Task. Trials are two-alternative forced choices. The scored tokens are “A” and “B” at the final prompt position. A pair is consistent when every stored presentation of it is answered correctly. Exact Clopper–Pearson 95% intervals are reported for pair-level outcomes. Presentation-level rates are descriptive because the four presentations share a numerical pair.

	Unpatched	Aligned donor	Swapped donor	Donor letter
Llama, 6 heads, 64 trials	34/64	62/64	3/64	61/64
four-way pairs	0/16	14/16	0/16	—
Qwen, 14 heads, 32 trials	17/32	31/32	2/32	30/32
two-way pairs	1/16	15/16	0/16	—

Table 1: Stored held-out confirmations. Llama pairs require all four presentations; Qwen pairs require both slot orders. Counts are taken from llama_orientation_swapped_summary.json and qwen_gate1_orientation_swapped_summary.json. Llama's aligned target margin has mean 3.19; the swapped-donor target margin has mean −3.05 (pair-level SEM 0.25). Qwen's swapped-donor target margin has mean −10.50 (SEM 1.15).

The orientation-swapped donor exchanges the two operands' slots in a digits-only prompt, so its correct letter is the opposite of the target's while the two numbers stay the same.

Intervention. For a frozen head set we overwrite those heads' slices of the input to the attention output projection at the final prompt position, using the donor run, and leave every other head and position untouched. Llama-3.1-8B-Instruct uses $S^* = \{$L14H4, L14H5, L14H24, L15H4, L15H8, L17H24$\}$. Qwen2.5-7B-Instruct uses the 14 heads in Appendix A.

What is not in this version. This manuscript reports only quantities present in the JSON and JSONL files listed in Appendix A. It does not report a five-block battery, Gemma, a de-biasing split, a BOS rerun, or selection stability. Stage-3C operand-span K/V patches are reported separately below; the earlier summary-only ablation file remains unused.

4 Two Stored Confirmations

Table 1 is the confirmatory contrast. With an aligned donor, Llama reaches 14/16 four-way consistent pairs (exact 95% interval $[61.7\%, 98.4\%]$) and Qwen reaches 15/16 ($[69.8\%, 99.8\%]$). Qwen's frozen set recovers a fraction 0.541 of the held-out layer-band margin effect (10.99 against a band effect of 20.32). That misses a preregistered recovery target of 75%.

The random-set file stored beside these runs, null_distributions_comparison.json, contains 30 uniform and 30 layer-matched 6-head draws for Llama, not 1,000. In that file the frozen set's recorded recovery is 0.962, the best layer-matched draw is 0.302, and no draw matches the frozen set. The file's empirical_p field is 0 because it counts matches without the $(k{+}1)/(N{+}1)$ correction; with that correction, 0 matches in 30 draws is 1/31. This file does not support a claim of $p \leq 0.01$ from a 1,000-draw null. No stored file contains Qwen's 100-draw null.

5 The Swapped Donor Is Followed

On the orientation-swapped donor, stored summaries report Llama target accuracy of 3/64 and donor-letter output on 61/64 presentations, and Qwen target accuracy of 2/32 and donor-letter output on 30/32 presentations. We treat these presentation-level counts as descriptive; pair-clustered trial logs are not available in the current tree for these contrasts, so no presentation-level confidence intervals are reported. The donor and target contain the same two numbers. A patch that repaired the reading of the mixed-notation operand would have left the target's letter correct.

A second stored Qwen file, e1_qwen_summary.json, records swapped-donor tracking on four blocks of 32 trials each. It is a summary, not a trial log. Tracking is 30/32 on the failing numeric condition, 19/32 on competent numbers, 16/32 on word length, and 23/32 on alphabetical order. The preregistered rule for this kind of battery requires the per-domain rates and no single label. These four rates are that report. They are not ordered into a magnitude scale, and they are not evidence about Llama or Gemma.

6 Relabeled-Option Transfer in Llama

The same-alphabet orientation swap in §5 flips both the correct letter and the slot of the larger number simultaneously. To test whether donor-selected position can affect a target that uses different answer labels, we preregistered a relabeled-option intervention in Llama-3.1-8B-Instruct across 16 held-out pairs (64 presentation trials; document RELABELED-OPTION-TRANSFER-PREREG-2026-09-24, with raw per-trial records stored in relabeled_option_trials.jsonl).

In this design, clean donors present options as (1) and (2), while corrupted targets retain (A) and (B). If the patch acts only by directly injecting a literal donor answer token, its effect should be mismatched to the target's answer alphabet. If it transfers donor-selected option position in this task, aligned patching should restore target (A)/(B) choice consistency, and swapped donor patching should steer choices toward the target token associated with the donor's selected position. This contrast distinguishes output-level donor-position transfer from direct literal-token emission; it does not distinguish an abstract internal code from other mediated computations.

Arm	Trial Acc.	Pair Rate	Pair CI	Margin	Donor Δz
Unpatched target	37/64 (57.8%)	0/16	[0.0%, 20.6%]	$+0.37 \pm 0.09$	-10.73
Relabeled aligned	64/64 (100.0%)	16/16	[79.4%, 100.0%]	$+1.93 \pm 0.11$	-9.98
Relabeled swapped	3/64 (4.7%)	0/16	[0.0%, 20.6%]	-1.48 ± 0.08	-9.93
Relabeled scrambled*	63/64 (98.4%)	15/16	[69.8%, 99.8%]	$+1.88 \pm 0.10$	-10.00

Table 2: Relabeled-option transfer in Llama-3.1-8B-Instruct. Donors use (1)/(2) and targets use (A)/(B). Trial accuracies are descriptive; pair-level inference uses $N = 16$, and margin SEMs use pair means. *The scrambled arm is cyclically collinear with target slots, so it is not an independent activation-noise control.

Table 2 reports the outcomes. Clean donor baseline competence is 64/64 (100%, exceeding the 85% floor). Under aligned patching, target choice consistency reaches 16/16 four-way consistent pairs (64/64 trials, pair-level 95% CI [79.4%, 100.0%]), while literal donor tokens are suppressed ($\Delta z_{\text{donor_literal}} = -9.98$). Under swapped donor patching, the target choice follows the donor's selected slot on 61/64 presentations; descriptively, 13/16 pairs have all four donor-slot matches and three pairs have three of four. The target margin is inverted to -1.48 ± 0.08, and target accuracy is 3/64 (0/16 pairs fully correct). The preregistered decision branch was met using its presentation-level threshold, although the protocol also declares pairs as the inferential unit; this design mismatch limits inferential certainty. The three target-correct trials were zero-margin logit ties resolved by the default "B" tie-breaker.

These results satisfy Branch 1 of the preregistered decision rules (ACCEPT_ABSTRACT_SLOT_ROUTING). We interpret that frozen label narrowly: prompt-final patching of S^* transfers the donor-selected option position across the tested label alphabets, without directly emitting the donor's literal answer token. It does not establish an abstract representation that generalizes to new label pairs, items, models, or prompt formats.

A separately preregistered chat-format sensitivity attempt reused the same 16 pairs and tested donors with Llama chat rendering and duplicated-BOS tokenization. The required BOS contract passed, but clean aligned-donor accuracy was only 43/64 (67.2%), below the frozen 85% competence floor. The run stopped before target preflight and patching, so it provides no estimate of cross-label transfer under chat formatting and is not an independent replication (Appendix A).

7 Native Ablation Does Not Support Selective Cross-Notation Computation

The earlier aggregate necessity file lacks trial-level records and BOS provenance, so we ran a separately frozen Stage-2 test on 16 newly generated numeric pairs. Their values are disjoint from surviving Stage-1 raw records; broader discovery-value disjointness cannot be verified. The same cohort was evaluated intact and with the six heads in S^* zeroed at the prompt-final attention-output input; 30 layer-matched six-head controls (3/2/1 heads in

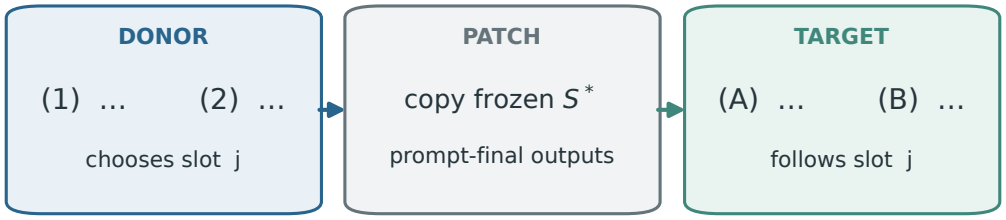


Figure 1: Tested donor-position intervention. Prompt-final activations from the frozen set S^* are copied from a donor using (1)/(2) labels into a mixed-notation target using $(A)/(B)$. The schematic summarizes the tested output-level transfer in Llama; it does not assert a general numerical router or identify how the comparison is computed.

layers 14/15/17) were evaluated on the 64 mixed-format presentations. For each pair and format, the primary loss is the mean intact correct-minus-incorrect A/B logit margin minus the mean ablated margin over its four factorial presentations. Digits/digits accuracy was 63/64 (98.4%), above the preregistered 80% competence floor. The exact tokenization used one leading BOS per prompt, and the complete raw run contains 2,176 condition/prompt records.

On the preregistered pair-clustered signed-margin estimand, the mean mixed-format knockout loss was 0.318 (95% pair-bootstrap CI $[0.180, 0.465]$; one-sided exact sign-flip $p = 0.00049$). No matched-null set had a larger mean loss (0/30; corrected empirical $p = 1/31 = 0.0323$). These two criteria indicate that the frozen intervention affects mixed-format answer margins in this cohort and that its effect is unusual relative to the sampled matched nulls.

The registered selectivity test fails, however: same-format mean loss was 2.371, and the mixed-minus-same contrast was -2.053 (95% pair-bootstrap CI $[-2.305, -1.771]$). Accuracy was 37/64 intact versus 42/64 under knockout on mixed-format prompts, while same-format accuracy was 63/64 versus 64/64. Therefore the positive margin loss is not a mixed-format-specific accuracy impairment. Under the frozen decision rule, the result is NATIVE_EFFECT_NOT_ESTABLISHED because all gates, including mixed-format selectivity, were required.

This outcome rejects the proposed selective cross-notation role for this set on this cohort. It does not show that the heads are irrelevant to every numerical task, nor does it identify what computation explains their much larger same-format margin effect. The study used one Llama checkpoint, 16 pairs, and prompt-final head-output ablation; it did not localize QK versus V/OV pathways. No downstream mechanistic or breadth stage was run because the preregistered selectivity gate failed. Full protocol, raw outputs, hashes, and adjudication are listed in Appendix A.

An outcome-informed follow-up then froze 24 additional pairs and measured the same S* knockout and 30 layer-matched six-head nulls on both formats. The same-minus-mixed pair-loss contrast replicated at 2.191 (95% pair-bootstrap CI $[1.956, 2.413]$; one-sided exact sign-flip $p = 5.96 \times 10^{-8}$); same-format loss was 2.451 (95% CI $[2.212, 2.673]$), versus mixed-format loss 0.260 (95% CI $[0.188, 0.336]$). No same-format null exceeded S* (0/30; corrected empirical $p = 0.0323$). Same-format accuracy was 96/96 both intact and under knockout; mixed-format accuracy increased from 53/96 to 57/96. Because this follow-up was prompted by the first cohort, it is reported as an outcome-informed replication and is not pooled with it. The replication confirms a format-conditioned margin effect, while the accuracy/margin dissociation leaves its computational interpretation open. The original Stage-2 gate for a larger mixed-format-than-same-format effect remains failed. No operand-path or broader mechanism test is established by these results.

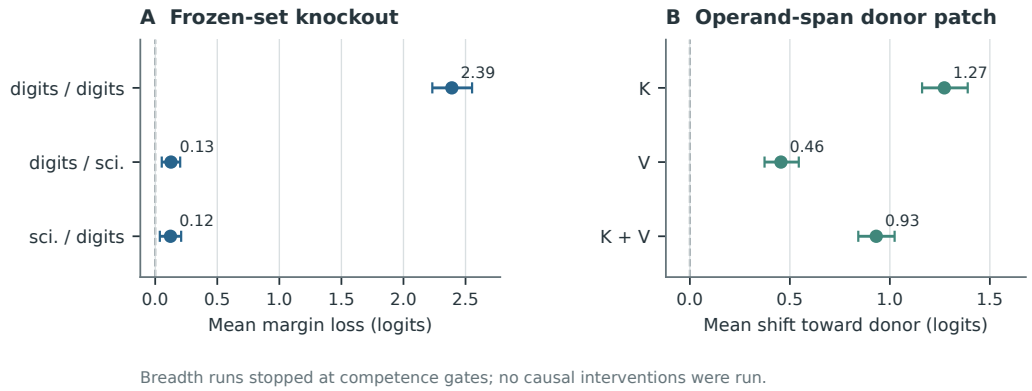


Figure 2: A bounded intervention effect; breadth testing stopped at competence gates. (A) Stage-3C mean signed-margin loss under knockout by format (digits = canonical-digit inputs; sci. = scientific-notation inputs); (B) aggregate signed-margin shifts under operand-span K, V, and joint K/V donor patches. Points show means and bars show 95% pair-bootstrap intervals. Each Stage-3C intervention exceeded all 30 matched-null means. Six query heads share five K/V groups. Stage 4 (2/8 cells), the separate development-only Stage 4A screen (1/8), and the chat-format donor preflight (43/64, below its 85% floor) stopped before causal breadth testing.

8 A Bounded Operand-Linked Effect Does Not Generalize Yet

A preregistered, outcome-informed Stage-3C follow-up used a new 16-pair cohort and 192 prompts balanced across digits/digits, digits/scientific, and scientific/digits; it is not pooled with earlier cohorts. Exact-context no-op-hook parity passed (maximum signed-margin change 0), and same-format accuracy was 64/64. The frozen S^* knockout reduced the equal-weight correct-answer margin by 0.880 (95% pair-bootstrap CI $[0.805, 0.962]$; one-sided sign-flip $p = 1.53 \times 10^{-5}$), exceeding all 30 group-matched null sets (0/30; corrected empirical $p = 0.0323$). The per-condition mean losses were 2.390 for digits/digits, 0.127 for digits/scientific, and 0.123 for scientific/digits. Counterfactual donor patches at operand spans shifted margins toward the donor answer in K-only (1.273), V-only (0.455), and joint K/V (0.932) arms; each exceeded all 30 matched nulls. The set's six query heads map to five shared K/V groups under grouped-query attention, so these interventions do not isolate six independent head pathways. This supports a task- and checkpoint-bounded operand-linked causal effect, dominated by canonical digits; it does not establish a general comparison algorithm.

A separate held-out breadth attempt used 24 pairs and 768 prompts, crossing digits with scientific, comma-grouped, zero-padded, and word forms under two templates. Exact-context parity passed, but only the two comma-format cells met the preregistered 80% intact-accuracy floor (comma: 90.6–92.7%; scientific: 50.0–63.5%; zero-padded: 53.1%; words: 49.0–50.0%). The run stopped before any knockout or matched-null intervention. It therefore provides no causal generalization estimate; selecting the comma cells after observing the gate would not rescue the broad claim. Full protocols and raw records are listed in Appendix A.

Figure 2 summarizes the Stage-3C effect estimates and marks that breadth runs stopped at competence gates. The effect intervals are pair-bootstrap intervals from the frozen adjudication; no causal breadth interventions were run.

9 Discussion

What the patch transfers. The aligned and swapped donors agree on a limited statement. Prompt-final patching of these frozen sets is sufficient to move held-out answers, and the moved answer matches the donor's letter when the slots are exchanged. The sets pass a sufficiency check on the stored confirmations. The donor-recovery analysis does not support the previously claimed $p \leq 0.01$ specificity result from a 1,000-draw null, because that

null is not the file on disk. Separately, the Stage-2 knockout exceeds all 30 preregistered matched-null means (empirical $p = 0.0323$) but fails its mixed-format selectivity gate.

Resolution of the slot-vs-letter confound in Llama. In the same-alphabet contrast (§5), exchanging the operands flipped both the correct letter and the slot of the larger number, leaving open whether the patch copied a letter or a slot-bound choice. The relabeled-option experiment (§6) resolves this confound for the tested Llama setup: clean donors lacking "A" and "B" options reliably steer target "A"/"B" outputs to the donor's selected slot with complete suppression of literal donor tokens ($\Delta z \approx -10$). However, this dissociation is established strictly for the tested Llama architecture and numerical prompts; it does not demonstrate universal routing across arbitrary alphabets or models (e.g., Qwen was not evaluated with relabeled options), nor does it imply that S^* computes relational magnitude or that individual heads are modularly necessary.

Sufficiency, bounded native role, and pathway. The donor-patching results remain sufficiency tests (Heimersheim & Nanda, 2024). Stage 3C adds a matched-null-specific native margin effect and operand-span K/V donor-patch effects for this frozen set and task. The much smaller cross-notation knockout effects and failed Stage-4 competence gate sharply limit the interpretation: these data do not show that the heads implement a general comparator, nor that the same causal pathway operates across number formats. Shared K/V groups also prevent attributing the effect to six independently isolated query-head pathways.

10 Limitations

The donor-transfer results use two models, but all native-knockout and operand-path evidence is from one Llama checkpoint. Qwen's frozen set misses its preregistered recovery target, and its four-block follow-up is a summary rather than a trial log. The Llama random-set file contains 30 draws. Relabeled-option transfer has not been tested in Qwen or across broader option vocabularies. The relabeled-option standalone result has not been reproduced under the original chat-rendered, double-BOS format: the same-cohort sensitivity attempt failed its donor-competence gate at 43/64 (below the preregistered 85% floor) and stopped before target preflight or patching. It therefore provides no chat-format transfer estimate. Stage 3C uses 16 pairs and interventions on shared K/V groups; it does not isolate six independent head pathways, establish an algorithm, or show comparable effects across formats. Stage 4 tested 24 new pairs, but failed the intact-competence gate in six of eight cells and ran no causal interventions. A separate 12-pair development screen (Stage 4A) tested explicit notation-semantic wording and also failed: only one of eight format-template cells reached 80

11 Conclusion

On the stored Llama and Qwen confirmations, sparse prompt-final head patches transfer a donor-selected answer position; a relabeled-option intervention shows this output-level transfer across the tested label alphabets in the Llama one-BOS prompt setup. Two Stage-2 cohorts show a replicated matched-null-specific margin effect that is much larger on canonical digits than on mixed notation. A later preregistered Llama study supports a bounded native margin role and operand-linked K/V pathway, again with much smaller cross-notation effects. The held-out four-format study failed its competence gate before causal testing. The evidence supports cross-label donor-position transfer and a narrow operand-linked effect in one Llama task; it does not support a portable routing code, broad sparse numerical circuit, or generalization across number formats.

AI Use Statement

Generative AI tools supported literature review, research planning, experimental methodology, code development and execution, result interpretation, and manuscript preparation. AI-assisted analysis included a sensitivity run that stopped at a preregistered competence gate before interventions. Synthetic fixtures were used only for software diagnostics, not

as empirical data. The authors reviewed the work and take responsibility for the paper's claims.

Reproducibility Statement

- Models: Llama-3.1-8B-Instruct and Qwen2.5-7B-Instruct. Exact model revisions and environment details are recorded in the referenced run manifests.
- Artifacts: Stored summaries and the relabeled-option raw trial log are named in Appendix A. Chat-format sensitivity attempt 004's frozen protocol, manifest, donor preflight records, partial summary, and execution report are retained at relabeled_option_transfer/runs/attempt_004_chat_template/; it stopped at the donor-competence gate and has no target or patch outcomes. Complete Stage-2 and Stage-2B run records are retained. Stage-3C records are in native_necessity/runs/attempt_006/; Stage-4 baseline/parity records and stop adjudication are in native_necessity/runs/attempt_007/; the separate development-only Stage-4A screen and its raw records are in native_necessity/runs/attempt_008/. The accompanying anonymous CAND042_ICLR2027_reproducibility_supplement.zip includes these records and the relevant scripts, but not model weights. Stage 2B is outcome-informed and is not pooled with Stage 2; Stage 3C is a separate, outcome-informed cohort. Stages 4 and 4A ran no causal interventions; Stage 4A met its 80
- Excluded evidence: Summary-only or artifact-missing older runs (including the five-block battery, Gemma, de-biasing, and stability analyses) and figures based on those excluded runs are not used. The earlier summary-only necessity file is not cited; the Stage-2 and Stage-3C raw runs are reported separately. Figures 1 and 2 use the bounded claims and stored adjudications described in the text and appendix.
- Environment: Llama same-alphabet on CPU, Qwen on MPS, and Llama relabeled-option on CUDA (NVIDIA A40, PyTorch 2.4.1, transformers 4.48.3). Stage 3C and Stage 4 ran on secure RTX 4090 pods with the pinned environment recorded in their manifests. Stage 4A ran on a secure A40 after an availability amendment; the pinned model revision and BF16 precision were unchanged.

References

Ephraim Atta-Duncan. Same quantity, different answer: Numerical representation invariance in language models. arXiv preprint arXiv:2609.25009, 2026. URL https://arxiv.org/abs/2609.25009.

Arya Bhardia, Julian Ramirez, Siddhanta Verma, and Karen Mkrtchyan. Mechanistic analysis of universality: Numerical comparison circuits across transformer architectures. In Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pp. 951–967, San Diego, California, United States, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-srw.84. URL https://aclanthology.org/2026.acl-srw.84/.

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, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.723. URL https://aclanthology.org/2025.emnlp-main.723/.

Rahul Chowdhury, Timothy A. Rupprecht, Senhao Cao, Jiahao Liu, Octavia Camps, David Bau, Pu Zhao, and Yanzhi Wang. Mechanistic interpretability of structure-aware numerical reasoning in LLaMA 3.1 8B. arXiv preprint arXiv:2608.18419, 2026. doi: 10.48550/arXiv.2608.18419. URL https://arxiv.org/abs/2608.18419.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, and Jacob Hilton. Training verifiers to solve math word problems. arXiv preprint arXiv:2110.14168, 2021. URL https://arxiv.org/abs/2110.14168.

Arthur Conmy, Augustine 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. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/34e1dbe95d34d7ebaf99b9bcaeb5b2be-Abstract-Conference.html.

Andrea Gregor de Varda, Sana Pandey, Pengrui Han, Jacob Andreas, and Evelina Fedorenko. Shared circuits predict whether LLMs generalize across formats in arithmetic reasoning. arXiv preprint arXiv:2609.04463, 2026. URL https://arxiv.org/abs/2609.04463.

Joshua Engels, Eric J. Michaud, Isaac Liao, Wes Gurnee, and Max Tegmark. Not all language model features are one-dimensionally linear. arXiv preprint arXiv:2405.14860, 2024. doi: 10.48550/arXiv.2405.14860. URL https://arxiv.org/abs/2405.14860.

Michael Hanna, Ollie Liu, and Alexandre Variengien. How does GPT-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model. In Advances in Neural Information Processing Systems, volume 36, pp. 76033–76060, 2023. doi: 10.52202/075280-3322. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/efbba7719cc5172d175240f24be11280-Abstract.html.

Stefan Heimersheim and Neel Nanda. How to use and interpret activation patching. arXiv preprint arXiv:2404.15255, 2024. URL https://arxiv.org/abs/2404.15255.

Tom Lieberum, Matthew Rahtz, J'anos Kr'am'ar, Neel Nanda, Geoffrey Irving, Rohin Shah, and Vladimir Mikulik. Does circuit analysis interpretability scale? evidence from multiple choice capabilities in chinchilla. arXiv preprint arXiv:2307.09458, 2023. URL https://arxiv.org/abs/2307.09458.

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

Sharath Naganna, Tanvir Ahmed Sijan, and Uddipta Kalita. Are arithmetic heuristic neurons form-invariant? A mechanistic analysis of symbols, text, and code in LLMs. arXiv preprint arXiv:2607.16693, 2026. doi: 10.48550/arXiv.2607.16693. URL https://arxiv.org/abs/2607.16693. Under review.

Gustaw Opielka, Hannes Rosenbusch, and Claire Stevenson. Causality $\neq$ invariance: Function and concept vectors in LLMs. In The Fourteenth International Conference on Learning Representations, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/86b3697c4eb7792c951831636bfdacd5-Abstract-Conference.html.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. Are NLP models really able to solve simple math word problems? In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 2080–2094, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.168. URL https://aclanthology.org/2021.naacl-main.168/.

Gregory Polyakov, Christian Hepting, Carsten Eickhoff, and Seyed Ali Bahrainian. Interpretability analysis of arithmetic in-context learning in large language models. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pp. 1758–1777, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.92. URL https://aclanthology.org/2025.emnlp-main.92/.

Georgios Spithourakis and Sebastian Riedel. Numeracy for language models: Evaluating and improving their ability to predict numbers. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 2104–2115, Melbourne, Australia, 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1196. URL https://aclanthology.org/P18-1196/.

Michal Stefánik, Timothee Mickus, Marek Kadlčik, Bertram Højer, Michal Spiegel, Raúl Vázquez, Aman Sinha, Josef Kuchař, Philipp Mondorf, and Pontus Stenetorp. Language models learn universal representations of numbers and here's why you should care. In Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 30663–30681, San Diego, California, United States, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.1415. URL https://aclanthology.org/2026.acl-long.1415/.

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, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.435. URL https://aclanthology.org/2023.emnlp-main.435/.

Luxshan Thavarasa and Sivasuthan Sukumar. Test, then route: How language models execute in-context conditional rules across models and languages. arXiv preprint arXiv:2608.04183, 2026. URL https://arxiv.org/abs/2608.04183.

Eric Todd, Millicent L. Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau. Function vectors in large language models. In International Conference on Learning Representations, 2024. URL https://openreview.net/forum?id=AwyxtyMwaG. arXiv:2310.15213.

Eric Wallace, Yizhong Wang, Sujian Li, Sameer Singh, and Matt Gardner. Do NLP models know numbers? probing numeracy in embeddings. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 5307–5315, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1534. URL https://aclanthology.org/D19-1534/.

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://openreview.net/forum?id=NpsVSN6o4ul.

Sarah Wiegreffe, Oyvind Tafjord, Yonatan Belinkov, Hannaneh Hajishirzi, and Ashish Sabharwal. Answer, assemble, ace: Understanding how LMs answer multiple choice questions. In The Thirteenth International Conference on Learning Representations, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/c248154176c08147e82c0b30961604f7-Abstract-Conference.html.

Hugh Mee Wong, Rick Nouwen, and Albert Gatt. When models decide and when they bind: A two-stage computation for multiple-choice question answering. In Findings of the Association for Computational Linguistics: ACL 2026, pp. 22804–22821. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.1144. URL https://aclanthology.org/2026.findings-acl.1144/.

Mengge Xue, Zhenyu Hu, Liqun Liu, Kuo Liao, Shuang Li, Honglin Han, Meng Zhao, and Chengguo Yin. Strengthened symbol binding makes large language models reliable multiple-choice selectors. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 4331–4344. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.237. URL https://aclanthology.org/2024.acl-long.237/.

Zhen Yang, Ping Jian, and Chengzhi Li. Option symbol matters: Investigating and mitigating multiple-choice option symbol bias of large language models. In Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pp. 1902–1917. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.naacl-long.95. URL https://aclanthology.org/2025.naacl-long.95/.

Fengting Yuchi, Li Du, and Jason Eisner. LLMs know more about numbers than they can say. In Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 659–673, Rabat, Morocco, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.eacl-short.47. URL https://aclanthology.org/2026.eacl-short.47/.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In The Twelfth International Conference on Learning Representations, 2024. URL https://openreview.net/forum?id=Hf17y6u9BC.

Fangwei Zhu, Damai Dai, and Zhifang Sui. Language models encode the value of numbers linearly. In Proceedings of the 31st International Conference on Computational Linguistics, pp. 693–709, Abu Dhabi, UAE, 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.47/.

A Stored records

The body uses the files listed below under cand_042_head_routing_circuit/.

- results/llama_orientation_swapped_summary.json. Heads: L17H24, L14H5, L15H8, L14H4, L15H4, L14H24. Unpatched 34/64 trials and 0/16 four-way pairs; aligned donor 62/64 and 14/16; swapped donor target-correct 3/64, donor-letter 61/64.
- cross_model_extension/results/qwen_gate1_orientation_swapped_summary.json. Fourteen heads: L18H14–20 and L20H21–27. Unpatched 17/32 and 1/16 pairs; aligned donor 31/32 and 15/16; swapped donor target-correct 2/32, donor-letter 30/32.
- cross_model_extension/results/qwen_gate1_primary_summary.json. Frozen-14 recovery fraction 0.5408575273985772 relative to the held-out band.
- results/null_distributions_comparison.json. Thirty uniform draws and thirty layer-matched draws. Recorded frozen-set recovery 0.9622093023255814. Maximum layer-matched draw 0.3023255813953488. The file's empirical_p is 0.
- followups/results/e1_qwen_summary.json. Swapped-donor tracking: failing numbers 30/32, competent numbers 19/32, word length 16/32, alphabetical order 23/32. Summary statistics only.
- relabeled_option_transfer/results/relabeled_option_summary.json (with raw per-trial log relabeled_option_trials.jsonl). Relabeled donor (1)/(2) options vs target (A)/(B) on Llama S^*. Unpatched 37/64 trials and 0/16 four-way pairs; aligned donor 64/64 and 16/16; swapped donor target-correct 3/64, donor-slot tracking 61/64; literal donor token margin ≈ -10.
- Chat-format sensitivity attempt 004 (same cohort as attempt 002; no pooling): relabeled_option_transfer/runs/attempt_004_chat_template/. Its two-BOS preflight passed, but clean donor accuracy was 43/64 (67.2%), below the frozen 85% floor. The runner stopped before unpatched-target or patch arms; this is not a transfer estimate. The directory retains its protocol, manifest, preflight, raw donor records, summary, log, and execution report.
- Stage-2 native-necessity run: NATIVE_NECESSITY_PREREGISTRATION_2026-09-24.md; attempt-specific native_necessity/runs/attempt_002/manifest.json, stimuli.json, preflight.json, and results/raw_trials.jsonl, results/adjudication.json, results/execution_summary.json. The execution report and chronology, resource, and execution amendments are retained in native_necessity/runs/attempt_002/. The raw log contains 2,176 unique condition/prompt records. Attempt 001 stopped after 16 intact prompts under its outcome-blind resource gate; its partial records were neither scored nor pooled. The frozen attempt-002 verdict is NATIVE_EFFECT_NOT_ESTABLISHED: mixed-format margin loss exceeds all 30 matched nulls, but the preregistered mixed-minus-same selectivity contrast is negative.
- Stage-2B outcome-informed independent-cohort replication: native_necessity/runs/attempt_003/STAGE2B_PREREGISTRATION.md, manifest.json, stimuli.json, preflight.json, COHORT_AUDIT.json, results/raw_trials.jsonl, results/adjudication.json, results/execution_summary.json, and EXECU-

TION_REPORT.md. The 24-pair run has 6,144 unique condition/prompt records. Its frozen verdict, REVERSE_SELECTIVE_EFFECT_REPLICATED, confirms a larger same-format than mixed-format signed-margin effect; it is not pooled with Stage 2 and does not establish an accuracy benefit or operand-computation mechanism.

- Stage-3C native and operand-path follow-up (run directory: native_necessity/runs/attempt_006/). The 16-pair cohort has 192 prompts balanced across digits/digits, digits/scientific, and scientific/digits. Exact-context no-op-hook parity passed (maximum margin difference 0), and same-format accuracy was 64/64. The frozen knockout's equal-weight mean signed-margin loss was 0.8799 (95% pair-bootstrap CI $[0.8050, 0.9619]$; one-sided sign-flip $p = 1.53 \times 10^{-5}$), with 0/30 matched nulls exceeding it (corrected empirical $p = 0.0323$). Mean losses by condition were 2.3896, 0.1270, and 0.1230. Operand-span donor-patch shifts were 1.2725 for K, 0.4554 for V, and 0.9320 for joint KV; each exceeded all 30 group-matched nulls. The six query heads map to five shared K/V groups, so this does not isolate six individual pathways. The result is bounded to this Llama checkpoint, task, and cohort. The run directory contains the frozen protocol, manifest, stimuli, raw trials, adjudication, and execution report.
- Stage-4 held-out format/template breadth attempt (run directory: native_necessity/runs/attempt_007/). The 24-pair, 768-prompt baseline passed exact-context identity-hook parity but only 2/8 cells met the 80% intact-accuracy gate: comma 90.6–92.7%, scientific 50.0–63.5%, zero-padded 53.1%, and words 49.0–50.0%. The preregistered runner stopped before any knockout or matched-null intervention; this attempt provides no causal breadth estimate. The run directory contains the protocol, manifest, stimuli, baseline/parity raw trials, adjudication, and report.
- Stage-4A development-only competence screen (run directory: native_necessity/runs/attempt_008/). On 12 fresh pairs, identity-hook parity passed exactly (maximum margin difference 0; all six sites fired), but only 1/8 format-template cells met the 80% floor: comma/terse 94/96; comma/explicit 63/96; scientific/terse 50/96; scientific/explicit 48/96; words/terse 63/96; words/explicit 49/96; zero-padded/terse 58/96; zero-padded/explicit 51/96. The frozen runner stopped before causal interventions. Pair-clustered descriptive intervals, raw trials, amendments, and run report are retained in the run directory. This development-only screen is not pooled with attempt 007 and provides no causal breadth estimate.

The quantitative panel in Figure 2 is read from the retained Stage-3C adjudication and the competence-gate summaries. The plotting script is included in the submission source bundle; it does not run inference.