BEYOND THE FITTED ANSWER: CAUSAL DISSECTION AND DOWNSTREAM CONSEQUENCES OF LEARNED ACTIVATION INTERVENTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Activation interventions are a widely used method to associate language model (LM) representations with interpretable variables. However, success on the objective used to fit an intervention neither establishes that the intervention produced the intended downstream consequences, nor does it identify which internal components caused the behavioural changes. We investigate these questions by constructing a Theory of Mind (ToM) belief-tracking dataset comprised of short narratives and associated questions about agents' beliefs and their consequences. Rank-16 activation patches are trained under the objectives of 1) transferring the location answer from a source narrative and 2) implementing a consistent remapping of locations. Evaluations of the activation patches are carried out by measuring the differences in matched target rules, held-fixed downstream consequence tests, and with unfitted principal component analysis (PCA) controls. Answer accuracy is shown to be insufficient to establish a coherent intervention. On the consequence test set, a PCA intervention followed by a fixed output relabelling achieves 100% accuracy on the target question under the alternative mapping, yet fails to answer all follow-up questions correctly for any narrative. By contrast, holding each patch constant across questions and subsequent observations, we find that the learned remapping led to the LMs behaving as if the underlying events had been rewritten, showing high agreement with follow on consequences, including those not trained for. Nevertheless, comparisons with PCA controls reveal no additional accuracy benefit from fitting for the intended source-answer transfer objective on these consequence-based evaluations. To test the causal contribution of attention keys to the behavioural difference between learned and PCA interventions, we interchange the attention keys induced by the two interventions. Queries and values evolve within the recipient. Exchanging the PCA-induced attention keys with the learned patch attention keys led to increased agreement with the trained patches' outputs by up to 75.4 percentage points, averaged over multiple seeds, whereas the reciprocal exchanges favoured PCA outputs. On paired colour-lookup questions, performance also exceeds an exact upper bound achievable by any single table-independent transformation of the PCA colour answer. Together, these results establish the tested downstream scope of learned remapping and provide causal evidence that attention keys induced by the learned patch contribute to the behavioural differences between learned and PCA interventions under the tested multilayer intervention setting.

1 INTRODUCTION

An activation intervention is a technique that is used to directly modify a neural network's internal hidden states during inference, primarily to steer or alter its output behaviour. Such interventions are thus commonly used as a tool to link a language model's (LM) internal representations to interpretable variables (Geiger et al., 2022; 2024). A common view in the relevant literature is that an intervention has found such a variable if it has succeeded on its fitting objective. This leaves researchers with two open questions. Firstly, what does the intervention actually change? If the edit changes a model's

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

answer about where an agent thinks an object may be, is only the agent's belief changed? Or is the event rewritten entirely? Does it alter how the answer is computed? Each of these possibilities can agree on the same answer but predict differing world states and answers to subsequent questions. Secondly, which specific internal components contributed to the altered behaviour, beyond what an unfitted intervention already produced? We use downstream consequences to establish the behaviour that requires explanation, then intervene on components of the patched computation.

We make use of distributed alignment search (DAS), a method which learns a low-rank subspace wherein swapping the activations between a base narrative and a source narrative produces a specified answer (Geiger et al., 2022; 2024). We fit two objectives under the same conditions: (1) an intended objective, which copies the source narrative's location answer, and (2) an alternative objective, which applies a fixed permutation to the six possible locations in the dataset. The targets are locations, so each hypothesis about what a patch has changed yields testable predictions for later questions, which we check while holding the patch fixed. An unfitted principal component analysis (PCA) basis of the same rank is used as the baseline for what fitting adds. Our mechanistic question concerns the learned–PCA behavioural difference, including consequences absent from fitting.

Main result. Exchanging the critical-token attention keys induced by learned and PCA patches substantially transfers their downstream behaviours in both directions for the tested models. Keys are the only directly exchanged component; queries and values evolve within the recipient, and attention outputs are not restored. Natural base keys instead recover the affected base answers. Our finding is a causal contribution of keys to the learned intervention's effects, conditional on the patched computation. Effects span many blocks, and narrower regions and their complements contribute asymmetrically.

Three complementary tests. Table 2 summarises the experimental settings. An eight-question evaluation (Section 5) establishes what the learned patch changes and what it preserves. We then exchange only the attention keys on 120 fresh narratives (Section 6) to test if behaviour transfers in both directions without supplying target values or restoring attention outputs. A further replication on 108 different narratives repeats it in both models with 300-update fits of both the pair-swap and a six-cycle. We report fidelity, correctness and preservation, where fidelity is agreement with the actual output of the run that donated keys.

2 Related Work and Scope

Control tasks and causal alignment. Hewitt & Liang (2019) use random labels assigned consistently by word type in control tasks and selectivity experiments to assess how much probe learning can contribute to supposed representational success. Arbitrary token targets and alignment robustness are tested by Wu et al. (2023) using boundless DAS. Arora et al. (2024) fit alternative output labels and subtract control-task log-odds effects from the original task effects to extend control-task evaluation to causal interventions. Our alternative objective builds upon this body of work with forecasts of event-level consequences.

Interpreting and evaluating interventions. Examining the implications of successful interventions on internal computations is motivated by debates about subspace patching illusions (Makelov et al., 2024; Wu et al., 2024) and work on steering non-identifiability (Venkatesh & Kurapath, 2026). RAVEL (Huang et al., 2024) investigates whether interventions can change a target attribute while also preserving others. MIB (Mueller et al., 2025) benchmarks causal localisation, while Li et al. (2026) attempt to identify input regions where causal abstractions are faithful.

Belief representations and the attention-key comparison. Prakash et al. (2026) examine distributed belief retrieval through a lookback mechanism, while belief decoding and causal steering are investigated by Zhu et al. (2024). Bortoletto et al. (2025) investigate brittle belief representations. More recently, Steele et al. (2026a;b) propose shared value-slot and routing accounts of belief and discourse representations. Our contribution is instead a reciprocal dissection of the learned–PCA intervention difference. Matched nulls and natural base donors connect the downstream consequences of this difference to critical-token keys, while a colour relabelling bound constrains an output-level explanation. The novelty is this controlled comparison supported by downstream consequences.

3 MATCHED INTERVENTION OBJECTIVES

3.1 INTERVENTION AND TARGET RULES

Fix a model, a site s, a rank r and a training procedure $\mathcal{A}$, including its data and optimisation budget. For column-vector base and source activations $h_b, h_s \in \mathbb{R}^d$, let $U \in \mathbb{R}^{r\times d}$ have orthonormal rows, $UU^\top = I_r$. We apply

$$I_U(h_b, h_s) = h_b + \Pi_U(h_s - h_b), \qquad \Pi_U = U^\top U. \tag{1}$$

The source replaces selected components of the base. Complementary activation components are preserved. DAS learns U such that the patched model gives a specified answer.

Two target rules. Let y_s be the source narrative's answer. The source answer transfer rule transfers y_s unchanged, $f^*(y_s) = y_s$; the alternative rule applies $m_2(y_s)$, where $m_2 = (0\ 1)(2\ 3)(4\ 5)$, swapping the six locations in pairs. We fit a separate basis $U_{f,\omega}$ for each rule f and seed ω. A separate replication compares 300-update fits of m_2 with those of the six-cycle $m_1 = (0\ 1\ 2\ 3\ 4\ 5)$, which moves each location to the next and the last to the first.

Measuring success. Each objective is measured against its own target. We report two measures, candidate-ranking success and next-token success, that is, the target ranks first among the six locations and the target is the most likely next token over the entire vocabulary.

In Qwen, an exact candidate-score tie is resolved differently by candidate order and vocabulary argmax, explaining the small prefilled intended-score difference (Appendix A.3).

It is possible for the alternative target to match the base answer in the case of an intervention that changes nothing. We therefore report a 'clean opportunity' subset. These are pairs where the unpatched model answers both narratives correctly and both targets differ from the base answer.

4 EXPERIMENTAL DESIGN

Task and counterfactual targets. Let $\mathcal{G}$, $\mathcal{O}$ and $\mathcal{L}$ denote the sets of agents, objects and locations, respectively. $\mathcal{G}^{\leq k}$ contains agent sequences of length at most k, where sequence length represents belief depth and the empty sequence ϵ represents the world state.

Synthetic theory of mind (ToM) style narratives allow tracking of world locations and corresponding nested beliefs as $\mathrm{Loc} : \mathcal{G}^{\leq k} \times \mathcal{O} \to \mathcal{L}$: $\mathrm{Loc}(\epsilon, o)$ is object o's actual location, $\mathrm{Loc}(a, o)$ agent a's belief, and $\mathrm{Loc}(ab, o)$ what a believes b believes. The five-shift task uses two agents, maximum belief depth $k = 2$, 20 objects and six locations. Paired narratives share a query and only differ by one move or tell event, the critical event, and the location it names is the critical location. The token naming the critical location is the critical token. Evaluation pairs change the queried answer while distractor events concern other objects and do not change the queried object's location, leaving associated beliefs unchanged.

Models and matched fitting. Frozen models, Qwen2.5-72B-Instruct (Yang et al., 2024), and Mistral-Small-24B-Instruct-2501 (Mistral AI, 2025) run in BF16 on identical 1,000 training and 1,500 evaluation pairs (Table 1). Each model fits rank-16 orthonormal bases for both objectives at the fourth decoder block's complete event span, using shared seeds. Each basis is optimised with AdamW under matched settings for the paired objectives. Every training pair is visited exactly once during each primary 1,000-update fit. The separate 300-update mapping replication samples training pairs with replacement (Appendix D). The paired objectives share the shuffled order and PCA initialisation. Our rank-16 and fourth block design decisions are detailed and justified in Appendix A.

Response interfaces and development. Fitting uses the models' standard chat format and scores the candidate answers at the start of the model's reply. At this position, the candidate answers carry little probability mass, so candidate-ranking accuracy does not necessarily imply correct next-token generation. To counter this, we also evaluate each saved basis after prefilling the model's reply with `Answer:` to encourage an immediate location answer. This requires no extra training and keeps the materialised event patch identical across formats. Candidate probability mass, candidate-ranking accuracy and global next-token accuracy are reported separately.

Table 1: Matched ordered semantic inputs and three fit seeds per objective. Native response wrappers and numerical execution differ

Model	Train	Primary	Disc. / check
Qwen72	1000	1500	120 / 120
Mistral24	1000	1500	120 / 120

Table 2: **Three settings answer different questions.** Models share stories within each study; study populations and operators are not pooled. A separate 108-story cohort repeats the key-only setting for two mappings with 300-update fits. The 48-story fixed relabelling bound does not transfer automatically to either key only cohort.

Setting / question	Population per model	Operation	Readouts / interpretation
Eight question evaluation What consequences change?	120 discovery + 120 checking stories.	One fixed rank-16 event patch reused across eight queries; no component exchange.	Direct answer, world, other agent, search, updates, irrelevant object and report: behavioural scope.
Fixed value colour test Can keys change later colour answers at fixed values?	72 fixed stories, matched across models; 48 distinct location stories primary.	Exchange keys at natural target values V_T; restore recipient attention outputs through the critical token.	Direct location, two observer colours, two nonobserver colours: conditional key effect and a fixed-recoding bound.
Key only confirmation Does reciprocal transfer persist without a value transplant?	All 120 fresh stories primary: 96 distinct + 24 collision cases.	Reciprocal learned/PCA keys only; no value transplant or output restoration. Downstream Q/V evolve.	Same five types of readout, on new cores: actual endpoint fidelity, correctness and preservation.

Five-shift evaluation. We have five equally weighted environments with varying templates, choice order, lexical support, distractors or all four jointly, with 60 calibration and 300 held-out pairs each. We use these for evaluation.

PCA and execution controls. Controls for the contribution of supervised fitting are provided by unfitted rank-16 PCA bases. Learned and PCA patches use the original source event; remapped source events provide separate full-event controls. PCA is computed on source-minus-base activation differences from the training pairs. PCA construction and execution checks are detailed in Appendix A.

Three experiments with separate populations. Table 2 summarises the three experiments. The behavioural evaluation reuses each model's six bases (two objectives $\times$ three seeds) on a *discovery* set, used for comparing candidate accounts of intervention effects, and a separate *checking* set for evaluating their forecasts across eight question types. The following views are covered by the checking set: direct belief, world location, another agent's belief, search, observed and unobserved updates, an irrelevant object and truthful reporting. Search and truthful report questions follow explicit task conventions.

The two component experiments (fixed-value and key-only) use different narratives and are never pooled. A separate replication repeats the key-only experiment on 108 further narratives for both m_1 and m_2, with three fits of each trained for 300 updates rather than 1,000.

Uncertainty. The key-only experiment and the mapping replication each use 20,000 resamples of whole narratives, after averaging over the affected answers (Section 6.1) and the three fits within each narrative, per fit estimates are also reported. All intervals are pointwise and conditional on the saved fits.

Table 3: All-pair success over three fixed fits. C/G denotes candidate ranking/global next token percentages, gaps are percentage points. Each objective is scored against its own target. Intervals resample pairs within five equally weighted families. Original boundary mean candidate masses are $2.85/3.23 \times 10^{-5}$ (Qwen intended/alternative) and $.00325/.00299$ (Mistral); with `Answer:`, they are approximately $1/1$ and $.950/.943$. Candidate mass and global correctness are separate diagnostics.

Model	Interface	Intended C/G	Alternative C/G	Global gap [95% CI]
Qwen72	Original	95.67 / 0.00	93.73 / 0.00	+0.00 [+0.00, +0.00]
Qwen72	Answer:	98.13 / 98.16	95.73 / 95.73	+2.42 [+1.67, +3.18]
Mistral24	Original	93.73 / 0.00	83.02 / 0.00	+0.00 [+0.00, +0.00]
Mistral24	Answer:	95.09 / 95.18	87.11 / 87.20	+7.98 [+6.73, +9.27]

5 Behavioural Consequences Beyond the Fitted Answer

5.1 Target success depends on the answer format

On the evaluation pairs (Table 3), candidate-ranking success for the intended and alternative fits is 95.67% and 93.73% in Qwen, and 93.73% and 83.02% in Mistral. Using the same patch and the `Answer:` prefill, next-token success is 98.16% and 95.73% in Qwen, and 95.18% and 87.20% in Mistral. The alternative rule also performing well means that target success does not by itself establish which variable a patch captures. Thus we hold the patch fixed in order to test its effects beyond the target question.

5.2 A fixed patch predicts untrained consequences

Alternative patches, held fixed on the checking set, match event-rewrite forecasts on 97.22% of questions in Qwen and 90.45% in Mistral. The intended patches match every question in Qwen and an average of 99.65% across the three seeds in Mistral (Table 4). Every base narrative, source narrative and transformed narrative question is answered correctly on both the discovery and checking sets by the unpatched models when the prefill is used, so they can do the task without intervention.

The answers show what the alternative patches change. All 120 nonobserver and unrelated-object answers are left unchanged, and the world, search and report answers coherently follow the permuted location in most narratives. Relabelling every answer would have changed the preserved answers too, and where a selective belief edit and an event rewrite predict different world locations, the world answer is mostly consistent with the event rewrite. The report questions are not used for fitting; however, several accounts predict the same report answer.

5.3 PCA separates ordinary transfer from learned remapping

PCA matches every intended-objective forecast on the checking set in both models. Checking accuracy is not increased by fitting. Evaluation on the five-shift pairs shows intended learned-minus-PCA differences of -0.18 and $+0.93$ points in Qwen and Mistral respectively (Table 4). On the shared 1,069-pair clean-opportunity set with the `Answer:`prefill, intended PCA transfer is perfect in both models; the learned patchestrail by 0.20 and 0.36 percentage points in Qwen and Mistral, respectively (Appendix A.5). For the alternative objective, learned patches gain 59.72 and 52.95 points over PCA on the checking set in Qwen and Mistral respectively. PCA receives no signal for the permuted target, so its 37.50% agreement comes solely from the three question types that should stay unchanged or be overwritten by later observations (3/8). The informative result is therefore the learned patches' coherent downstream pattern.

Ordinary PCA is consistent on every checking narrative but never gives the alternative target, while the opposite is true for output-permuted PCA (Appendix B). Learned alternative patches are consistent

Table 4: Answer: global next token percentages and learned minus PCA percentage point differences on all pairs or all checking stories. Three fixed fits, intervals cluster pairs/stories.

Model	Panel / objective	PCA	Learned	Difference [95% CI]
Qwen72	Five shifts / Intended	98.33	98.16	-0.18 [-0.42, +0.07]
Qwen72	Checking / Intended	100.00	100.00	+0.00 [+0.00, +0.00]
Qwen72	Five shifts / Alternative	0.40	95.73	+95.33 [+94.49, +96.11]
Qwen72	Checking / Alternative	37.50	97.22	+59.72 [+58.19, +61.04]
Mistral24	Five shifts / Intended	94.24	95.18	+0.93 [+0.27, +1.64]
Mistral24	Checking / Intended	100.00	99.65	-0.35 [-0.87, +0.00]
Mistral24	Five shifts / Alternative	1.31	87.20	+85.89 [+84.42, +87.27]
Mistral24	Checking / Alternative	37.50	90.45	+52.95 [+50.28, +55.59]

Figure 1: **Native key-only causal test.** Keys are induced by fixed rank-16 patches using the same natural base/source donor pair. At the critical token, the full intervention covers all eight KV groups in blocks 6-40 (Mistral) or 6-80 (Qwen). Values and queries evolve endogenously, attention outputs are not restored. Agreement with actual endpoint outputs includes errors.

Recipient computation	Replace critical-token keys	Compare output with
PCA patch P	$K_P \leftarrow K_M$ (addition)	Actual learned endpoint M
Learned patch M	$K_M \leftarrow K_P$ (removal)	Actual PCA endpoint P

Own keys and signed-permutation nulls supply matched controls; natural K_B tests donor content.

in 119 of 120 narratives in every Qwen fit and in 110, 112 and 110 in the three Mistral fits. However, output-permuted PCA has higher target accuracy on the five-shift evaluation (98.33% versus 95.73% in Qwen and 94.24% versus 87.20% in Mistral).

6 Dissecting the Learned–PCA Difference with Key Exchanges

In Section 5 we showed that the learned alternative patches and unfitted PCA patches lead to different downstream answers. The learned patches propagate the permuted location whereas PCA patches transfer the original location. We now investigate whether the attention keys that each patch induces causally contribute to this difference.

6.1 Reciprocal key-only exchange

Let P and M denote runs with the PCA and learned remapping patches. Let B, S and T denote the base, source and remapped-source narratives, with the critical location in T obtained by applying m_2 to that in S. The critical event is the move or tell event whose span is patched, and the critical location is the location it names. The transformed T is S with its critical locations changed by m_2. Both patches use the original B/S activation pair. In *addition*, the recipient P receives the keys of M, giving the *learned-key hybrid*; in *removal*, the recipient M receives the keys of P. We compare each exchange with the recipient's own keys (the self control) and with a matched null, and score fidelity: agreement with the actual next-token output of the run that donated the keys (M for addition, P for removal). Keys from the natural base narrative B provide a separate control for key content (Figure 1).

The matched null applies a fixed signed coordinate permutation to the donor-minus recipient key displacement within each block and KV group, then adds it to the recipient keys. This preserves the requested displacement norm, but changes its direction (Appendix C.1).

The swap acts only on the keys at the critical token (after the key projection and before RoPE), in all eight key-value heads and across blocks 6-80 in Qwen and 6-40 in Mistral. No values are transplanted and no attention outputs are restored here, so queries, values and residual states can change through later computation.

The same exchanged keys serve five questions: the direct location question and two colour questions each for the observer and the nonobserver. The colour questions use two lookup tables, placed after

Figure 2: **Native key-only effects in fresh stories.** Gain over the matched key null in agreement with the other endpoint's actual outputs, averaged across direct and two observer-colour answers. Large points and pointwise 95% intervals average the three fixed fits within each of 120 cores before 20,000 paired core resamples; small points show each fit. Endpoint fidelity includes copied errors. Absolute correctness and preservation appear in Table 5.

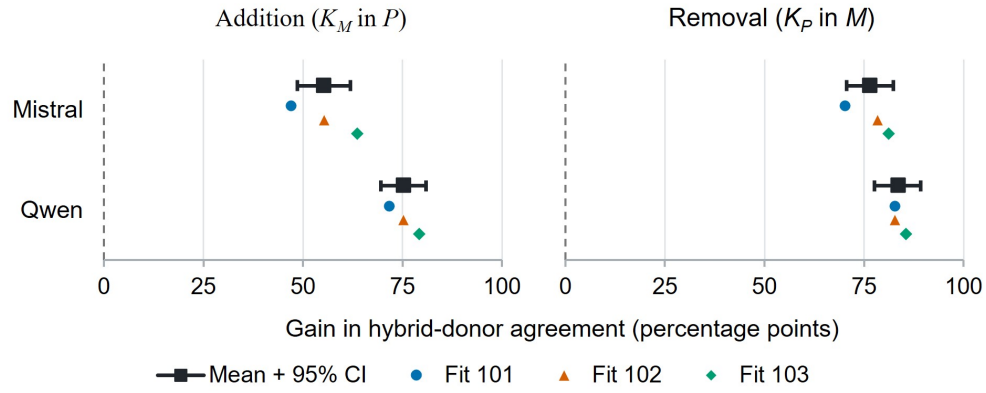

Table 5: **Absolute exchange performance (%).** Fidelity matches actual M outputs for addition and P outputs for removal; it can copy errors. Correctness uses the remapped-target oracle for addition and source oracle for removal. The first two metrics cover 1,080 affected answers (direct and two observer colours); preservation requires both nonobserver answers, and the last column requires all five answers correct, each over 360 core fit patterns. All 120 cores and three fixed fits are included.

Model	Exchange	Endpoint fidelity	Symbolic correctness	Nonobserver pair intact	All five correct
Mistral	Addition	75.4	54.3	100.0	36.7
Mistral	Removal	99.9	99.9	100.0	99.7
Qwen	Addition	89.8	73.9	100.0	61.7
Qwen	Removal	100.0	100.0	100.0	100.0

the critical event, that assign colours to locations. Each question asks for the colour of the location an agent believes the object is in. The observer witnesses the critical event, so the direct and observer answers should change (the *affected* answers). The nonobserver answers should remain unaffected.

6.2 Broad key exchanges transfer behaviour in both directions

On all 120 key-only narratives, addition raises agreement with the actual outputs of M over the matched null by 75.4 (95% CI $[69.5, 80.8]$) percentage points and 55.3 ($[48.6, 61.9]$) in Qwen and Mistral respectively. For removal, agreement with the actual outputs of P is raised by 83.7 ($[77.6, 89.2]$) and 76.6 ($[70.6, 82.3]$) points in Qwen and Mistral. Both directions outperform the self and null controls in every fit (Figure 2, Table 16). Absolute control and exchange fidelities are shown in Table 17.

For addition, the learned-key hybrid also copies errors, so fidelity exceeds correctness. In Qwen, 97.18% of M's incorrect affected answers are reproduced, while Mistral sits at 82.67%. Both observer colours are correct in 79.44% of narrative-fit evaluations in Qwen, and in 60.28% in Mistral.

Keys from the natural base narrative reproduce all five correct answers from B in both recipient runs and every fit, including all 96 narratives with distinct locations. The outcome therefore depends on which run supplies the keys.

Narrower block exchanges exhibit a distributed and asymmetric pattern. With block bands selected on previously examined narratives and confirmed on fresh narratives, neither the selected band nor its complement recovers most of the full-range addition effect. Each recovers at most 11.1 of 75.4

Table 6: **Reciprocal key effects across mappings and models.** The same 108 stories in both models, with three 300 update fits per mapping. Rates are percentages of 972 affected answers or, in the last two columns, of 324 story fit patterns, scoring follows Table 5. Intervals resample stories after averaging fits.

Mapping	Exchange	Gain over null (pp; 95% CI)	Endpoint fidelity	Symbolic correct	Nonobserver preserved	All five correct
Qwen2.5-72B						
m_1 (cycle)	Addition	67.3 [60.4, 74.0]	87.7	64.8	100.0	52.8
m_1 (cycle)	Removal	79.5 [72.3, 86.2]	100.0	100.0	100.0	100.0
m_2 (swap)	Addition	69.9 [62.4, 77.0]	89.3	68.3	100.0	56.8
m_2 (swap)	Removal	80.0 [72.7, 86.9]	100.0	100.0	100.0	100.0
Mistral-Small-24B						
m_1 (cycle)	Addition	48.0 [40.9, 55.3]	74.7	40.7	99.4	26.2
m_1 (cycle)	Removal	72.4 [65.0, 79.5]	100.0	100.0	100.0	100.0
m_2 (swap)	Addition	47.8 [41.0, 54.6]	74.3	45.9	99.7	30.9
m_2 (swap)	Removal	71.3 [64.2, 78.0]	99.9	99.9	100.0	99.7

percentage points in Qwen and 7.0 of 55.3 in Mistral. For removal, either region alone recovers a much larger portion of the effect, at least 51.2 of 83.7 points in Qwen and 51.1 of 76.6 in Mistral (Appendix C.3, Table 19).

The effect recurs for a second fitted mapping. In the separate replication, both exchanges outperform the matched nulls for both mappings in every fit of both models. The gains are also positive when averaged over fits on the 54 narratives where mappings require different targets (Table 6, Appendix D). Addition increases agreement with M by 67.3 and 69.9 points respectively for m_1 and m_2 in Qwen, and 48.0 and 47.8 points in Mistral. Removal increases agreement with P by 79.5 and 80.0 points in Qwen and 72.4 and 71.3 points in Mistral. The effect therefore extends beyond the pair-swap to a second mapping fitted with the same training exposure.

Especially for Mistral, addition remains incomplete. For m_1 and m_2, respectively, all five answers are correct in 52.8% and 56.8% of Qwen's narrative–fit evaluations, compared with 74.1% and 77.5% in the learned-patch runs. The corresponding rates for Mistral are 26.2% and 30.9%, compared with 53.1% and 59.6% in the learned-patch runs.

Addition preserves both nonobserver answers in 99.38% and 99.69% of patterns for m_1 and m_2 in Mistral. Both colour tables commute with the six-cycle m_1, so the fixed-value experiment's 50% recoding bound does not apply to that mapping (Appendix D).

Key exchanges with fixed target values. We also exchange keys while fixing the critical-token values to the natural target values V_T and restoring attention outputs. This supplies target information through attention values. Across 48 primary narratives, addition increases agreement with M over unchanged keys by 92.6 percentage points in Qwen and 73.8 points in Mistral. Removal increases agreement with P by 98.1 and 94.4 points, respectively. In every fit, the learned-key hybrid answers both observer colour questions correctly on more than 24 of the 48 primary stories. This exceeds the exact 50% bound for any fixed transformation that uses only the PCA colour answer and applies the same rule to both tables. The lookup construction and results are detailed in Appendix E.1, Table 26.

7 Discussion and Limitations

Contribution and implications. We connect the downstream effects of learned remapping to a causal role for attention keys, going beyond success on the fitted answer. Experimental results show reciprocal key exchanges transfer substantial parts of the behavioural difference between the learned and PCA patches, without transplanting attention values or restoring attention outputs. Natural base-narrative keys recover the affected base answers; we therefore conclude that the outcome depends on which run is supplying the keys rather than an exchange simply forcing remapped answers.

Together, these results evidence a component-level explanation of the tested intervention effects. The fixed-value experiment adds complementary evidence: paired colour accuracy exceeds the bound for any fixed relabelling that uses only the PCA colour answer and applies the same rule to both tables. Separating fidelity, correctness and preservation also reveals what a successful transfer does not guarantee: key exchanges largely preserve the tested nonobserver information but can reproduce incorrect answers. Our experiments therefore establish both the downstream scope of the learned interventions and a causal contribution from attention keys, without equating behavioural transfer with correctness or unique semantic identification.

Scope of the findings. The key exchange spans 75 and 35 blocks in Qwen and Mistral respectively in a synthetic data setting. Rank and patch site remain fixed throughout the experiments, and we have not tested whether the 24B and 72B results are robust to the rank. We do not claim that keys alone encode location information, only that they play a part.

AI Use Statement

Generative-AI assistants were used to aid debugging of initial prototype experimental code before being updated and reviewed by the authors. It was also used for some grammatical and writing polish, particularly localised to appendix details. The authors take full responsibility for the paper and its contents.

Reproducibility Statement

All experimental code with relevant seeds and datasets are made publicly available at `https://anonymous.4open.science/r/Beyond-the-fitted-Scope---Causal-keys-230B/`.

Ethics Statement

Synthetic datasets are used with publicly released model weights. No human participants or private data was involved. We use ToM datasets as proxy to measure downstream consequences, we make no claims about human cognition and our results should not be interpreted as evidence that models possess human like mental states.

References

Aryaman Arora, Dan Jurafsky, and Christopher Potts. CausalGym: Benchmarking causal interpretability methods on linguistic tasks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14638–14663. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.785. URL https://aclanthology.org/2024.acl-long.785/.

Matteo Bortoletto, Constantin Ruhdorfer, Lei Shi, and Andreas Bulling. Brittle minds, fixable activations: Understanding belief representations in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 22521–22543, 2025.

Atticus Geiger, Zhengxuan Wu, Hanson Lu, Josh Rozner, Elisa Kreiss, Thomas Icard, Noah D. Goodman, and Christopher Potts. Inducing causal structure for interpretable neural networks. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 7324–7338, 2022.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah Goodman. Finding alignments between interpretable causal variables and distributed neural representations. In *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pp. 160–187. PMLR, 2024. URL https://proceedings.mlr.press/v236/geiger24a.html.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th*

International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 2733–2743. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1275. URL https://aclanthology.org/D19-1275/.

Jing Huang, Zhengxuan Wu, Christopher Potts, Mor Geva, and Atticus Geiger. RAVEL: Evaluating interpretability methods on disentangling language model representations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8669–8687, 2024.

Puyin Li, Jiyuan Tan, Ahmad Jabbar, Thomas Icard, and Atticus Geiger. Bucketing the good apples: A method for diagnosing and improving causal abstraction. *arXiv preprint arXiv:2605.02234*, 2026.

Aleksandar Makelov, Georg Lange, 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.

Mistral AI. Mistral-Small-24B-Instruct-2501: Model card. https://huggingface.co/mistralai/Mistral-Small-24B-Instruct-2501, 2025. Accessed September 8, 2026.

Aaron Mueller, Atticus Geiger, Sarah Wiegreffe, Dana Arad, Iván Arcuschin, Adam Belfki, Yik Siu Chan, Jaden Fried Fiotto-Kaufman, Tal Haklay, Michael Hanna, Jing Huang, Rohan Gupta, Yaniv Nikankin, Hadas Orgad, Nikhil Prakash, Anja Reusch, Aruna Sankaranarayanan, Shun Shao, Alessandro Stolfo, Martin Tutek, Amir Zur, David Bau, and Yonatan Belinkov. MIB: A mechanistic interpretability benchmark. In *Proceedings of the 42nd International Conference on Machine Learning*, pp. 45069–45108, 2025.

Nikhil Prakash, Natalie Shapira, Arnab Sen Sharma, Christoph Riedl, Yonatan Belinkov, Tamar Rott Shaham, David Bau, and Atticus Geiger. Language models use lookbacks to track beliefs. In *International Conference on Learning Representations*, 2026.

Oliver Steele, Jiangtao Wen, and Yuxing Han. Belief-reality separation lives in routing over a shared value slot in language models. *arXiv preprint arXiv:2607.11945*, 2026a. URL https://arxiv.org/abs/2607.11945v2. Version 2.

Oliver Steele, Jiangtao Wen, and Yuxing Han. One mechanism for many mental spaces: a shared router over a value slot in language models. *arXiv preprint arXiv:2607.10248*, 2026b. URL https://arxiv.org/abs/2607.10248v2. Version 2.

Sohan Venkatesh and Ashish Mahendran Kurapath. On the non-identifiability of steering vectors in large language models. *arXiv preprint arXiv:2602.06801*, 2026.

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, pp. 78205–78226. Curran Associates, Inc., 2023. doi: 10.52202/075280-3418. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/f6a8b109d4d4fd64c75e94aaf85d9697-Paper-Conference.pdf.

Zhengxuan Wu, Atticus Geiger, Jing Huang, Aryaman Arora, Thomas Icard, Christopher Potts, and Noah D. Goodman. A reply to Makelov et al. (2023)'s "Interpretability Illusion" arguments. *arXiv preprint arXiv:2401.12631*, 2024.

An Yang et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Wentao Zhu, Zhining Zhang, and Yizhou Wang. Language models represent beliefs of self and others. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

A MATCHED-DATA CROSS-MODEL COMPARISON

A.1 MODELS, FITTING AND EXECUTION

What is matched. We compare Qwen2.5-72B-Instruct and Mistral-Small-24B-Instruct-2501 using the same ordered 1,000 training pairs, 1,500 five-shift evaluation pairs, 120 discovery and checking narratives each. Each model has a rank-16 basis for each objective and three seeds. The same training order and PCA initialisation is used within each seed for each of the two objectives. A separate 300-update mapping replication uses sampling with replacement, as described in Appendix D

Both models use their native tokenisers and chat templates. Patching is done at the fourth decoder block for each model. We therefore match semantic inputs, but not token sequences, activations or numerical execution. The comparison does not isolate the effects of model size or architecture.

Fitting. Language-model (LM) weights remain frozen in unquantised BF16. The orthonormal basis uses FP32 parameters, AdamW with learning rate 10^{-3} and zero weight decay, conditional six-choice cross entropy, and gradient clipping at norm 1.

PCA construction. PCA pools full event-span activation differences from all 1,000 training pairs. Two power iterations and a seed $\omega + 16$ is used for centred randomised PCA.

Competence and technical checks. Both clean endpoints are correct on 1,369/1,500 and 1168/1500 pairs for Qwen and Mistral pairs under six-candidate ranking, and on 1,460 and 1,368 pairs with the prefill.

A.2 LAYER AND RANK CHOICE

The fourth block is chosen from preliminary 7B source-transfer studies. A full-prefix interchange sweep over layers 4, 8 and 12 gives the strongest source transfer at layer 4 (Table 7), so we make this design choice.

Table 7: Full-prefix interchange layer sweep in two 7B models on 100 unique pairs. Ctf. is unconditional transfer of the source answer. Conditional rates condition on both clean base and source endpoints being correct; the number of such pairs is in parentheses.

Model	Layer	Base acc.	Source acc.	Ctf.	Changed	Conditional ctf.
Qwen2.5-7B	4	0.800	0.780	0.690	0.670	0.846 (65)
Qwen2.5-7B	8	0.800	0.780	0.570	0.600	0.662 (65)
Qwen2.5-7B	12	0.800	0.780	0.440	0.510	0.492 (65)
Mistral-7B	4	0.860	0.840	0.860	0.840	1.000 (73)
Mistral-7B	8	0.860	0.840	0.520	0.510	0.562 (73)
Mistral-7B	12	0.860	0.840	0.250	0.230	0.233 (73)

Rank choice. We use rank 16 as a fixed intervention budget based on preliminary source-transfer experiments in 7B models. Comparisons of ranks 1, 4, 16 and 64 shows that rank 16 approximately match the full patch effect. In the DAS pilots at layer 4, each evaluated using 200 examples, ranks 16 and 64 achieve source-answer transfer rates of 98.0% and 96.5% in Qwen2.5-7B, and 100% at both ranks in Mistral-7B. PCA patches fitted on the event-sentence span, rank 16 is within 0.2 percentage points of the full patch in both models, and rank 64 adds no benefit (Table 8). We use rank 16 for both objectives and for learned and PCA patches.

Table 8: Source answer transfer at layer 4.

Model	Rank 4	Rank 16	Rank 64	Full patch
Qwen2.5-7B	72.8	96.4	96.2	96.6
Mistral-7B	80.5	99.5	99.5	99.5

A.3 RESPONSE INTERFACE AND SCORING

Candidate-ranking success compares the six location-token scores. Comparing the vocabulary argmax with the target token ID gives us a measure of global next token success. It's possible for the next token to fall outside the candidate vocabulary even when the target ranks first among the other candidates. The `Answer:` prefill changes the response boundary without changing the materialised event patch or refitting its basis.

Documented Qwen tie. Paired Qwen response format studies result in agreement between the global and six choice decisions on 8999/9000 of the prefill observations. Candidate order and vocabulary argmax tie breakers do not agree on the same answer, explaining the 98.13% candidate-ranking and 98.16% global intended rates in Table 3.

A.4 QUESTION TYPES, FIXED PATCHES AND CANDIDATE ACCOUNTS

Both agents are privy to the location of two objects at the start of discovery and checking narratives, only one of these agents can witnesses the critical move. Seven question types are available for discovery: direct belief, world location, the other agent's belief, belief-guided search, a later observed move, a later unobserved move and an unrelated object. The checking set adds truthful reporting of the observer's belief to the other agent, giving eight question types.

For each story and basis, one materialised BF16 patch is reused across questions and subsequent events. Learned and PCA patches use the original source event. A naturally transformed event is a separate full event control.

Candidate accounts. Table 9 gives six candidate accounts of a fixed patch's effects. b and s are the critical locations in the base and source narratives, i the shared initial location, n the later location, and d the unrelated object's location. The mapping π is the identity for the intended objective and the pair-swap for the alternative, with $t = \pi(s)$.

Table 9: Candidate predictions for the seven follow-up questions. The direct belief target is t for every account and is not repeated in the columns. Observed and unobserved refer to the later move; report uses the explicit truthful-report convention.

Account	World	Other	Search	Observed	Unobserved	Irrelevant	Report
Selective belief edit	b	i	t	n	t	d	t
Event rewrite	t	i	t	n	t	d	t
Broad state write	t	t	t	n	t	d	t
Persistent direct override	b	i	b	t	t	d	b
Cleared direct override	b	i	b	n	b	d	b
Global output recoding	t	$\pi(i)$	t	$\pi(n)$	t	$\pi(d)$	t

A.5 COMPLETE MATCHED-DATA RESULTS

Opportunity-matched diagnostics. Model opportunity requires both clean endpoints correct under the specified scorer/interface and both objective targets different from the base answer. Model-specific populations are not directly comparable between models. Means retain equal weight for all five environments.

Table 10: Original opportunity populations and objective success gaps. Each objective is scored against its own target.

Model	Score	Population	n	Intended/alternative (%)	Gap (pp), 95% CI
Qwen72	C	Model	1094	99.64 / 94.19	+5.44 [+4.39, +6.56]
Qwen72	C	Shared	892	99.71 / 95.42	+4.29 [+3.33, +5.30]
Qwen72	G	Model	0	NE	NE
Qwen72	G	Shared	0	NE	NE
Mistral24	C	Model	934	99.72 / 83.87	+15.85 [+14.20, +17.69]
Mistral24	C	Shared	892	99.71 / 83.96	+15.75 [+13.98, +17.57]
Mistral24	G	Model	0	NE	NE
Mistral24	G	Shared	0	NE	NE

Table 11: This table is the same as above but with 10 but uses the prefill..

Model	Score	Population	n	Intended/alternative (%)	Gap (pp), 95% CI
Qwen72	C	Model	1162	99.69 / 96.37	+3.32 [+2.61, +4.09]
Qwen72	C	Shared	1069	99.80 / 96.50	+3.30 [+2.63, +4.05]
Qwen72	G	Model	1162	99.69 / 96.37	+3.32 [+2.61, +4.09]
Qwen72	G	Shared	1069	99.80 / 96.50	+3.30 [+2.63, +4.05]
Mistral24	C	Model	1079	99.61 / 88.18	+11.43 [+10.11, +12.78]
Mistral24	C	Shared	1069	99.64 / 88.48	+11.16 [+9.88, +12.53]
Mistral24	G	Model	1079	99.62 / 88.25	+11.37 [+10.02, +12.71]
Mistral24	G	Shared	1069	99.64 / 88.51	+11.13 [+9.82, +12.51]

Checking overview. Table 12 reports aggregate view, complete-story and report counts for every fit.

Table 12: Prefill global next token and fixed event rewrite forecast agreement with the prefill.

Model	Objective	Seed	All views	Joint stories	Report
Qwen72	Intended	101	960/960	120/120	120/120
Qwen72	Intended	102	960/960	120/120	120/120
Qwen72	Intended	103	960/960	120/120	120/120
Qwen72	Alternative	101	894/960	106/120	106/120
Qwen72	Alternative	102	949/960	117/120	117/120
Qwen72	Alternative	103	957/960	119/120	120/120
Mistral24	Intended	101	960/960	120/120	120/120
Mistral24	Intended	102	960/960	120/120	120/120
Mistral24	Intended	103	950/960	117/120	117/120
Mistral24	Alternative	101	846/960	92/120	97/120
Mistral24	Alternative	102	847/960	94/120	97/120
Mistral24	Alternative	103	912/960	105/120	109/120

Table 13: Qwen checking results per view.

Objective	View	Learned 101	102	103	PCA 101/102/103
Intended	Direct	120	120	120	120 / 120 / 120
Intended	World	120	120	120	120 / 120 / 120
Intended	Other agent	120	120	120	120 / 120 / 120
Intended	Search	120	120	120	120 / 120 / 120
Intended	Observed update	120	120	120	120 / 120 / 120
Intended	Unobserved update	120	120	120	120 / 120 / 120
Intended	Irrelevant object	120	120	120	120 / 120 / 120
Intended	Report	120	120	120	120 / 120 / 120
Alternative	Direct	107	118	119	0 / 0 / 0
Alternative	World	107	118	119	0 / 0 / 0
Alternative	Other agent	120	120	120	120 / 120 / 120
Alternative	Search	107	118	120	0 / 0 / 0
Alternative	Observed update	120	120	120	120 / 120 / 120
Alternative	Unobserved update	107	118	119	0 / 0 / 0
Alternative	Irrelevant object	120	120	120	120 / 120 / 120
Alternative	Report	106	117	120	0 / 0 / 0

Table 14: Mistral checking results.

Objective	View	Learned 101	102	103	PCA 101/102/103
Intended	Direct	120	120	118	120 / 120 / 120
Intended	World	120	120	118	120 / 120 / 120
Intended	Other agent	120	120	120	120 / 120 / 120
Intended	Search	120	120	119	120 / 120 / 120
Intended	Observed update	120	120	120	120 / 120 / 120
Intended	Unobserved update	120	120	118	120 / 120 / 120
Intended	Irrelevant object	120	120	120	120 / 120 / 120
Intended	Report	120	120	117	120 / 120 / 120
Alternative	Direct	95	97	108	0 / 0 / 0
Alternative	World	95	97	110	0 / 0 / 0
Alternative	Other agent	120	120	120	120 / 120 / 120
Alternative	Search	100	99	112	0 / 0 / 0
Alternative	Observed update	120	120	120	120 / 120 / 120
Alternative	Unobserved update	99	97	113	0 / 0 / 0
Alternative	Irrelevant object	120	120	120	120 / 120 / 120
Alternative	Report	97	97	109	0 / 0 / 0

B Retrospective Target and Scope Diagnostics

An observed answer as a consistency anchor. Let $\widehat{z}_{g,\omega} = \widehat{y}_{g,\omega,\mathrm{direct}}$ be the patched direct-belief answer and let $t_g = m_2(y_{s,g})$ for the alternative objective. Write $E(g, v; z)$ for the existing event forecast with event location z. It predicts z for the world, search, unobserved-overwrite and report views, the initial value for the nonobserver, the new value for an observed overwrite, and the distractor's location for the irrelevant object. Define

$$X_{g,\omega} = \mathbf{1}\{\widehat{z}_{g,\omega} = t_g\}, \tag{2}$$

$$K_{g,\omega} = \mathbf{1}\{\widehat{z}_{g,\omega} \in \mathcal{L}\} \prod_{v \neq \mathrm{direct}} \mathbf{1}\{\widehat{y}_{g,\omega,v} = E(g, v; \widehat{z}_{g,\omega})\}. \tag{3}$$

Here $\mathcal{L}$ is the six-location vocabulary. The direct answer is excluded from K. X measures requested target correctness, K measures joint behavioural consistency on the other seven views. Their product is exactly the original eight-target correct indicator. The intended diagnostic substitutes $t_g = y_{s,g}$.

For the alternative objective, the other seven views are jointly consistent with the direct answer ($K = 1$) on 119 of 120 stories in every Qwen fit and on 110, 112 and 110 stories in the three Mistral fits.

Table 15: Target correctness and scope consistency using the alternative checking panel.

	Direct target correct (%)	Other seven jointly consistent (%)	All eight targets correct (%)
PCA (both models)	0.00	100.00	0.00
PCA + output permutation (both)	100.00	0.00	0.00
Learned alternative: Qwen72	95.56	99.17	95.00
Learned alternative: Mistral24	83.33	92.22	80.83

C KEY-ONLY INTERVENTIONS AND STORY CONFIRMATION

This extension asks whether the learned-versus-PCA key difference matters without supplying natural target values or restoring attention outputs in two models of the same story task. It is separate from the fixed-V_T, restored-prefix experiment in Appendix E.2. Here every declared scientific condition runs without a competence or effect-size gate.

C.1 OPERATOR, COHORTS AND ESTIMATION

What is intervened on. Let B, S be natural base/source executions and $T = \pi(S)$ the remapped target. The original fitted intervention replaces a rank-16 component of the complete event after block 4, using only B/S activations. P denotes the PCA-control endpoint and M the remapping-trained endpoint. At the critical value token, we transplant post-projection, pre-RoPE keys in all eight KV groups of the selected blocks. The full path is blocks 6–40 for Mistral 24B and 6–80 for Qwen 72B. No query, value or attention-output row is directly replaced or restored. Subsequent queries, values and residual states may change endogenously: "native" does not mean numerically fixed. Unselected keys come from the live computation. Thus this tests a causal contribution under the original activation intervention, not necessity in the unmodified model.

In each recipient frame, four arms use its own keys (self), the other patched endpoint's keys (other), a signed-permutation null, or natural B keys. We call $M \to P$ *key addition* and $P \to M$ *key removal*; these labels denote the direction of key exchange. For recipient keys K_R and opposite-endpoint donor keys K_D, the null applies a fixed signed coordinate permutation $S_{\ell,g}$ to the donor-minus-recipient displacement within each block ℓ and KV group g:

$$K^{\text{null}}_{\ell,g} = K_{R,\ell,g} + S_{\ell,g}\big(K_{D,\ell,g} - K_{R,\ell,g}\big). \tag{4}$$

Table 16: All four full-path primary contrasts: percentage-point changes in actual opposite-endpoint fidelity across three affected answers per core. All 120 cores, 20,000 paired core draws."Add" and "remove" denote the direction of key exchange.

Contrast	Fit mean [95% CI]	Fit 101	Fit 102	Fit 103
Story: Mistral				
Add−self	55.28 [48.61, 61.85]	46.94 [39.17, 55.00]	55.28 [47.78, 62.78]	63.61 [56.67, 70.28]
Add−null	55.28 [48.61, 61.85]	46.94 [39.17, 55.00]	55.28 [47.78, 62.78]	63.61 [56.67, 70.28]
Remove−self	79.81 [73.80, 85.46]	75.00 [67.78, 81.94]	79.17 [72.22, 85.56]	85.28 [79.17, 90.83]
Remove−null	76.57 [70.56, 82.31]	70.28 [62.50, 77.78]	78.33 [71.11, 85.28]	81.11 [74.44, 87.22]
Story: Qwen				
Add−self	75.37 [69.54, 80.83]	71.67 [64.72, 78.33]	75.28 [68.06, 81.94]	79.17 [72.50, 85.28]
Add−null	75.37 [69.54, 80.83]	71.67 [64.72, 78.33]	75.28 [68.06, 81.94]	79.17 [72.50, 85.28]
Remove−self	85.56 [79.72, 90.83]	85.56 [79.17, 91.39]	84.44 [78.06, 90.28]	86.67 [80.83, 92.50]
Remove−null	83.70 [77.59, 89.17]	82.78 [76.11, 89.17]	82.78 [75.83, 89.17]	85.56 [79.17, 91.39]

C.2 STORY TRANSFER, ERRORS AND PRESERVATION

The full-path effects in Table 16 are positive against both controls in each story model and fit. Natural $B/S/T$ and P are correct throughout the five-view panel, learned M sometimes acquires the wrong value or produces inconsistent consequences. Consequently, addition's endpoint agreement is higher

Table 17: Full-path agreement with actual opposite-endpoint outputs (%).

Model	Direction	Self	Signed null	Exchanged keys
Qwen	Addition	14.4	14.4	89.8
Qwen	Removal	14.4	16.3	100.0
Mistral	Addition	20.1	20.1	75.4
Mistral	Removal	20.1	23.3	99.9

than target correctness (Table 18). Removal is much more complete: Mistral reproduces the source on 1079/1080 affected answers across the fixed fits, and Qwen on 1080/1080. The single Mistral miss is a direct answer in the $B = T$ stratum; observer pairs remain correct.

Table 18: Narrative other-key arms, all 120 cores. Affected answers compare actual M IDs on addition and P IDs on removal, symbolic correctness targets T and S, respectively.

Model	Direction	Actual endpoint	Oracle answers	Complete oracle
Mistral	Addition	75.37 [70.28, 80.28]	54.26 [47.41, 61.02]	36.67 [28.89, 44.72]
Mistral	Removal	99.91 [99.72, 100.00]	99.91 [99.72, 100.00]	99.72 [99.17, 100.00]
Qwen	Addition	89.81 [87.13, 92.31]	73.89 [68.06, 79.35]	61.67 [54.72, 68.61]
Qwen	Removal	100.00 [100.00, 100.00]	100.00 [100.00, 100.00]	100.00 [100.00, 100.00]

C.3 Development-selected localisation

Localisation uses the 12 exposed development cores and fit 101. Eight contiguous, nearly equal bands partition the full path, assigning remainder blocks earliest. Every band alone and every leave-one-band-out mask is tested in both frames, all four arms and all five views.

Each direction nominates the band maximising other minus null affected-ID fidelity minus nonobserver disagreement with the actual recipient.

Table 19: Frozen localisation confirmation, all 120 cores and all three fits. Each entry is other-minus-matched-null actual-endpoint fidelity. Both development directions nominated the same band, duplicates are evaluated once in both frames. All masks retain eight KV groups.

Model	Mask	Blocks	Addition	Removal
Mistral	Nominated	6–10	6.20 [3.15, 9.81]	51.11 [44.26, 58.06]
Mistral	Complement	11–40	7.04 [3.52, 11.11]	58.33 [51.67, 64.72]
Mistral	Spatial control	23–27	0.00 [0.00, 0.00]	1.11 [0.56, 1.76]
Qwen	Nominated	45–53	4.44 [1.94, 7.41]	51.20 [44.35, 57.96]
Qwen	Complement	6–44, 54–80	11.11 [6.76, 16.02]	58.43 [50.74, 65.93]
Qwen	Spatial control	15–23	0.00 [0.00, 0.00]	8.61 [5.00, 12.78]

D Robustness Across Two Fitted Mappings

This appendix repeats the key-only exchange for a six-cycle m_1 and the pair-swap m_2: first in Qwen with earlier fits, then in Mistral with new fits on the same stories.

D.1 Qwen2.5-72B: 300-update fits

Question and separate fitting cohort. Does the reciprocal key effect recur for a different fitted objective? We compare a six-cycle m_1 with a pair-swap m_2. In one-line notation, their images are $[m_1(0), \ldots, m_1(5)] = [1, 2, 3, 4, 5, 0]$ and $[m_2(0), \ldots, m_2(5)] = [1, 0, 3, 2, 5, 4]$, in the location order box, basket, shelf, drawer, cabinet, closet. Both use the same Qwen72B 300-update fitting cohort, rank-16, ordered 1,000-pair training pool and saved fits 101-103.Each fit trains one PCA-initialised orthonormal basis at the block-4 event span with model weights frozen in BF16. Optimisation

samples one cached pair with replacement at each of 300 single-pair updates, exposing 252, 265 and 269 distinct pairs in seeds 101–103. The two mappings share the schedule within each seed. These fits form a separate cohort from the 1,000-update fits, so we do not pool their rates or attribute differences across cohorts to training duration.

Table 20: Paired gains in actual endpoint agreement (percentage points), over the three affected answers.

Mapping	Exchange	Other − null	Other − self	Other − null fits 101/102/103
m_1 (cycle)	Addition	67.3 [60.4, 74.0]	67.3 [60.4, 74.0]	67.3/66.7/67.9
m_1 (cycle)	Removal	79.5 [72.3, 86.2]	79.6 [72.3, 86.3]	78.1/79.3/81.2
m_2 (swap)	Addition	69.9 [62.4, 77.0]	69.9 [62.4, 77.0]	68.2/75.3/66.0
m_2 (swap)	Removal	80.0 [72.7, 86.9]	80.6 [73.3, 87.3]	78.4/82.7/79.0

Reciprocal effects and absolute outcomes. All primary mean and per-fit key effects are positive relative to both self and null controls (Table 20). Addition's absolute fidelity is 87.7% for m_1 and 89.3% for m_2, while target correctness is 64.8% and 68.3%. Removal reaches 100% actual-P agreement and source correctness for both mappings. Both nonobserver answers remain correct on every core, fit, mapping and native arm; no invalid global predictions occur. Natural $B/S/T$ endpoints are correct on all five questions, and P is source-correct throughout. The unexchanged M endpoints achieve 736/972 and 758/972 target-correct affected answers and 240/324 (74.1%) and 251/324 (77.5%) fully correct patterns for m_1/m_2, respectively. Addition produces 630/972 and 664/972 correct affected answers, and 171/324 and 184/324 fully correct patterns.

Table 21: Input-defined secondary populations, without filtering the primary 108-core cohort. Gains are other-minus-null endpoint agreement. Disagreement is the common 54-core subset with unequal mapping targets, and overlaps the three structural strata. $B = T$ and distinct memberships depend on the mapping.

Map	Population	n	Addition gain	Removal gain	Addition correct (%)
m_1	Distinct	72	76.1 [71.1, 80.9]	94.3 [91.5, 96.8]	72.4
m_1	$B = S$	18	0.0 [0.0, 0.0]	0.0 [0.0, 0.0]	0.0
m_1	$B = T$	18	99.4 [98.1, 100.0]	100.0 [100.0, 100.0]	99.4
m_1	Disagreement	54	67.3 [57.4, 76.7]	81.3 [71.0, 90.3]	65.2
m_2	Distinct	72	79.8 [73.8, 85.2]	95.1 [91.5, 98.0]	77.5
m_2	$B = S$	18	0.0 [0.0, 0.0]	0.0 [0.0, 0.0]	0.0
m_2	$B = T$	18	100.0 [100.0, 100.0]	100.0 [100.0, 100.0]	100.0
m_2	Disagreement	54	69.1 [58.4, 79.2]	78.6 [68.1, 88.1]	68.9

Natural-base specificity and scores. Natural-B keys recover both the actual natural-B global IDs and the symbolic B answers on all 972 affected answers and all 324 complete patterns, in both recipient frames and under both mappings. This also holds in each fixed fit and structural/disagreement subset. Because the 72 distinct cases separate B from S and T, recovery cannot be explained entirely by the collision cases.

Output recoders and claim boundary. The two colour tables commute with the six-cycle m_1. Its table-conjugate forward colour recoders therefore answer both observer questions correctly on every story when applied to the source-correct PCA outputs, but corrupt every nonobserver pair. For m_2, each table-specific fixed forward recoder gives zero jointly correct observer pairs. The 50% bound in Appendix E.1 is established for the fixed-value experiment's balanced pair-swap panel.

D.2 MISTRAL-SMALL-24B: 300-UPDATE FITS

Design and initialisation. After the Qwen result, we fixed a Mistral replication of both mappings on the same semantic pilot and confirmation stories. These stories had been used for Qwen but not

for Mistral, so they are not a second independent sample. The earlier Mistral pair-swap fits used 1,000 updates and are not matched to the 300-update Qwen suite, so we fit six new interventions: m_1 and m_2 at each of seeds 101–103, with 300 updates each. The ordered 1,000-pair training pool and the sampling-with-replacement schedule match the Qwen suite; within each seed, the two mappings share the schedule and PCA initializer. The Mistral patch, optimiser and scoring recipe remain model specific, with frozen pretrained parameters, rank-16 and the block-4 event span. Sources, fitting, both evaluation designs and the analysis were fixed before fitting, and all nine saved bases are used without selection on losses or outcomes. Evaluation uses the same five consumers and four donor arms as Qwen, with all eight KV groups in blocks 6-40, endogenous queries and values, and no attention-output restoration.

Table 22: New Mistral 300-update fits: actual-endpoint fidelity gains in percentage points, all 108 cores and all three fits. Addition targets actual M outputs; removal targets actual P. Intervals are pointwise and conditional on the fixed fits.

Map	Contrast	Fit-mean [95% CI]	Fit 101	Fit 102	Fit 103
m_1	Add − null	48.0 [40.9, 55.3]	49.7 [41.7, 57.7]	47.5 [39.5, 55.9]	46.9 [39.2, 54.9]
m_1	Add − self	48.0 [40.9, 55.3]	49.7 [41.7, 57.7]	47.5 [39.5, 55.9]	46.9 [39.2, 54.9]
m_1	Remove − null	72.4 [65.0, 79.5]	74.7 [66.7, 82.1]	71.9 [63.6, 79.6]	70.7 [62.3, 78.7]
m_1	Remove − self	73.4 [65.9, 80.3]	75.3 [67.3, 83.0]	72.5 [64.2, 80.2]	72.2 [64.2, 79.9]
m_2	Add − null	47.8 [41.0, 54.6]	50.9 [42.9, 59.0]	46.6 [38.3, 54.9]	46.0 [38.0, 54.0]
m_2	Add − self	47.8 [41.0, 54.6]	50.9 [42.9, 59.0]	46.6 [38.3, 54.9]	46.0 [38.0, 54.0]
m_2	Remove − null	71.3 [64.2, 78.0]	73.8 [66.0, 81.2]	71.0 [63.0, 78.7]	69.1 [60.8, 77.2]
m_2	Remove − self	73.5 [66.3, 80.1]	75.6 [67.9, 82.7]	70.7 [63.0, 78.1]	74.1 [66.0, 81.8]

Reciprocal effects and completeness. All primary and per-fit intervals in Table 22 have positive lower endpoints. Addition's null-relative gains are 48.0 points for m_1 and 47.8 for m_2; removal gives 72.4 and 71.3. Addition fidelity is 74.7% and 74.3%, but target correctness is only 40.7% and 45.9%, and all five answers are correct in 26.2% and 30.9% of story–fit patterns. The unexchanged learned endpoints are themselves target-correct on 60.0% and 66.2% of affected answers and fully correct on 53.1% and 59.6% of patterns. Removal recovers 100% and 99.9% of actual-P answers and source correctness; the sole m_2 error is a direct answer in fit 101. Natural $B/S/T$ and P are correct on all five questions, and no invalid global prediction occurs.

Errors, preservation and specificity. Addition copies most learned-endpoint errors while repairing few and damaging some correct answers. Summed over fits, 334 of the 389 wrong affected m_1 answers are copied, 4 are repaired and 51 become another wrong answer, while 191 correct answers become wrong; for m_2 the counts are 279 of 329 copied, 3 repaired, 47 otherwise wrong and 200 damaged.

Table 23: Mistral secondary populations. Effects are other-minus-null A3 endpoint fidelity.

Map	Population	n	Addition gain	Removal gain	Addition correct (%)
m_1	Distinct	72	49.2 [41.8, 56.6]	84.0 [78.4, 89.2]	38.4
m_1	$B = S$	18	0.0 [0.0, 0.0]	0.0 [0.0, 0.0]	0.0
m_1	$B = T$	18	91.4 [82.1, 100.0]	98.8 [96.3, 100.0]	90.7
m_1	Disagreement	54	46.3 [36.2, 56.6]	74.1 [63.4, 84.0]	37.0
m_2	Distinct	72	47.1 [40.9, 53.4]	81.9 [77.0, 86.6]	44.1
m_2	$B = S$	18	0.0 [0.0, 0.0]	0.0 [0.0, 0.0]	0.0
m_2	$B = T$	18	98.8 [96.3, 100.0]	100.0 [100.0, 100.0]	98.8
m_2	Disagreement	54	49.8 [39.7, 59.9]	73.9 [63.8, 82.9]	47.9

Populations, scores and recoding. The effects remain positive on the distinct stories and on the 54-story mapping-disagreement subset (Table 23); on $B = S$, endpoint agreement can be perfect while the gain and target correctness are zero. The paired $m_1 - m_2$ null-relative differences are 0.2 $[-5.7, 6.1]$ points for addition and 1.1 $[-3.1, 5.1]$ for removal, which do not establish equivalence.

Scope. Mistral's addition effects are smaller than Qwen's, while both models show strong removal. This is a descriptive comparison on the same stories; the model, path length and PCA provenance

differ, and no cross-model test is made. Together the two models support recurrence of a reciprocal native key contribution for both mappings in this story domain.

E Colour Lookups and the Fixed-Value Key Experiment

E.1 Colour lookup task and the fixed-recoding bound

For the ordered locations (box, basket, shelf, drawer, cabinet, closet), table g_1 assigns (red, blue, green, yellow, black, white), and table g_2 assigns (blue, green, yellow, black, white, red). Each table appears after the unchanged critical event. The question asks which colour the table assigns to the location where the named agent believes the object is. Observer answers are $g_j(x)$ for natural event value x; nonobserver answers are $g_j(I)$ for the unchanged initial location I. Naming both agents prevents always using the newly mentioned location from satisfying the entire task. The lookup instruction was never part of fitting, and its colour alphabet is disjoint from the location alphabet. On 12 development stories, unpatched Mistral answered all 216 natural questions correctly: the base, source and transformed stories, each with the direct and other-agent location questions and both colour tables for each agent. Natural competence in both models on the fixed-value panel is reported in Appendix E.2.

Table 24: The two post-event lookup tables. The same colour answer is relabelled by the same function h in both tables when testing a table-independent recoder.

	box	basket	shelf	drawer	cabinet	closet
g_1	red	blue	green	yellow	black	white
g_2	blue	green	yellow	black	white	red

The paired tables allow a direct test against fixed output recoding. The induced colour permutations $h_j = g_j \pi g_j^{-1}$ have disjoint swap matchings, so either fixed recoder h_1 or h_2, applied to the source colours, completes no pair; these are the two output recoders of Appendix E.2. More generally, suppose the PCA colour answers are the source colours and the source locations are balanced. Let h be any single table-independent function of the PCA colour answer. Index locations and colours by $i \in \{0, \ldots, 5\}$, so that $g_1(i) = i$, $g_2(i) = i + 1 \pmod 6$, and $\pi = m_2 = (0\ 1)(2\ 3)(4\ 5)$. Answering both observer questions correctly for source i requires

$$h(i) = \pi(i), \qquad h(i+1) = \pi(i) + 1 \pmod 6. \tag{5}$$

Success for source $i + 1$ would instead require $h(i + 1) = \pi(i + 1)$. For the pair-swap, $\pi(i + 1) \neq \pi(i) + 1 \pmod 6$ for every i. Adjacent sources therefore cannot both succeed. Successful sources form an independent set of the six-cycle, which has size at most three. With balanced sources, joint correctness is at most $3/6 = 50\%$.

The bound is attained: $h(i) = i + 1 \pmod 6$ succeeds for the three even-indexed sources. Enumerating all $6^6 = 46,656$ functions gives the same maximum. On the 48 balanced primary stories, no such function can therefore answer both observer questions correctly on more than 24 stories.

E.2 Key exchange with fixed natural values

Design. This experiment tests whether changing the critical-token keys between the learned endpoint M and its PCA comparator P remains effective when the critical-token values are fixed at the natural target values V_T. The primary population contains the 48 stories whose base, source and transformed stories are distinct; the other 24 are collision cases outside the primary analysis. The five views are the direct location question and two observer and two nonobserver colour questions (Appendix E.1); the direct answer and the observer colours form the three affected answers, and the nonobserver colours test preservation.

Each learned basis and seed-matched PCA comparator builds a fourth-block patch from natural B/S activations, reused across the five views; natural $T = \pi(S)$ supplies downstream references only. Post-projection keys before RoPE and values at the critical token are replaced in all eight KV groups of blocks 6–40 in Mistral (exact saved PCA initializer) and 6–80 in Qwen (reconstructed PCA), and

Table 25: Four fixed-value fidelity contrasts on the 48 distinct stories.

Model	Effect	101 [CI]	102 [CI]	103 [CI]	Fit mean [CI]
Mistral	Add: changed − unchanged	0.674 [0.521, 0.812]	0.646 [0.472, 0.806]	0.896 [0.812, 0.965]	0.738 [0.627, 0.840]
Mistral	Add: changed − null	0.646 [0.493, 0.785]	0.604 [0.438, 0.757]	0.861 [0.778, 0.931]	0.704 [0.597, 0.803]
Mistral	Remove: changed − unchanged	0.938 [0.868, 0.993]	0.910 [0.833, 0.972]	0.986 [0.958, 1.000]	0.944 [0.898, 0.981]
Mistral	Remove: changed − null	0.868 [0.778, 0.944]	0.910 [0.840, 0.965]	0.917 [0.840, 0.979]	0.898 [0.843, 0.947]
Qwen	Add: changed − unchanged	0.882 [0.764, 0.972]	0.938 [0.854, 1.000]	0.958 [0.875, 1.000]	0.926 [0.875, 0.970]
Qwen	Add: changed − null	0.882 [0.764, 0.972]	0.938 [0.854, 1.000]	0.958 [0.875, 1.000]	0.926 [0.875, 0.970]
Qwen	Remove: changed − unchanged	0.965 [0.910, 1.000]	0.979 [0.944, 1.000]	1.000 [1.000, 1.000]	0.981 [0.961, 0.998]
Qwen	Remove: changed − null	0.951 [0.889, 1.000]	0.951 [0.896, 0.993]	1.000 [1.000, 1.000]	0.968 [0.938, 0.991]

The primary changed/unchanged/null key contrasts fix absolute natural V_T; B/B uses natural B values. Fidelity compares actual destination global IDs within each model, including errors. Means first average the three fits within core. Intervals are pointwise 95% intervals from 20,000 whole-core draws, fits are not independent bootstrap units. The affected set contains three answers per story: direct and both observer colours.

recipient attention outputs are restored through that token. Queries are not replaced directly, so later computation can respond to the exchange. In each recipient P or M and each fit, the keys are the recipient's own, the other endpoint's, a signed-permutation null or natural K_S and K_T, all with V_T; a sixth condition uses natural K_B with V_B. Addition in P scores agreement with actual M outputs, and removal in M scores agreement with actual P outputs, over the three affected answers. Each is compared with unchanged keys and with the null, giving four effects. Paired differences are averaged over fits within story and resampled over the 48 primary stories (20,000 draws).

Effects. All four primary effects are positive in every fit in both models, with positive pointwise lower bounds (Table 25). Mistral's fit-average addition gains are 0.738 over unchanged keys and 0.704 over null; removal gains are 0.944 and 0.898. Qwen's corresponding gains are 0.926, 0.926, 0.981 and 0.968. Qwen addition obtains 137/144, 141/144 and 141/144 faithful affected answers in fits 101–103, compared with 10/144, 6/144 and 3/144 for both unchanged keys and null; removal obtains 144/144 in every fit in both models. Every exchange condition preserves both nonobserver answers on all 72 stories in both models, and natural base keys and values reproduce every complete base pattern and its oracle in both recipients and every fit.

Colour lookups and fixed output recoding. On the primary stories, target-correct observer pairs under addition are 38/39/44 out of 48 in Mistral and 44/46/48 in Qwen; actual M gives 37/34/46 and 42/46/47. Actual PCA, unchanged keys, the key null and both fixed recoders give zero in every fit, and the recoders also corrupt all nonobserver pairs, whereas the exchange conditions preserve them (Table 26). Both premises of the bound in Appendix E.1 hold, so no single table-independent recoding of the PCA colour answers can solve more than 24 of the 48 primary pairs; addition exceeds this bound in every fit. The bound does not cover table-, question- or story-dependent programs. Fidelity and correctness also differ here: addition copies some of M's erroneous observer pairs (7/8/1 in Mistral and 4/1/0 in Qwen across fits), and its mean correctness gain over M is only 0.028 $[-0.035, 0.097]$ in Mistral and 0.021 $[0, 0.049]$ in Qwen.

Scope. These results support a causal increment from these particular learned key changes, relative to their PCA comparators, with values held at natural V_T and attention outputs restored.

Table 26: Complete observer-pair target correctness in all 48 distinct cores.

Model	Output	Correct 101/102/103	Fit mean [CI]
Mistral	Actual P	0/0/0	0.000 [0.000, 0.000]
Mistral	Actual M	37/34/46	0.812 [0.722, 0.896]
Mistral	K_M/V_T in P	38/39/44	0.840 [0.764, 0.910]
Mistral	$h_1(P)$	0/0/0	0.000 [0.000, 0.000]
Mistral	$h_2(P)$	0/0/0	0.000 [0.000, 0.000]
Mistral	Best fixed colour map	24/24/24	Exact panel maximum
Qwen	Actual P	0/0/0	0.000 [0.000, 0.000]
Qwen	Actual M	42/46/47	0.938 [0.896, 0.972]
Qwen	K_M/V_T in P	44/46/48	0.958 [0.924, 0.986]
Qwen	$h_1(P)$	0/0/0	0.000 [0.000, 0.000]
Qwen	$h_2(P)$	0/0/0	0.000 [0.000, 0.000]
Qwen	Best fixed colour map	24/24/24	Exact panel maximum

The primary changed/unchanged/null key contrasts fix absolute natural V_T; B/B uses natural B values. Fidelity compares actual destination global IDs within each model, including errors. Means first average the three fits within core. Intervals are pointwise 95% intervals from 20,000 whole-core draws; fits are not independent bootstrap units. Every denominator is 48 pairs per fit. A pair is correct only when both colours equal the T oracle. Recoders act on actual PCA outputs. The best-map row enumerates all 6^6 functions or uses the balanced-source theorem only when its actual-output premise is verified; it has no bootstrap interval. Table-aware programs are outside this family.