CONTENT-DEPENDENT RECOVERY FROM TASK-RELEVANT ACTIVATION DISRUPTION IN LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

When prompts are modified or rewritten, large language models (LLMs) need to determine which information in the prompt remains valid in order to continue completing the specified task. Prior work has examined how models process such changed information, but what happens to the model when its internal computation is disrupted remains largely a black box. We gradually attenuate internal activations at task-related token positions, select intervention strengths from the resulting performance curves, and write activations recorded during normal runs back into the perturbed model. Experimental results across five open-weight models and four reasoning tasks show that reasoning performance drops more when task-related information is disrupted, while nearby layers can show similar effects. When activations carrying the correct information are written back, the models can recover their reasoning performance. We further find that recovery in reasoning does not simply correspond to internal activations returning toward their normal state: incorrect information can also move perturbed activations closer to the normal state while producing substantially weaker recovery of the receiver's correct answer. This pattern is observed across all five models, while component analyses show model-dependent effects, with the clearest effects occurring at different attention and feed-forward sites.

1 INTRODUCTION

When a prompt contains an initial fact followed by revisions, a model must determine which information remains valid in order to answer correctly. We refer to the currently valid information needed to determine the answer as the *task state*. Studies by Lyu et al. (2025) and Kim et al. (2026) examine how models distinguish current facts from superseded information. The evaluation proposed by Atri et al. (2026) tests whether models preserve or revise information across successive interactions. We place the relevant records in a single prompt and examine the internal computation used to answer the final question.

Activation interventions provide a direct way to test how models use this information. Studies by Feng & Steinhardt (2024), Dai et al. (2024), and Dai et al. (2026) show that internal activations associate entities with attributes and can alter relational predictions. Work by Hendel et al. (2023) and Todd et al. (2024) further shows that activations extracted from demonstrations can influence later predictions. These findings motivate our central question. After task-related activations are weakened enough to reduce reasoning performance, what information must replacement activations carry to restore the correct answer?

Our central objective is to determine how source content affects performance restoration after task-relevant activation disruption. Selective vulnerability in Alzheimer's disease (Mathys et al., 2024) motivates comparing disruption across internal locations. Recall improvements after targeted stimulation, with larger gains at milder baseline impairment (Benussi et al., 2022), motivate testing restoration across levels of performance loss. This biological connection is a functional analogy. We attenuate task-related activations and select two settings from measured performance curves. Activation patching supplies states recorded from source runs (Heimersheim & Nanda, 2024). Varying source content and wording tests how supplied information supports restoration, with position and

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.

random-vector controls assessing location and content effects. Appendix A details the evidence motivating these experimental choices.

Answer recovery and activation similarity are evaluated separately. Candidate-ranking accuracy and answer margin measure agreement with the receiver's correct answer. The clean-reference nearest-neighbor ratio measures standardized distance to a finite set of unperturbed calibration activations. Paired cosine distance measures angular separation from the unperturbed activation of the same prompt. Both diagnostics use a fixed 128-dimensional projection of the final-layer residual vector at the last prompt token after the intervened forward pass. It also follows the findings of Makelov et al. (2024), Méloux et al. (2025), and Grant et al. (2026), who show that causal interventions can recruit alternative computations or produce activations that differ from unperturbed inference.

RQ1. How does weakening activations at task-related token positions affect reasoning performance compared with interventions at alternative token positions or nearby layers?

RQ2. How does source content affect recovery of the receiver's correct answer across performance settings and replacement windows?

RQ3. How do the clean-reference nearest-neighbor ratio and paired cosine distance relate to recovery of the receiver's correct answer?

Our main contribution is to show how disrupting task-related information affects reasoning performance and how the information carried by activations influences recovery. Reasoning performance drops more under task-related disruption than under the tested alternative-position controls, while calibrated nearby-layer interventions produce similar effects. Writing back activations carrying correct information restores reasoning performance. Incorrect information can also move perturbed activations closer to their normal state while producing substantially weaker recovery than correct information. Component analyses further show that intervention effects vary across models, with the clearest effects occurring at different attention and feed-forward sites

2 Related Work

2.1 Maintaining valid information and measuring internal disruption

Research on evolving information addresses several levels of the problem. FactTrack from Lyu et al. (2025) maintains an explicit time-aware record of facts to identify inconsistencies in story outlines. The benchmark developed by Kim et al. (2026) evaluates how language models adapt as new information arrives. The evaluation introduced by Atri et al. (2026) examines whether information remains consistent across multiple turns. These studies establish the importance of retaining valid information while replacing obsolete facts. Our constructed tasks place these decisions within one prompt so that internal interventions can be applied to the positions associated with the relevant records. Mechanistic studies examine how contextual associations are implemented. Binding-vector interventions from Feng & Steinhardt (2024) can change entity–attribute associations. The analyses of Dai et al. (2024) identify compact structure associated with binding, while Dai et al. (2026) provide a cell-based account of relational information. The work of Gur-Arieh et al. (2026) distinguishes positional and lexical mechanisms for retrieving bound entities from context. These findings motivate our comparison between task-related token positions and alternative context positions. Targeted internal interventions have also been used to study impaired language performance. He et al. (2026) identify neurons associated with Alzheimer's disease transcripts and scale their down-projection contributions. Amplifying these contributions produces graded impairment, while attenuation largely preserves performance and improves some outcomes. Our intervention attenuates residual activations at selected token positions during inference. The comparisons reported by Chen et al. (2025) and Laptev et al. (2025) provide additional context for studying component-level effects.

2.2 Restoring reasoning through internal interventions

The studies of Hendel et al. (2023) and Todd et al. (2024) show that activations obtained from demonstrations can carry information that affects later predictions. These studies primarily encode the task or function to be performed. Within each base case, our source conditions keep the task and label inventory fixed while varying which state supports the answer. This design compares

activations from prompts that support the current answer with activations from prompts that support a different answer. Recovery can also follow parameter modification or retraining. Li et al. (2024) use findings about compositional reasoning to edit attention modules. Wang et al. (2025) recover language performance after targeted damage by retraining intact experts, with recovery depending on the training content. Our experiments keep model parameters fixed and replace activations during the disrupted run. We compare exact sources, paraphrased sources, incorrect-state sources, question-position controls, and random-vector controls at two selected levels of performance loss. Activation interventions differ in the representational object they modify. Learned subspace replacement aligns internal representations with specified causal variables, while directional addition shifts activations along a contrast derived from paired examples (Wu et al., 2023; Rimsky et al., 2024). Our experiment replaces complete residual vectors at aligned token positions to test the effect of source content under a fixed replacement operator. The resulting comparison concerns full-vector source-state patches across the tested performance settings.

2.3 Activation similarity and correct-answer recovery

Geometric analyses show that internal activations can reflect structured information. The controlled-sequence study of Shai et al. (2024) identifies belief-state geometry in the residual stream, while Park et al. (2025) characterize categorical and hierarchical organization. Such measurements describe relations in activation space, but they do not by themselves establish whether a particular activation is used to produce a correct answer. Sharma et al. (2026) distinguish decodability from causal use in a controlled transformer setting. The framework of Heimersheim & Nanda (2024) links changes in internal activations to changes in model outputs. Its interpretation depends on the intervention and the measured outcome. Makelov et al. (2024) show that subspace interventions can activate alternative computational paths. Méloux et al. (2025) show that different internal explanations can support the same output in controlled settings. Grant et al. (2026) show that replacement activations can differ from those observed during unperturbed inference. These findings motivate evaluating answer correctness and activation distance as separate outcomes.

3 Methods

3.1 Experimental overview

The experiment proceeds through site selection, strength calibration, and evaluation. Site selection uses a dedicated prompt set to identify layer–token combinations affected by attenuation. Calibration selects intervention strengths and activation-replacement layers. Evaluation then uses these fixed settings. Graded attenuation measures performance loss. Component interventions compare internal locations. Source-controlled replacement measures performance restoration. Figure 1 illustrates source-state replacement through a passage memory-update case. Exact and paraphrased sources restore the correct answer, while an incorrect-state source shifts the top-ranked candidate to the answer supported by that source.

3.2 Models and task construction

Table 1 summarizes the five open-weight models. The Qwen series is represented by Qwen2.5 32B Instruct from Yang et al. (2024) and Qwen3 14B from Yang et al. (2025). We evaluate Gemma 3 12B instruction-tuned from Gemma Team et al. (2025) and Gemma 4 12B instruction-tuned from Gemma Team et al. (2026). Phi-4 14B from Abdin et al. (2024) completes the model set. ChatGPT generated the initial synthetic task content and the authors edited it before instantiation. Appendix B.1 describes the deterministic rules and audit.

Four short-task families share a common binary update rule while varying the semantic relation and answer type. They track the current custodian, authoritative author, current value, and active assignment. These correspond to identity binding, source binding, memory updating, and task-state continuity in the released identifiers. Two passage-level extensions add person-to-unit-to-code lookup and signed-correction resolution. Each item contains a context, a question, one short correct answer, and typed alternatives. A base case fixes the template and entity structure. Three variants change the valid state. Each target variant is paired with a source record. Table 2 states the deterministic

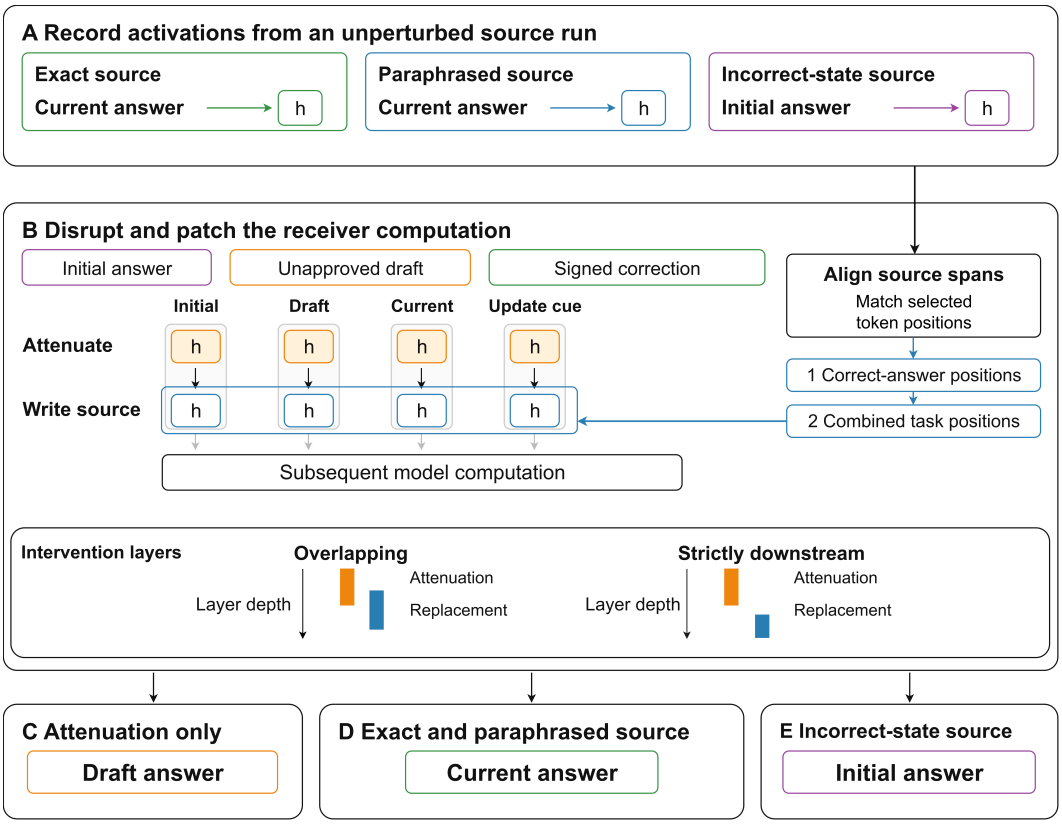

Figure 1: Source-state replacement in a passage memory-update example. The signed correction defines the current answer. Panel A records unperturbed source activations h. Panel B illustrates attenuation followed by aligned replacement at shared sites in an overlapping window; its inset compares the two window types. Model weights remain fixed. Panels C–E show the highest-scoring candidates from separate runs under the indicated conditions.

Table 1: Models with publicly available weights. Architecture values follow the experiment manifests.

Model	Layers	Hidden	Heads	Weight identifier
Qwen2.5 32B	64	5120	40	`Qwen/Qwen2.5-32B-Instruct`
Qwen3 14B	40	5120	40	`Qwen/Qwen3-14B`
Gemma3 12B IT	48	3840	16	`google/gemma-3-12b-it`
Gemma4 12B IT	48	3840	16	`google/gemma-4-12B-it`
Phi4 14B	40	5120	40	`microsoft/phi-4`

answer rules. All five models rank the correct answer first on the 120 discovery prompts, yielding 600 correct model–prompt pairs.

Separate manifests govern discovery, calibration, and evaluation. The later short-task evaluation contains 20 base cases per family and a separately worded source record for each of three variants. Each passage family contains 10 calibration base cases and 20 evaluation base cases. Exact and digit-normalized audits find no shared context–question text between the later evaluation sets and calibration. Three historical short-task families reuse earlier wording under new identifiers. The main text-disjoint restoration analysis combines 10 passage model–task groups from the common five-model protocol with 5 short-task transfer pairs contained in the earlier transfer protocol. The complete short-task attenuation analysis covers all 20 model–task pairs. Appendix B.4 reports the split audit.

Table 2: Deterministic rules used to derive the unique correct answer.

Family	Correct-answer rule	Type
Identity binding	Return the initial custodian after a revoked transfer and the named replacement after an activated transfer.	Person
Source binding	Return the original author after a cancelled amendment and the amendment author after approval.	Person
Memory update	Return the initial value after a voided update and the replacement value after activation.	Value
Task-state continuity	Return the opening assignment after a withdrawn order and the ordered assignment after confirmation.	Task
Passage source binding	Follow the queried person to one current unit, then follow that unit to one response code.	Code
Passage memory update	Return the signed authoritative correction. The initial and draft codes are superseded.	Code

3.3 TARGET SELECTION AND PERFORMANCE CALIBRATION

Each prompt is formatted with the model's chat template. Token groups are then defined over the formatted input. Correct-answer positions contain accepted answer strings in the context. Incorrect-candidate positions contain competing answer strings. Update-cue positions contain phrases indicating whether a revision is active. Question-overlap positions contain words shared with the question. Question positions cover the complete question. The combined task-related group is the union of the first four groups. These labels correspond to the implementation keys `gold_context`, `competitor_context`, `state_cue`, `query_overlap`, `question`, and `state_evidence`.

During discovery, each candidate site pairs a residual layer with one token group and is screened at a fixed attenuation strength of $\alpha = 0.70$. Candidates are ranked primarily by the decrease in correct-answer margin caused by attenuation. Complete-answer scores validate the ranking. The top 16 sites per family are fixed before calibration. A second screen evaluates the normalized self-attention input, value-projection output, attention output after the output projection, feed-forward gate, feed-forward up projection, feed-forward down projection, and normalized writeback.

For prompt i, let $h_{i,\ell,t} \in \mathbb{R}^{d_\ell}$ be the residual-stream vector at layer ℓ and token t. Let $S_{\text{int}} \subseteq \mathcal{L} \times \mathcal{R}$ contain the selected layer–group pairs. If $T_{i,r}$ is the token set assigned to group r, define

$$m_{i,\ell,t} = \mathbf{1}\left[t \in \bigcup_{r:(\ell,r)\in S_{\text{int}}} T_{i,r}\right], \qquad \widetilde{h}_{i,\ell,t} = (1 - \alpha m_{i,\ell,t}) h_{i,\ell,t}, \tag{1}$$

where $\alpha \in [0, 1]$. The union mask scales the complete residual vector at every selected layer and token.

Calibration selects the measured point closest to 50 percent accuracy inside the first downward crossing of that threshold. The first later point at or below 10 percent defines the low-accuracy setting. When no such point exists, the lowest observed post-transition point is retained. The search uses increments of 0.005 and refines the crossing to 0.001 when needed. Evaluation uses the fixed calibration value. The attenuation coefficient specifies computational intervention strength. The two selected settings represent behaviorally defined levels of performance loss within each model and task.

For an answer sequence $y = (y_1, \ldots, y_{|y|})$, the final experiments use the mean conditional log probability across answer tokens

$$S(y \mid x, c) = \frac{1}{|y|} \sum_{j=1}^{|y|} \log p_\theta^{(c)}(y_j \mid x, y_{<j}). \tag{2}$$

After removing accepted aliases from the typed competitor set $C(x)$, condition c is evaluated by the answer margin and candidate-ranking accuracy

$$M_i^{(c)} = S(y_i^+ \mid x_i, c) - \max_{y^- \in C(x_i)} S(y^- \mid x_i, c), \qquad A_c = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}[M_i^{(c)} \geq 0]. \tag{3}$$

Answer-margin reduction is $M_i^{(\text{clean})} - M_i^{(\text{attenuation})}$. Replacement effects are measured from the run in which attenuation remains active. Margins are measured in nats per answer token.

3.4 Position controls and component interventions

The matched nearby-layer control preserves token positions and residual-site count while shifting attenuation to neighboring nontarget layers. Calibration evaluates eight offset–seed configurations with strength multipliers in $\{0.50, 0.75, 1.00, 1.25, 1.50\}$. Candidate strengths are clipped to the interval $[0.01, 0.95]$. Selection minimizes the predefined matching score J among candidates satisfying its fixed tolerances. All 20 selected controls use multiplier 1.00. The layer configuration is selected to match the target deficit, so this comparison assesses nearby-layer interventions after calibration for comparable performance loss. Five alternative token-position controls apply the same intervention at the same layer and strength to early context, incorrect-candidate positions, noncritical context, task-irrelevant context, or context outside protected task groups. Each control requests the target token count and uses all eligible positions when the requested count cannot be reached. Token counts match in all 1,690 rows for the early-context, outside-protected-group, and noncritical-context controls. Matching rates are 47.7 percent for incorrect-candidate positions and 89.2 percent for task-irrelevant context, with fewer control positions in the remaining rows. These contrasts combine location and count differences. Appendix C.2 specifies nearby-layer matching, and Appendix C.3 reports realized position-control counts. Component interventions use sites selected before evaluation.

3.5 Source-controlled performance restoration

Performance restoration begins from a calibrated attenuation setting. Calibration selects a single-layer, contiguous, or bridge replacement window from the candidates defined in Appendix E.2. At the fixed layers, aligned activations from a source run replace receiver activations at correct-answer positions and combined task-related positions. Within each base case, the entity and answer-label inventory remains fixed. A exact source preserves the current answer and original wording. A paraphrased source preserves the current answer while changing the surrounding wording. A incorrect-state source makes another answer current. The question-position control changes the replacement location. Five norm-matched random controls change the vector content. Performance restoration is measured as an improvement in candidate-ranking accuracy or answer margin relative to the attenuation-only condition.

Let L_p denote the selected replacement layers and $T_{i,\text{state}}$ the combined task-related positions in receiver i. Define $p_{i,\ell,t} = \mathbf{1}\{\ell \in L_p,\ t \in T_{i,\text{state}}\}$. For source condition c, let $a_{i,\ell,t}^{(c)}$ denote the source vector aligned to receiver position t, and let $h_{i,\ell,t}^{(c,\text{pre})}$ denote the receiver activation immediately before replacement, after any attenuation at that layer. The effective source-state patch is

$$h_{i,\ell,t}^{(c)} = \big(1 - p_{i,\ell,t}\big) h_{i,\ell,t}^{(c,\text{pre})} + p_{i,\ell,t} a_{i,\ell,t}^{(c)}. \tag{4}$$

Alignment uses evenly spaced source indices rounded to the nearest integer. Equal-length sequences use direct correspondence. When the receiver group contains one token, alignment uses the mean source vector. Appendix E.2 specifies the mapping. For the combined task-related group, alignment operates on the ordered union of selected token positions. Resampling this sequence can associate positions with different semantic roles when source and receiver spans have different lengths. The reported paraphrased-source results measure recovery under this union-index alignment. The implementation writes correct-answer positions first and combined task-related positions second. Since the combined group contains all correct-answer positions, the second assignment determines the final vectors throughout the replacement mask. Equation 4 expresses this effective operation. Attenuation executes before replacement at each shared layer. An overlapping window replaces attenuated vectors at selected positions. A strictly downstream window begins after the final attenuated layer.



Figure 2: Reasoning performance under graded activation attenuation. Each panel traces the four short-task families over 21 measured strengths. Markers show observed values and lines connect adjacent strengths without smoothing.

Confirmatory inference clusters the three prompt variants from each base case. It uses 10,000 cluster-bootstrap replicates and two-sided sign-flip tests. Random seeds are averaged within each prompt before clustering. Benjamini–Hochberg false-discovery-rate adjustment (BH-FDR) is applied within each fixed analysis family. Exact-text clustering, paired source-and-receiver correctness, alternative answer reducers, and free-generation checks form separate sensitivity analyses.

3.6 Activation-distance analyses

Two sets of activation measurements address different questions. The first measures the standardized distance between the mean activation at incorrect-candidate positions and the mean activation over the combined task-related group after attenuation. The second set measures the final-layer residual vector at the last prompt token after the intervened forward pass. A fixed Gaussian map projects this vector into 128 dimensions. The clean-reference nearest-neighbor ratio uses coordinate scales fitted on unperturbed calibration projections. Paired cosine distance compares the unstandardized projections for the intervened and unperturbed runs of the same prompt. Appendix E.4 defines both diagnostics. These downstream measurements are analyzed separately from receiver-answer accuracy and margin.

4 Results

4.1 RQ1. Attenuating task-related positions reduces reasoning performance

Figure 2 shows 21 attenuation strengths for each of the 20 model–task pairs on text-disjoint evaluation prompts. Equal-weight mean accuracy decreases from 95.75 percent without intervention to 51.00 percent at an attenuation strength of 0.20 and 15.50 percent at a strength of 1. Eighteen curves contain at least one upward step between adjacent strengths. We therefore use the measured curves to describe performance loss and select the transition setting from the first downward crossing.

Attenuation at task-related positions causes greater answer-margin reduction than the five alternative token-position controls in 167 of 200 model–task–strength comparisons. The aggregate difference is positive for every control and ranges from 0.168 to 0.208. Table 3 reports the comparisons. Appendix C.2 reports the matching quality of the nearby-layer control. All three fixed tolerances hold in 19 of 20 evaluation pairs, and residual-site and token counts match in all 20 pairs. The

Table 3: Answer-margin reduction at task-related positions relative to alternative position controls. Positive counts report model–task–strength comparisons.

Position control	Mean difference	95% interval	Positive comparisons
Early context	0.208	0.139 to 0.281	34/40
Outside protected groups	0.180	0.124 to 0.241	35/40
Incorrect-candidate positions	0.173	0.105 to 0.247	31/40
Task-irrelevant context	0.170	0.112 to 0.231	34/40
Noncritical context	0.168	0.109 to 0.234	33/40

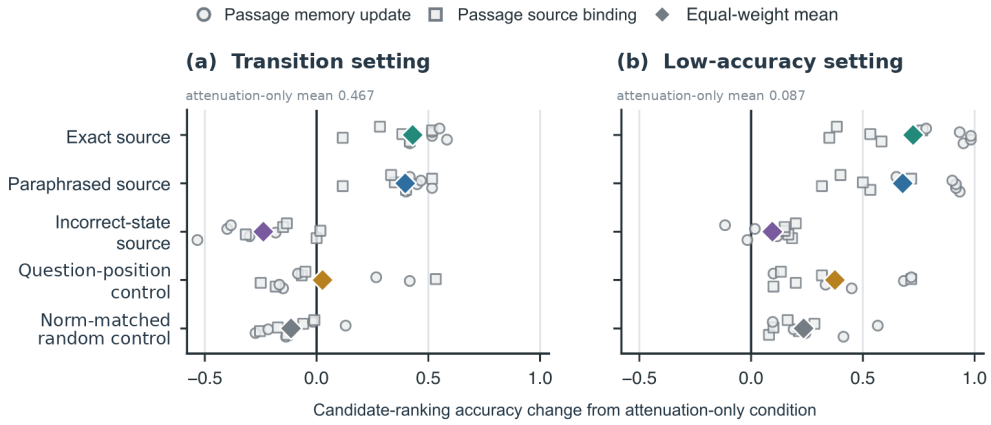

Figure 3: Change in candidate-ranking accuracy relative to the attenuation-only condition. Gray markers show model–task groups. Diamonds show equal-weight means. The vertical line marks no change.

target-minus-control contrast for Qwen2.5 task-state continuity is -1.547 nats per answer token, with a 95 percent interval from -2.364 to -0.809 and a BH-FDR-adjusted p-value of 0.046. The target exceeds the matched control in 0 of 20 pairs after correction. Together with the position controls, these results distinguish sensitivity to task-related token locations from performance loss reproduced by calibrated nearby-layer interventions. Component-level support varies across models. The target exceeds all four controls in 16 of 40 evaluated units, with the strongest support in Gemma4.

4.2 RQ2. Source content determines the strength of performance restoration

Figure 3 summarizes changes in passage-task candidate-ranking accuracy relative to the attenuation-only condition. At the calibration-selected transition settings, accuracy in the passage evaluation rises from 46.7 percent under attenuation to 89.7 percent after replacement with exact-source activations. Paraphrased-source activations reach 86.3 percent. Incorrect-state and random-vector controls reach 22.8 percent and 35.2 percent. Under incorrect-state replacement, selects the source-specified answer in 62.0 percent of 600 passage-task predictions across 10 model–task groups. The receiver's correct answer ranks first in 22.8 percent of predictions. The remaining 15.2 percent select the third candidate.

The text-disjoint analysis combines 10 passage groups with five short-task transfer pairs from the earlier protocol. Both correct-source conditions exceed incorrect-state and random-vector controls on accuracy and answer margin in all 15 groups after BH-FDR correction. Gains over attenuation alone are supported for margin in all 15 groups and accuracy in 14 groups. Four short-task transfer pairs show severe performance loss at their fixed calibration strengths.Appendix Table 12 gives the complete contrast counts, and Appendix C.1 reports the realized accuracies. Both correct-source conditions exceed incorrect-state sources and random controls on accuracy and answer margin in every group after correction within the 11 overlapping-window groups and the 4 strictly downstream Gemma3 groups.

At the low-accuracy setting, mean passage accuracy rises from 8.7 percent to 81.2 percent with exact sources and 76.5 percent with paraphrased sources. Incorrect-state sources yield 18.2 percent

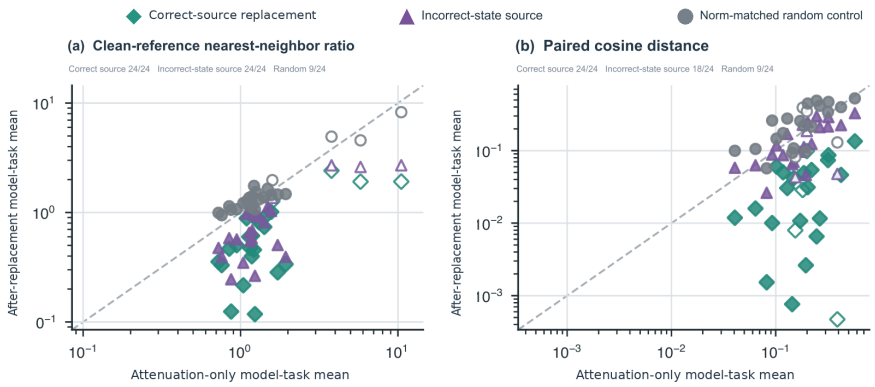

Figure 4: Activation distances in the historical 24-group analysis. The left panel shows the clean-reference nearest-neighbor ratio, and the right shows paired cosine distance. Axes are logarithmic; diagonals indicate unchanged distance after replacement. Fifteen groups have calibration–evaluation text overlap.

accuracy. Norm-matched random replacement yields 32.2 percent, close to the 33.3 percent uniform-choice reference for the three-candidate passage tasks. Both correct-source conditions improve accuracy and margin in all 10 passage groups after correction. Question-position replacement reaches 49.3 percent passage accuracy at transition and 46.2 percent at the low-accuracy setting. In Gemma3 passage source binding, its accuracy is 93.3 percent, compared with 78.3 percent for exact-source replacement at task-related positions. Question-position replacement therefore provides an alternative intervention location with model-dependent efficacy.

4.3 RQ3. Activation distance and receiver-answer recovery

In the historical 24-group analysis shown in Figure 4, incorrect-state replacement reduces the nearest-neighbor ratio in all 24 groups after BH-FDR correction. Paired cosine distance decreases in 18 groups, with 15 decreases supported after correction. Correct-source replacement yields higher receiver-answer performance, while both source types can reduce distance. The distance reductions occur under source conditions with different receiver-answer outcomes. Source-content comparisons therefore distinguish movement toward unperturbed reference activations from recovery of the receiver's correct answer. Three historical short-task families have complete calibration–evaluation text overlap, covering 15 groups in this analysis. Their repeated prompts contribute to the low unperturbed nearest-neighbor ratios. The separate passage summary gives ratios of 5.415 under attenuation, 1.878 after correct-source replacement, and 2.343 after incorrect-state replacement. These passage results provide a text-disjoint descriptive comparison within the historical protocol. The 15-group behavioral evaluation in Section 4.2 uses a separate protocol.

5 Discussion and Conclusion

In this study, we use graded activation attenuation and source-controlled replacement to examine changes in reasoning performance and recovery after internal computation is disrupted across five language models. We find that incorrect information can also move perturbed activations closer to their unperturbed states while yielding substantially weaker recovery of the receiver's correct answer than correct information, indicating that proximity to normal activation states does not directly correspond to reasoning recovery. This pattern occurs across all five models, while component analyses reveal model-dependent intervention effects at attention and feed-forward sites. The study is limited by its primary reliance on candidate-answer ranking within fixed entity and answer-label inventories and by calibration–evaluation text overlap in part of the historical distance analysis. Future work will extend this experimental framework to large speech models to compare how correct and incorrect information affect reasoning recovery in speech and text models, and whether the relationship between activation changes and reasoning recovery is consistent across the two.

REPRODUCIBILITY STATEMENT

The appendices expand on the experimental rationale, methodological details, and analyses summarized in the main text. Appendix A explains the evidence motivating the experimental design. Appendix B documents task construction and quality audits, together with token-group definitions, answer scoring, and target selection. Appendix C specifies performance calibration and control procedures. Appendix D presents complete performance and component results, including token-group distance analyses. Appendix E details the source-controlled restoration experiments and their relationship to activation distances, with additional sensitivity analyses and qualitative examples. The experiment package for regenerating the tables and figures will be released upon publication.

ETHICS STATEMENT

This study examines publicly available language models through synthetic and nonclinical language tasks.

AI USE STATEMENT

ChatGPT was used to generate the initial synthetic task content and to assist with language polish editing. The authors reviewed and revised all generated material.

REFERENCES

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, et al. Phi-4 Technical Report. Technical Report MSR-TR-2024-57, Microsoft Research, 2024. URL https://www.microsoft.com/en-us/research/publication/phi-4-technical-report/.

Yash Kumar Atri, Steven L. Johnson, and Thomas Hartvigsen. Evaluating Temporal Consistency in Multi-Turn Language Models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 45964–45982. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.2133. URL https://aclanthology.org/2026.acl-long.2133/.

Alberto Benussi, Valentina Cantoni, Mario Grassi, Lucie Brechet, Christoph M. Michel, Abhishek Datta, Chris Thomas, Stefano Gazzina, Maria Sofia Cotelli, Marta Bianchi, Enrico Premi, Yasmine Gadola, Maria Cotelli, Marta Pengo, Federica Perrone, Maria Scolaro, Silvana Archetti, Eino Solje, Alessandro Padovani, Alvaro Pascual-Leone, and Barbara Borroni. Increasing Brain Gamma Activity Improves Episodic Memory and Restores Cholinergic Dysfunction in Alzheimer's Disease. *Annals of Neurology*, 92(2):322–334, 2022. doi: 10.1002/ana.26411.

Lei Chen, Joan Bruna, and Alberto Bietti. Distributional Associations vs In-Context Reasoning: A Study of Feed-forward and Attention Layers. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/533a89d7c2980218d82dbe3ea85f77ae-Abstract-Conference.html.

Qin Dai, Benjamin Heinzerling, and Kentaro Inui. Representational Analysis of Binding in Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17468–17493. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.967. URL https://aclanthology.org/2024.emnlp-main.967/.

Qin Dai, Benjamin Heinzerling, and Kentaro Inui. Cell-Based Representation of Relational Binding in Language Models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 47464–47524. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.2194. URL https://aclanthology.org/2026.acl-long.2194/.

Jiahai Feng and Jacob Steinhardt. How Do Language Models Bind Entities in Context? In *International Conference on Learning Representations*, 2024. URL https://proceedings.

iclr.cc/paper_files/paper/2024/hash/9d1b7fc578c0d2d6431fc26d736ecaf3-Abstract-Conference.html.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, et al. Gemma 3 Technical Report, 2025. URL https://arxiv.org/abs/2503.19786. arXiv preprint.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, et al. Gemma 4 Technical Report, 2026. URL https://arxiv.org/abs/2607.02770. arXiv preprint.

Satchel Grant, Simon Jerome Han, Alexa R. Tartaglini, and Christopher Potts. Addressing Divergent Representations from Causal Interventions on Neural Networks. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/133e588e1429f9f1e25b215da145580e-Abstract-Conference.html.

Yoav Gur-Arieh, Mor Geva, and Atticus Geiger. Mixing Mechanisms: How Language Models Retrieve Bound Entities In-Context. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/2eeff35664016c7f0f8aa704f0d9a83e-Abstract-Conference.html.

Rui He, Ercong Nie, Hong Jiang, Iris E. Sommer, Philipp Homan, and Wolfram Hinzen. Activation-Guided Neuron Intervention to Induce Alzheimer's-Related Computational Language Phenotypes in a Large Language Model, 2026. URL https://arxiv.org/abs/2608.03067. arXiv preprint.

Stefan Heimersheim and Neel Nanda. How to Use and Interpret Activation Patching, 2024. URL https://arxiv.org/abs/2404.15255. arXiv preprint.

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. URL https://aclanthology.org/2023.findings-emnlp.624/.

Jiyeon Kim, Hyunji Lee, Dylan Zhou, Sue Hyun Park, Seunghyun Yoon, Trung Bui, Franck Dernoncourt, Sungmin Cha, and Minjoon Seo. Can Large Language Models Keep Up? Benchmarking Online Adaptation to Continual Knowledge Streams. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 42240–42272. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1956. URL https://aclanthology.org/2026.acl-long.1956/.

Daniil Laptev, Nikita Balagansky, Yaroslav Aksenov, and Daniil Gavrilov. Analyze Feature Flow to Enhance Interpretation and Steering in Language Models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 32593–32616. PMLR, 2025. URL https://proceedings.mlr.press/v267/laptev25a.html.

Zhaoyi Li, Gangwei Jiang, Hong Xie, Linqi Song, Defu Lian, and Ying Wei. Understanding and Patching Compositional Reasoning in LLMs. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 9668–9688. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.576. URL https://aclanthology.org/2024.findings-acl.576/.

Zhiheng Lyu, Kevin Yang, Lingpeng Kong, and Dan Klein. FactTrack: Time-Aware World State Tracking in Story Outlines. 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. 2825–2848. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.naacl-long.144. URL https://aclanthology.org/2025.naacl-long.144/.

Aleksandar Makelov, Georg Lange, Atticus Geiger, and Neel Nanda. Is This the Subspace You Are Looking for? An Interpretability Illusion for Subspace Activation Patching. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/70b8505ac79e3e131756f793cd80eb8d-Paper-Conference.pdf.

Hansruedi Mathys, Carles A. Boix, Leyla Anne Akay, Ziting Xia, Jose Davila-Velderrain, Ayesha P. Ng, Xueqiao Jiang, et al. Single-Cell Multiregion Dissection of Alzheimer's Disease. *Nature*, 632 (8026):858–868, 2024. doi: 10.1038/s41586-024-07606-7. URL https://www.nature.com/articles/s41586-024-07606-7.

Maxime Méloux, François Portet, Silviu Maniu, and Maxime Peyrard. Everything, Everywhere, All at Once: Is Mechanistic Interpretability Identifiable? In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/51157713f1f1f572a05647e1aef1b1ed-Abstract-Conference.html.

Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. The Geometry of Categorical and Hierarchical Concepts in Large Language Models. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/be7430d22a4dae8516894e32f2fcc6db-Abstract-Conference.html.

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. URL https://aclanthology.org/2024.acl-long.828/.

Adam S. Shai, Sarah E. Marzen, Lucas Teixeira, Alexander Gietelink Oldenziel, and Paul M. Riechers. Transformers Represent Belief State Geometry in Their Residual Stream. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/8936fa1691764912d9519e1b5673ea66-Abstract-Conference.html.

Aryan Sharma, Cutter Dawes, and Shivam Raval. Dissociating Decodability and Causal Use in Bracket-Sequence Transformers, 2026. URL https://arxiv.org/abs/2604.22128. arXiv preprint.

Eric Todd, Millicent Li, Arnab Sen Sharma, Aaron Mueller, Byron Wallace, and David Bau. Function Vectors in Large Language Models. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/4ae163cb8788970e53b4fd9578141139-Abstract-Conference.html.

Yifan Wang, Jingyuan Sun, Jichen Zheng, Yunhao Zhang, Chunyu Ye, Jixing Li, Chengqing Zong, and Shaonan Wang. Bridging Brains and Models: MoE-Based Functional Lesions for Simulating and Rehabilitating Aphasia, 2025. URL https://arxiv.org/abs/2508.04749. arXiv preprint.

Zhengxuan Wu, Atticus Geiger, Thomas Icard, Christopher Potts, and Noah D. Goodman. Interpretability at scale: Identifying causal mechanisms in Alpaca. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/f6a8b109d4d4fd64c75e94aaf85d9697-Abstract.html.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, et al. Qwen2.5 Technical Report, 2024. URL https://arxiv.org/abs/2412.15115. arXiv preprint.

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

A Evidence motivating the experimental design

The connection to Alzheimer's disease concerns selective vulnerability and responsiveness to intervention. These observations motivate experimental questions about the location of disruption and

recovery across levels of performance loss. The correspondence is a functional analogy. Our attenuation and replacement operators are defined as computational interventions during language-model inference.

Vulnerability can differ across internal locations. A single-cell study of six human brain regions identified neuronal populations depleted in specific regions in Alzheimer's disease (Mathys et al., 2024). This observation motivates comparing the effects of disruption across model locations. Our experiments implement this comparison through task-related token positions and alternative-position controls. Nearby-layer controls assess whether calibrated interventions at neighboring layers produce comparable performance loss. The model locations are defined by the computational task and intervention protocol.

Intervention responses can depend on baseline impairment. A randomized crossover study of 60 patients with Alzheimer's disease reported improved verbal recall after a single session of gamma-frequency alternating-current stimulation targeting the precuneus (Benussi et al., 2022). Its exploratory predictor analysis associated milder baseline impairment with larger immediate and delayed recall gains. This observation motivates measuring performance restoration at different levels of model impairment. We select transition and low-accuracy settings from measured performance curves. The attenuation coefficient specifies computational intervention strength, and the selected settings are defined by task performance. The clinical evidence concerns short-term recall-test improvement following stimulation.

The replacement procedure follows the computational method of activation patching. This method overwrites selected activations with states recorded from another run and measures the resulting output change (Heimersheim & Nanda, 2024). Our source runs supply task information through recorded activations. Exact sources retain the current answer and wording. Paraphrased sources preserve the current answer while changing the wording, enabling a test of recovery across surface changes. Incorrect-state sources support a different answer and test the effect of changing the supplied state. Random-vector controls assess the contribution of structured source content. Question-position controls assess an alternative replacement location.

These source comparisons address the information used by perturbed model computation. Recovery is measured relative to attenuation alone through candidate-ranking accuracy and answer margin. Activation-distance measurements characterize a separate outcome. Together, these measurements test content-dependent recovery in the evaluated language models. The biological findings motivate the functional questions, while activation patching supplies the replacement method.

B Experimental Details

B.1 Task construction and reproducible specification

ChatGPT generated the initial synthetic task content. The authors edited the sentence-level language before instantiation. Recorded provenance begins with the edited templates because the original generation prompt and model version were not recorded. Deterministic Python functions instantiate each record from fixed vocabularies and assignment rules. The same rules derive the valid answer and typed alternatives. Every fixed label follows this deterministic process.

The four short-task families isolate related state operations. Identity binding tracks custody after a transfer. Source binding tracks authorship after an amendment. Memory updating identifies the operative value after a revision. Task-state continuity identifies the active assignment after a later order. Two passage families add longer contexts. The multi-document task composes person-to-unit and unit-to-code records. The conflict-update task resolves an initial value through an unapproved draft and a later signed correction. Table 2 records each decision rule.

The first four families share a binary override rule. Let z_0 be the initial state, z_1 the proposed replacement, and $a \in \{0, 1\}$ the operative-status flag. The answer is

$$f(z_0, z_1, a) = \begin{cases} z_1, & a = 1, \\ z_0, & a = 0. \end{cases} \tag{5}$$

Activation, approval, and confirmation set $a = 1$. Revocation, cancellation, and withdrawal set $a = 0$. For the multi-document task, the answer is $f_{\mathrm{doc}}(p) = C(U(p))$, where U maps a person to

the current unit and C maps that unit to its response code. For the conflict-update task, the signed correction is the final authority, so $f_{\text{upd}} = v_{\text{signed}}$.

Table 4 summarizes the fixed files. The discovery file contains 10 base cases and three state variants for each short-task family. A base case fixes the template and entity structure. Its variants change the valid state. This yields 120 prompts shared by all five models. Component localization uses 15 unperturbed-correct discovery prompts from each family, or 60 prompts in total. The union of historical calibration inputs contains 90 distinct prompts before model-specific correctness filtering. The text-disjoint short-task evaluation contains 20 base cases with three target variants per family. Every target has a separately worded source record with the same valid state. The passage calibration contains 10 base cases per family and the passage evaluation contains 20.

Table 4: Inventory of distinct experimental prompts. Historical calibration reports the union before model-specific correctness filtering.

Stage	Families	Bases per family	Variants	Targets	Sources
Discovery	4	10	3	120	0
Historical calibration union	4	8, 8, 8, 6	3	90	0
Text-disjoint short-task evaluation	4	20	3	240	240
Passage calibration	2	10	3	60	60
Passage evaluation	2	20	3	120	120

Within each later calibration and evaluation base case, `original`, `cf1`, and `cf2` encode three different valid answers. Twenty discovery groups in identity and source binding use two distinct answers because the second altered context returns to the original answer. These variants supply incorrect-state sources. They may change the proposed state and activation status. Record order and filler selection can also differ.

A exact source preserves the valid state and the corresponding original wording. A paraphrased source preserves the entity labels, answer labels, and deterministic task result while changing the surrounding wording. The deterministic audit verifies the rule outcome separately on every target and source. A source prompt can preserve the valid answer even when the model does not predict that answer. Source-run accuracy is therefore reported separately and paired-correct analyses retain cases in which both source and receiver runs answer correctly.

The executed candidate filter first collects answers from variants of the same base case in their stored order, then appends the row's distractor list. Normalization applies case folding, replaces each run of characters outside lowercase ASCII letters and digits with a space, and trims surrounding whitespace. Deduplication retains the first occurrence of each normalized string. The filter removes all normalized accepted-answer aliases and applies the declared answer-type predicate. It retains the first three type-compatible alternatives, using the filtered candidate list when the type-compatible list is empty. Thus, truncation follows stored appearance order. The correct string occurs in 67 of the 120 raw discovery distractor lists, and the runtime scoring sets exclude accepted correct-answer aliases in every audited prompt.

The task artifact supplies fixed JSONL files and construction scripts. It also contains the token-group and intervention implementation. The audit checks 1,050 distinct executed prompts after deduplicating model-specific calibration copies. A separate generated calibration set supports reproduction of the short-task generator and lies outside these executed inputs.

B.1.1 Complete short-task examples

All four examples use the system instruction "Use only the supplied record. Return the shortest answer with no explanation." The user wrapper places the context after *Records*, the question after *Query*, and ends with *Response*.

Identity binding.

> The crimson satchel 780 begins under the registered custody of Adrian Darrow. Beatrix Kendall is mentioned only in a nonbinding note. The authorized transfer identifies Callum

> Radford as the replacement custodian. The transfer is activated, replacing the earlier custodian.
>
> Who is the current custodian of the crimson satchel 780?

The correct answer is Callum Radford. The alternatives are Adrian Darrow, Beatrix Kendall, and Delia Ashford.

Source binding.

> For report-1180, Adrian Darrow issues the original notice with code aurora. Beatrix Kendall discusses basilisk without filing a change. A formal amendment is authored by Callum Radford and specifies citadel. The amendment receives approval and becomes the authoritative notice.
>
> Who authored the currently authoritative notice for report-1180?

The correct answer is Callum Radford. The alternatives are Adrian Darrow, Beatrix Kendall, and Delia Ashford.

Memory update.

> The active value in register-980 is initially aurora. A training annotation mentions basilisk but changes nothing. A signed update sets the value to citadel. That update is activated and overwrites the initial value.
>
> What is the current value in register-980?

The correct answer is citadel. The alternatives are aurora, basilisk, and drizzle.

Task-state continuity.

> Team unit-580 starts with the active task to catalog permits batch 700. A planning example about inspect ledgers batch 701 is not a switch. The coordinator later orders the team to route receipts batch 702. The order is confirmed and becomes the active task.
>
> What task is currently active for team unit-580?

The correct answer is route receipts batch 702. The alternatives are catalog permits batch 700, inspect ledgers batch 701, and verify dispatches batch 703.

B.1.2 Complete passage examples

Multi-document source binding. The system instruction is "Answer with only the requested short code. Use the records in the supplied dossier."

> Document 1. A regional bulletin summarizes weather disruption and an earlier maintenance window.
>
> Document 5. A planning note mentions several provisional options that were never formally adopted.
>
> Document 6. A regional bulletin summarizes weather disruption and an earlier maintenance window.
>
> Document 3. The compliance appendix discusses formatting rules, retention dates, and unrelated travel approvals.
>
> Personnel record. Julian Reed currently serves with the Meridian Desk.
>
> Operations register. The Meridian Desk is assigned response code Juniper-42.
>
> Personnel record. Owen Price currently serves with the Orchard Response Team.
>
> Document 7. The quarterly logistics note describes delayed equipment checks and a revised archival schedule.
>
> Document 2. A background memorandum records routine staffing changes that do not alter operational ownership.
>
> Personnel record. Leila Hart currently serves with the Northbridge Unit.
>
> Document 4. The finance attachment lists historical totals but contains no current assignment decision.

> Document 8. A planning note mentions several provisional options that were never formally adopted.
> Operations register. The Orchard Response Team is assigned response code Lumen-58.
> Operations register. The Northbridge Unit is assigned response code Willow-64.
> Using the personnel record and the operations register together, what response code belongs to Julian Reed?

The correct answer is Juniper-42. The alternatives are Willow-64 and Lumen-58.

Long-context conflict update. The system instruction is "Return only the current authoritative short code. Later signed corrections override drafts and initial records."

> Document 5. A regional bulletin summarizes weather disruption and an earlier maintenance window.
> Document 1. The quarterly logistics note describes delayed equipment checks and a revised archival schedule.
> Document 8. A regional bulletin summarizes weather disruption and an earlier maintenance window.
> Document 6. The finance attachment lists historical totals but contains no current assignment decision.
> Archive reminder. Historical entries remain visible for audit purposes even after a signed correction.
> Document 7. The compliance appendix discusses formatting rules, retention dates, and unrelated travel approvals.
> Document 3. A background memorandum records routine staffing changes that do not alter operational ownership.
> Document 4. The compliance appendix discusses formatting rules, retention dates, and unrelated travel approvals.
> Document 10. A planning note mentions several provisional options that were never formally adopted.
> Document 2. A background memorandum records routine staffing changes that do not alter operational ownership.
> Initial intake note. Case-4107 was provisionally recorded under Willow-64.
> Document 9. A planning note mentions several provisional options that were never formally adopted.
> Follow-up discussion. A draft worksheet proposed changing Case-4107 to Lumen-58, but the proposal was not approved.
> Signed correction issued later by the records officer. Replace every earlier code for Case-4107 with Cobalt-31. This signed correction is authoritative.
> What is the currently authoritative code for Case-4107?

The correct answer is Cobalt-31. The alternatives are Willow-64 and Lumen-58.

These tasks model current-state selection in longer conversations and document workflows. Reassignment and provenance tracking form one pair of operations. Superseding updates and persistent instructions form another. Controlled construction isolates the state governing each answer. The passage versions add multi-hop lookup among irrelevant records. The six families form a rule-governed mechanistic diagnostic. Naturally occurring requests fall outside this evaluation.

B.2 Task quality audit

Table 5 reports the fixed audit. For the 1,248 short-task and passage records covered by an executable grammar, the released audit reconstructs each correct answer from the parsed task state and compares it with the stored label. The 120 discovery records are checked against their earlier source manifest as a provenance test. Structural checks cover the schema, identifiers, and rendered messages. Semantic checks evaluate aliases and typed candidates. Pairing checks verify source records and state variants. Split checks operate on exact and digit-normalized text. Surface checks cover spacing, quotation marks, and sentence termination. These automated tests establish structural well-formedness and rule consistency. Human naturalness was not independently rated.

Table 5: Deterministic quality checks on the fixed task artifact. File-record counts include repeated model-specific calibration inputs.

Check	Result	Scope
Required schema and in-file prompt identifiers	1,368 passed	1,368 records
Distinct rendered prompts fixed by hash	1,050	1,050 prompts
Runtime target candidate sets valid	630 passed	630 targets
Surface-format lint flags	0	1,368 records
Rule-derived answer agrees with stored answer	1,248 passed	1,248 records
Target–reworded-source state and metadata agree	420 passed	420 pairs
Variant groups contain at least two answers	210 passed	210 groups
Variant groups contain three answers	190 passed	210 groups
Audited file comparisons with exact overlap	0 rows	22 comparisons
Audited file comparisons with digit-normalized overlap	0 rows	22 comparisons

The constructed passages reuse a small bank of irrelevant document sentences. All 180 target passages contain at least one repeated filler sentence after document numbers are removed. The governing person-to-unit-to-code path remains unchanged. The signed-update rule also remains unchanged. Linguistic diversity is narrower than in naturally occurring documents. The passage suite tests the selected state operations under controlled stress conditions.

Saved execution outputs contain 2,181 distinct model–suite–phase–prompt records. Each record has nonempty correct-answer and combined task-related position counts. Applying the supplied token-group code and five tokenizers to the rendered prompts reproduces the stored character spans and token indices. The human-review templates contain 400 unannotated rows. Reported quality results come from automated surface checks together with structural and rule-consistency checks.

The 22 file-level split comparisons consist of two shared passage-family comparisons and 20 short-task comparisons. Each of the five model-specific calibration files is compared with the later short-task evaluation file separately for all four families. The passage comparisons use shared task files and are counted once per family. Thus, the total is $2 + 5 \times 4 = 22$. These comparisons concern the audited files paired with the later evaluation sets. Table 6 separately reports the executed historical calibration–evaluation overlap.

Of the 210 variant groups, 190 contain three distinct answers. The remaining 20 discovery groups in identity and source binding contain two distinct answers by construction.

B.3 Token groups, scoring, and target selection

Each rendered prompt stores the context and question together with accepted answers and distractors. Correct-answer positions contain accepted aliases found in the context. Incorrect-candidate positions contain runtime competitor answers found there. Update-cue positions contain cue phrases located in the context and question. Question-overlap positions contain question content words at their occurrences in the context and question. Question positions cover the complete question span. The combined task-related group is the union of correct-answer, incorrect-candidate, update-cue, and question-overlap positions. Tokenizer offsets map every intersecting character span to token positions.

For an answer sequence $y = (y_1, \ldots, y_{|y|})$, the run manifest selects a mean-token or sum reducer

$$S_\rho(y \mid x, c) = \frac{1}{d_\rho(|y|)} \sum_{j=1}^{|y|} \log p_\theta^{(c)}(y_j \mid x, y_{<j}), \qquad d_\rho(n) = \begin{cases} n, & \rho = \text{mean}, \\ 1, & \rho = \text{sum}. \end{cases} \tag{6}$$

The final short-task and passage experiments use the mean reducer. Earlier sensitivity runs also report the sum reducer. Runtime filtering removes accepted correct-answer aliases from the typed competitor set $C(x)$. For condition c, the answer margin and candidate-ranking accuracy are

$$M_i^{(c)} = S_\rho(y_i^+ \mid x_i, c) - \max_{y^- \in C(x_i)} S_\rho(y^- \mid x_i, c), \qquad A_c = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}[M_i^{(c)} \geq 0]. \tag{7}$$

The correct answer appears first in the candidate list, so a numerical tie is scored as correct. Answer-margin reduction is $D_i^M = M_i^{(\text{clean})} - M_i^{(\text{attenuation})}$. A replacement effect is measured relative to the run in which attenuation remains active.

Residual candidates are selected on discovery prompts. The initial screen covers six prespecified token groups. It separates the combined task-related group from correct-answer and incorrect-candidate positions. Additional groups capture update cues, question overlap, and the full question. The screen evaluates residual sites and component sites. Individual attention heads are also evaluated where supported. Ranking uses correct-answer margin reduction as the first signal. Effect consistency and complete-answer validation provide the remaining signals.

The stored localization aggregates contain 5,712 model–family–group–layer–site candidate records for Qwen2.5. Qwen3 and Phi4 each contain 4,992. Gemma3 and Gemma4 each contain 2,976. The top 16 validated residual candidates from each family enter the distance and cue analyses. A fixed manifest records their layers before performance calibration. The permitted token group covers correct-answer or combined task-related positions. A shared cross-view site set serves as the fallback.

The discovery screen samples every second layer between approximately 15 percent and 95 percent of model depth. It covers 26 layers in Qwen2.5. Qwen3 and Phi4 each contribute 16 layers. Both Gemma models contribute 20 layers. A later fixed component screen tests value projections, attention outputs, feed-forward gates, feed-forward up projections, feed-forward down projections, and normalized writeback. Candidate identities and token groups remain fixed during evaluation.

Selection follows six steps.

1. Render discovery prompts and verify alignment between token-group text and tokens.
2. Evaluate each retained candidate at the fixed screening strength $\alpha = 0.70$ and measure its answer-margin reduction.
3. Aggregate results separately for every model–family pair and each group–layer–site combination. Retain candidates with consistent positive effects across variants.
4. Rank candidates by the stored causal score. Keep the top 16 per family and a shared set spanning multiple views.
5. Validate candidates through attenuation and by replacing attenuated activations with activations from unperturbed runs. Fix the selected residual sites for each token group.
6. Select transition and low-accuracy settings on calibration prompts, then apply the fixed target and strength to evaluation prompts.

Discovery fixes target identities. Calibration then selects attenuation strengths and replacement windows.

B.4 Prompt duplication audit

The audit normalizes rendered user text and compares exact calibration strings with evaluation strings within each model and family. Table 6 reports the overlap. Identity binding and both passage families use distinct texts. In the other three historical short-task families, calibration contains every unique evaluation string. Some strings occur under a different sample identifier. Identifier-based clustering places those duplicate texts in different bootstrap units. A sensitivity analysis instead merges equal normalized texts into one bootstrap unit.

The later short-task evaluation contains 20 base cases per family with three variants. Its prompts have no overlap with the historical calibration set used to select targets and replacement windows. This holds for exact text and after digit normalization. A further grouping analysis removes identifier distinctions within the compact synthetic grammar. The main restoration summary uses the 10 passage comparisons and the 5 short-task transfer pairs listed in Table 8 in Appendix C.1. The short-task portion is a partial transfer analysis, while the complete attenuation curves cover all 20 short-task model–task pairs. The final passage transfer experiment has no exact overlap between calibration and evaluation in any of the 10 model–task split comparisons. Digit normalization gives the same result.

Table 6: Exact text overlap between calibration and evaluation. Later evaluation texts are absent from the intervention-selection sets.

Family	Exact overlap	Model–task comparisons	Interpretation
Identity binding	0%	5	Text disjoint
Memory update	100%	5	Repeated text in historical analysis
Source binding	100%	5	Repeated text in historical analysis
Task-state continuity	100%	5	Repeated text in historical analysis
New short-task full curve	0%	20	Text disjoint from selection
Passage tasks	0%	10	Held out across five models

C Performance Calibration and Controls

C.1 Performance calibration

For prompt i, let $h_{i,\ell,t} \in \mathbb{R}^{d_\ell}$ be the residual vector at layer ℓ and token t. The fixed intervention set $S_{\text{int}} \subseteq \mathcal{L} \times \mathcal{R}$ contains layer–group pairs and $T_{i,r}$ contains positions assigned to group r. The joint mask and multiplicative attenuation are

$$m_{i,\ell,t} = \mathbf{1}\left[t \in \bigcup_{r:(\ell,r)\in S_{\text{int}}} T_{i,r}\right], \qquad \widetilde{h}_{i,\ell,t} = (1 - \alpha m_{i,\ell,t}) h_{i,\ell,t}, \tag{8}$$

where $\alpha \in [0, 1]$.

For the final common protocol, let $G_0 = \{0, 0.05, \ldots, 1\}$ be the coarse calibration grid. The first downward crossing is

$$j^\star = \min\{j : A_{\text{cal}}(\alpha_j) > 0.5,\ A_{\text{cal}}(\alpha_{j+1}) \le 0.5\}. \tag{9}$$

The search refines the interval $[\alpha_{j^\star}, \alpha_{j^\star+1}]$ at 0.005 spacing and at 0.001 spacing when needed. Let $G^\star$ contain all evaluated points in this crossing interval. The transition candidates and selected strength are

$$B_{\text{tr}} = \arg\min_{\alpha \in G^\star} |A_{\text{cal}}(\alpha) - 0.5|, \qquad \alpha_{\text{tr}} = \min B_{\text{tr}}. \tag{10}$$

The minimum selects the smaller strength when several points are equally close. A coarse curve with no crossing contributes its closest observed calibration point and carries a status label.

Let G_{eval} contain all strengths evaluated after transition refinement. The low-accuracy setting is the first later evaluated strength at or below 0.10, the low-accuracy setting is defined when the evaluated grid contains at least one strength greater than α_{tr}.

$$\alpha_{\text{low}} = \min\{\alpha \in G_{\text{eval}} : \alpha > \alpha_{\text{tr}},\ A_{\text{cal}}(\alpha) \le 0.10\}. \tag{11}$$

When this set is empty, the protocol selects the post-transition strength with the lowest observed calibration accuracy. Ties are resolved by the smaller strength.

Earlier distance and cue analyses use five stored candidates. They choose the candidate inside the declared 30 percent to 70 percent calibration range that is closest to 50 percent, independent of crossing order. The accuracy-recovery runner evaluates nine values from 0.160 to 0.200 in increments of 0.005. The passage runner evaluates 11 values from 0.200 to 0.250. Rounding the stored keys to approximately 0.01 retains five candidates per run.

Table 7 reports seven selected settings for earlier distance, cue, and sensitivity analyses. Table 8 reports the fixed transition settings for the passage and short-task text-disjoint evaluations.

The passage evaluation and short-task transfer evaluation differ in realized attenuation severity. All 10 passage groups retain evaluation accuracy between 38.3 and 58.3 percent at their calibration-selected strengths. The short-task transfer accuracies are

Table 7: Selected settings from the refined strength grid in earlier distance, cue, and sensitivity analyses.

Protocol	Model and task	α	Calibration accuracy	Evaluation accuracy	Text disjoint
Distance and cue	Gemma3 identity	0.230	0.375	0.300	Yes
Distance and cue	Gemma3 continuity	0.170	0.417	0.433	No
Distance and cue	Gemma4 continuity	0.190	0.444	0.444	No
Distance and cue	Qwen2.5 source	0.175	0.625	0.600	No
Distance and cue	Qwen2.5 continuity	0.135	0.500	0.467	No
Accuracy recovery	Gemma3 continuity	0.170	0.500	0.533	No
Passage sensitivity	Gemma4 memory	0.205	0.600	0.517	Yes

Table 8: Fixed calibration-selected attenuation strengths for passage evaluation and short-task transfer. Calibration accuracy and evaluation accuracy are candidate-ranking accuracies measured on their respective splits.

Model	Task	α	Calibration	Evaluation
Passage tasks				
Qwen3	Memory update	0.245	0.500	0.417
Qwen3	Source binding	0.200	0.533	0.533
Gemma3	Memory update	0.190	0.433	0.483
Gemma3	Source binding	0.170	0.533	0.400
Gemma4	Memory update	0.205	0.667	0.483
Gemma4	Source binding	0.195	0.433	0.483
Phi4	Memory update	0.310	0.533	0.450
Phi4	Source binding	0.240	0.533	0.450
Qwen2.5	Memory update	0.160	0.500	0.583
Qwen2.5	Source binding	0.165	0.433	0.383
Text-disjoint short-task transfer				
Gemma3	Source binding	0.215	0.500	0.033
Gemma3	Continuity	0.170	0.500	0.517
Gemma4	Continuity	0.200	0.500	0.033
Phi4	Source binding	0.360	0.533	0.150
Qwen2.5	Continuity	0.160	0.417	0.100

Model	Task	Evaluation accuracy
Gemma3	Source binding	3.3%
Gemma4	Task-state continuity	3.3%
Phi4	Source binding	15.0%
Qwen2.5	Task-state continuity	10.0%
Gemma3	Task-state continuity	51.7%

The five short-task pairs therefore measure transfer of a calibration-selected strength across template distributions. Their realized evaluation accuracies define the operating regime, and four pairs exhibit severe performance loss after transfer.

C.2 Original and matched controls

The original control grid uses five seeds and contains 200 summaries for 40 model–task–strength groups. Four controls support the comparison. One randomizes token positions. Another uses an alternative same-layer component selected by a fixed priority rule. The remaining two select the nearest available component by activation change or output KL divergence.

The target produces positive answer-margin reduction in 33 of 40 model–task–strength groups and exceeds every original control in 16. Against random-token intervention, its mean additional reduction is 0.182. The 95 percent interval spans 0.127 to 0.243 and differences are positive in 36 groups. Against the output-KL-nearest control, mean additional reduction is 0.051. The interval spans 0.018 to 0.085 and differences are positive in 22 groups. Intervals for the alternative same-layer component and activation-change-nearest control include zero.

The initial task-related-token control fell outside at least one matching tolerance in all 20 model–task pairs. The matched nearby-layer control preserves token positions and residual-site count while shifting attenuation to neighboring nontarget layers. Calibration evaluates eight offset–seed configurations with strength multipliers in $\{0.50, 0.75, 1.00, 1.25, 1.50\}$. Candidate strengths are clipped to the interval $[0.01, 0.95]$. Selection minimizes the predefined matching score J among candidates satisfying its fixed tolerances. All 20 selected controls use multiplier 1.00. The layer configuration is selected to match the target deficit, so this comparison assesses nearby-layer interventions after calibration for comparable performance loss.

For matching quantity k, let T_k and C_k be its target and candidate-control means on calibration data. The relative gap and selection score are

$$g_k = \frac{|T_k - C_k|}{\max(|T_k|, 10^{-5})}, \qquad J = 0.50 g_{\Delta M} + 0.30 g_{\Delta H} + 0.20 g_{K_{\text{out}}}. \tag{12}$$

Here $\Delta M_i = M_i^{(\text{clean})} - M_i^{(\text{attenuation})}$ is answer-margin reduction. The activation-change budget ΔH averages $\alpha \|h_{i,\ell,t}^{(\text{clean})}\|_2$ over affected sites and positions. Let T_x be the set of nonpadding prompt-token positions, with cardinality $|T_x|$. Output divergence averages unperturbed-to-intervention KL over this set

$$K_{\text{out}} = \frac{1}{|T_x|} \sum_{t \in T_x} \mathrm{KL}\Big(p_{\theta,t}^{(\text{clean})} \,\|\, p_{\theta,t}^{(\text{intervention})} \Big). \tag{13}$$

Candidates must satisfy $g_{\Delta M} \le 0.25$, $g_{\Delta H} \le 0.15$, and $g_{K_{\text{out}}} \le 0.25$. Calibration selects the passing candidate with the smallest J. A model–task pair with no passing candidate retains its closest candidate and records the failed tolerance. Selection is fixed before evaluation.

Table 9 reports matching. After BH-FDR correction, the Qwen2.5 continuity contrast is supported. The target-minus-control estimate is -1.547. The 95 percent interval spans -2.364 to -0.809 and the BH-FDR-adjusted p-value is 0.046. The control causes greater answer-margin reduction for this model–task pair. Remaining pairs give similar target and control estimates.

Table 9: Matching quality of the nearby-layer control. Counts show model–task pairs meeting each fixed criterion.

Criterion	Calibration	Evaluation
Answer-margin gap at most 25%	20/20	20/20
Activation-change gap at most 15%	20/20	20/20
Output-KL gap at most 25%	20/20	19/20
All three criteria	20/20	19/20
Exact residual-site and token-count match	20/20	20/20
Target exceeds control after BH-FDR	Not applicable	0/20

C.3 Alternative token-position controls

Each position control uses the same model, prompt, intervention type, and attenuation strength as the target. The procedure requests the target mask size. A smaller eligible pool contributes every available position and records the realized count. Exact count matching holds for all 1,690 early-context rows. It also holds for every outside-protected-group row and noncritical-context row. Matching holds for 806 of 1,690 incorrect-candidate rows and 1,508 of 1,690 task-irrelevant rows.

The noncritical-context profile uses a fixed seed. It excludes correct-answer positions, incorrect-candidate positions, question positions, update cues, and question overlap. The early-context profile takes the first eligible positions. The outside-protected-group profile excludes correct-answer and combined task-related positions, the question, and the answer prefix. The task-irrelevant profile draws from context words outside inferred task groups and incorrect-candidate positions. Recorded fallback pools supply positions when needed.

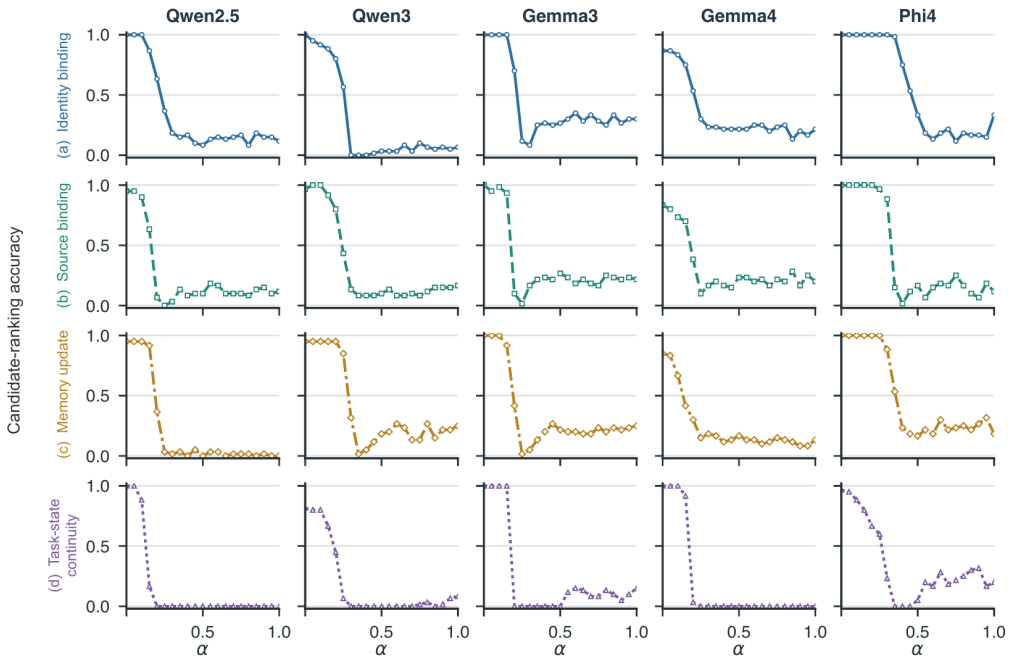

Figure 5: Complete performance curves for 20 short-task model–task pairs. Columns identify models. Rows show identity binding, source binding, memory update, and task-state continuity. Every marker is one measured strength from 0 to 1 in increments of 0.05.

D Complete Performance and Component Results

Figure 5 gives the complete response curve for every short-task model–task pair. In the original manifest analysis, accuracy falls from 99.83 percent without intervention to 56.36 percent at the selected transition settings and 2.71 percent at the low-accuracy settings. The corresponding answer margins are 18.718, 4.899, and -9.770. Each line connects 21 measured strengths without smoothing or prompt-count weighting.

D.1 Residual-stream replacement and component scale

Component-level support varies across models. The target exceeds all four controls in 16 of 40 evaluated units. Gemma4 provides the most consistent all-controls evidence, while the Qwen3 feed-forward-gate effect varies across the evaluated units. The counts below distinguish this criterion from the representative-site contrasts in Table 10. The representative-site contrast for Qwen2.5 is -0.172.

Model	Units exceeding all four controls
Gemma4	7/8
Qwen3	3/8
Gemma3	2/8
Phi4	0/8
Qwen2.5	4/8

Figure 6 summarizes residual-stream replacement and representative component effects. For Qwen2.5, the mean residual, complete-attention, and complete-feed-forward effects are 34.476, 0.058, and 0.180. Its correct-answer, combined task-related, and question-position effects are 49.391, 40.063, and 0.194. For Qwen3, the corresponding triplets are 30.838, 0.352, 0.469 and 48.186, 32.819, 0.434. Gemma3 yields 31.101, 0.362, 0.255 and 48.259, 32.236, 1.399. Gemma4 yields 22.381, 0.669, 3.509 and 32.926, 23.968, 0.311. Phi4 yields 13.967, 0.111, 0.320 and 20.807, 15.453, 0.085.

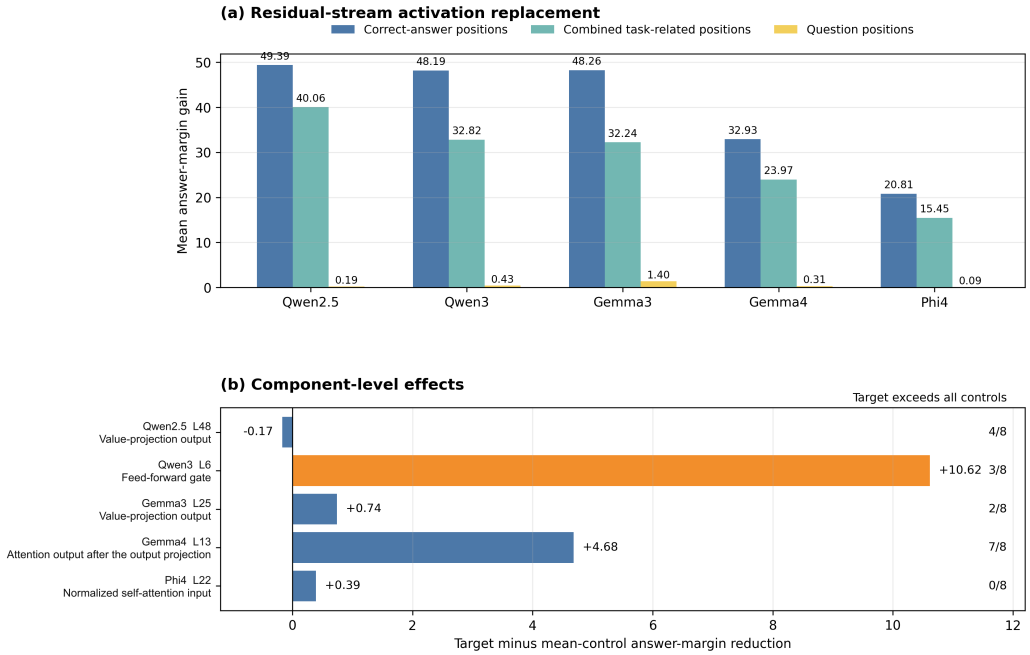

Figure 6: Residual-stream activation replacement and component-level effects. Panel a compares three replacement locations on one answer-margin scale. Panel b shows target-minus-mean-control answer-margin reduction at one fixed representative component site per model. The all-controls fraction counts task–component units in which the target exceeds every control.

Table 10: Strength of component-level evidence. The contrast compares the target with the mean of four controls. All-controls counts task–component units in which the target exceeds every control.

Model	Representative site	Target minus control	All controls	Summary
Qwen3 14B	Feed-forward gate	10.616	3/8	Task dependent
Gemma3 12B IT	Value projection	0.741	2/8	Two units pass
Gemma4 12B IT	Attention output	4.682	7/8	Seven units pass
Phi4 14B	Self-attention input	0.389	0/8	Zero units pass
Qwen2.5 32B	Value projection	-0.172	4/8	Four units pass

Correct-answer and combined task-related positions produce larger effects than question-position replacement across all five models. These values use the answer-margin scale declared in the run manifest. Residual interventions replace vectors. Complete-module interventions replace attention or feed-forward outputs. Raw effects are compared within each intervention class because intervention locations, tensor shapes, and replacement operations differ across classes.

For component candidate o, intervention arm b, and attenuation α, define answer-margin reduction and the target-minus-control contrast as

$$\begin{aligned} D_{o,i}^{(b)}(\alpha) &= M_i^{(\text{clean})} - M_i^{(\text{attenuated}(o,b,\alpha))}, \\ C_o(\alpha) &= \frac{1}{N}\sum_{i=1}^{N}\left[D_{o,i}^{(\text{target})}(\alpha) - \frac{1}{|K_o|}\sum_{k\in K_o} D_{o,i}^{(k)}(\alpha)\right]. \end{aligned} \tag{14}$$

A positive C_o means that target attenuation causes greater answer-margin reduction than the mean of the four controls. The self-attention input intervention acts on normalized activations entering attention. The value intervention acts on value-projection outputs. The attention-output intervention acts after the output projection and before residual addition. Numerical comparisons remain within each intervention class.

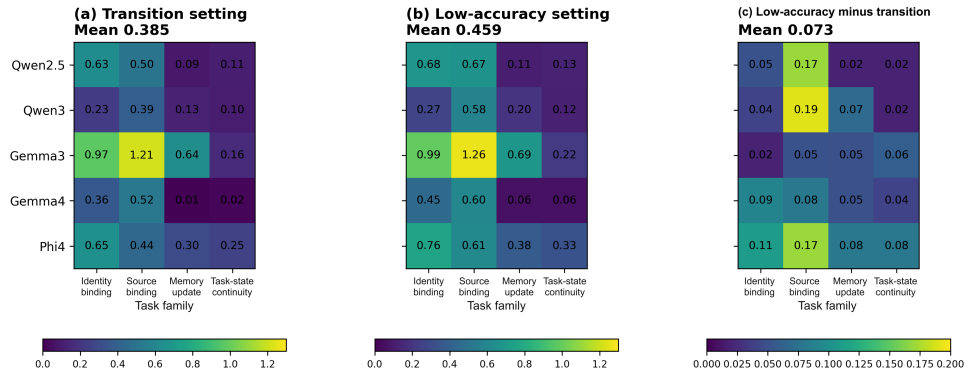

Figure 7: Change in standardized distance between incorrect-candidate and combined task-related activations. Larger values indicate a greater reduction in distance after attenuation. The final panel subtracts the transition decrease from the low-accuracy decrease.

Seven-point component sweeps cover attenuation values from 0.1 to 0.7. The Qwen3 feed-forward-gate contrast and Gemma4 attention-output contrast are largest at high attenuation. Their control contrasts and downstream propagation results follow the same strength pattern. Gemma3 value-projection effects remain below the Qwen3 and Gemma4 representative effects across the sweep. Phi4 and Qwen2.5 remain below the fixed representative-site criterion.

D.2 Distances between token-group means

The primary descriptive analysis averages aligned tokens within each prompt and group, then forms group centroids over prompts in one model–task–condition group. For condition q, define

$$v_{i,r,\ell}^{(q)} = \frac{1}{|T_{i,r}|} \sum_{t \in T_{i,r}} h_{i,\ell,t}^{(q)}, \qquad c_{r,\ell}^{(q)} = \frac{1}{N_r} \sum_{i=1}^{N_r} v_{i,r,\ell}^{(q)}. \tag{15}$$

The six-group set $\mathcal{R}$ supplies a shared center. Each centered centroid is normalized as

$$\bar{c}_{\ell}^{(q)} = \frac{1}{|\mathcal{R}|} \sum_{r \in \mathcal{R}} c_{r,\ell}^{(q)}, \qquad \mu_{r,\ell}^{(q)} = \frac{c_{r,\ell}^{(q)} - \bar{c}_{\ell}^{(q)}}{\max\{\|c_{r,\ell}^{(q)} - \bar{c}_{\ell}^{(q)}\|_2, 10^{-12}\}}. \tag{16}$$

The incorrect-candidate group is contained within the combined task-related group, so the groups are not disjoint. For sampled layers $\mathcal{L}_g$, their regularized centered cosine distance and attenuation-induced change are

$$D_{cs,\ell}^{(q)} = 1 - \left(\mu_{c,\ell}^{(q)}\right)^{\top} \mu_{s,\ell}^{(q)}, \qquad \Delta D_{cs} = \frac{1}{|\mathcal{L}_g|} \sum_{\ell \in \mathcal{L}_g} \left(D_{cs,\ell}^{(\text{attenuation})} - D_{cs,\ell}^{(\text{clean})} \right). \tag{17}$$

A negative ΔD_{cs} marks reduced separation. Secondary analyses measure group silhouette, group mixing, and linear-probe separability.

Figure 7 displays the centroid-level decrease $-\Delta D_{cs}$ for every model–task pair. Its mean is 0.385 at transition and 0.459 at the low-accuracy setting. The low-accuracy decrease exceeds the transition decrease in all 20 pairs and the mean difference is 0.073.

A prompt-level analysis changes the aggregation level. It computes the incorrect-candidate to combined-task-related cosine distance within each prompt and layer, then averages over sampled layers and base templates. The raw version uses projected vectors directly. Centering subtracts the mean unperturbed group projection within each model–task–layer evaluation group. Whitening uses the covariance of those unperturbed projections. Its ridge is the larger of $0.10\,\mathrm{tr}(\Sigma)/d$ and 10^{-5}. Both transforms are fitted and evaluated in the same group. Unit rescaling leaves cosine unchanged for nonzero vectors a and b

$$\cos\left(\frac{a}{\|a\|_2}, \frac{b}{\|b\|_2}\right) = \cos(a, b). \tag{18}$$

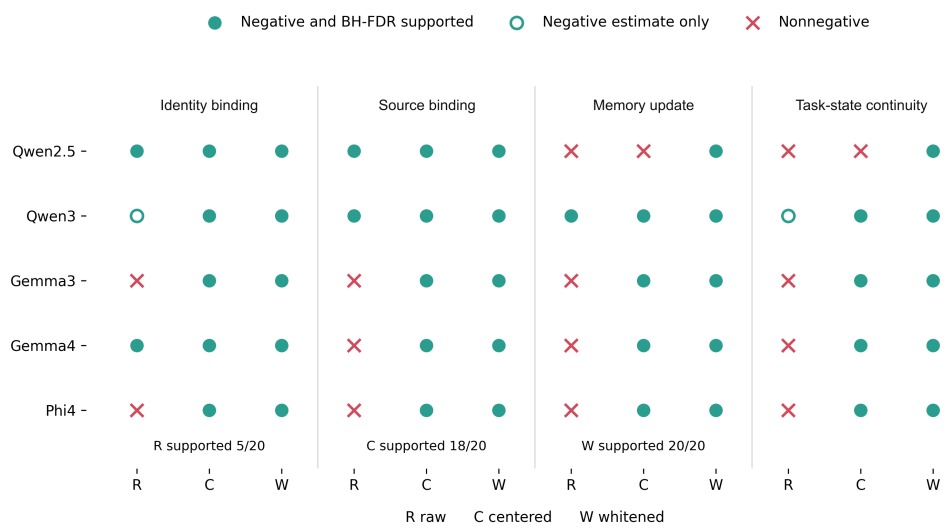

Figure 8: Prompt-level preprocessing sensitivity at the transition setting. R denotes raw projected vectors. C denotes centered vectors. W denotes vectors whitened with unperturbed covariance from the same model–task–layer group.

The incorrect-candidate group is a subset of the combined task-related group, so their centroids share token contributions. The coordinate center also weights group centroids equally and includes the union group. These constructions couple the compared representations and weight positions according to their membership in overlapping groups. Distance reductions depend on the coordinate system. Raw projected distances decrease in 7 of 20 pairs, with 5 reductions supported after correction. Centering yields supported reductions in 18 pairs, and covariance whitening yields supported reductions in all 20 pairs. These results characterize the specified overlapping groups under the corresponding offline transformations.

At transition, linear-probe accuracy declines in 17 of 20 model–task pairs. Group silhouette decreases in 5 and mixing increases in 5. At the low-accuracy setting, probe accuracy declines in 16 pairs. Silhouette decreases in 8 and mixing increases in 5. All three directions coincide in 3 pairs at each setting. Standardized relations between the predefined groups decline consistently. Broader dedifferentiation measures vary by pair.

E Performance Restoration Experiments

E.1 Prompt cues and the original downstream activation probe

State-reactivation cues name the current or final state. Evidence-focus cues point to the latest text that establishes it. Neutral and distractor cues provide prompt-level controls. The neutral cue asks for careful use of context, while the distractor cue emphasizes earlier evidence.

The original downstream probe keeps upstream attenuation active and inserts an unperturbed group or task-state vector at a later layer. The incorrect-state source changes the written content. The question-position control changes the write location. A norm-matched random vector changes the content while preserving its norm. Figure 9 summarizes these probes. Prompt cues improve accuracy in 16 of 40 tests and 2 intervals exclude zero. They improve answer margin in 24 tests and 11 intervals exclude zero. Downstream activation replacement improves accuracy in 2 tests without interval support. It improves answer margin in 18 tests and 12 intervals exclude zero. These counts use 95 percent identifier-cluster bootstrap intervals and precede FDR adjustment.

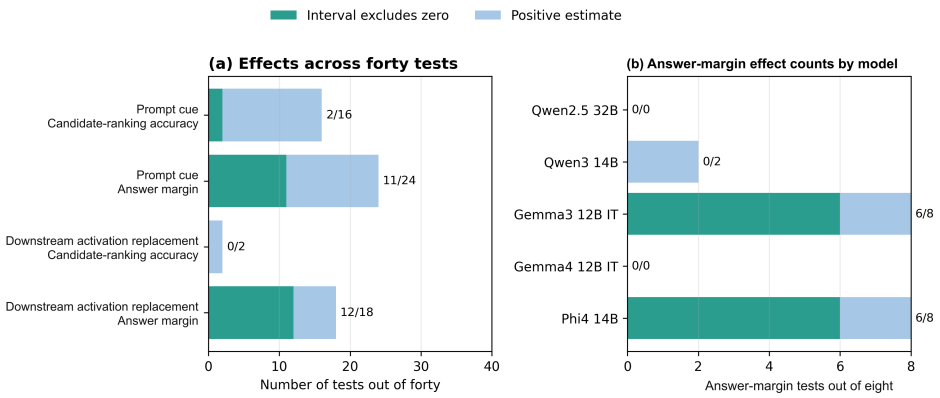

Figure 9: Original prompt-cue and downstream activation-replacement probes. Panel a counts positive and interval-supported effects across forty tests. Panel b resolves the eight downstream answer-margin tests within each model.

E.2 Content-dependent performance restoration after performance loss

The performance-restoration experiment selects a calibration setting that reduces candidate-ranking accuracy. Let L be the number of layers and let $\ell_{\max}$ be the final attenuated layer. The late and middle layers are

$$\ell_{\text{late}} = \min\{L-1, \max(\ell_{\max}+2, \text{round}(0.80(L-1)))\}, \tag{19}$$

$$\ell_{\text{mid}} = \min\{L-1, \max(\ell_{\max}+1, \text{round}(0.65(L-1)))\}. \tag{20}$$

Calibration evaluates four replacement-layer candidates. The single-layer candidate is $\{\ell_{\text{late}}\}$. For $k \in \{3,5\}$, the contiguous candidate is $\{\max(0,\ell_{\text{late}}-k+1),\ldots,\ell_{\text{late}}\}$. The bridge candidate is the sorted unique set $\{\ell_{\max},\ell_{\text{mid}},\ell_{\text{late}}\}$. Eligible candidates improve answer margin over attenuation and produce a larger effect with correct current information than with a incorrect-state source. All four candidates meet this criterion in all 24 model–task calibration groups. Calibration selects one candidate per group and fixes it before evaluation on separate prompts.

The source-controlled replacement copies aligned activation sequences at correct-answer and combined task-related positions. Exact sources preserve the target answer and original wording. Paraphrased sources preserve the entity labels, answer labels, and deterministic answer while changing surrounding wording. Incorrect-state sources encode an alternative valid state and may change more than one textual factor. Question-position controls write source activations at question positions. Norm-matched random controls use five seeds.

Let $\mathcal{L}_p$ be the fixed replacement-layer set. The ordered replacement groups are $r_1 =$ correct-answer positions and $r_2 =$ combined task-related positions. Let $T_{i,r_q} = (t_1,\ldots,t_{n_{r_q}})$ be receiver positions for group r_q. A source span contains n_s residual vectors at positions $(s_1,\ldots,s_{n_s})$. For $n_{r_q} > 1$, the alignment index is

$$\pi_q(j) = 1 + \text{round}\left(\frac{j-1}{n_{r_q}-1}(n_s-1)\right). \tag{21}$$

Equal-length spans use direct correspondence. When $n_{r_q} = 1$, the aligned source vector is the mean over the source span. If $g(i,c)$ denotes the source paired with receiver i under source condition c, define

$$a^{(c,q)}_{i,\ell,t_j} = \begin{cases} h^{(\text{clean})}_{g(i,c),\ell,s_{\pi_q(j)}}, & n_{r_q} > 1, \\ \dfrac{1}{n_s}\sum_{k=1}^{n_s} h^{(\text{clean})}_{g(i,c),\ell,s_k}, & n_{r_q} = 1. \end{cases} \tag{22}$$

Let $h^{(c,\text{pre})}_{i,\ell,t}$ denote the receiver activation immediately before replacement in the current forward pass. It incorporates the effects of earlier interventions and any attenuation already applied at layer ℓ.

Table 11: Condition means for content-dependent performance restoration. Values give equal weight to each model–task group. Margin is measured in nats per answer token.

Short tasks			Passage tasks		
Condition	Accuracy	Margin	Condition	Accuracy	Margin
Unperturbed	0.951	12.036	Unperturbed	0.985	4.822
Attenuation only	0.469	-1.000	Attenuation only	0.467	-0.375
Exact source	0.909	8.443	Exact source	0.897	3.408
Question-position control	0.484	0.576	Paraphrased source	0.863	3.130
Incorrect-state source	0.298	-12.256	Incorrect-state source	0.228	-4.365
Norm-matched random	0.399	-1.966	Norm-matched random	0.352	-1.235
			Question-position replacement	0.493	0.436

The ordered replacement is

$$\begin{aligned} h_{i,\ell,t}^{(c,0)} &= h_{i,\ell,t}^{(c,\mathrm{pre})}, \\ h_{i,\ell,t}^{(c,q)} &= \big(1 - p_{i,\ell,t}^{(q)}\big) h_{i,\ell,t}^{(c,q-1)} + p_{i,\ell,t}^{(q)} a_{i,\ell,t}^{(c,q)}, \qquad q \in \{1,2\}, \\ h_{i,\ell,t}^{(c)} &= h_{i,\ell,t}^{(c,2)}. \end{aligned} \tag{23}$$

Attenuation hooks execute before replacement hooks. Correct-answer replacement executes before combined task-related replacement. Since $T_{i,r_1} \subseteq T_{i,r_2}$ and both steps use the same replacement layers, the second assignment supplies every final vector in the combined mask. Equation 23 therefore reduces to the effective single-mask replacement in Equation 4. The intermediate correct-answer assignment records the implementation order and has no effect on the final residual vector at that layer. Union-index alignment determines the source vectors written at all selected positions. At an overlapping layer, replacement directly overwrites attenuated residual vectors at the selected positions. Attenuation remains active elsewhere. A strictly downstream window satisfies $\min \mathcal{L}_p > \max \mathcal{L}_{\mathrm{int}}$. An overlapping window satisfies $\mathcal{L}_p \cap \mathcal{L}_{\mathrm{int}} \neq \varnothing$.

For source vector $h_j^{(\mathrm{source})}$, a norm-matched random vector is

$$r_j = \frac{\|h_j^{(\mathrm{source})}\|_2}{\max\{\|\epsilon_j\|_2, 10^{-12}\}} \epsilon_j, \qquad \epsilon_j \sim \mathcal{N}(0, I). \tag{24}$$

Let I_b contain the three prompt variants in base-case cluster b. Condition margins are averaged within the cluster. The random condition first averages its five seeds within each prompt

$$\bar{M}_b^{(c)} = \frac{1}{|I_b|} \sum_{i \in I_b} M_i^{(c)}, \qquad \bar{M}_b^{(\mathrm{random})} = \frac{1}{|I_b|} \sum_{i \in I_b} \frac{1}{S} \sum_{s=1}^{S} M_{i,s}^{(\mathrm{random})}. \tag{25}$$

For B base cases, recovery and control contrasts are

$$\begin{aligned} \Delta M_c &= \frac{1}{B} \sum_b \left(\bar{M}_b^{(c)} - \bar{M}_b^{(\mathrm{attenuation})} \right), \\ \Gamma M_c &= \frac{1}{B} \sum_b \left(\bar{M}_b^{(c)} - \bar{M}_b^{(\mathrm{incorrect})} \right), \\ \Omega M_c &= \frac{1}{B} \sum_b \left(\bar{M}_b^{(c)} - \bar{M}_b^{(\mathrm{random})} \right). \end{aligned} \tag{26}$$

The superscript $\mathrm{incorrect}$ denotes the incorrect-state source condition. The ΓM_c and ΩM_c contrasts apply to the same-wording and paraphrased-source conditions. Accuracy contrasts use the same clustering and averaging order.

With identifier-based clustering, the correct-source conditions exceed the incorrect-state source in all 24 model–task evaluation groups after BH-FDR correction. They also exceed the random-vector control in every group. Correct-source conditions improve answer margin over attenuation in all 24

Table 12: Supported restoration contrasts in 15 text-disjoint evaluation groups. The groups comprise 11 overlapping-window groups and 4 strictly downstream Gemma3 groups. Counts refer to receiver-answer accuracy and answer margin.

Contrast	Accuracy	Margin
Exact source minus attenuation	14/15	15/15
Paraphrased source minus attenuation	14/15	15/15
Exact source minus different state	15/15	15/15
Paraphrased source minus different state	15/15	15/15
Exact source minus random mean	15/15	15/15
Paraphrased source minus random mean	15/15	15/15

groups and accuracy in 22. Merging repeated short-task prompts into exact-text clusters preserves positive correct-source over incorrect-state accuracy and margin contrasts in all 24 groups. Both contrasts retain adjusted support in 20. Correct-source over attenuation accuracy is positive in all 24 groups and supported in 10. The answer-margin counterpart is positive in all 24 and supported in 20.

The earlier fully disjoint subset contains five identity-binding model–task groups and four passage model–task groups. Correct-source conditions exceed incorrect-state sources on both metrics in all nine groups with adjusted support. Correct-source over attenuation accuracy is positive in all nine and supported in eight. Its answer-margin contrast is positive and supported in all nine.

The final distinct-text evaluation contains 15 model–task groups. Passage tasks at the transition setting contribute 10 groups. Short-task transfer contributes 5. Both correct-source conditions exceed the incorrect-state source in all 15 groups after correction. Exact sources exceed the five-seed random mean in every group for accuracy and margin. Paraphrased sources do the same. Both correct-source conditions improve answer margin over attenuation in all 15 groups and accuracy in 14. These counts persist in the paired source-and-receiver correctness subset and after removing identifier distinctions.

At late layers, question tokens can carry information obtained from context. Exact-source replacement at the combined task-related positions outperforms question-position replacement in most groups. Gemma3 passage source binding reverses this ordering. Accuracy reaches 0.933 under question-position replacement and 0.783 under correct-source replacement at the predefined task-related positions. Useful state information can therefore appear at multiple token locations. The source-content comparison remains supported by the advantage of correct sources over incorrect-state sources.

E.3 Qualitative success and failure cases

Table 13 gives a fixed audit slice across all six families using Qwen2.5 32B. Short-task rows use the first original evaluation target in source order and attenuation strength 0.50. Passage rows use the first unused original target and its fixed transition setting. The passage memory-update row skips the two prompts in Table 14. Each block follows one receiver through unperturbed scoring, attenuation, and three source conditions.

Table 13: A fixed qualitative audit slice spanning six task families. Margin is the correct score minus the strongest competing score in nats per answer token. The passage memory-update example retains the correct top-ranked answer under attenuation and illustrates resistance to the tested disruption.

Condition	Top-ranked candidate	Correct	Margin
Short identity binding			
Unperturbed	Callum Radford	1	9.427
Attenuation only	Delia Ashford	0	-2.831
Exact source	Callum Radford	1	8.363
Paraphrased source	Callum Radford	1	7.772
Incorrect-state source	Delia Ashford	0	-12.190
Short source binding			

Condition	Top-ranked candidate	Correct	Margin
Unperturbed	Callum Radford	1	8.761
Attenuation only	Adrian Darrow	0	-2.811
Exact source	Adrian Darrow	0	-1.442
Paraphrased source	Adrian Darrow	0	-1.113
Incorrect-state source	Adrian Darrow	0	-13.693
Short memory update			
Unperturbed	citadel	1	13.469
Attenuation only	aurora	0	-5.802
Exact source	citadel	1	13.328
Paraphrased source	citadel	1	9.684
Incorrect-state source	drizzle	0	-20.832
Short task-state continuity			
Unperturbed	route receipts batch 702	1	1.361
Attenuation only	catalog permits batch 700	0	-9.696
Exact source	route receipts batch 702	1	1.385
Paraphrased source	route receipts batch 702	1	0.318
Incorrect-state source	verify dispatches batch 703	0	-11.971
Passage source binding			
Unperturbed	Atlas-09	1	4.972
Attenuation only	Kestrel-17	0	-3.589
Exact source	Kestrel-17	0	-4.677
Paraphrased source	Kestrel-17	0	-4.063
Incorrect-state source	Cobalt-31	0	-1.894
Passage memory update			
Unperturbed	Quartz-73	1	5.521
Attenuation only	Quartz-73	1	3.885
Exact source	Quartz-73	1	5.573
Paraphrased source	Quartz-73	1	5.719
Incorrect-state source	Willow-64	0	-16.445

Table 14 retains two detailed passage memory-update records. In Case-4101, the initial record assigns Atlas-09 and an unapproved proposal names Cobalt-31. A later signed correction makes Lumen-58 authoritative. Attenuation ranks Cobalt-31 first. Exact and paraphrased-source replacement restore Lumen-58. The incorrect-state source ranks Atlas-09 first. The question-position control and displayed random-vector replacement leave Cobalt-31 first. In Case-4119, the signed correction is Willow-64. Exact-source replacement restores Willow-64. Reworded replacement leaves a competitor first. Question-position replacement ranks the correct answer first in this case.

Table 14: Representative restoration success and partial failure from fixed transition outputs.

Condition	Top-ranked candidate	Correct	Margin	Intervention
Case-4101. Successful recovery in Qwen3 14B memory update				
Unperturbed	Lumen-58	1	4.783	No intervention
Attenuation only	Cobalt-31	0	-2.767	$\alpha = 0.245$
Exact source	Lumen-58	1	2.956	Layers 27–31
Paraphrased source	Lumen-58	1	2.140	Layers 27–31
Incorrect-state source	Atlas-09	0	-12.862	Layers 27–31
Question-position control	Cobalt-31	0	-6.633	Layers 27–31
Norm-matched random seed 1	Cobalt-31	0	-6.963	Layers 27–31
Case-4119. Partial failure in Gemma3 12B IT memory update				

Condition	Top-ranked candidate	Correct	Margin	Intervention
Unperturbed	Willow-64	1	9.555	No intervention
Attenuation only	Lumen-58	0	-7.438	$\alpha = 0.190$
Exact source	Willow-64	1	7.752	Layers 34–38
Paraphrased source	Lumen-58	0	-5.808	Layers 34–38
Incorrect-state source	Quartz-73	0	-5.094	Layers 34–38
Question-position control	Willow-64	1	6.297	Layers 34–38
Norm-matched random seed 1	Lumen-58	0	-4.962	Layers 34–38

The experiment ranks a fixed set of answer candidates. The random rows display seed 1. Statistical comparisons average all five random seeds within each prompt.

E.4 Distance to unperturbed activations

Let $a_i^{(c)}$ be the final-layer residual vector at the last prompt token. A fixed Gaussian map P gives the 128-dimensional projection $z = Pa_i^{(c)}$. Let $u_1, \ldots, u_n$ be the unperturbed calibration projections, with $n \geq 2$. For coordinate k, let s_k be their sample standard deviation, computed with denominator $n - 1$. The coordinate scales are

$$\widehat{\sigma}_k = \begin{cases} s_k, & s_k > 10^{-6}, \\ 1, & s_k \leq 10^{-6}, \end{cases} \qquad W = \operatorname{diag}(\widehat{\sigma}^{-1}).$$

Define the scaled pairwise distance as

$$\delta(z, u) = \frac{\|W(z - u)\|_2}{\sqrt{128}}.$$

The clean-reference nearest-neighbor ratio is

$$R_{\mathrm{NN}}(z) = \frac{\min_{1 \leq j \leq n} \delta(z, u_j)}{\max\left\{ \operatorname{median}_{1 \leq j \leq n} \min_{\substack{1 \leq k \leq n \\ k \neq j}} \delta(u_j, u_k), 10^{-6} \right\}}. \tag{27}$$

Reference neighbors exclude the same sample index. Lower values indicate proximity to the finite unperturbed calibration reference set.

For the same prompt i, define $z_i^{(c)} = Pa_i^{(c)}$ and $z_i^{(\mathrm{clean})} = Pa_i^{(\mathrm{clean})}$. Paired cosine distance compares these unstandardized projections

$$D_{\mathrm{pair},i}^{(c)} = 1 - \frac{\left(z_i^{(c)}\right)^\top z_i^{(\mathrm{clean})}}{\|z_i^{(c)}\|_2 \|z_i^{(\mathrm{clean})}\|_2}. \tag{28}$$

Using an equal-weight mean over model–task groups, the short-task nearest-neighbor ratio is 0.054 without intervention and 1.204 under attenuation. Correct-source replacement reduces it to 0.530. Incorrect-state replacement reduces it to 0.649. Random replacement yields 1.292. Passage ratios are 1.408 without intervention and 5.415 under attenuation. Correct-source replacement reduces the ratio to 1.878. Incorrect-state replacement reduces it to 2.343. Random replacement yields 4.933. Repeated calibration text in three historical short-task families produces unusually small unperturbed distances.

Incorrect-state replacement reduces the nearest-neighbor ratio relative to attenuation in all 24 model–task groups after BH-FDR correction. It reduces paired cosine distance in 18 groups and 15 reductions remain supported after correction. Across the two metrics and 24 groups, source activation replacement reduces divergence in 42 of 48 comparisons and 39 remain supported. Every correct-source condition remains farther from the unperturbed calibration reference than its paired unperturbed condition. The same holds for every incorrect-state condition. Replacement recovers part of the measured geometry while residual displacement remains in all groups.

Figure 4 presents these two diagnostics and is attached only to the data and equations above. The figure does not display the layerwise Mahalanobis or gate-flip quantities defined next.

E.4.1 Additional layerwise displacement diagnostics

An archived comparison uses a different activation source and coordinate system. At diagnostic layer ℓ, activations at the combined task-related positions are averaged before a layer-specific map P_ℓ projects them into 128 dimensions

$$z_{i,\ell}^{(c)} = P_\ell \left(\frac{1}{|T_{i,\text{state}}|} \sum_{t \in T_{i,\text{state}}} h_{i,\ell,t}^{(c)} \right). \tag{29}$$

The implementation returns an undefined value when the product of the two projected vector norms is at most 10^{-12}. This threshold was not reached in the historical 24-group analysis. Let μ_ℓ and Σ_ℓ be the mean and covariance of unperturbed projected vectors within the same model–task evaluation group and layer, with $d = 128$. The Mahalanobis diagnostic is

$$d_{\text{Mah},i,\ell}^{(c)} = \sqrt{\left(z_{i,\ell}^{(c)} - \mu_\ell\right)^\top \left(\Sigma_\ell + \lambda_\ell I\right)^{-1} \left(z_{i,\ell}^{(c)} - \mu_\ell\right)}, \qquad \lambda_\ell = \max\left\{0.10 \frac{\operatorname{tr} \Sigma_\ell}{d}, 10^{-5}\right\}. \tag{30}$$

For the diagnostic gate projection, let $g_{i,\ell,k}^{(c)}$ indicate whether mean captured coordinate k over combined task-related positions is positive. The captured vector is the feed-forward gate output in four models and the fused gate–up projection in Phi4. The gate-flip rate is

$$F_{\text{gate},i,\ell}^{(c)} = \frac{1}{d_g} \sum_{k=1}^{d_g} \mathbf{1}\left[g_{i,\ell,k}^{(c)} \neq g_{i,\ell,k}^{(\text{clean})}\right]. \tag{31}$$

The archived comparison uses positions outside the combined task-related group and question, selected close to target-position activation norms. A pairing audit found unequal affected-position counts in 549 of 1,023 prompt pairs across 10 model–task groups. Residual-site counts differ in 294 pairs across 5 groups. The non-state arm draws specifications from a broader target list than the target arm. These quantities therefore describe displacement under unequal count matching. The matched nearby-layer control supplies the primary location comparison in the main analysis.

In the archived summaries, target attenuation produces larger paired cosine shift in 19 of 20 model–task groups after correction. Mahalanobis distance and gate-flip rate show the same count. Target attenuation reduces the activation norm more in all 20 groups.

E.5 Passage tasks and performance restoration across performance settings

The two passage families cover conflicting updates and multi-document source binding. Each family contains 10 calibration base cases and 20 evaluation base cases. Every base case has three variants. These runs use the same answer reducer, top 16 residual targets, token groups, sequence alignment, hook order, and cross-layer replacement procedure as the short-task experiment. Calibration selects one fixed replacement-layer candidate for each of 10 model–task groups.

Table 15 reports transition and low-accuracy means with equal weight for each model–task group. At transition, exact-source replacement raises accuracy from 0.467 to 0.897. Paraphrased-source replacement reaches 0.863. Incorrect-state and random-vector controls reach 0.228 and 0.352. Both correct-source conditions improve accuracy in all 10 groups and nine gains remain supported after correction. Their answer-margin gains remain supported in every group. Both correct-source conditions exceed the incorrect-state source in accuracy and margin throughout the evaluation.

At the low-accuracy settings, attenuation yields mean accuracy 0.087. Accuracy reaches 0.812 with exact-source replacement and 0.765 with paraphrased-source replacement. Incorrect-state and random-vector controls reach 0.182 and 0.322. Both correct-source conditions improve accuracy and margin in all 10 endpoint groups after correction.

Paraphrased source prompts preserve entity labels, answer labels, and the valid state determined by the task rule while changing surrounding language. Table 16 reports standalone unperturbed accuracy for both passage families. Mean standalone accuracy of the paraphrased source runs is 91.7 percent across the 10 passage model–task groups. Memory update reaches 100 percent in all five models. The paired source-and-receiver correctness analysis retains only cases with correct target and source runs.

Table 15: Passage-task performance at calibration-selected transition and low-accuracy settings under the common five-model protocol. Accuracy is scored against the receiver answer. Uniform selection among the three candidates gives expected accuracy $1/3$. Margins are mean log-probability differences in nats per answer token.

	Transition setting		Low-accuracy setting	
Condition	Accuracy	Margin	Accuracy	Margin
Attenuation only	0.467	-0.375	0.087	-4.653
Exact source	0.897	3.408	0.812	2.636
Paraphrased source	0.863	3.130	0.765	2.473
Incorrect-state source	0.228	-4.365	0.182	-4.059
Norm-matched random mean	0.352	-1.235	0.322	-1.401
Unperturbed source at question positions	0.493	0.436	0.462	-0.293

Table 16: Standalone unperturbed accuracy of paraphrased source runs with the intended task-state meaning.

Model	Memory update	Source binding
Qwen3	1.000	0.750
Gemma3	1.000	0.867
Gemma4	1.000	1.000
Phi4	1.000	0.567
Qwen2.5	1.000	0.983
Equal-weight model–task mean	1.000	0.833

Recovery remains positive in all 15 text-disjoint model–task groups under both source conditions. Accuracy remains supported in 14 groups and answer margin in all 15.

Let P be the concatenation of stored context and question fields. Let $V(P)$ be its lowercase alphanumeric token set. Wording overlap is

$$J(P, P') = \frac{|V(P) \cap V(P')|}{|V(P) \cup V(P')|}. \tag{32}$$

Mean Jaccard overlap between original and reworded passage prompts is 0.363 and values range from 0.323 to 0.451. In the text-disjoint short-task evaluation, the mean is 0.330 and values range from 0.269 to 0.417. These values quantify surface change. Fixed entity and answer labels provide an additional alignment constraint.

The 15 text-disjoint groups comprise 11 overlapping-window groups and 4 strictly downstream groups. Overlapping windows provide local repair at layer–token sites shared by attenuation and replacement. The strictly downstream groups all use Gemma3 and begin replacement after the final attenuated layer. In each subset, both correct-source conditions exceed incorrect-state sources and random controls on accuracy and answer margin in every group after correction. The combined evidence establishes content-dependent answer recovery, with strictly downstream recovery demonstrated in the four Gemma3 groups.

In an earlier Gemma4 passage-memory sensitivity protocol, unperturbed and attenuated accuracy remain 1.000 and 0.517 whether scoring uses the designated pair or all typed competitors. Mean and sum reducers preserve these accuracies. Their scales yield unperturbed and attenuated margins of 4.511 and -0.522 under the mean reducer, compared with 19.053 and -1.229 under the sum reducer. In free generation, unperturbed accuracy remains 1.000 at all three temperatures. Attenuated accuracy is 0.167 at temperature 0, 0.157 at 0.3, and 0.160 at 0.7. The attenuation effect persists across the tested decoding settings and competitor definitions.

Across the broader sensitivity grid, attenuation produces positive damage in all 36 model–task–decoding groups and all 48 model–task–competitor groups. These comparisons establish stability of the attenuation effect across output settings. They do not evaluate source-controlled recovery under full free generation.

E.6 Low-accuracy prompt cues and association across analysis levels

For prompt i and cue u, the low-accuracy analysis compares the cued run with the no-cue run at the same selected strength

$$a_i^{(c)} = \mathbf{1}[M_i^{(c)} \geq 0], \qquad \Delta a_i^{(u)} = a_i^{(u)} - a_i^{(\text{no cue})}, \tag{33}$$

$$\Delta M_{i,\text{std}}^{(u)} = \frac{M_i^{(u)} - M_i^{(\text{no cue})}}{\max\{s_{\text{clean}}, 10^{-8}\}}, \tag{34}$$

where s_{clean} is the sample standard deviation of unperturbed no-cue margins in the corresponding model–task group. State-reactivation and evidence-focus cues are tested separately. The three variants of a base case remain in the same bootstrap cluster. Equivalence requires the full unadjusted 95 percent interval to fall strictly inside -0.05 to 0.05 for accuracy or -0.20 to 0.20 for standardized margin.

Low-accuracy equivalence holds in 34 of 80 external cue tests. These comprise 25 of 40 accuracy tests and 9 of 40 standardized-margin tests. Gemma4 identity binding passes all four cue–metric combinations. Phi4 source binding also passes all four. Internal activation replacement yields positive accuracy and answer-margin effects in all 10 passage model–task endpoint groups after BH-FDR correction.

At transition, the association between token-group distance and recovery yields 20 of 80 correlations with intervals excluding zero. The median absolute Spearman correlation is 0.178. Leave-one-template-out prediction yields positive R^2 in 21 of 80 comparisons. At the low-accuracy setting, 19 correlations exclude zero and the median absolute correlation is 0.132. Positive R^2 occurs in 18 comparisons. A single activation-distance measure explains little of the recovery variation across individual examples.

E.7 Complementary evidence from interventions and activation distances

Fixed task-related interventions measure the consequences of activation attenuation through performance curves. The nearby-layer control compares intervention locations at matched performance loss. Alternative token-position controls compare different parts of the prompt. Component interventions measure vulnerability at specific attention and feed-forward locations. Qwen3 targets exceed every control in 3 of 8 evaluated task–component units. Gemma4 reaches 7 of 8 and Gemma3 reaches 2 of 8. Phi4 reaches 0 of 8 and Qwen2.5 reaches 4 of 8.

The distance analyses describe changes after attenuation and replacement. Sample-level correlations quantify their association with recovery. Source-controlled replacement directly compares the information supplied by different prompts. Correct sources outperform incorrect-state and random-vector controls across five models on passage tasks. Gemma3 also supports recovery from strictly downstream windows. Replacements at predefined task-related and question positions show that useful information can appear at several token locations. Together, these analyses provide converging evidence for the central result.