QUERY TRANSFER CAN REVISE RULES WITHOUT PRESERVING THE TASK

Anonymous authors
Paper under double-blind review

ABSTRACT

When a model follows a revised instruction, the unchanged requirements must still hold. If the destination country changes, the budget and distance ranking still apply. We copy attention queries from a final-rule restatement into a prompt that must resolve the revision. These queries govern how the model reads its context. Query transfer improves revision in Llama-3.1-8B-Instruct and Qwen2.5-7B-Instruct, as direct restatement also does. Paired scoring nevertheless exposes correct answers lost within the aggregate gains. A separate last-token query intervention tests source-score reconstruction between two prompts that both restate the final rule. In Qwen, adding record states to query transfer brings a monitored answer margin closer to its source value but lowers selection accuracy. Most new errors copy an incorrect source answer. Correct-source counterexamples reveal a second failure because the monitored margin omits an eligible competitor in the requested category. Queries from prompts without records also shift scores toward changed requirements, but these effects do not consistently improve complete execution. Complete-candidate scoring locates the missed ranking comparison, and paired evaluation identifies the successful answers an intervention fails to preserve.

1 INTRODUCTION

Changing one requirement should leave the others in force. A travel assistant may need to choose the closest Canadian destination within a budget. If the country changes to France, the budget and ranking rule still apply. Choosing a French destination satisfies only part of the request. A useful account of instruction revision must explain how the model changes the requested criterion and preserves the remaining ones.

Attention queries offer a way to study this process. They determine which contextual states a model reads. Sen Sharma et al. (2026) show that queries in selected heads can carry filtering predicates between inputs. We extract queries from a prompt that clearly restates the final rule and insert them into a run that must resolve a revision. This tests whether the transferred state helps execute the new requirement together with the unchanged budget and ranking constraints.

Two forms of evaluation can obscure the outcome. A targeted answer-score comparison may recover the desired category preference yet miss a ranking error. Across inputs, net accuracy combines repaired answers with newly broken ones. These ambiguities matter when the intervention is used to infer what a state represents and how it supports task execution. We study both in tasks with exact ground truth for every requirement.

Figure 1 illustrates the score-level ambiguity in a separate compatibility experiment. Both prompts restate the final rule, either before or after the records. We test whether transferring the corresponding record states improves query transfer. D and C are affordable French destinations, and D is closer. Queries alone select D. Adding record states brings the D–F score difference closer to the source run but changes the answer to C. Because F is a Canadian alternative, the monitored comparison misses the French competitor that causes the error.

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.

Our evaluation connects rule transfer to task preservation through three findings.

- Query transfer improves revision, but aggregate gains conceal new errors among originally correct answers. Paired outcomes expose this loss of preservation alongside the repaired failures.
- Better source-score reconstruction can accompany worse selection in Qwen. Full-candidate scoring distinguishes copied source errors from correct-source counterexamples caused by an omitted competitor.
- Record-free queries influence each revised requirement and also affect the others. Comparing these score changes with complete answers separates rule-directed influence from successful execution.

Direct restatement provides the behavioral baseline throughout. Comparing it with query transfer shows which errors arise from the internal intervention and which are shared with an effective prompt-level change.

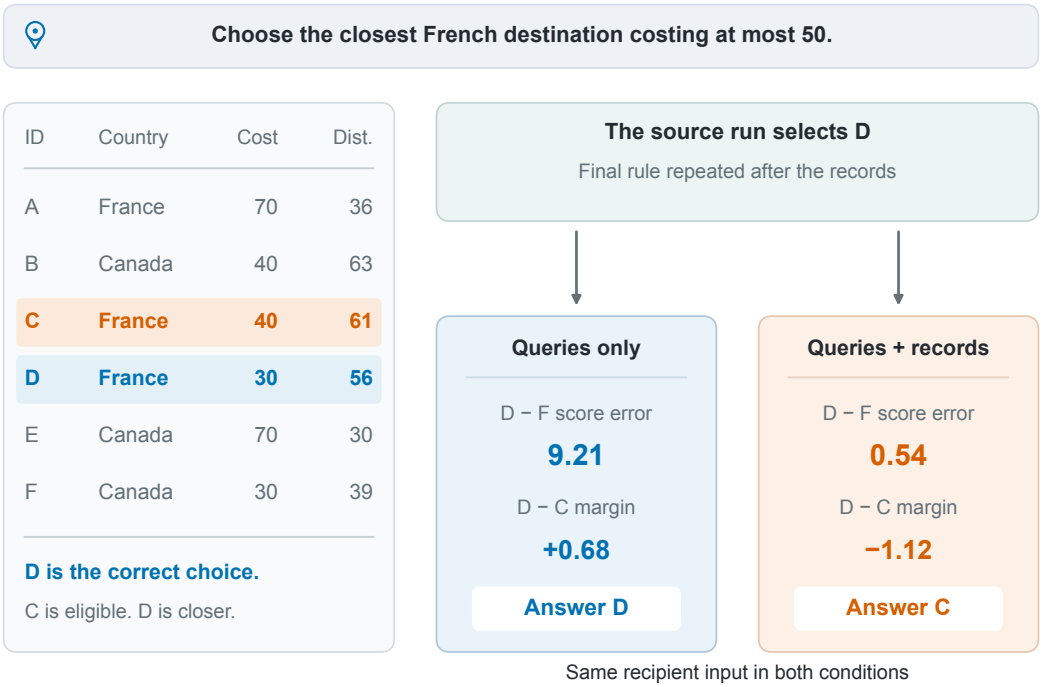

Figure 1: Adding record states improves reconstruction of the source D–F margin but changes the answer from D to C in this recorded Qwen example.

2 Related Work

Transferable task representations. Function vectors identify attention-head outputs that carry an input–output task across contexts (Todd et al., 2024). Task vectors provide a related account of in-context learning as compression into an internal representation (Hendel et al., 2023). Contrastive activation addition changes behavior through residual-stream differences (Rimsky et al., 2024). These approaches establish that internal representations can causally transfer a computation. Causal abstraction formalizes the correspondence between states and task variables (Geiger et al., 2021). Distributed alignment search learns candidate correspondences (Geiger et al., 2024). Their intervention sites and extraction procedures differ from replacing attention queries. Nonlinear instruction-vector effects (Bigoulaeva et al., 2026) and collateral damage from steering (Nguyen et al., 2026) further motivate testing complete task behavior alongside a targeted score.

Sen Sharma et al. (2026) identify sparse heads whose queries encode filtering predicates. Their answer-free inputs and ordering controls distinguish predicate transfer from answer transfer. They

evaluate causal success against all answer candidates and distinguish later filtering from eager evaluation in record states. We use this distinction to study explicit revision with retained threshold and ranking requirements. We test query transfer at the shared response suffix and use a last-token intervention to examine query–record compatibility. We compare a two-answer reconstruction score with complete-answer correctness and paired preservation.

Instruction and constraint evaluation. IFEval uses verifiable instructions to support reproducible scoring (Zhou et al., 2023). FollowBench evaluates progressively added constraints (Jiang et al., 2024). ComplexBench studies their composition (Wen et al., 2024). Our selection tasks provide exact ground truth for category, eligibility, and ranking. Work on evolving conversations identifies failures to integrate later information (Laban et al., 2026). Instruction hierarchies instead specify precedence across sources of different privilege (Wallace et al., 2024). Our histories occur within one user message. The last stated rule determines the required answer. DisentQA separates answers supported by parametric and contextual knowledge (Neeman et al., 2023). Here, the records and final rule explicitly determine a single correct selection.

Context placement can affect retrieval (Liu et al., 2024). Prompt repetition can also help beyond a length-matched control (Leviathan et al., 2025). These findings motivate separating restatement content from suffix coordinates. Direct restatement also provides a practical reference for judging whether an internal intervention offers better revision or preservation.

Causal interventions and their evaluation. Causal tracing links internal activations to factual predictions (Meng et al., 2022). Circuit studies develop faithfulness, completeness, and minimality criteria (Wang et al., 2023). Automated discovery searches for computational subgraphs that preserve a specified behavior (Conmy et al., 2023). Patching results depend on the output metric, corruption, and ablation scheme (Zhang & Nanda, 2024; Heimersheim & Nanda, 2024; Miller et al., 2024). Faithfulness can also distinguish circuits with similar component overlap (Hanna et al., 2024). Self-repair (McGrath et al., 2023) and interacting mediators (Vaidyanathan et al., 2026) explain why recovered effects need not form an additive account. Our evaluation tests whether source-score reconstruction tracks full task execution and whether accuracy gains preserve individual correct answers. Related concerns arise in knowledge editing. RippleEdits checks consequences for related facts (Cohen et al., 2024). MQuAKE tests whether edited facts support multi-hop answers (Zhong et al., 2023). Our task requires a revised criterion to remain compatible with the other requirements of the same selection.

3 METHOD

3.1 SELECTION WITH REVISED AND RETAINED REQUIREMENTS

A task specifies a category, an eligibility threshold, and a ranking rule over the eligible records. For example, the model must choose the French destination with cost at most 50 and the shortest distance. An independent solver determines the unique correct record ID. We use four instruction histories on each table. In *change*, the category is revised. In *no update*, it remains in force. In *keep original*, a further instruction confirms it. In *initial final*, the final category is requested from the start. The threshold and ranking rule remain fixed. These histories occur within one user prompt.

Let record r have category $c(r)$, threshold attribute $u(r)$, and ranking attribute $v(r)$. The rule specifies a target category c^* and threshold τ. For an at-most threshold and minimum ranking, the feasible set and correct answer are

$$\mathcal{E} = \{r : c(r) = c^*,\ u(r) \le \tau\}, \qquad y = \arg\min_{r \in \mathcal{E}} v(r). \tag{1}$$

The at-least and maximum variants change the corresponding comparison. This decomposition locates a selection error. A record can have the requested category and still exceed the budget. Even an eligible record can lose to another eligible candidate on distance.

We distinguish two forms of preservation. Within a revised input, the answer must still satisfy the unchanged threshold and ranking requirements. Across histories, applying the intervention should not break a control response that was already correct. The first concerns the composition of requirements in one answer. The second concerns paired behavior across input conditions.

3.2 REPLACING THE QUERIES USED TO READ THE CONTEXT

Let B be the original prompt and A the same prompt with the complete final rule repeated after the records. We call A the *source* of the transferred state and B its *recipient*. Both contain the same records and request the same answer. We first run A and save its projected attention queries. During the run on B, we replace queries at the shared suffix. This suffix contains the response-format instruction and answer cue. The replacement covers all query heads and layers at these prompt positions.

We replace queries before rotary positional encoding and apply the recipient's coordinates. For a head of width d_h, let q_t be a projected query and $R(p_t)$ its rotary transform (Vaswani et al., 2017; Su et al., 2024). If $\pi(t)$ maps a recipient suffix token to the source, the intervention uses

$$\widetilde{q}_t^I = R(p_t^B) q_{\pi(t)}^A, \qquad P_{tj}^I \propto \exp\Big((\widetilde{q}_t^I)^\top R(p_j^B) k_j^I / \sqrt{d_h} \Big). \tag{2}$$

Here P_{tj}^I is an attention weight and k_j^I is the current key after earlier replacements. We call this operation *query transfer*. The two prompts share the prefix through the records, so their record K/V are identical under causal attention. Transfer reuses the source queries but recomputes suffix keys, values, and attention at recipient coordinates. It also omits the added rule tokens. Similar revision accuracy is therefore expected; paired differences test what the transfer preserves.

3.3 SEPARATING QUERY CONTENT FROM THE STATES BEING READ

To test query–record compatibility, we use a second prompt pair. Both repeat the complete final rule. One places it before the records and the other after them. These placements can change the records' keys and values. Keys govern attention selection. Values supply the information read. We compare query-only transfer with joint transfer of queries and mapped record keys and values (K/V). Both directions between the prompts are tested.

For this compatibility comparison, queries are replaced at the final prompt token in the later half of the layers. The record K/V replacement covers every record token at every layer. Source and recipient tokens are matched by their positions within the identical record text. Keys are replaced before rotary encoding and use recipient coordinates, as in Eq. 2. Values require no positional rotation. Nonrecord keys and values are computed in the recipient. This comparison holds the query intervention fixed and asks what changes when it reads record states from the same source.

The reconstruction metric monitors a cross-category decision. Let z be the best eligible record in the other category and let ℓ_r be the first-answer-token logit for ID r. For condition J, define

$$m_J = \ell_y^J - \ell_z^J, \qquad e_J = |m_J - m_A|. \tag{3}$$

Here A is the source, and lower e_J means closer reconstruction of its margin. This metric can detect a failure to recover the source comparison. It does not constrain the scores of the other candidates. In Figure 1, the monitored pair is D and F. Raising C's logit can leave that pair's difference unchanged and still make the answer wrong.

We compare this reconstruction error with the complete candidate margin

$$m_{\text{all}} = \ell_y - \max_{r \neq y} \ell_r. \tag{4}$$

Its sign determines whether the correct ID outranks every other valid ID. Complete decoded output supplies the behavioral test, because the model can also generate an invalid answer. We retain source correctness in this comparison. Matching an incorrect source can reproduce its mistake even when the whole candidate ordering is recovered. A new error under a correct source exposes a different failure in what the intervention or its metric preserves.

3.4 TESTING WHICH REQUIREMENTS THE QUERIES INFLUENCE

To identify rule-related effects, we extract queries from prompts without records. The final restatement switches category, swaps at-most with at-least, or swaps minimum with maximum ranking. Earlier instructions remain unchanged. Removing records preserves the other tokens' absolute positions. For input i, let y and y_k be the independently solved answers to the original rule and variant

k. We measure the directional effect

$$D_{ik} = (\ell_{y_k} - \ell_y)_{Q_k} - (\ell_{y_k} - \ell_y)_{Q_{\text{faithful}}}, \tag{5}$$

where ℓ is the recipient's first-answer-token logit. Q_k and Q_{faithful} transfer queries from the altered and original rule, respectively. A positive effect means the change favors the corresponding answer. Complete generation tests whether the model actually executes the altered rule.

The same changed queries are also scored against the answers to the other rule variants. These comparisons reveal whether their effect is confined to the intended requirement. A category-directed change, for example, can also alter the margin between the nearest and farthest eligible records. The record-free source has no table from which to compute either answer, so a directional effect cannot require reading the recipient's particular records in the source run. Cyclic relabeling separately tests the remaining association with output-letter preferences.

3.5 Scoring complete answers and paired preservation

An answer is jointly correct only when its complete decoded output is exactly the correct record ID after removing surrounding whitespace. Extra prose and invalid IDs count as errors. Separate scores check the requested category and the unchanged requirements. The latter requires eligibility and the best rank within the selected record's category. These scores distinguish partial compliance from a correct selection.

We classify incorrect answers in the order of invalid output, wrong category, threshold violation, and suboptimal rank. The categories are mutually exclusive. Ranking errors contain only eligible records in the requested category, separating failures of the revised requirement from failures of the retained requirements.

For preservation across histories, a new error occurs when the intervention breaks an originally correct *no update* or *keep original* response. If C^0_{ih} and C^I_{ih} indicate joint correctness for table i, history h, and intervention I, we use

$$H_i(I) = \mathbf{1}\left[\exists h \in \{\text{no update}, \text{keep original}\} : C^0_{ih} = 1,\ C^I_{ih} = 0\right]. \tag{6}$$

Each table is counted once. We also report repairs and new errors in every history. Their difference is the net gain in correct answers. Two interventions can reach the same accuracy through different transitions. Paired scoring makes the lost correct answers visible within an aggregate improvement.

4 Experimental setup

We evaluate Llama-3.1-8B-Instruct (Grattafiori et al., 2024) and Qwen2.5-7B-Instruct (Qwen et al., 2024). Each table contains three records in each of two categories. The generator uses travel, shipping, and candidate-shortlisting vocabulary. Requirements appear together or in separate sentences. Each rule has a unique answer. Appendix A provides a recorded prompt and its complete responses.

Direct final-rule restatement is the behavioral baseline. A position-matched control removes the added rule content and retains the other tokens' source coordinates. Attention-edge blocking reproduces this control. Random changes are matched to the norm of each actual replacement delta over the concatenated projection at each layer and position. The record-state test requires better reconstruction than both query-only transfer and random record changes. The first comparison tests the contribution of source record states. The second checks whether an equally large perturbation explains that contribution.

Interventions replace projected states during prompt processing. Greedy decoding uses the resulting computation and cached states. Inference uses FP32 eager attention and a fixed prefill shape. Appendix D gives the settings.

The main query-transfer test uses the recipient's original positions. We select a capability subset using the unmodified model's answers to *no update* and *initial final*. Both must be correct. This gives 33 Llama and 63 Qwen tables from the full pool of 96. We also report the complete pool. Record-free sources retain the instruction and output format.

Compatibility and position-matched tests share a separate pool of 96 tables. The rule-variant tests use the 86 tokenizer-compatible tables from the main query-transfer pool. Each table remains the

unit of inference across its histories and variants. Directional effects average histories within each table. Binary comparisons use paired exact tests. Continuous intervals use table-level bootstrap resampling, and directional effects use sign tests. Primary tests receive Holm correction within families and globally. Error-set analyses are descriptive. Appendix B gives the full specification.

5 RESULTS

5.1 QUERY TRANSFER IMPROVES REVISION AND CAN BREAK RETAINED REQUIREMENTS

The *change* history tests whether query transfer helps execute a revised rule. The other histories test preservation of the original requirements.

Query transfer raises revision accuracy from 16/33 to 32/33 in Llama's capability subset and from 31/63 to 48/63 in Qwen's (Figure 2). Both gains survive global Holm correction ($p = 1.53 \times 10^{-4}$ and 9.96×10^{-4}). In the full cohorts, query transfer and direct restatement each reach 63/96 for Llama and 69/96 for Qwen. Equal aggregate accuracy nevertheless conceals different changes to individual answers.

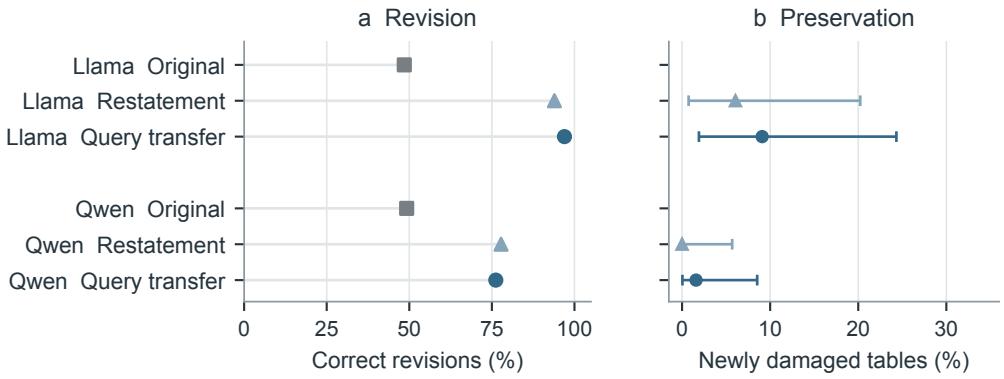

Figure 2: Query transfer and restatement improve revision. Preservation intervals are exact 95% binomial intervals for newly damaged tables.

Query transfer newly damages retention on 6/96 Llama tables and 3/96 Qwen tables, the same counts as direct restatement. The corresponding exact 95% intervals are 2.3–13.1% and 0.6–8.9%. These are selection failures: every generated response is a valid single ID. Paired tests find no preservation advantage over restatement.

Equal revision accuracy does not imply equivalent behavior. Query transfer and restatement select different IDs on 5/96 Llama and 7/96 Qwen change inputs. Each has two Llama successes and three Qwen successes that the other misses. Relative to the original answers, query transfer repairs 33 Llama errors and introduces six, versus 34 repairs and seven new errors for restatement (Figure 3). The comparison identifies gains and losses hidden by net accuracy.

Norm-matched random query controls give 34–35 correct revisions of 96 in Llama and 41–42 in Qwen. Source queries yield 63 and 69. The revision gain depends on the transferred direction rather than perturbation magnitude, even though it does not ensure preservation.

The restated content also contributes beyond suffix position. In both models, restatement improves correction over the position-matched control after global adjustment. The causal mask keeps the preceding record states unchanged. The added rule acts through later reading and computation.

5.2 CLOSER SOURCE SCORES CAN ACCOMPANY A WORSE ANSWER

Matching record K/V reduces Qwen's mean source-margin error by 1.274 logits for after-to-before transfer and 2.453 logits for before-to-after transfer relative to queries alone (Figure 4, left). The

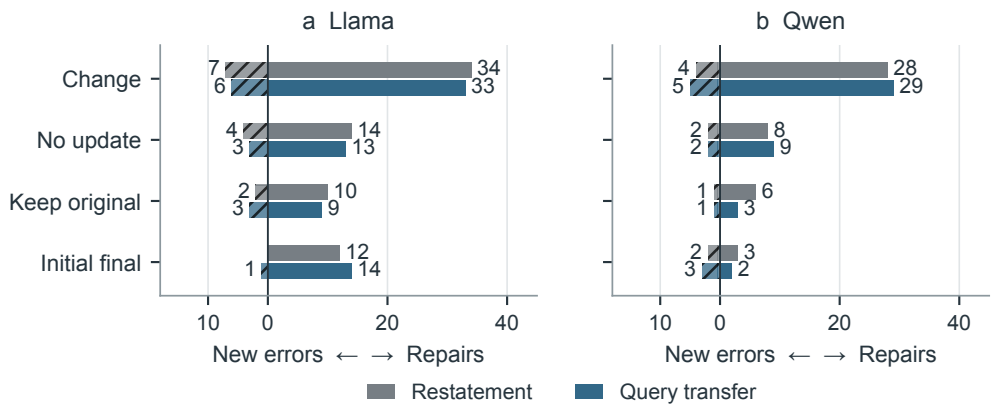

Figure 3: Paired full-cohort outcomes distinguish repaired answers from new errors under restatement and native-position query transfer.

paired 95% intervals are [0.999, 1.557] and [2.170, 2.742]. Both directions also outperform norm-matched random record changes. Llama instead increases error by 0.285 and 0.201 logits. Both prompts state the final rule, so this comparison tests compatibility between transferred queries and record states.

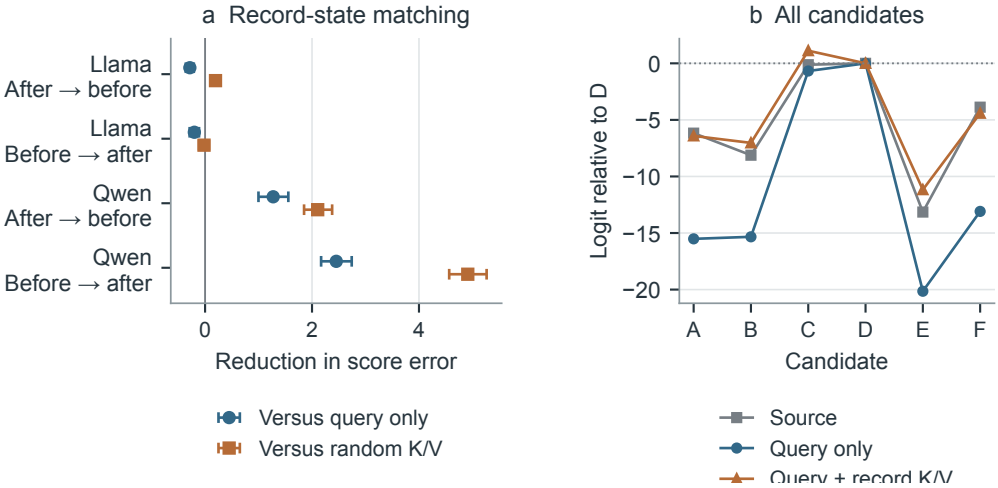

Figure 4: Record-state transfer improves Qwen's source-margin reconstruction but reverses the D–C ranking in the Figure 1 example. Directions refer to source and recipient rule placement. Horizontal intervals are paired table-bootstrap 95% intervals.

Better reconstruction need not improve the decision. In the Figure 1 example, adding record K/V reduces the D–F reconstruction error from 9.21 to 0.54 logits but reverses the D–C margin from $+0.68$ to -1.12 (Figure 4, right). C and D both satisfy the category and budget requirements, but D is closer. The omitted competitor therefore changes the answer from correct D to incorrect C.

For Qwen's before-record recipient, joint transfer repairs 18 query-only errors but creates 28, reducing correctness from 315/384 to 305/384 table–history conditions. Of the new errors, 23 copy an incorrect source answer. These are faithful transfers of a source mistake. Three arise under a correct source, including two with improved reconstruction. Both reversals occur among the 63 correct-source, correct-query change conditions (Figure 1; Appendix E.3). They expose an omitted competitor even when the source decision is correct.

Table 1 separates correct-source errors from copied mistakes. A reversal requires lower reconstruction error and a new incorrect answer. Qwen has two in after-to-before transfer and none in the reverse direction. Llama has two in each direction despite worse average reconstruction. The metric's blind spot and its average improvement are distinct properties.

For Qwen's after-record recipient, adding record states instead raises correctness from 306/384 to 327/384. All nine new errors match the incorrect source answer. Llama gains correct answers in this direction despite worse reconstruction. Reconstruction and selection can disagree in either direction.

Table 1: Paired outcomes when record K/V is added to query transfer. Each direction contains 384 table–history conditions. The lower panel restricts the analysis to correct sources.

Model	Source → recipient	Q correct	Joint correct	Repairs	New errors	Copied
Llama	After → before	282	266	14	30	27
Llama	Before → after	251	262	24	13	10
Qwen	After → before	315	305	18	28	23
Qwen	Before → after	306	327	30	9	9

Model	Source → recipient	Source correct	Q also correct	New errors	Reversals
Llama	After → before	267	246	3	2
Llama	Before → after	257	230	3	2
Qwen	After → before	300	283	3	2
Qwen	Before → after	326	295	0	0

5.3 Directional rule effects do not ensure joint execution

Record-free source queries shift scores toward each changed requirement in both models (Figure 5). All directional tests survive global Holm correction. For Llama's threshold reversal, the intended margin increases by 0.258 logits (95% interval [0.212, 0.302]), yet complete execution rises only from 60/344 to 65/344 table–history conditions. Qwen's threshold margin increases by 3.924 logits while execution rises from 30/344 to 49/344. Each transplant also changes margins for other requirements (Figure 6). A directional score effect can thus coexist with poor execution of the full rule.

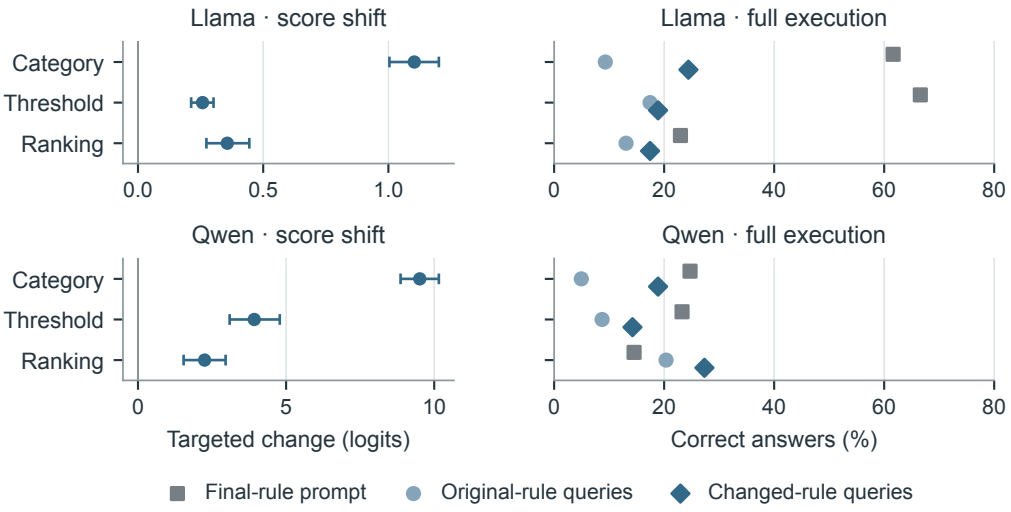

Figure 5: Record-free queries shift the targeted margins without consistent gains in complete execution. Intervals use table-level bootstrap resampling.

Position-matched source controls show the same separation. Relocating the source records changes row order and answer ID. Transferred queries can improve category accuracy as threshold and

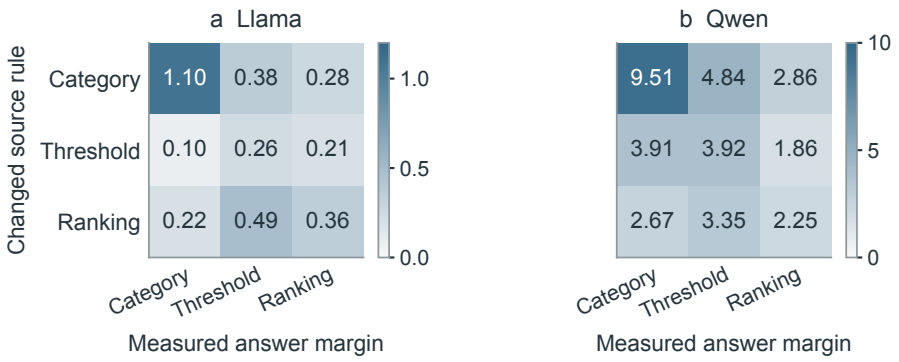

Figure 6: Changing one source requirement also affects margins associated with other requirements. Cells show mean logit shifts on separate model scales.

ranking performance decline (Appendix C). Neither relocated nor record-free sources establish a complete-execution gain in these comparisons.

Record-free sources also retain output-letter preferences. During cyclic relabeling, their input stays fixed as the recipient's correct ID traverses A–F. Qwen's source and recipient letter-margin shifts remain correlated. The association is strongest for ranking. Rule-directed effects coexist with these answer-level preferences.

6 Discussion

Query transfer can repair a revised answer without preserving every requirement across the task. The revision experiment exposes this gap through the correct answers lost alongside the repairs. The separate compatibility experiment explains why an improved internal score can also mislead. Reconstruction can propagate a source mistake or leave an unmonitored competitor free to overtake a correct source answer. Complete execution depends on the comparisons that the reconstruction score omits.

Complete-candidate scoring and paired outcomes address different parts of this problem. The first checks the comparisons that determine the current answer. The second records which previously correct responses an intervention damages. Reporting them together separates targeted influence, complete execution, and preservation across inputs. This distinction also makes direct restatement a useful baseline for interpreting an internal intervention.

Earlier instructions contribute to execution as well. Removing them helps some conflicting-rule cases in Qwen and damages faithful-rule execution. The effect can reverse across histories (Appendix G).

Scope. The experiments use suffix-query replacement for revision and a last-token query intervention for record compatibility in controlled selection tasks. Threshold reversal changes the eligible set, so cross-rule effect sizes combine representational and task differences. Record-free sources retain answer-letter preferences. The present comparisons identify rule-directed influence rather than independently composable representations of each requirement.

7 Conclusion

Query transfer improves rule revision, yet successful source-score reconstruction can accompany incorrect selection. Full-candidate comparisons reveal the missed ranking decisions. Paired outcomes identify the correct answers lost during revision.

AI USE STATEMENT

Generative AI tools assisted manuscript writing and translation.

REPRODUCIBILITY STATEMENT

Appendix A gives the task construction and a complete input–output example. Appendix D specifies the models and intervention settings. Appendix B describes the statistical analysis.

REFERENCES

Irina Bigoulaeva, Jonas Rohweder, Subhabrata Dutta, and Iryna Gurevych. Patches of Nonlinearity: Instruction Vectors in Large Language Models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12209–12262. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.559.

Roi Cohen, Eden Biran, Ori Yoran, Amir Globerson, and Mor Geva. Evaluating the Ripple Effects of Knowledge Editing in Language Models. *Transactions of the Association for Computational Linguistics*, 12:283–298, 2024. doi: 10.1162/tacl_a_00644.

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

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

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

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

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

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

Roee Hendel, Mor Geva, and Amir Globerson. In-Context Learning Creates Task Vectors. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 9318–9333. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-emnlp.624.

Yuxin Jiang, Yufei Wang, Xingshan Zeng, Wanjun Zhong, Liangyou Li, Fei Mi, Lifeng Shang, Xin Jiang, Qun Liu, and Wei Wang. FollowBench: A Multi-level Fine-grained Constraints Following Benchmark for Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4667–4688. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.257.

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. LLMs Get Lost In Multi-Turn Conversation. In *International Conference on Learning Representations*, pp. 54738–54778, 2026.

Yaniv Leviathan, Matan Kalman, and Yossi Matias. Prompt Repetition Improves Non-Reasoning LLMs. arXiv preprint arXiv:2512.14982, 2025.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638.

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

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, pp. 17359–17372, 2022. doi: 10.52202/068431-1262.

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

Ella Neeman, Roee Aharoni, Or Honovich, Leshem Choshen, Idan Szpektor, and Omri Abend. DisentQA: Disentangling Parametric and Contextual Knowledge with Counterfactual Question Answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10056–10070. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.559.

Tam Nguyen, Tu Anh Nguyen, Sina Alemohammad, and Richard G. Baraniuk. Minimizing Collateral Damage in Activation Steering. arXiv preprint arXiv:2605.01167, 2026.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 Technical Report. arXiv preprint arXiv:2412.15115, 2024.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering Llama 2 via Contrastive Activation Addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.828.

Arnab Sen Sharma, Giordano Rogers, Natalie Shapira, and David Bau. LLMs Process Lists With General Filter Heads. In *International Conference on Learning Representations*, pp. 63227–63263, 2026.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with Rotary Position Embedding. *Neurocomputing*, 568:127063, 2024. doi: 10.1016/j.neucom.2023.127063.

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*, pp. 17282–17333, 2024.

Sankaran Vaidyanathan, David Arbour, Aaron Mueller, Scott Niekum, and David Jensen. The Curse of Multiple Mediators: Hidden Interaction Effects in Activation Patching. arXiv preprint arXiv:2606.27510, 2026.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Eric Wallace, Kai Xiao, Reimar Leike, Lilian Weng, Johannes Heidecke, and Alex Beutel. The Instruction Hierarchy: Training LLMs to Prioritize Privileged Instructions. arXiv preprint arXiv:2404.13208, 2024.

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.

Bosi Wen, Pei Ke, Xiaotao Gu, Lindong Wu, Hao Huang, Jinfeng Zhou, Wenchuang Li, Binxin Hu, Wendy Gao, Jiaxin Xu, Yiming Liu, Jie Tang, Hongning Wang, and Minlie Huang. Benchmarking Complex Instruction-Following with Multiple Constraints Composition. In *Advances in Neural Information Processing Systems*, volume 37, pp. 137610–137645, 2024. doi: 10.52202/079017-4371.

Fred Zhang and Neel Nanda. Towards Best Practices of Activation Patching in Language Models: Metrics and Methods. In *International Conference on Learning Representations*, pp. 1651–1678, 2024.

Zexuan Zhong, Zhengxuan Wu, Christopher D. Manning, Christopher Potts, and Danqi Chen. MQuAKE: Assessing Knowledge Editing in Language Models via Multi-Hop Questions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 15686–15702. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.971.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-Following Evaluation for Large Language Models. arXiv preprint arXiv:2311.07911, 2023.

A TASK CONSTRUCTION AND EXAMPLES

A.1 A RANKING ERROR UNDER IMPROVED RECONSTRUCTION

The following input underlies Figure 1. It illustrates how a two-answer margin misses a within-category competitor. The final country is France, the cost limit is 50, and the selected eligible distance must be minimal. D is the unique answer. C meets the country and cost requirements but has a larger distance. F is the best eligible Canadian record.

```
Which record has country=Canada?
Additional requirements: cost at most 50; among qualifying records select the lowest distance.
Change only the country requirement to France. Keep the threshold and ranking requirements
↪  unchanged.
Records:
A: country=France; cost=70; distance=36
B: country=Canada; cost=40; distance=63
C: country=France; cost=40; distance=61
D: country=France; cost=30; distance=56
E: country=Canada; cost=70; distance=30
F: country=Canada; cost=30; distance=39
Respond with exactly one record ID (A, B, C, D, E, or F). Do not add any other text.
Answer:
```

The recipient inserts the sentence below immediately before `Records:`. The source inserts it after the records and before the response-format instruction.

```
Final complete requirements: country=France; cost at most 50; select the qualifying record
↪  with the lowest distance.
```

Table 2: Complete decoded outputs and two answer margins for the overview example.

Condition	Output	$D - F$	$D - C$
Donor, rule after records	D	3.87620	0.12257
Recipient, rule before records	D	15.50130	8.63461
Recipient with donor Q	D	13.08863	0.67902
Recipient with donor Q and record K/V	C	4.41463	-1.11589

The monitored margin $m = \ell_D - \ell_F$ omits the competing French record C. Adding record states reduces its reconstruction error from 9.21 to 0.54 logits but reverses the D–C margin from 0.68 to -1.12. The answer changes from D to C. Appendix E.3 gives the frequency of such reversals.

A.2 HISTORIES AND ANSWER CONSTRUCTION

For every table, the original instruction specifies a category, threshold comparison, and ranking direction. The *change* instruction changes only the category and explicitly retains the other requirements. The *no update* prompt omits that change. The *keep original* instruction confirms the initial category. The *initial final* prompt asks for the final category from the start. The generator gives each category two threshold-eligible records and one ineligible record. Ranking values determine a unique best eligible record. Requirements can be packed into one sentence or separated into an initial category question and an additional-requirements line.

The semantic counterfactuals alter the final restatement only. A category flip chooses the other category. A threshold flip reverses at-most to at-least, or the reverse. A ranking flip reverses lowest to highest, or the reverse. Each variant is solved independently and must change the correct ID. Earlier requirements remain in the prompt. The final restatement defines ground truth for these counterfactual tasks.

The threshold flip leaves one eligible record in the requested category. Faithful, category, and ranking variants leave two. Cross-axis accuracy therefore reflects both the intervention and the eligible set. Record rotation shifts the six rows and reassigns IDs A–F. It changes row order and answer identity together. Record-free prompts remove the record tokens but retain the remaining absolute positions.

B STATISTICAL ANALYSIS

Statistical inference uses tables as the sampling unit. Histories and variants retain their shared table identity. Table-level percentile bootstrap intervals use 10,000 resamples. Seeds are derived from the comparison labels.

For binary correction gains, an exact one-sided binomial test considers the originally-wrong/intervention-correct and originally-correct/intervention-wrong pairs. This is the directional exact McNemar comparison. Preservation compares paired damage indicators from Eq. 6. Damage rates also have exact binomial intervals. No noninferiority margin was specified for query transfer versus restatement.

For directional semantic effects, the four histories are averaged per table. Differences above .01 logits count as positive, those below $-.01$ as negative, and the remainder as ties. The one-sided sign test excludes ties from its binomial denominator. Record-state comparisons use the same tolerance after averaging four histories. Each record-state hypothesis is an intersection-union test requiring improvement over query-only transfer and the mean of the three random-record controls. Its raw p value is the maximum of those two directional sign-test values.

The four hypothesis families were specified separately. We report Holm adjustment within families and across all 20 primary hypotheses. History deletion, random-margin comparisons, semantic selectivity, letter correlations, and error-set examples are descriptive analyses.

Table 3: All 20 primary comparisons, including the global multiplicity sensitivity analysis.

Model	Hypothesis	n	Raw p	Family p	Global p
Llama	Record K/V, before ← after	96	1	1	1
Llama	Record K/V, after ← before	96	1	1	1
Qwen	Record K/V, before ← after	96	1.04×10^{-11}	3.11×10^{-11}	1.56×10^{-10}
Qwen	Record K/V, after ← before	96	4.38×10^{-23}	1.75×10^{-22}	7.88×10^{-22}
Llama	Restatement versus position gap	40	9.54×10^{-7}	4.77×10^{-6}	1.05×10^{-5}
Llama	Restated Q, relocated donor	40	0.274	0.823	1
Llama	Restated Q, record-free donor	40	1	1	1
Qwen	Restatement versus position gap	66	2.98×10^{-8}	1.79×10^{-7}	3.58×10^{-7}
Qwen	Restated Q, relocated donor	66	0.0717	0.287	0.574
Qwen	Restated Q, record-free donor	66	0.339	0.823	1
Llama	Category-direction effect	86	1.12×10^{-24}	5.62×10^{-24}	2.14×10^{-23}
Llama	Threshold-direction effect	86	2.15×10^{-14}	6.46×10^{-14}	3.45×10^{-13}
Llama	Ranking-direction effect	86	6.75×10^{-9}	1.35×10^{-8}	9.45×10^{-8}
Qwen	Category-direction effect	86	1.29×10^{-26}	7.75×10^{-26}	2.58×10^{-25}
Qwen	Threshold-direction effect	86	7.62×10^{-16}	3.05×10^{-15}	1.30×10^{-14}
Qwen	Ranking-direction effect	86	6.75×10^{-9}	1.35×10^{-8}	9.45×10^{-8}
Llama	Native Q correction gain	33	1.53×10^{-5}	6.10×10^{-5}	1.53×10^{-4}
Llama	Native Q preservation advantage	33	1	1	1
Qwen	Native Q correction gain	63	1.11×10^{-4}	3.32×10^{-4}	9.96×10^{-4}
Qwen	Native Q preservation advantage	63	1	1	1

B.1 COHORT OVERLAP AND CAPABILITY SELECTION

The record-state and position-matched transfer comparisons share the same 96-table pool. Native-position query transfer uses a second pool of 96 tables. Semantic transfer excludes 10 second-pool tables before model outcomes because the two tokenizers cannot jointly preserve the required positions. Its 86 tables are also used by history deletion. Native-position query transfer uses all 96 tables. Statistical comparisons retain these shared input groups.

Capability selection requires correct original *no update* and *initial final* answers. The record/position pool yields 40 Llama and 66 Qwen tables for the position-matched transfer tests. The second full pool yields 33 and 63 for native-position transfer. The 86-table semantic intersection yields 31 and 53. Intervention outcomes never enter this filter.

C JOINT EXECUTION AND RETAINED REQUIREMENTS

C.1 NATIVE-POSITION TRANSFER ACROSS EVERY HISTORY

History columns count correct complete answers. Harm counts tables with a new retention error under Eq. 6. Each table is counted once even if both retention histories are damaged.

Table 4: Correct selections and newly introduced retention errors across instruction histories.

Model	Subset	Condition	n	Change	No upd.	Keep	Initial	Harm
Llama	Full	Original	96	36	54	58	58	0
Llama	Full	Restatement	96	63	64	66	70	6
Llama	Full	Q transfer	96	63	64	64	71	6
Llama	Full	Random Q 1	96	34	51	57	57	14
Llama	Full	Random Q 2	96	35	49	59	53	18
Llama	Full	Random Q 3	96	34	55	58	53	14
Llama	Capable	Original	33	16	33	30	33	0
Llama	Capable	Restatement	33	31	31	32	33	2
Llama	Capable	Q transfer	33	32	31	31	32	3
Llama	Capable	Random Q 1	33	17	29	30	28	5
Llama	Capable	Random Q 2	33	16	24	28	29	9
Llama	Capable	Random Q 3	33	13	28	29	26	5
Qwen	Full	Original	96	45	78	76	80	0
Qwen	Full	Restatement	96	69	84	81	81	3
Qwen	Full	Q transfer	96	69	85	78	79	3
Qwen	Full	Random Q 1	96	41	74	74	80	9
Qwen	Full	Random Q 2	96	42	79	74	80	4
Qwen	Full	Random Q 3	96	41	79	77	77	4
Qwen	Capable	Original	63	31	63	56	63	0
Qwen	Capable	Restatement	63	49	63	59	61	0
Qwen	Capable	Q transfer	63	48	62	57	60	1
Qwen	Capable	Random Q 1	63	27	59	58	61	4
Qwen	Capable	Random Q 2	63	28	62	57	62	1
Qwen	Capable	Random Q 3	63	29	62	57	59	2

C.2 RESTATEMENT CONTENT AND DONOR RECORDS

The transfer controls compare a donor with final restatement to its matched donor without restatement. A relocated donor changes row order and answer ID together. A record-free donor contains no data that uniquely determines the correct record. Neither transformation guarantees that the donor is free from output-letter preferences.

Table 5: Complete joint-correction and retention-damage counts for the matched donor comparisons.

Model	Subset	Condition	n	Correct	Harm
Llama	Full	Position gap	96	32	10
Llama	Full	Relocated Q, no restatement	96	41	22
Llama	Full	Record-free Q, no restatement	96	42	19
Llama	Full	Restatement	96	64	6
Llama	Full	Relocated Q, restatement	96	51	23
Llama	Full	Record-free Q, restatement	96	28	27
Llama	Capable	Position gap	40	19	6
Llama	Capable	Relocated Q, no restatement	40	20	11
Llama	Capable	Record-free Q, no restatement	40	22	10
Llama	Capable	Restatement	40	39	1
Llama	Capable	Relocated Q, restatement	40	23	14
Llama	Capable	Record-free Q, restatement	40	13	15
Qwen	Full	Position gap	96	33	4
Qwen	Full	Relocated Q, no restatement	96	30	34

Model	Subset	Condition	n	Correct	Harm
Qwen	Full	Record-free Q, no restatement	96	34	27
Qwen	Full	Restatement	96	66	10
Qwen	Full	Relocated Q, restatement	96	40	40
Qwen	Full	Record-free Q, restatement	96	34	31
Qwen	Capable	Position gap	66	25	3
Qwen	Capable	Relocated Q, no restatement	66	21	28
Qwen	Capable	Record-free Q, no restatement	66	24	22
Qwen	Capable	Restatement	66	50	6
Qwen	Capable	Relocated Q, restatement	66	28	32
Qwen	Capable	Record-free Q, restatement	66	27	26

Relocated sources improve category correctness while reducing satisfaction of the unchanged requirements. In the capability subsets, joint correctness rises from 20/40 to 23/40 for Llama and from 21/66 to 28/66 for Qwen when restatement is added to the source. These gains accompany retention damage on 14/40 and 32/66 tables.

C.3 Counterfactual execution and answer-letter effects

The following counts use all 86 compatible tables and all four histories. Each semantic axis has 344 outputs per model. Native prompts keep the earlier history and explicitly state the counterfactual rule at the end. Faithful-Q answers are scored against the counterfactual answer for the same comparison.

Table 6: Joint counterfactual answers on all compatible table-history inputs.

Model	Changed rule	Native	Faithful Q	Changed Q	Outputs
Llama	Category	212	32	84	344
Llama	Threshold	229	60	65	344
Llama	Ranking	79	45	60	344
Qwen	Category	85	17	65	344
Qwen	Threshold	80	30	49	344
Qwen	Ranking	50	70	94	344

The record-free donor and recipient can share answer-letter preferences even though the donor input cannot uniquely identify the correct record. In Qwen, table-level correlations between donor and recipient letter-margin shifts are .487 for category, .673 for threshold, and .798 for ranking. The corresponding Llama correlations are $-.029$, .165, and .168. These correlations measure shared answer-letter preferences. Qwen's ranking effect is negative on average for the counterfactual-B subgroup and positive for the other answer IDs. A near-zero mean donor margin does not remove this coupling.

Within-axis comparisons to the mean of three random-query controls support positive directional effects except that Llama's threshold interval crosses zero. The random norm is matched to the complete counterfactual-donor transplant into the recipient, rather than only the counterfactual-minus-faithful semantic delta. Cross-axis selectivity is mixed. The absence of clear ranking selectivity in some comparisons limits the claim of independent abstract rule factors.

C.4 History deletion and its faithful control

History deletion removes the requirements before `Records:`, leaving the subsequent token IDs and their absolute positions unchanged. It changes the context available when record states are computed.

Table 7: History deletion counts include the faithful-rule control.

Model	Subset	Final rule	Original	Deleted	Outputs
Llama	Full	Faithful	235	212	344
Llama	Full	Category	212	210	344
Llama	Full	Threshold	229	199	344
Llama	Full	Ranking	79	78	344
Llama	Capable	Faithful	119	99	124
Llama	Capable	Category	106	99	124
Llama	Capable	Threshold	53	59	124
Llama	Capable	Ranking	30	17	124
Qwen	Full	Faithful	278	185	344
Qwen	Full	Category	85	187	344
Qwen	Full	Threshold	80	217	344
Qwen	Full	Ranking	50	126	344
Qwen	Capable	Faithful	195	127	212
Qwen	Capable	Category	59	128	212
Qwen	Capable	Threshold	32	117	212
Qwen	Capable	Ranking	19	85	212

For Qwen, faithful-rule violations of the threshold within the target category increase from 34 to 119 after deletion. In ranking reversal, 144 outputs remain eligible but suboptimal and 74 violate eligibility. Llama retains 162 eligible-but-suboptimal and 104 ineligible ranking outputs. These errors persist after removing earlier instructions. The pooled counts describe correlated histories.

D Intervention implementation

D.1 Execution settings

Table 8: Settings for the experiments reported in the main text.

Setting	Value
Models	Llama-3.1-8B-Instruct; Qwen2.5-7B-Instruct
Precision and attention	FP32; eager attention; TF32 disabled
Prefill shape	Fixed 320-token computation with masked padding
Generation	Greedy; at most 64 new tokens; model EOS termination
Prompt formatting	Model chat template; one user message; generation prompt added
Revision and rule-variant Q	Pre-RoPE projections; all layers and query heads; 34 shared prompt tokens
Compatibility Q	Pre-RoPE projections; all query heads in the later half of layers; final prompt token
Compatibility record K/V	Pre-RoPE key and value projections; all record tokens and layers
Position convention	Recipient RoPE coordinates for mapped record and suffix tokens
Random controls	Three seeds; replacement-delta L2 matched per layer and prompt position
Hardware	NVIDIA A800, 80 GB

Append-only restatement cannot change preceding records under causal attention. Fixed 320-token computation shapes keep the numerical prefix comparison consistent across prompt lengths.

Invalid or capped responses count as errors.

D.2 Choice of query positions

On the capability subsets used to select query positions, full-suffix transfer gives 17/19 correct Llama answers and final-position transfer gives 16/19. Both give 23/33 in Qwen. The revision experiment uses the full suffix to include query effects distributed across these positions.

E Error classification and counterfactual controls

E.1 The complete selection rule

Equation 1 defines the feasible set and unique optimum. The exclusive error classification below applies the same rule to the full generated answer.

We classify an output first by exact joint correctness. Among the remaining answers, invalid complete output, wrong category, threshold violation within the target category, and eligible-but-suboptimal choice are distinct outcomes. This order prevents counting a wrong-category answer again as a ranking error in the target category. Figure 7 uses all 96 native-position transfer tables, including tables excluded from the capability subset.

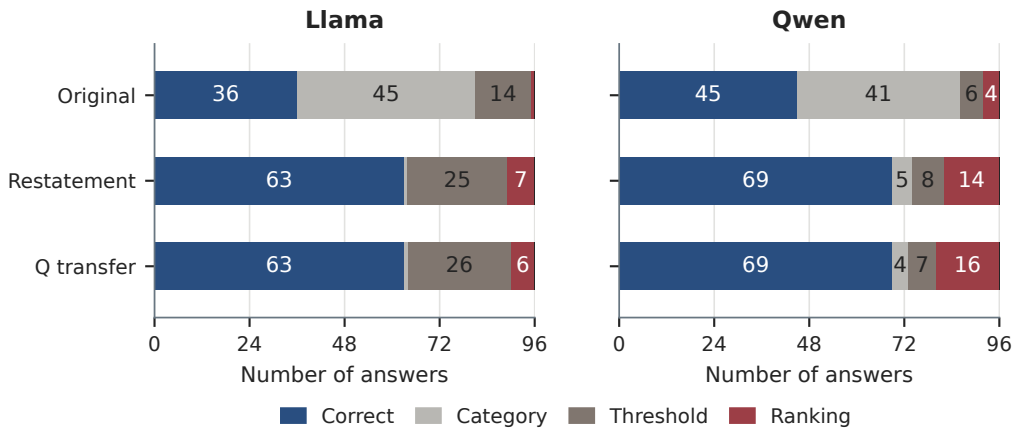

Figure 7: Disjoint error categories show which parts of the full revision rule fail before and after the two interventions.

E.2 A scalar reconstruction metric leaves competitors unconstrained

Equations 3 and 4 distinguish the monitored cross-category margin from the complete candidate margin. Figure 4 plots all six candidates in the recorded example. Subtracting D's logit preserves every within-condition pairwise margin and candidate ordering. The sign of the complete margin determines the best valid ID, but complete generation and preservation on other histories still require their own recorded outcomes.

E.3 Source correctness and new selection errors

We separate faithful copying of an incorrect source from new errors under a correct source. A new error occurs when query-only transfer is correct and adding record K/V makes the answer incorrect. A reconstruction–selection reversal additionally requires a decrease in the monitored source-margin error.

Table 1 groups results by source correctness and restatement placement. Both Qwen reversals occur on different tables in the change history, among 63 conditions with a correct source and a correct query-only answer. The remaining new Qwen error increases reconstruction error. Adding record states can therefore improve reconstruction while introducing a selection error under a correct source.

E.4 Cross-axis effects and shared answer preferences

To inspect selectivity, each counterfactual donor is evaluated against each counterfactual answer margin. Define y_j as the answer to semantic variant j. The entry in donor row i and measured

column j is

$$E_{ij} = \mathbb{E}_{\text{table}}\left[\mathbb{E}_{\text{history}}\left\{(\ell_{y_j} - \ell_y)_{Q_i} - (\ell_{y_j} - \ell_y)_{Q_{\text{faithful}}}\right\}\right]. \tag{7}$$

The diagonal recovers the primary directional effects. Off-diagonal entries reveal other answer margins changed by the same transplant (Figure 6). Color scales differ across models because their logit-effect magnitudes differ.

Answer-letter correlations compare faithful-to-counterfactual margin changes in the record-free source and recipient after averaging histories within each table (Figure 8). The source input is invariant to record relabeling, so shared letter preferences can contribute even without record information.

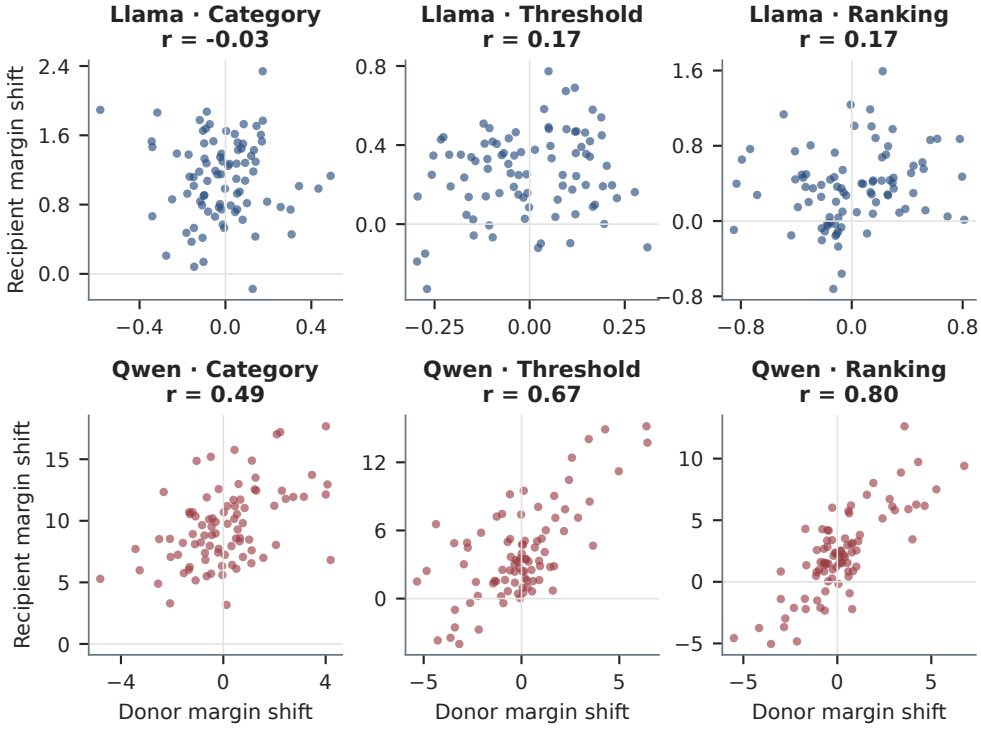

Figure 8: Record-free donor and recipient margin changes reveal model- and rule-dependent coupling to answer-letter preferences.

E.5 The effect of deleting history depends on the history

Figure 9 retains all four histories for each final-rule variant. A cell reports the number of correct answers after deletion minus the number before deletion, with 86 paired tables. In particular, the Qwen category effect in the change history reverses the direction suggested by pooling histories.

F Intervention operators and reproducible scoring

F.1 Replacement at recipient coordinates

At each patched layer, the source query is projected before rotary encoding and substituted at the mapped recipient token. Revision and rule-variant tests use the shared suffix at every layer. Compatibility tests use the final prompt token in the later half of the layers. The recipient's position ID is then applied. Nonpatched positions and the causal attention mask remain those of the recipient. Hidden states produced by earlier patched layers propagate normally to later layers. Earlier

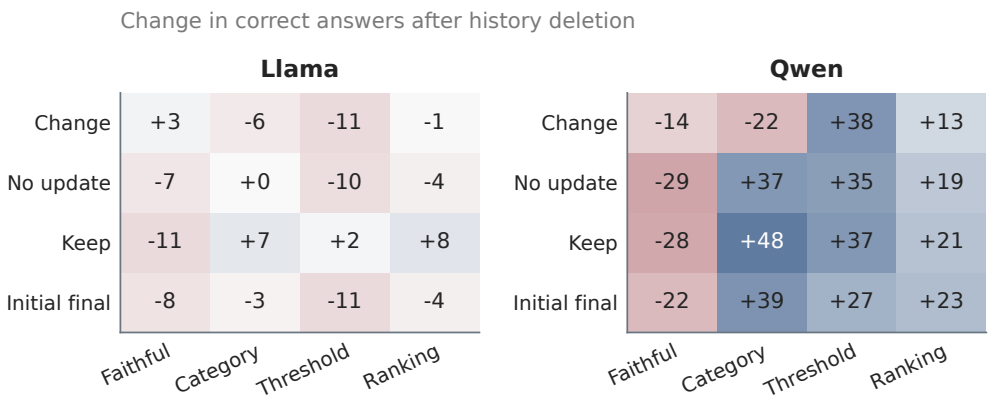

Figure 9: The full history-by-rule breakdown exposes opposite-signed effects concealed by pooled history-deletion accuracy.

replacements alter the states entering later layers, so the operator changes a coupled sequence of computations.

For a norm-matched random query, let q_B be the recipient's query tensor at one layer and token and let $\delta = q_A - q_B$. We sample a random direction u of matching shape and replace q_B by

$$q_B + \|\delta\|_2 u / \|u\|_2. \tag{8}$$

The norm is taken over the concatenated query projection rather than separately within each head. Random directions test whether perturbation magnitude explains the effect. Record-state controls use the replacement norm for each specified tensor group.

F.2 Complete-answer scoring

Leading and trailing whitespace are removed from generated responses. No other normalization is applied. The symbolic rule then assigns each complete output to the mutually exclusive categories above. Capability membership uses only original no-update and initial-final answers. Table-level aggregation and paired inference follow Appendix B.

G EARLIER INSTRUCTIONS AND CONFLICTING FINAL RULES

Earlier instructions may help represent the records or compete with the final rule. Deleting requirements before the records removes both influences while preserving all remaining token IDs and absolute positions.

For Qwen, history deletion improves pooled category, threshold, and ranking counterfactual accuracy. The same deletion substantially damages execution when the final rule agrees with the earlier requirements. Llama shows no corresponding pooled counterfactual improvement. Even within Qwen, category reversal becomes less accurate in the *change* history, reversing the pooled trend. The benefit depends on which earlier instruction is removed and how it relates to the final rule. Figure 10 reports the full tradeoff.

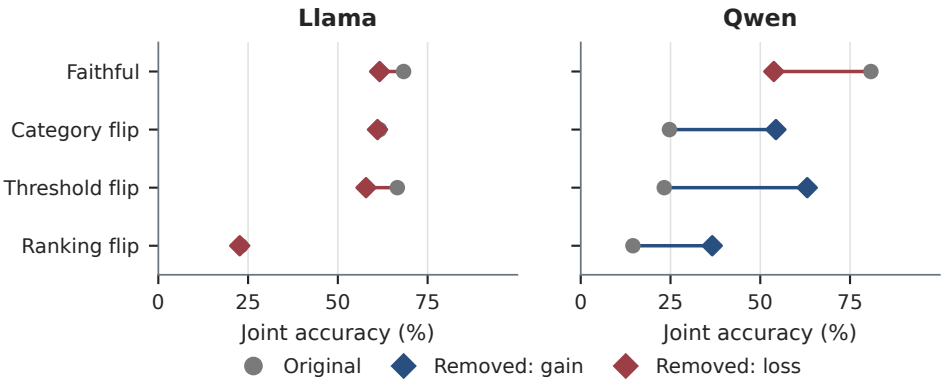

Figure 10: Removing earlier requirements has different effects across final-rule variants and models, and can damage faithful execution.

Deletion changes the context used to compute record states as well as possible instruction conflict. Its benefit on some counterfactual rules and damage on faithful rules show why removing earlier requirements is not a general repair.