AFFECT ON DEMAND: CAUSAL RETRIEVAL OF CHARACTER AFFECT IN A LANGUAGE MODEL

Anonymous authors
Paper under double-blind review

ABSTRACT

Language models can use a character's earlier success or failure after several neutral sentences, but behavioral success leaves open a mechanistic question: is the relevant affect-linked state carried continuously, or retrieved from the earlier episode when it becomes useful? We distinguish these possibilities with a prospectively frozen four-stage intervention pipeline in Qwen2.5-1.5B-Instruct. A held-out representational confirmation first finds both persistent maintenance and cue-dependent reinstatement (maintenance $= 1.648$ diagnostic-TRAIN SD; retrieval-dependent index $= 0.289$; both Holm-adjusted $p < 2.1 \times 10^{-5}$), showing that decoding alone does not select a single mechanism. We then intervene on source-span key/value states. Opposite-affect source patching at a development-selected layer produces positive cue-specific causal restoration ($\mathrm{CRE} = 0.0397$, 95% CI $[0.0371, 0.0424]$) and strong source-over-bridge specificity (0.0593, $[0.0566, 0.0620]$). Head-level development localizes a two-KV-head pathway (L12H0, L13H0); protected confirmation supports perturbation-relative necessity, matched-head specificity, and restoration rescue. Finally, an update experiment shows selective retrieval after irrelevant suppression ($R = 0.186$, $[0.127, 0.245]$), while a contradictory update does not produce the prospectively frozen outdating signature. Across levels of analysis, the evidence supports a hybrid mechanism: affect-linked information remains decodable across discourse, yet later use depends causally on retrieving source-bound states through a compact pathway rather than simply reading out a continuously overwritten scalar state.

1 INTRODUCTION

When a character succeeds or fails early in a story, a language model may later need that information after several neutral sentences have intervened. Where is the relevant state in the meantime? Behavioral accuracy cannot distinguish at least three internal strategies: the model may continuously maintain an affect-linked state across intervening text, preserve a source-bound trace and retrieve it when demanded, or combine both.

This distinction matters because *decodability is not mechanism*. A linear probe can show that information is available to a readout without establishing that the decoded state is used by the model. Conversely, causal interventions can reveal dependence on internal components, but localization depends on the intervention, metric, and selection procedure (Vig et al., 2020; Meng et al., 2022; Wang et al., 2023; Zhang and Nanda, 2023). Recent mechanistic work has identified entity-tracking circuits (Prakash et al., 2024), multi-step factual-recall pathways (Geva et al., 2023), sparse retrieval heads for long-context factuality (Wu et al., 2025), and query-time "lookback" computations for character beliefs (Prakash et al., 2026). In parallel, emotion-focused studies show that affective information is linearly structured, localized, and steerable in language-model representations (Tak et al., 2025; Avsian et al., 2025). What is not yet established is how affect-linked information about a *particular earlier episode* is carried across neutral discourse and causally recruited when it becomes relevant again.

We address this question with **Affect on Demand**, a sequential experiment program that separates four inferential levels: representational signatures, causal source intervention, pathway localization, and dynamic updating. Development choices are frozen before protected confirmation, and inference is performed over held-out scenario families rather than individual sentences. The resulting design

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.

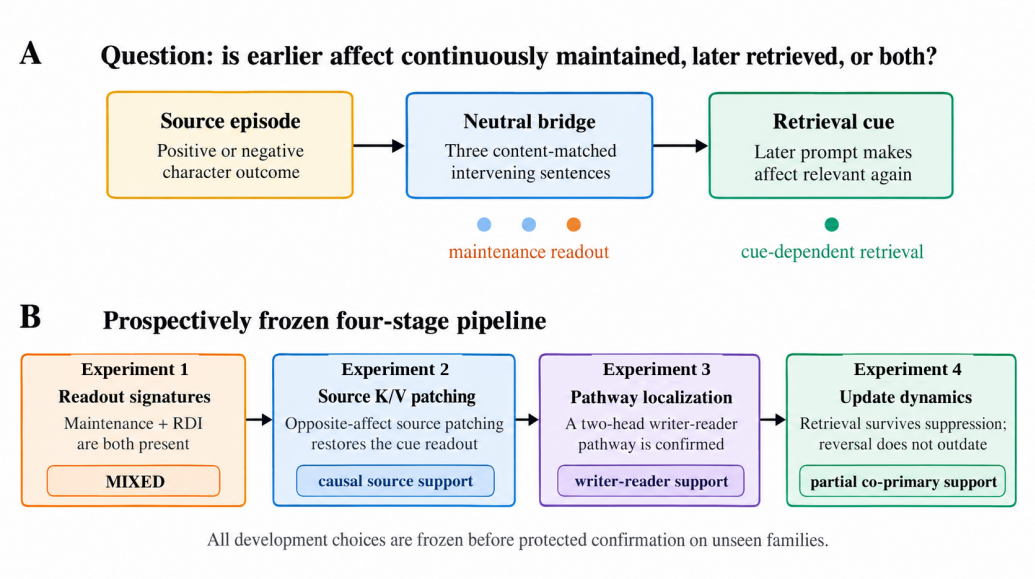

Figure 1: Overview of the task and frozen experimental program. **A**: A positive or negative source episode is followed by a matched neutral bridge and a later retrieval cue; readouts distinguish maintained separation from cue-dependent reinstatement. **B**: The paper advances through a prospectively frozen four-stage pipeline from representational confirmation to source intervention, pathway localization, and update dynamics.

asks not merely *whether* affect is decodable, but *when* it is available, *where* causal information enters the computation, and *which* K/V components carry the source-to-cue effect.

Across four experiments, the evidence converges on a hybrid retrieval mechanism. **First**, sustained maintenance and cue-dependent reinstatement coexist, so the representational trajectory alone does not identify a pure maintenance or pure retrieval strategy. **Second**, patching source-span K/V states causally shifts the later cue representation, whereas matched bridge patching does not. **Third**, the source effect concentrates in a compact two-head K/V pathway that survives protected necessity, matched-control specificity, and restoration tests. **Fourth**, the earlier orientation can be recovered after an irrelevant suppressing update, while contradiction does not behave like a simple overwrite operation. Together, these results move the claim from "affect is encoded" to a stronger mechanistic statement: source-bound affect information makes a causal, cue-specific contribution through a sparse retrieval pathway.

Our main contributions are: **(i)** a prospectively frozen test that separates maintenance from cue-dependent reinstatement at the family level; **(ii)** a source-versus-bridge K/V intervention that converts a decoding result into causal evidence; **(iii)** a development-to-confirmation procedure that localizes and independently tests a compact writer-reader pathway; and **(iv)** an update experiment that distinguishes recovery after suppression from contradiction-driven rewriting. Figure 1 summarizes the experimental logic.

2 Related Work

State tracking and retrieval. Language models can track changing entity states, but the algorithms that implement state tracking vary across training regimes and tasks (Kim and Schuster, 2023; Li et al., 2025). Mechanistic studies have identified reusable entity-tracking circuits (Prakash et al., 2024), multi-stage information flow for factual recall in which later representations query earlier subject information (Geva et al., 2023), sparse retrieval heads that causally support long-context factuality (Wu et al., 2025), and lookback mechanisms that recover information relevant to a character's beliefs at query time (Prakash et al., 2026). Our setting asks a complementary question: whether an affect-

linked consequence tied to one discourse participant is continuously maintained, retrieved from its source episode, or both.

Causal mechanistic interpretability. Causal mediation, activation patching, and circuit-level interventions are widely used to move beyond correlational probes (Vig et al., 2020; Meng et al., 2022; Wang et al., 2023; Conmy et al., 2023). At the same time, patching conclusions can vary with corruption scheme, metric, and intervention choice (Zhang and Nanda, 2023); causal abstractions provide a broader framework for relating internal interventions to computational hypotheses (Geiger et al., 2021). We therefore separate development from protected confirmation, freeze selection rules before opening confirmation outcomes, and pair the selected pathway with matched controls and restoration tests.

Affective representations in language models. Recent work finds linearly recoverable, localized, and manipulable affective structure in language models (Tak et al., 2025; Avsian et al., 2025). Those studies primarily characterize the geometry or causal manipulability of emotion-related representations. Our target is different: an affect-linked *episodic consequence* embedded in a discourse, followed by neutral intervening text and a later retrieval demand. The central question is therefore not whether affective structure exists, but how information from an earlier character-specific episode contributes to a later computation.

3 Experimental Framework

3.1 Model and Readout

All experiments use Qwen2.5-1.5B-Instruct, a 1.54B-parameter causal transformer with 28 layers and grouped-query attention, at a fixed model revision (Qwen et al., 2024). Model weights are never updated. We fit an L2 logistic probe to distinguish positive from negative source consequences using diagnostic families only. Candidate hidden-state indices are $\{4, 8, 12, 16, 20, 24, 28\}$ and $C \in \{.01, .1, 1, 10, 100\}$. Standardization is fit on TRAIN only; a single layer/C pair is chosen on VALIDATION by maximizing the minimum balanced accuracy across bridge-like and cue-like positions, with frozen tie-breaks. TEST is opened once. Downstream readouts are standardized in diagnostic-TRAIN SD units.

The diagnostic gate requires balanced accuracy $\geq .65$, a bootstrap 95% CI lower bound $> .55$, and permutation $p \leq .05$ at both bridge-like and cue-like sites. It passed (bridge-like BA $= .9875$, 95% CI $[.958, 1]$; cue-like BA $= 1.0$, 95% CI $[1, 1]$; both $p < 10^{-4}$). Lexical-identity, binding, factual-retrieval, tokenizer, structural, and scenario-independence controls were also required before the protected core could be opened (Appendix C).

3.2 Family-Level Confirmatory Inference

The family, not the individual sentence, is the inferential unit. Unless otherwise noted, protected confirmations use $n = 128$ unseen families. Confidence intervals use 20,000 family bootstrap draws; directional tests use 100,000 family sign flips/permutations. Frozen co-primary families of tests use Holm correction. No imputation or outcome-based exclusion is allowed. The complete decision rules and seeds are listed in Appendix B.

3.3 Sequential Design

The program advances only when the preceding frozen gate authorizes it. Experiment 1 is a fresh representational confirmation. Experiment 2 then tests whether source-span K/V states causally influence later affect retrieval. Experiment 3 localizes a sparse writer-reader pathway on development families and confirms it via ablation, matched controls, and rescue. Experiment 4 tests whether the recovered state is maintained, suppressed and retrieved, or out-dated by a contradictory update. Development choices never use protected confirmation outcomes.

Table 1: Confirmatory safeguards and the alternative explanations they target. All choices were fixed before the corresponding protected outcomes were opened.

Stage	Potential artifact	Frozen safeguard / observed check
1	Probe tracks superficial identity cues	Lexical-identity effects equivalent to zero; binding and factual controls pass
1	Probe works only at one discourse position	Diagnostic gate passes at both bridge-like and cue-like sites
2	Any aligned patch shifts the cue readout	Source patch is contrasted with neutral-bridge patching; self-patch shift $= 0$
3	Selected heads reflect generic layer sensitivity	Matched nonselected heads at the same layers; protected selected-minus-control specificity
3	Ablation effect is not reversible	Original source K/V restored after ablation in the frozen rescue test
4	Sign is chosen after observing reversal behavior	Outdating orientation and co-primary tests frozen before confirmation

3.4 Controls and confirmatory safeguards

The central mechanistic claims depend on three separations: affect must be decodable without collapsing onto lexical identity; source interventions must have effects that are not reproduced by patching an intervening bridge; and pathway localization must survive a matched-head comparison rather than merely selecting a high-variance site. We therefore placed the strongest controls *before* protected confirmation wherever possible and treated them as gates rather than post-hoc robustness analyses.

For Experiment 1, the diagnostic probe had to generalize at both bridge-like and cue-like positions before the protected core could be opened. It achieved balanced accuracy .9875 at the bridge-like site (95% CI [.958,1]) and 1.0 at the cue-like site (95% CI [1,1]). A binding control confirmed that the readout tracked the target character rather than a distractor (target-minus-distractor mean $= .0736$, 95% CI [.0627,.0836]), and an independent factual-retrieval control reached balanced accuracy 1.0. Crucially, lexical-identity effects were prospectively required to be equivalent to zero: the 90% CIs for affect-retrieval ($[-.0512, .0145]$) and neutral-retrieval ($[-.0519, .0544]$) both fell wholly inside the frozen $[-.20, .20]$ equivalence interval. These checks make a simple lexical-cue account of the protected maintenance/RDI effects unlikely.

The intervention stages add orthogonal controls. Experiment 2 contrasts source patching with bridge patching and includes self-patching as an implementation check; the maximum absolute self-patch shift was zero in the locked result artifact. Experiment 3 freezes the selected pathway on development families, then compares its perturbation effect against matched nonselected heads at the same layers and restores the original source K/V after ablation. Experiment 4 freezes both the update conditions and the sign convention before confirmation, preventing a contrary-direction update effect from being re-labeled after the fact. Table 1 summarizes how each safeguard targets a distinct alternative explanation; exact gate definitions remain in Appendix C.

4 Results

4.1 Experiment 1: maintenance and retrieval-dependent signatures coexist

The protected core crosses positive/negative histories with affect-retrieval and neutral-retrieval cues while holding the neutral bridge fixed within family (Figure 1). We define maintenance as the signed positive-minus-negative probe separation averaged over the last two bridge positions and pre-cue position. Retrieval-dependent reinstatement is

$$\mathrm{RDI} = [(\mathrm{cue} - \mathrm{pre})_{\mathrm{affect}} - (\mathrm{cue} - \mathrm{pre})_{\mathrm{neutral}}].$$

A smallest effect size of interest (SESOI) of 0.20 diagnostic-TRAIN SD is fixed prospectively. A quantity is PRESENT when its 95% CI is above zero and its Holm-adjusted directional $p \leq .05$; it is EQUIVALENT when its 90% CI lies wholly inside $[-.20, .20]$.

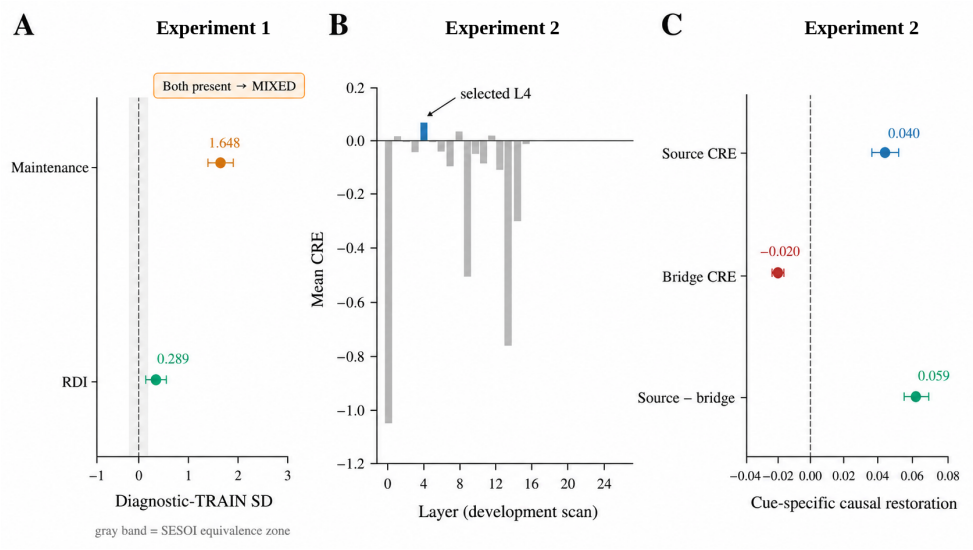

Figure 2: From descriptive signatures to causal source intervention. **A**: Experiment 1 confirms that both maintenance and retrieval-dependent reinstatement are present. **B**: On development families, the layer scan selects layer 4 by the frozen CRE rule. **C**: On protected confirmation families, opposite-affect source K/V patching produces positive cue-specific causal restoration, whereas bridge patching does not. Points show family means; horizontal bars show 95% family-bootstrap CIs.

Both signatures are PRESENT. Maintenance is large (1.648, 95% CI $[1.543, 1.754]$, Holm $p = 2.00 \times 10^{-5}$), and RDI is also positive (0.289, $[0.237, 0.342]$, Holm $p = 2.00 \times 10^{-5}$). Neither is equivalent to zero. Under the frozen classification table, PRESENT+PRESENT is labeled MIXED; this label denotes joint presence of both signatures, not failed or unstable evidence. The result therefore does not justify choosing a purely maintenance-based or purely retrieval-based mechanism from representational evidence alone, and the frozen rule authorizes causal testing in Experiment 2.

4.2 EXPERIMENT 2: SOURCE K/V STATES HAVE A CUE-SPECIFIC CAUSAL EFFECT

Experiment 2 patches source-span K/V projections from an opposite-affect donor into the matched recipient, in both affect directions. For each family, the signed causal restoration effect is

$$\mathrm{CRE} = \Delta_{\mathrm{cue}} - \Delta_{\mathrm{pre}},$$

so positive values indicate that the intervention shifts the affect readout more strongly at retrieval than immediately before the cue. Development ($n = 64$ families) evaluates every layer and freezes the layer with the largest positive mean CRE, using cue shift, absolute pre-cue shift, and shallower depth as predeclared tie-breaks. This selects layer 4. Confirmation uses 128 unseen families and includes bridge-span patching, self-patching, and same-affect donor controls.

Source patching yields positive cue-specific restoration (CRE $= 0.03975$, 95% CI $[0.03710, 0.04239]$, Holm $p = 2.00 \times 10^{-5}$). In contrast, bridge-span patching produces a negative CRE (-0.01957, $[-0.02009, -0.01905]$). The source-over-bridge specificity contrast is 0.05932 ($[0.05663, 0.06197]$, Holm $p = 2.00 \times 10^{-5}$; Figure 2). Self-patching changes the readout by at most 10^{-5} and passes its technical control. The frozen disposition is therefore CAUSAL SOURCE SUPPORT. This establishes a causal role for the source representation under the intervention, beyond the descriptive decodability shown in Experiment 1.

4.3 EXPERIMENT 3: A COMPACT K/V PATHWAY PASSES ABLATION, SPECIFICITY, AND RESCUE TESTS

Experiment 3 asks whether the source effect is distributed uniformly or concentrated in a small writer-reader pathway. On 64 development families, we patch each K/V head separately at each layer

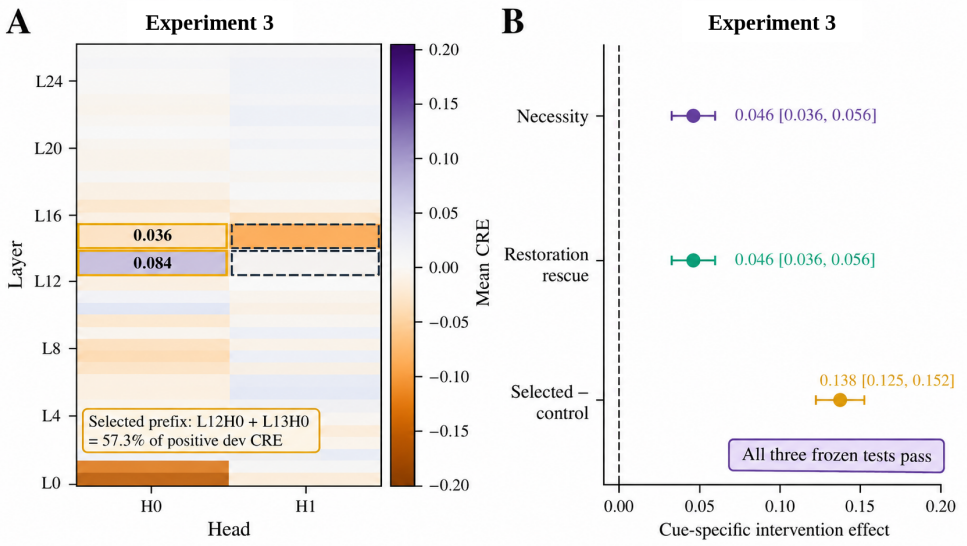

Figure 3: Compact pathway localization in Experiment 3. **A**: Head-level development scan of mean CRE identifies a sparse writer-reader pathway; the frozen prefix L12H0+L13H0 accounts for 57.3% of all positive development CRE, while L12H1+L13H1 serve as matched controls. **B**: In protected confirmation, necessity, matched-control specificity, and restoration rescue are all positive with 95% family-bootstrap CIs excluding zero. Rescue restores the original source K/V after ablation and is therefore a restoration test rather than a claim of full sufficiency.

and rank positive-CRE heads. Before confirmation, we freeze the smallest ranked prefix reaching 50% of total positive CRE (minimum 1, maximum 8 heads). The selected pathway contains **L12H0** and **L13H0**, with development mean CREs 0.0839 and 0.0361, respectively. The matched nonselected control pathway is L12H1 and L13H1.

For protected confirmation, selected source-position K/V vectors are replaced by development-derived neutral baselines. *Necessity* measures the cue-specific clean-minus-ablation effect; *specificity* subtracts the corresponding matched-control-head effect; *rescue* restores each form's original source K/V after the ablation operation. Rescue is therefore a restoration test, not a claim of standalone sufficiency.

All three protected tests pass: necessity $= 0.04602$ (95% CI $[0.03595, 0.05592]$), rescue $= 0.04602$ (same CI), and selected-minus-control specificity $= 0.13836$ ($[0.12488, 0.15156]$); all Holm-adjusted $p = 3.00 \times 10^{-5}$ (Figure 3). The frozen disposition is WRITER-READER SUPPORT. The data therefore localize a compact K/V pathway whose perturbation changes the cue-specific readout, whose effect exceeds matched nonselected heads, and whose original signal can be restored after ablation.

4.4 EXPERIMENT 4: RETRIEVAL SURVIVES SUPPRESSION, BUT REVERSAL DOES NOT OUTDATE THE OLD STATE

The final experiment introduces four frozen update conditions: MAINTAIN, REVERSE-OUTDATE, IRRELEVANT-SUPPRESS, and NEUTRAL-UPDATE. Each family is scored at source end, update end, pre-cue, and cue end. Two estimands are co-primary. O_{pre} asks whether a contradictory update shifts the pre-cue representation toward the new orientation relative to the original source. R asks whether the original orientation is selectively reinstated at the cue after an irrelevant suppressing update, relative to a neutral update. Holm correction is applied across these two tests. O_{update}, maintenance M, and suppression S are secondary.

Retrieval after suppression is supported: $R = 0.186$ (95% CI $[0.127, 0.245]$, Holm $p = 2.00 \times 10^{-5}$). The outdating co-primary is not supported: $O_{\text{pre}} = -0.271$ ($[-0.466, -0.076]$, Holm $p = .996$ under the prespecified positive-direction test). The immediate update estimand is also negative, $O_{\text{update}} = -0.976$ ($[-1.058, -0.894]$). Meanwhile, maintenance remains positive ($M = 0.771$,

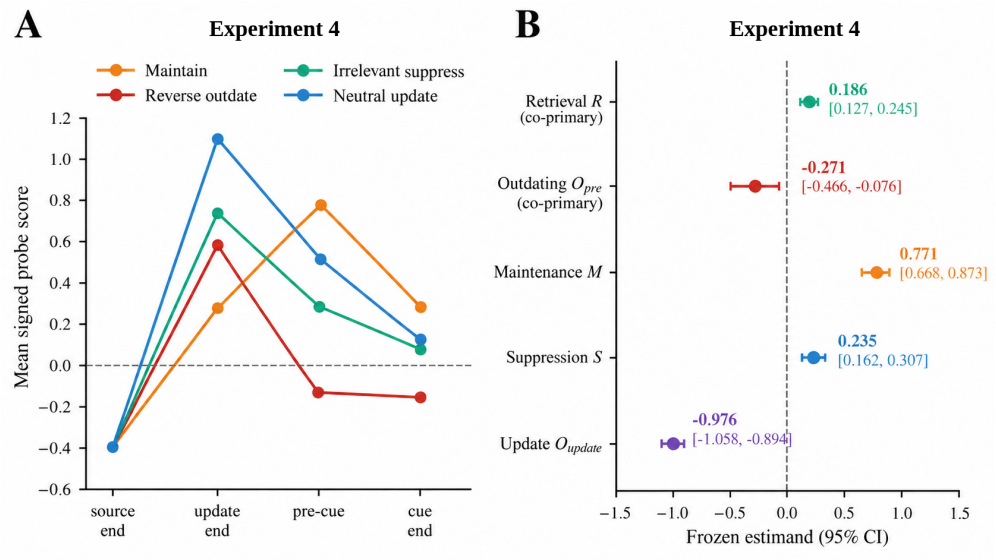

Figure 4: Update dynamics in Experiment 4. **A**: Condition-wise trajectories show that irrelevant suppression reduces the pre-cue readout while still allowing cue-time recovery, whereas reversal does not flip the state into a clean outdating pattern. **B**: Family-level frozen estimands with 95% family-bootstrap CIs. Retrieval after irrelevant suppression (R) is positive, while the co-primary outdating effect O_{pre} is negative under the prospectively fixed sign convention.

Table 2: What each protected stage establishes. Human-readable labels summarize the frozen executable dispositions; exact strings appear in Appendix B.

Stage	Confirmatory question	Frozen disposition
1	Are maintenance and/or retrieval-dependent signatures present?	MIXED: both present
2	Does source K/V causally affect cue-specific retrieval beyond bridge patching?	Causal source support
3	Does a frozen sparse K/V pathway pass necessity, specificity, and rescue?	Writer-reader support
4	Do reversal-outdating and retrieval-after-suppression co-primary tests pass?	Partial co-primary support

$[0.668, 0.873]$), and the suppression contrast is positive ($S = 0.235$, $[0.162, 0.307]$). The frozen disposition is therefore PARTIAL CO-PRIMARY SUPPORT: the model can recover the earlier affect-linked orientation after irrelevant suppression, but the reversal manipulation does not produce the prospectively predicted outdating of the old orientation.

4.5 CROSS-FAMILY CONSISTENCY AND CONVERGENCE ACROSS INTERVENTION LEVELS

Aggregate confidence intervals establish the confirmatory decisions, but they do not show how strongly those means depend on a small subset of families. Figure 5 therefore visualizes every protected-family effect for the three intervention stages. These sign counts are descriptive—they are not additional hypothesis tests—but they make the structure of the locked results transparent.

Experiment 2 shows the cleanest family-wise separation: source CRE is positive in all 128 confirmation families, bridge CRE is negative in all 128, and the paired source-minus-bridge contrast is positive in all 128. Thus the source-specificity result is not driven by a few extreme narratives. In Experiment 3, selected-pathway necessity and restoration rescue are positive in 112/128 families, while the matched control necessity effect is positive in 0/128; the selected-minus-control specificity contrast is positive in 126/128. The protected mean effects therefore coincide with a broad family-level shift away from the matched control pathway.

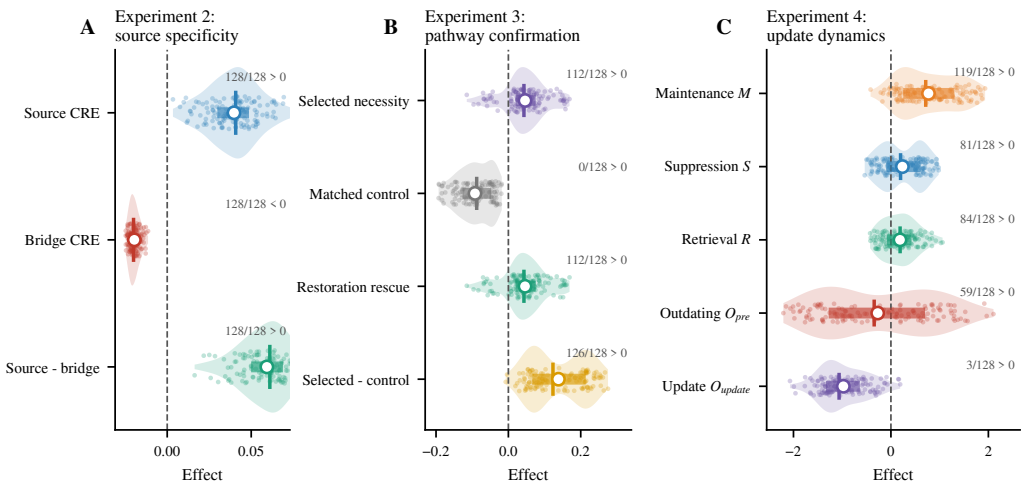

Figure 5: Protected-family effect distributions for the intervention stages. **A**: Experiment 2 source CRE, bridge CRE, and their paired difference. **B**: Experiment 3 selected-pathway necessity, matched-control necessity, restoration rescue, and selected-minus-control specificity. **C**: Experiment 4 maintenance, suppression, retrieval-after-suppression, and outdating/update estimands. White circles mark means, thick bars mark interquartile ranges, and small points show individual protected families. Sign counts are descriptive summaries and are not used for the confirmatory decisions.

Experiment 4 is intentionally more heterogeneous because its conditions manipulate the evolving discourse state rather than a single internal component. Maintenance remains positive in 119/128 families. Retrieval after irrelevant suppression is positive in 84/128, whereas the immediate outdating estimand O_{update} is positive in only 3/128. The co-primary O_{pre} is mixed at the family level (59/128 positive) and negative on average under the frozen orientation, consistent with the formal partial-support disposition. This pattern is useful mechanistically: the robust source and pathway effects coexist with substantially richer variability once the narrative introduces competing update information.

4.6 What the four-stage sequence identifies

The sequence of tests matters because no single result carries the full mechanistic claim. Experiment 1 establishes that affect-linked information is available across the bridge and changes again at retrieval, but the coexistence of those signatures leaves the mechanism underdetermined. Experiment 2 identifies a causal source dependency: replacing K/V states at the original episode changes the later cue-specific readout more than replacing matched bridge states. Experiment 3 then narrows that dependency to a frozen two-head pathway under necessity, matched-control specificity, and restoration. Finally, Experiment 4 tests whether the retrieved variable behaves like a state that is simply overwritten by new information. It does not: retrieval after irrelevant suppression is supported, whereas contradiction fails the frozen outdating criterion.

The resulting inference is therefore a conjunction rather than a label attached to one plot. Persistent decodability rules out a purely absent bridge state; cue-dependent reinstatement rules out an account in which nothing changes when retrieval becomes relevant; source patching establishes causal dependence on the earlier episode; pathway perturbation localizes a compact route for that dependence; and update manipulations show that accessibility and rewriting are separable. This layered structure is why we describe the mechanism as *hybrid retrieval*: maintained information and source-dependent retrieval coexist, and later discourse can modulate accessibility without behaving as a simple destructive overwrite.

5 DISCUSSION

The four experiments form an escalation of evidence from representation to mechanism. Experiment 1 shows that both maintained separation and cue-triggered reinstatement are present, making a decoding-only account underdetermined. Experiment 2 resolves that ambiguity causally: replacing source-span K/V states with an opposite-affect donor changes the later cue-specific readout, while bridge patching moves in the opposite direction. Experiment 3 then localizes this dependence to a compact two-head K/V pathway that passes necessity, matched-head specificity, and restoration tests on protected families. Experiment 4 extends the mechanism into a dynamic setting, showing that an earlier orientation can be recovered after irrelevant suppression.

The resulting picture is neither pure maintenance nor pure dormant storage. Some affect-linked separation remains available across the neutral bridge, but later demand also recruits information causally tied to the original source episode. This hybrid account connects naturally to query-time retrieval mechanisms reported for factual recall and belief tracking (Geva et al., 2023; Wu et al., 2025; Prakash et al., 2026), while extending them to affect-linked character histories. In this sense, "affect on demand" refers to a computational property of discourse processing: information established earlier can remain partially represented yet still require a source-dependent retrieval operation when a later cue makes it useful.

Experiment 4 sharpens the mechanism rather than merely adding a robustness check. The suppressing update reduces the pre-cue signal, after which the cue selectively recovers the earlier orientation. By contrast, the reversal manipulation does not produce the predicted positive outdating effect under the frozen sign convention. This pattern argues against a simple scalar-overwrite picture in the tested setting and is more consistent with persistence plus context-dependent retrieval or competition between traces. The distinction is mechanistically useful because overwrite and retrieval architectures make different predictions about what should remain recoverable after intervening information.

The evidence establishes a causal retrieval pathway for an affect-linked discourse variable, not a claim about subjective emotion. The readout measures consequence orientation in controlled text, and the selected heads are defined by intervention effects rather than by an assertion that all relevant computation is confined to two components. Within that scope, however, the central result is supported at three complementary levels: the signal is decodable, source-state interventions change its later expression, and a frozen sparse pathway mediates a substantial portion of the causal effect.

The study also illustrates a broader design principle for mechanistic interpretability. A development scan can be useful for finding an intervention site, but the scientific claim should be attached to a protected test of that frozen choice rather than to the scan itself. Here, layer 4 and the L12H0+L13H0 pathway are selected only on development families; causal source support and writer-reader support are assigned only after confirmation on unseen families. This separation reduces the risk that a visually compelling activation pattern is mistaken for an independently supported mechanism, while still allowing targeted circuit discovery within a feasible compute budget.

Mechanistic predictions. The hybrid account makes concrete predictions that distinguish it from two simpler alternatives. Under a pure-maintenance account, manipulating the original source after a stable bridge representation has formed should add little beyond what can be achieved by intervening on the currently maintained state; the Experiment 2 source-over-bridge contrast points in the opposite direction. Under a pure dormant-trace account, affect separation should collapse during neutral text and emerge only when the cue demands retrieval; Experiment 1 instead finds a substantial maintained component before the cue. A hybrid account predicts both phenomena at once: a bridge representation that preserves information above the equivalence region, together with a cue-time computation that remains causally sensitive to source-bound states.

The update experiment adds a second prediction. If contradictory information simply overwrites the earlier variable, reversal should move the representation toward the new state and reduce recoverability of the old orientation. The observed negative outdating estimands do not show that pattern, while retrieval after irrelevant suppression remains positive. This motivates a direct follow-up: patch the old and new source spans separately after contradiction and test whether they make dissociable contributions to the same cue, distinguishing destructive overwrite from coexistence or competition between addressable traces.

Reproducibility statement

The anonymous supplement contains the final experiment code, deterministic material generators, frozen prospective protocol, exact statistical table, final per-experiment disposition files, and provenance audit. The model name and immutable revision, candidate selection rules, development/confirmation splits, family-level estimands, bootstrap/permutation seeds, multiplicity correction, and advancement rules are stated in the paper and supplement. Result artifacts are SHA256-audited, and the supplement preserves the frozen analysis path needed to reproduce the reported results.

Ethics statement

This work uses synthetic text scenarios and an open-weight language model; it does not involve human participants, private records, or personal data. Mechanistic tools for affect inference could be misapplied to overinterpret model states or to support manipulative affective systems. We therefore avoid anthropomorphic claims, restrict conclusions to the tested model/task, and release intervention definitions and negative results alongside positive findings. Any downstream use for human affect assessment would require separate validation, consent, fairness, and privacy analysis.

AI use statement

Generative AI tools were used to assist with code checking and language editing. All AI-assisted code suggestions were independently reviewed and tested for correctness, and all manuscript edits were reviewed by the authors. The authors take full responsibility for the final content, code, and results.

References

Adar Avsian, Benjamin Reichman, and Larry Heck. Emotions where art thou: Understanding and characterizing the emotional latent space of large language models. In *Conference on Language Modeling (COLM)*, 2025.

Arthur Conmy, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

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

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of EMNLP*, pp. 12216–12235, 2023.

Najoung Kim and Sebastian Schuster. Entity tracking in language models. In *Proceedings of ACL*, pp. 3835–3855, 2023.

Belinda Z. Li, Zifan Carl Guo, and Jacob Andreas. (How) do language models track state? In *Proceedings of ICML*, volume 267, pp. 34429–34452, 2025.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, 2022.

Nikhil Prakash, Tamar Rott Shaham, Tal Haklay, Yonatan Belinkov, and David Bau. Fine-tuning enhances existing mechanisms: A case study on entity tracking. In *International Conference on Learning Representations*, 2024.

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

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Ala N. Tak, Amin Banayeeanzade, Anahita Bolourani, Mina Kian, Robin Jia, and Jonathan Gratch. Mechanistic interpretability of emotion inference in large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 13090–13120, 2025.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: A circuit for indirect object identification in GPT-2 small. In *International Conference on Learning Representations*, 2023.

Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. Retrieval head mechanistically explains long-context factuality. In *International Conference on Learning Representations*, 2025.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *ATTRIB: Workshop on Attributing Model Behavior at Scale at NeurIPS*, 2023.

A Prospective governance and material construction

The final program was executed in the fixed order Experiments $1 \to 2 \to 3 \to 4$. Protected confirmation outcomes were not available for downstream redesign, retuning, site selection, donor selection, or threshold changes. Earlier development/confirmation attempts that were exposed, invalidated, or retired for material/control/governance reasons are not treated as evidence in this paper. The fresh Experiment 1 material pool contained 320 diagnostic families, 144 control families, and a physically locked 128-family protected core. Public validation covered tokenizer alignment, identity-confound balance, scenario independence, structural anchors, binding readouts, numerical fixtures, and protected-outcome isolation before production.

Experiment 2 used 64 development and 128 confirmation families. Experiment 3 independently used 64 development and 128 confirmation families. Experiment 4 used 32 development and 128 confirmation families and crossed four update conditions: MAINTAIN, REVERSE-OUTDATE, IRRELEVANT-SUPPRESS, and NEUTRAL-UPDATE. Update spans were constructed to have within-family token-length spread at most two tokens.

B Exact statistical rules and results

Experiment 1 uses 20,000 bootstrap draws and 100,000 sign-flip/permutation draws for family-level inference. The maintenance/RDI bootstrap seed is 20260922 and permutation seed is 20260923. Diagnostic TEST uses an independent 20,000-draw bootstrap and 100,000-draw permutation test. The SESOI is 0.20 diagnostic-TRAIN SD with a 90% equivalence interval and 95% directional interval. Holm correction is applied across Maintenance/RDI. Experiment 2 uses the same family-level bootstrap/permutation defaults and Holm correction across source CRE and source-vs-bridge specificity. Experiment 3 applies Holm correction across necessity, specificity, and rescue. Experiment 4 applies Holm correction across co-primary O_{pre} and R.

C Experiment 1 diagnostic and control gates

The diagnostic gate passed before protected-core access. Bridge-like classification achieved balanced accuracy .9875 (95% CI [.9583, 1.0]) and cue-like classification achieved 1.0 (95% CI [1.0, 1.0]); both permutation $p = 9.9999 \times 10^{-6}$. The control gate also passed. Binding produced a mean target-minus-distractor contrast of .0736 (90% CI [.0645,.0821], 95% CI [.0627,.0836], $p = 9.9999 \times 10^{-6}$). Factual retrieval balanced accuracy was 1.0 (95% CI [1.0,1.0], paired-family permutation $p =$

Table 3: Final frozen family-level statistics.

Exp.	Quantity	Mean	95% CI low	95% CI high	Holm p
1	Maintenance	1.6476	1.5426	1.7545	2.00×10^{-5}
1	RDI	0.2891	0.2374	0.3418	2.00×10^{-5}
2	Bridge CRE	-0.0196	-0.0201	-0.0190	–
2	Source CRE	0.0397	0.0371	0.0424	2.00×10^{-5}
2	Source–bridge	0.0593	0.0566	0.0620	2.00×10^{-5}
3	Necessity	0.0460	0.0359	0.0559	3.00×10^{-5}
3	Rescue	0.0460	0.0359	0.0559	3.00×10^{-5}
3	Specificity	0.1384	0.1249	0.1516	3.00×10^{-5}
4	M	0.7708	0.6682	0.8726	–
4	O_{pre}	-0.2710	-0.4660	-0.0757	0.9959
4	O_{update}	-0.9762	-1.0578	-0.8936	–
4	R	0.1861	0.1268	0.2446	2.00×10^{-5}
4	S	0.2354	0.1623	0.3072	–

9.9999×10^{-6}). Lexical identity differences were equivalent around zero for affect retrieval (mean $-.0182$, 90% CI $[-.0512, .0145]$) and neutral retrieval (mean $.00138$, 90% CI $[-.0519, .0544]$), both within the frozen $[-.20, .20]$ interval. No protected-core rows were excluded or imputed.

The final Experiment 1 state table was Maintenance=PRESENT, RDI=PRESENT. The executable decision rule maps (PRESENT, EQUIVALENT) to H1-compatible, (EQUIVALENT, PRESENT) to H2-compatible, and (PRESENT, PRESENT) to MIXED; all three permit advancement. Thus the word "mixed" in the manuscript is a formal state-combination label and does not mean that the statistical evidence was weak.

D INTERVENTION DETAILS

D.1 EXPERIMENT 2 SOURCE-SPAN PATCHING

For a donor and recipient with opposite source consequences, K and V projection outputs at aligned source tokens are replaced at a candidate layer. The signed shift is oriented so that positive values indicate movement toward the donor's affect orientation. Development computes bidirectional effects for every layer and selects the largest positive mean CRE, with the fixed tie-break sequence described in the main text. Confirmation patches only the frozen layer. Bridge patching substitutes the aligned neutral bridge span rather than the source; self-patching verifies the implementation; same-affect donor rows are retained as controls.

D.2 EXPERIMENT 3 PATHWAY PERTURBATION

Development evaluates each K/V head at each layer with the same source-span patch logic. Positive-CRE heads are ranked descending by development CRE with shallower-layer/lower-head tie-breaks. L12H0 and L13H0 form the frozen smallest prefix exceeding 50% of total positive development CRE. During confirmation, selected source-position K/V vectors are replaced with head- and relative-position-specific neutral baselines estimated only from development materials. Matched controls use nonselected heads L12H1 and L13H1. Rescue repeats the ablation hook and then restores the original source K/V at the selected head positions, which makes it a restoration-after-perturbation test rather than an independent sufficiency manipulation.

D.3 Experiment 4 estimands

Let $s \in \{-1, +1\}$ denote the initial orientation and $\mathrm{new} = -s$. With probe scores at source end, update end, pre-cue, and cue end, the implementation defines

$$
\begin{aligned}
O_{\mathrm{update}} &= \mathrm{new}\,[r_{\mathrm{update}} - r_{\mathrm{source}}],\\
O_{\mathrm{pre}} &= \mathrm{new}\,[r_{\mathrm{pre}} - r_{\mathrm{source}}],\\
S &= s\,[r_{\mathrm{pre}}^{\mathrm{neutral}} - r_{\mathrm{pre}}^{\mathrm{irrelevant}}],\\
R &= s\,[(r_{\mathrm{cue}}^{\mathrm{irrelevant}} - r_{\mathrm{pre}}^{\mathrm{irrelevant}}) - (r_{\mathrm{cue}}^{\mathrm{neutral}} - r_{\mathrm{pre}}^{\mathrm{neutral}})],\\
M &= s\,r_{\mathrm{pre}}^{\mathrm{maintain}}.
\end{aligned}
$$

The sign convention was frozen before confirmation. Consequently, the negative observed values of O_{pre} and O_{update} are reported without sign reversal or post-hoc redefinition.

E Provenance and audit trail

The final production/confirmation chain completed through Experiment 4, and the provenance audit status is PASS. Several Experiment 2 confirmation attempts terminated before producing scientific outcomes because Git was unavailable on compute nodes. Before successful confirmation, provenance checking was replaced by atomic SHA256 freeze receipts; no hypotheses, materials, thresholds, selected layer, or statistical rules changed. Environment-specific job logs are omitted from the anonymous supplement to avoid revealing infrastructure provenance.

F Anonymous artifact map

The anonymous supplement contains final experiment code and utilities, deterministic material generators, prospective protocol/freeze documents, final statistical and disposition files, the provenance audit, and a README with model revision, dependencies, and execution order. Absolute user/home-directory paths are replaced by an environment variable; scientific logic is unchanged.