ONE TOKEN FLIPS THE VERDICT: HYPERGRAPH PROPAGATION AND STATE READOUT DISSOCIATE IN SELECTIVE BELIEF REVISION

Anonymous authors
Paper under double-blind review

ABSTRACT

Selective belief revision requires propagating evidence changes while preserving judgments whose justification remains unchanged. We show that evidence propagation and state readout dissociate in two 8B language models. We introduce Evidence–Intervention–Claim (EIC) to measure this separation by intervening on evidence and resolving each claim's supporting and opposing coverage before and after the change. Updating intermediate hypergraph sources improves multihop revision over matched StaticSource training by 9.01 points on FLDx2-formal and 34.34 on EntailmentBank across three Qwen seeds. Yet pairing a graph-recoverable partial endpoint with itself reduces Qwen's joint accuracy from 94.94% to 11.24%, despite unchanged endpoint evidence and a bitwise-identical graph residual. Changing only the preceding state token from partial to full raises second-state partial accuracy from 12.92% to 82.58%. Crossed interventions separate token, companion-evidence, and generated-history effects; a pre-specified prospective source-disjoint replication supports the same pattern. The dissociation also persists, though attenuated, when oracle graphs are replaced at inference by cross-fitted predicted graphs in an additional Qwen control. Training on the missing partial-preservation composition helps, but heavy augmentation causes severe state redistribution, whereas lower-dose augmentation improves preservation with substantially less collateral redistribution. Endpoint-factorized generation substantially restores partial preservation across all three Qwen seeds. Together, these results identify evidence propagation and state readout as separable computational problems and motivate isolating generation histories for endpoint-wise judgments that are semantically independent.

1 INTRODUCTION

Suppose a claim needs two premises and only one remains visible. Its justification is partial, whether paired with a complete view or a copy of itself. Two 8B models judge the first pairing correctly more than nine times in ten and the second fewer than once in eight, despite unchanged endpoint evidence, justification, and target. Partial endpoints occur in training, but partial-preservation pairings do not. Does this unseen composition disrupt evidence propagation or state readout?

Aggregate accuracy can conceal behavioral failures and shortcuts (Ribeiro et al., 2020; Geirhos et al., 2020). Evidence–Intervention–Claim (EIC) makes *evidence propagation* and *autoregressive state readout* separately testable: evidence interventions change registered support/opposition coverage, while a hypergraph branch supplies view-local residuals at the vocabulary head. Dynamic source updates improve multihop revision over matched StaticSource training by 9.01 points on FLDx2-formal and 34.34 on EntailmentBank across three Qwen seeds. Yet partial-diagonal pairing drops Joint from 94.94% to 11.24% (Qwen) and 91.01% to 4.49% (Llama), with a bitwise-identical graph residual.

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.

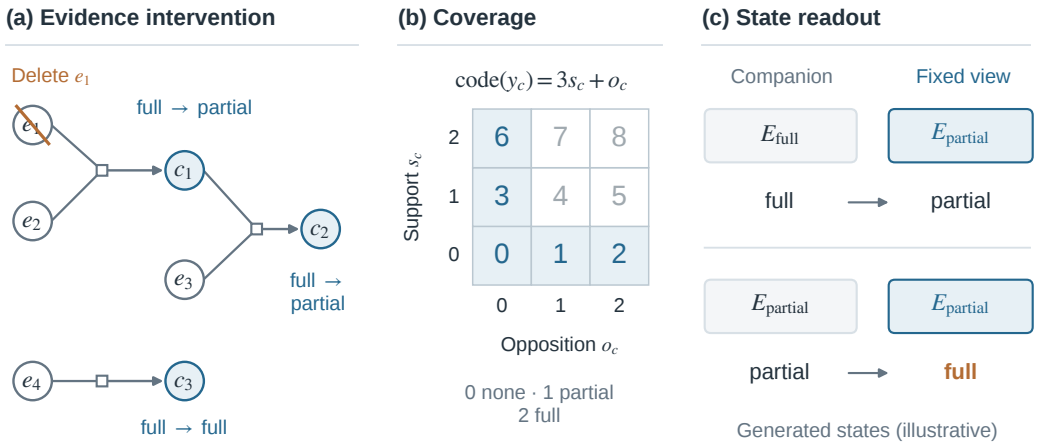

Figure 1: **Evidence determines the state; generated context can redirect its readout.** (a) Removing e_1 changes registered support for c_1, c_2 from full to partial while c_3 stays full; squares denote registered premise groups. (b) Support and opposition coverage define nine codes; shaded cells are the five states observed in the data. (c) Re-pairing a partial endpoint changes its second verdict despite fixed evidence, registered justification, and target (illustrative).

This coexistence motivates testing the decoder path directly. Crossed interventions separate effects of the preceding state token, companion evidence, and earlier history; endpoint-factorized generation then tests whether isolating histories improves preservation. Our contributions are threefold:

(1) Intervention framework. We introduce EIC to disentangle evidence propagation from autoregressive state readout in selective belief revision.

(2) Causal localization. We uncover a propagation–readout dissociation in two 8B language models: generated context and preceding state tokens can redirect endpoint judgments despite unchanged evidence and propagated residuals.

(3) Generation design. We demonstrate that endpoint-factorized generation isolates endpoint histories and substantially improves preservation, with measured original-task costs.

2 Related Work

Belief revision and consequence tracking. Classical belief revision formalizes belief change (Alchourrón et al., 1985); truth-maintenance systems track justifications (Doyle, 1979; de Kleer, 1986), and Belnap separates supporting from opposing information (Belnap, 1977). Belief-R scores update versus retention (Wilie et al., 2024); BeliefBank and rational-revision studies test consistency under new information (Kassner et al., 2021; Hase et al., 2023; 2024). Model editors make targeted factual changes (De Cao et al., 2021; Meng et al., 2022; 2023; Mitchell et al., 2022a;b), with related work addressing downstream consequences (Zhong et al., 2023; Cohen et al., 2024; Dong et al., 2025). EIC changes evidence at fixed inference-time weights, with a matched NoGraph baseline, then varies generation context at fixed warranted states.

Composition and generated context. Compositional generalization tests familiar components in novel combinations (Lake & Baroni, 2018; Keysers et al., 2020; Kim & Linzen, 2020), including multi-step composition (Dziri et al., 2023) and reversed relations (Berglund et al., 2024). Our partial diagonal pairs familiar endpoints in an unseen preservation composition, holding their warranted states fixed. Outputs also shift with prompt format and demonstrations (Sclar et al., 2024; Zhao et al., 2021; Lu et al., 2022; Min et al., 2022), and generated rationales can be unfaithful (Turpin et al., 2023). Our crossed interventions hold target evidence, residual, and output position fixed while independently varying companion evidence, earlier history, and the preceding state token, localizing the composition gap to readout.

Structured representations for evidence-conditioned reasoning. Structured architectures such as message passing and hypergraph networks organize relational dependencies among evidence, intermediate claims, and structured reasoning units (Gilmer et al., 2017; Feng et al., 2019; Chien et al., 2022), while reasoning-oriented methods model explicit inference structures or intermediate reasoning paths (Clark et al., 2020; Tafjord et al., 2021; Wei et al., 2022; Wang et al., 2023). Unlike prior structured reasoning approaches that primarily target improved inference, we use structured representations as a controlled interface for studying evidence-conditioned generation. Our hypergraph module keeps propagated residuals separate from autoregressive cache states, allowing evidence-dependent contributions to remain fixed while decoder context changes. Matched source-update controls distinguish intermediate-source propagation from target-side refinement effects. The contribution lies in enabling a controlled separation of evidence propagation and autoregressive state readout within the same model.

Table 1: Dataset coverage. A: affected claims; P: preserved claims; $M \subseteq A$: multihop affected claims. Scope cells give claims/source components (n/s) on the source-disjoint test set (Section 5).

Dataset	Train pairs	Test pairs	A: n/s	P: n/s	M: n/s
EntailmentBank	2532	374	221 / 186	973 / 186	33 / 33
FLDx2-formal	3209	629	384 / 249	705 / 384	135 / 98
SciFact	240	179	196 / 179	13 / 11	N/A

3 The EIC Framework

EIC separates the registered justification structure that defines target states from the view-visible structure supplied for propagation. A belief is a generated judgment of supporting and opposing coverage; it should change exactly when an evidence intervention changes its warranted state (Figure 1).

3.1 Task and registered dependencies

An annotated instance is $(x, \mathcal{H})$, with text input $x = (E_0, E_1, C, b)$: two evidence views E_v ($v \in \{0, 1\}$), an ordered claim set C, and an indicator $b \in \{0, 1\}$ identifying the before view. A registered hypergraph $\mathcal{H}$ defines the justification contract through groups $g = (S_g, t_g, p_g)$, each connecting evidence or claim premises S_g to target claim t_g with polarity $p_g \in \{+, -\}$ for support or opposition. Write $E^{\text{pre}} = E_b$ and $E^{\text{post}} = E_{1-b}$ for the two temporal endpoints. The target is $(y_c^{\text{pre}}, y_c^{\text{post}}) = (y_c(E^{\text{pre}}; \mathcal{H}), y_c(E^{\text{post}}; \mathcal{H}))$. Both endpoint states are generated; neither a gold prior state nor an affected-claim indicator is supplied. Propagation receives the view-visible graph $\mathcal{G}_v \subseteq \mathcal{H}$: only groups with all source identifiers available enter; groups with missing sources are omitted entirely.

We construct these instances from EntailmentBank (EB) (Dalvi et al., 2021), FLDx2-formal (FLD) (Morishita et al., 2023; 2024), and SciFact (Wadden et al., 2020). Registered proofs or rationale sets define the dependencies; evidence deletions receive a temporal direction independently of their labels, yielding both additions and deletions (Appendix B). Labels retain the scope of $\mathcal{H}$ when filtering removes groups from $\mathcal{G}_v$. Visible-graph replay can therefore return none for a partial target, exposing an information limit before learned errors are assessed. Source components connect shared identifiers across versions; training pairs whose component touches test are quarantined.

3.2 Dual-axis nine-state representation

A claim state records supporting and opposing coverage independently:

$$y_c(E; \mathcal{H}) = (s_c, o_c) \in \{0, 1, 2\}^2, \qquad \text{code}(y) = 3s_c + o_c, \tag{1}$$

Here s_c and o_c are support and opposition coverage; 0, 1, and 2 denote none, partial, and full. The nine states retain conflicting coverage: absent support does not imply opposition. Coverage is defined over registered $\mathcal{H}$, rather than filtered $\mathcal{G}_v$: a group contributes full when all premises are full, partial when some coverage remains but is incomplete, and none otherwise; each polarity

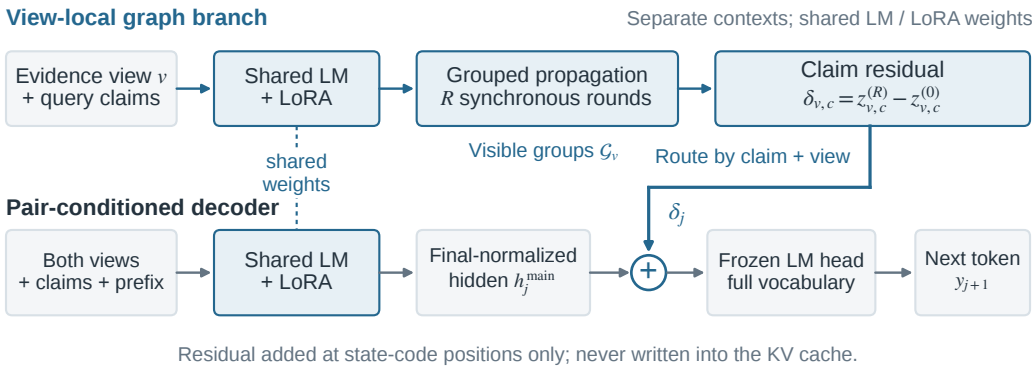

Figure 2: **Evidence hypergraphs condition state-token readouts.** Shared language-model (LM) weights with low-rank adaptation (LoRA) encode views and the paired prompt separately. View-local propagation supplies claim residuals at the vocabulary head, outside the key–value (KV) cache. LoRA and graph parameters learn by cross-entropy (CE); the base LM and head are frozen. Equation (2) specifies the claim update.

takes the maximum over alternative groups. Evidence is full when present and none when absent; claim premises inherit positive coverage, recomputed to a fixed point. Partial denotes incomplete justification, not confidence. The current datasets instantiate five codes: $\{0, 1, 2, 3, 6\}$.

The model outputs one line per claim, `cXXXX=before,after`, in a single free-running autoregressive completion, so each judgment conditions on the model's own preceding predictions.

3.3 Evaluation across correctness, behavior, and depth

Gold endpoint equality partitions claims into affected $A = \{c : y_c^{\text{pre}} \neq y_c^{\text{post}}\}$ and preserved $P = \{c : y_c^{\text{pre}} = y_c^{\text{post}}\}$. The multihop subset $M \subseteq A$ has registered intervention depth $d \geq 2$, the shortest path from intervened evidence through changed claims (alias edges count zero); d is a data annotation, distinct from neural rounds R. For $S \in \{A, P, M\}$, S-Joint is the macro-averaged indicator that both predicted endpoints $\hat{y}_c^{\text{pre}}, \hat{y}_c^{\text{post}}$ match the gold endpoints, averaging claims within pairs, pairs within source components, then components equally (Table 1 gives the counts). The behavior axis compares the predicted two-axis change $\hat{y}_c^{\text{post}} - \hat{y}_c^{\text{pre}}$ with its gold counterpart. On A, A-DeltaMatch is the rate of matching changes, A-NoChange of zero predicted changes, and A-WrongChange of nonzero but incorrect changes. On P, P-StayRate is the rate of unchanged predictions and P-SpuriousChange of changed predictions. These rates use the same source-macro averaging; retaining a wrong state earns P-StayRate but not Joint (Table 27; Appendix I.1). Malformed completions invalidate the pair and stay in every denominator.

4 Controlled Evidence Hypergraph Propagation

To separate evidence propagation from state readout, our hypergraph branch injects claim-specific residuals derived from view-visible evidence while keeping decoder context independently variable (Figure 2). This provides a controlled interface for testing the propagation–readout boundary.

4.1 View-local Evidence Representations

A shared language model (LM) with low-rank adaptation (LoRA) (Hu et al., 2022) encodes the joint pair and each view in three separate contexts. Each view lists evidence before claims without serializing graphs; evidence nodes use mean token states and claims use final-token states (Appendix B). Propagation receives $\mathcal{G}_v$; visible groups encode dependencies, not gold states or premise satisfaction.

4.2 Grouped Evidence Propagation

Grouped propagation learns dependency-conditioned transformations over visible evidence groups rather than executing symbolic proof rules. Let $z_s^{(r)} \in \mathbb{R}^D$ denote node s's D-dimensional state at round r (view index omitted). For $r = 0, \ldots, R-1$:

$$
\begin{aligned}
u_g^{(r)} &= \frac{1}{|S_g|} \sum_{s \in S_g} z_s^{(r)}, \\
m_g^{(r)} &= W_\uparrow \left[\mathrm{GELU}\left(W_\downarrow [z_{t_g}^{(r)}; u_g^{(r)}] \right) \odot s_{p_g} \right], \\
z_c^{(r+1)} &= z_c^{(r)} + \frac{1}{|\mathcal{G}_c|} \sum_{g \in \mathcal{G}_c} m_g^{(r)}, \qquad \mathcal{G}_c = \{ g \in \mathcal{G}_v : t_g = c \}.
\end{aligned}
\tag{2}
$$

Here $W_\downarrow, W_\uparrow$ are learned projections, s_{p_g} is a learned polarity vector, and GELU is the Gaussian error linear unit; $[\,;\,]$ concatenates and $\odot$ multiplies elementwise; empty source pools are zero, and nodes without incoming groups remain fixed. Appendix B specifies rank-16 projections and polarity vectors. Pooling is not symbolic conjunction or disjunction; all visible groups act from round one without truth-value gates.

4.3 Claim-aligned Controlled Readout Interface

After R rounds, each claim obtains a view-local residual representing the accumulated evidence-dependent contribution:

$$
\delta_{v,c} = z_{v,c}^{(R)} - z_{v,c}^{(0)}. \tag{3}
$$

At grammar-routed state-code positions, we add the claim/view residual to the main decoder's final-normalized hidden state, immediately before the model's original vocabulary head:

$$
p_\theta(\cdot \mid x, a_{<j}) = \mathrm{softmax}_{\mathcal{V}} \left(W_{\mathrm{LM}}(h_j^{\mathrm{main}} + \delta_j) \right), \tag{4}
$$

Here a_j is completion token j, θ denotes trainable parameters, and $\mathcal{V}$ is the full vocabulary; $h_j^{\mathrm{main}}, \delta_j \in \mathbb{R}^D$ and $W_{\mathrm{LM}} \in \mathbb{R}^{|\mathcal{V}| \times D}$; $\delta_j = \delta_{v,c}$ at routed positions and zero elsewhere (graph dependence omitted for readability). Head-local injection keeps residuals outside the transformer stream and KV cache, separating propagation from decoding and enabling the exact decomposition $W_{\mathrm{LM}}(h + \delta) = W_{\mathrm{LM}} h + W_{\mathrm{LM}} \delta$ at fixed prefixes. Generated tokens still affect later predictions. Greedy generation predicts both endpoints without gold states (Appendix B).

4.4 Propagation Controls

NoGraph omits the graph branch; True/Shuffled use registered/frozen randomized groups. All-evidence residual, Random, and Similarity use positively tagged evidence-only pools without claim dependencies; retrieval budgets count unique visible gold evidence ancestors, and similarity uses frozen GTE-Qwen2-1.5B-instruct embeddings. StaticSource replaces $u_g^{(r)}$ with $|S_g|^{-1} \sum_{s \in S_g} z_s^{(0)}$, retaining evolving targets, topology, polarity, parameters, rounds and gradients. Inference-time source freezing complements matched training. These controls separate structured propagation from added evidence exposure or repeated target refinement.

Label-free visible-graph replay audits information availability (Appendix G.1): EB reaches 96.97% M-Joint but 8.60% A-Joint; filtering can hide groups needed for partial coverage.

5 Experimental Design

Our experimental design follows three goals: measuring propagation benefits, localizing readout failures, and validating the propagation–readout dissociation under controlled interventions.

5.1 Models, Training, and Evaluation Scope

We evaluate Qwen3-8B-Base (Yang et al., 2025) and Llama-3.1-8B (Grattafiori et al., 2024) on tests sharing neither source components nor normalized evidence–claim endpoints with training. The

structure/round matrix and six-epoch CE training were fixed beforehand; all configurations report epoch 6. Seed 42 is used unless marked as three-seed repetitions (42/43/44). FLD has a primary test and three disjoint extensions: A for training/prefix–residual controls, B for readout/source-freezing replication, and C for prospective crossed attribution. Follow-up EB $R = 3$/FLD $R = 4$ were selected after the initial sweep and fixed before extension evaluation. Main graphs use registered gold dependencies; Section 6.8 tests predicted graphs. Intervals resample sources at fixed checkpoints; SDs measure training-run variation. Appendices A–B detail chronology, statistics, and implementation.

5.2 Propagation Evaluation and Source-update Controls

At $R = 2$, the six conditions in Section 4 retain all visible evidence in the decoder and differ only in the residual branch. TRUE trains separately at $R \in \{2, 3, 4, 5\}$ for each model/dataset. Matched three-seed Qwen STATICSOURCE training freezes source states at $R = 3$ (EntailmentBank) and $R = 4$ (FLD).

5.3 Controlled Readout Probes and Endpoint Recombination

Each endpoint's target depends only on its evidence and registered justification. We recombine the fixed multihop cohort into original, reversed, (E_0, E_0), and (E_1, E_1) pairs, retaining endpoint evidence, graph, and label; matched quartet batches score both positions and Joint. Training includes partial endpoints but no partial→partial transitions. FLD has 135 claims/98 sources; partial/full analysis uses 133/97. The *graph-recoverable cohort* retains 121 claims/89 sources whose partial labels match visible-graph replay, independently of neural predictions; twelve excluded labels receive none. This separates graph information limits from learned errors. Fixed-checkpoint probes test exact copying, single-view queries, one-token prefixes, and decoder/residual logits (Appendix D.1).

6 Results: Propagation and Readout Dissociate

6.1 Registered dependencies improve selective revision

At matched $R = 2$, the registered hypergraph improves A-Joint/P-Joint over shuffled and all-evidence residuals in all six model–dataset settings, and M-Joint over both retrieval controls in all four multihop settings (Table 2). Random-k and Similarity Top-k improve preservation but reduce M-Joint in all eight comparisons. Their 235 unchanged multihop predictions all assign full/full (83 distinct claims/65 sources): high preservation can mask missed changes. The comparison attributes gains to the registered graph package, whose grouping, polarity, and dependencies vary together. TRUE also reduces A-NoChange and P-SpuriousChange relative to NOGRAPH in all six settings (Table 27), but FLD multihop revision at $R = 2$ remains a specific weakness.

6.2 The multihop gain requires updated intermediate sources

At the matched EB $R = 3$ and FLD $R = 4$ configurations, the multihop gain requires updated intermediate sources. Across all three Qwen seeds, dynamic propagation improves over matched StaticSource training by 34.34 points on EntailmentBank (SD 4.63) and 9.01 on FLD (SD 3.83; Table 3). Matched StaticSource freezes sources at $z^{(0)}$ while retaining target updates and the remaining architecture. It retains most FLD affected/preserved accuracy (93.53/97.75 versus 95.83/98.76 mean A/P), localizing the larger gain to multihop revision. On the source-disjoint FLD extension B test, freezing sources within the same trained checkpoint lowers M-Joint from 79.38% to 37.37% (Qwen) and 81.70% to 41.24% (Llama). Matched training and fixed-checkpoint interventions establish the contribution of updated sources (Appendix C.1).

The gain is configuration-dependent: shallow FLD $R = 2$ misses changes in both depth strata, and the complete sweep is non-monotonic. A fixed-checkpoint $R = 2$ ablation nevertheless shows a large residual contribution: removing the learned graph residual lowers M-Joint by 42.35/41.84 points. Source-freezing effects at this shallow configuration have intervals spanning zero, separating the clear residual benefit from the source-update gains at deeper configurations (Appendix C.2).

Table 2: Main structure comparison with CE on the source-disjoint test set. All residual arms use matched $R = 2$; NoGraph uses $R = 0$. Values are source-macro Joint accuracy (%). Bold marks the best unrounded result within each model, dataset, and metric, including ties.

	EntailmentBank			FLDx2-formal			SciFact	
Condition	A	P	M	A	P	M	A	P
Qwen3-8B-Base								
NoGraph	92.47	98.43	66.67	92.03	94.55	**89.80**	42.74	18.18
Shuffled	93.28	98.01	69.70	88.49	94.82	78.57	39.11	27.27
All-evidence res.	93.28	98.72	63.64	88.35	96.26	78.57	37.71	18.18
Random-k	95.16	99.78	45.45	90.23	97.47	70.92	68.44	90.91
Similarity Top-k	94.89	**99.85**	42.42	90.70	97.76	69.90	69.55	**100.00**
True (ours)	**97.58**	99.78	**72.73**	**92.84**	**98.73**	83.16	**71.23**	**100.00**
Llama-3.1-8B								
NoGraph	93.82	98.55	78.79	90.03	95.18	82.65	38.27	31.82
Shuffled	93.82	98.06	78.79	86.41	96.22	76.53	38.27	18.18
All-evidence res.	91.40	98.27	63.64	90.56	94.40	**86.22**	34.92	13.64
Random-k	94.62	**99.88**	39.39	88.76	98.24	65.31	**65.92**	**81.82**
Similarity Top-k	94.62	99.85	39.39	87.15	97.72	62.24	65.08	**81.82**
True (ours)	**98.66**	99.78	**84.85**	**91.90**	**98.62**	71.94	**65.92**	**81.82**

A/P/M: affected, preserved, and affected multihop claims (SciFact has no M). All evidence remains visible to the main LM in every condition; the arms differ only in the residual branch. Random and Similarity use gold evidence-count budgets.

Table 3: **Updated sources improve propagation; decoder context redirects readout.** Top: three-seed Qwen mean multihop Joint (%); Gain averages paired per-seed differences. Bottom: second-state partial-minus-full logit margins, graph-recoverable FLD (121 claims/89 sources), True R4.

Source-update control	Dynamic	StaticSource	Gain
EB, $R = 3$, three-seed mean	94.95	60.61	+34.34
FLD, $R = 4$, three-seed mean	91.84	82.82	+9.01
Readout decomposition	**Original main**	**Diagonal main**	**Fixed residual**
Qwen	13.78	0.73	−5.90
Llama	16.73	1.61	−9.52

6.3 The same endpoint, re-paired, is read differently

We now hold an endpoint fixed and vary only its companion. On the graph-recoverable cohort (121 partial claims, 89 sources), the original pairs score 94.94% Joint for Qwen and 91.01% for Llama. Pairing the same partial endpoint with a copy of itself, so that the correct answer is partial at both positions, drops Joint to 11.24% and 4.49% (Figure 3a). The full diagonal, built from the same claims, stays at 98.88%/98.31%, and the full 133-claim/97-source cohort shows the same pattern (Appendix D). All three Qwen seeds reproduce the collapse (Table 22); EB concentrates failure in multihop targets despite the same missing composition outside that stratum (Appendix F). For the two reference checkpoints, the loss sits almost entirely in the second position: first-position partial accuracy is 88.20%/86.52%, second-position 15.73%/10.11%. A largely preserved first judgment is followed by an unreliable second readout.

Exact copying measures consistency. Exact-View Copy copies the first generated state only when both evidence/graph inputs match. This label-free rule reaches 88.20%/86.52% partial-diagonal Joint, equal to first-position accuracy by construction (Table 9). It restores consistency without a second evidence judgment. First-position quality still varies with propagation: Qwen NoGraph 46.63 < Similarity 51.12 < StaticSource 67.42 < True 88.20. On non-identical inputs, it leaves the model's predictions unchanged.

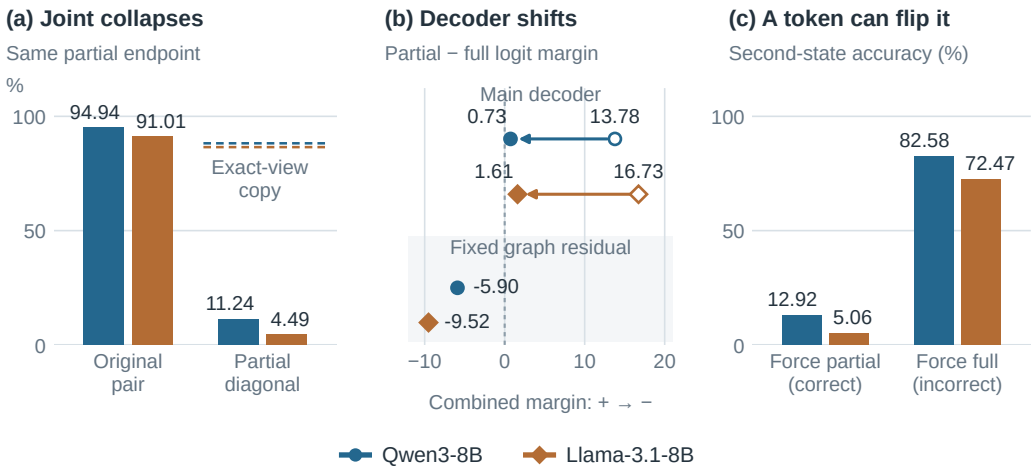

Figure 3: **The graph residual holds while one token redirects the verdict.** Graph-recoverable FLD: 121 claims/89 sources, TRUE $R = 4$; source-macro means. (a) Free-running Joint drops under partial-diagonal pairing. Dashed Exact-View Copy baselines reach 88.20/86.52%; full-diagonal Joint remains 98.88/98.31% (Qwen/Llama). (b) The main-decoder partial-minus-full margin shifts from original (open) to diagonal (filled), while the negative graph residual remains fixed; the combined margin reverses sign. (c) Conditional second-state accuracy after changing only the current first-state token. Forcing the incorrect full code improves the second judgment while making the first state wrong, leaving Joint incorrect. Paired-source effect intervals are reported in Section 6.5.

6.4 THE RESIDUAL DOES NOT MOVE; THE DECODER DOES

The view-local residual δ is bitwise identical across pairings (relative L_2 difference 0.000). Projecting the second-state readout onto partial-minus-full vocabulary-head rows, the main-decoder margin falls from 13.78 to 0.73 logits for Qwen and 16.73 to 1.61 for Llama. The fixed residual margins remain $-5.90/-9.52$, so the combined preference reverses (Table 3; Figure 3b). Pairing dependence therefore arises in the decoder path that sees both views and its prefix. The residual itself favors full: removing it at Qwen's fixed observed prefixes lifts partial accuracy from 15.46% to 37.11% but drops full accuracy from 98.97% to zero (133 claims/97 sources; Appendix D.2). Its contribution is state-dependent: unconditional removal trades partial recovery for complete loss of full-state accuracy.

6.5 A SINGLE PRECEDING STATE TOKEN CAN FLIP THE SECOND JUDGMENT

The local intervention fixes the paired prompt and generated history, replacing only the current claim's first state code. Forcing partial versus full changes second-state partial accuracy from 12.92% to 82.58% for Qwen and 5.06% to 72.47% for Llama (Figure 3c): effects of +69.66 pp [60.67,78.09] and +67.42 [58.43,76.40]. Full endpoints stay above 99.4% under either forced prefix. The correct partial prefix leaves the second-state collapse intact; the incorrect full prefix improves that judgment while making the pair wrong (Appendix D.1).

Crossed controls identify context-conditioned transition readout. We independently cross the companion evidence, the earlier generated history, and the current first-state token, holding the target evidence, residual, and output position fixed. Across four FLD checkpoints and two test cohorts, all 32 partial-target token effects are positive (19.31–92.27 pp), with positive family-adjusted lower bounds (Appendix E). Companion effects interact with the forced token, and earlier-history effects reach 30.95 points. Prospective extension C (36 claims/28 additional sources) gives 16/16 positive estimates and 15/16 positive simultaneous lower bounds; the remaining interval spans zero. Thus the state token is a strong causal route within a readout also conditioned on companion evidence and history.

The failure survives changes to the interface. Consistently permuting state codes leaves Qwen partial-diagonal Joint at 11.97%/6.93% on primary/extension B tests. Reversing output order also preserves the failure in both models, alongside high full-diagonal accuracy (Table 17). The failure survives both digit reassignment and temporal-order reversal. Configuration comparisons and the training-only transition reference appear in Appendices D.1–E.

6.6 Partial-preservation training redistributes errors

Training on the missing composition improves preservation and exposes a dosage-dependent redistribution of errors. Replacing 321 of 3,209 Qwen FLD training pairs with partial-preservation examples (Partial10) raises all-partial diagonal Joint from 9.77% to 79.12%. Matched Full10 uses full-preservation replacements and reaches only 9.39% (Appendix F). Both interventions lower original-task accuracy: M is 94.39% originally, 84.69% for Partial10, and 81.63% for Full10. The Partial10-minus-Full10 contrast is $+3.06$ pp $[-2.04, 8.67]$, showing M regression under both replacement controls. In contrast, Code 0 preservation is 7.20% versus 98.05%; 389/410 no-coverage endpoints are incorrectly emitted as partial/partial under Partial10. Replacing only 32 pairs (Partial1) reaches 44.35% partial preservation while retaining 95.72% Code 0 and 95.72% A accuracy (Table 18). Lower-dose augmentation delivers a better preservation–revision tradeoff. Missing composition exposes a failure localized by the crossed interventions to shared context.

6.7 Endpoint-factorized generation restores preservation

Endpoint-factorized readout retains both evidence views with separate generation histories and KV caches. Across all three Qwen seeds, it raises graph-recoverable partial Joint from 15.36 ± 7.14 to $90.64 \pm 1.72\%$. On all partial endpoints, diagonal Joint is 45.27%, close to Exact-View Copy's 44.64%; crucially, on non-identical, coverage-preserving inputs it improves Joint from 11.45% to 45.23%, where exact copying does not apply. Mean original A/P/M changes are $-1.92/-0.68/-2.72$ points (Table 22). The comparison measures the joint effect of separate histories and the changed grammar and objective. Gains on non-identical inputs motivate separate generation histories for independent endpoint judgments.

6.8 The dissociation persists under predicted graphs

The fixed Qwen $R = 4$ checkpoint retains both effects when oracle graphs are replaced at inference by cross-fitted predicted graphs on the same 133-claim/97-source FLD cohort. The dynamic-minus-frozen-source gain attenuates from 42.27 to 32.47 points; the preceding-token effect from 67.01 to 55.67. Partial-diagonal first/second accuracy remains sharply separated (75.26%/13.92%), versus 92.27% full-diagonal Joint. Thus the propagation/readout dissociation persists with imperfect learned structure in this fixed Qwen $R = 4$ checkpoint (Appendix G.2).

7 Conclusion

Successful evidence propagation does not guarantee stable state readout. In this setting, an unseen preservation composition exposes the dissociation. Causal interventions locate a strong route through shared autoregressive context, alongside contributions from companion evidence and earlier generation history. The pattern persists when oracle graphs are replaced with imperfect predicted graphs, suggesting that the dissociation is not an artifact of oracle graph access. Training on the missing composition improves preservation but redistributes errors, whereas endpoint-factorized generation substantially improves preservation with measured original-task costs. These results suggest that evidence propagation and state readout are distinct but interacting computational problems. These findings motivate controlling generation histories to reduce avoidable readout coupling while retaining propagation benefits for semantically independent endpoint judgments.

AI USE STATEMENT

We used generative AI tools to identify relevant literature and to assist with drafting and polishing portions of the manuscript. The authors developed the research ideas and methodology, designed and conducted the experiments, and analyzed and interpreted the results. We reviewed and edited all AI-assisted text and take full responsibility for the paper's content, claims, and conclusions.

ETHICS STATEMENT

This study uses source-disjoint tests from reasoning and fact-checking corpora. Support/opposition states describe registered evidence coverage, not calibrated probabilities or guaranteed truth. We disclose oracle inputs, distribution limits, and negative results.

REPRODUCIBILITY STATEMENT

Sections 3–5 define the task, model, and evaluation; appendices detail training, complete results, cohorts, and controls. Research artifacts bind data, configurations, predictions, and result hashes; scripts regenerate figures and tables from these records. Checkpoint selection, provenance, and batch-replay differences are documented. All quantitative claims use completed evaluations. We will release our code and constructed intervention datasets upon acceptance.

REFERENCES

Carlos E. Alchourrón, Peter Gärdenfors, and David Makinson. On the logic of theory change: Partial meet contraction and revision functions. *The Journal of Symbolic Logic*, 50(2):510–530, 1985. doi: 10.2307/2274239.

Nuel D. Belnap, Jr. A useful four-valued logic. In J. M. Dunn and G. Epstein (eds.), *Modern Uses of Multiple-Valued Logic*, volume 2 of *Episteme*, pp. 5–37. Springer, Dordrecht, 1977. doi: 10.1007/978-94-010-1161-7_2. URL https://doi.org/10.1007/978-94-010-1161-7_2.

Lukas Berglund, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. The reversal curse: LLMs trained on "A is B" fail to learn "B is A". In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2309.12288.

Eli Chien, Chao Pan, Jianhao Peng, and Olgica Milenkovic. You are AllSet: A multiset learning framework for hypergraph neural networks. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/pdf/2106.13264v4.

Peter Clark, Oyvind Tafjord, and Kyle Richardson. Transformers as soft reasoners over language. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pp. 3882–3890, 2020. doi: 10.24963/ijcai.2020/537. URL https://www.ijcai.org/proceedings/2020/537.

Roi Cohen, Eden Biran, Ori Yoran, Amir Globerson, and Mor Geva. Evaluating the ripple effects of knowledge editing in language models. *Transactions of the Association for Computational Linguistics*, 12:283–298, 2024. doi: 10.1162/tacl_a_00644. URL https://aclanthology.org/2024.tacl-1.16/.

Bhavana Dalvi, Peter Jansen, Oyvind Tafjord, Zhengnan Xie, Hannah Smith, Leighanna Pipatanangkura, and Peter Clark. Explaining answers with entailment trees. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7358–7370. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.585. URL https://aclanthology.org/2021.emnlp-main.585/.

Nicola De Cao, Wilker Aziz, and Ivan Titov. Editing factual knowledge in language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 6491–6506. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.522. URL https://aclanthology.org/2021.emnlp-main.522/.

Johan de Kleer. An assumption-based TMS. *Artificial Intelligence*, 28(2):127–162, 1986. doi: 10.1016/0004-3702(86)90080-9. URL https://doi.org/10.1016/0004-3702(86)90080-9.

Zilu Dong, Xiangqing Shen, Zinong Yang, and Rui Xia. ChainEdit: Propagating ripple effects in LLM knowledge editing through logical rule-guided chains. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13558–13571. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.665. URL https://aclanthology.org/2025.acl-long.665/.

Jon Doyle. A truth maintenance system. *Artificial Intelligence*, 12(3):231–272, 1979. doi: 10.1016/0004-3702(79)90008-0.

Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang (Lorraine) Li, Liwei Jiang, Bill Yuchen Lin, Sean Welleck, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena Hwang, Soumya Sanyal, Xiang Ren, Allyson Ettinger, Zaid Harchaoui, and Yejin Choi. Faith and fate: Limits of transformers on compositionality. In *Advances in Neural Information Processing Systems*, volume 36, pp. 70293–70332, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/deb3c28192f979302c157cb653c15e90-Abstract-Conference.html.

Yifan Feng, Haoxuan You, Zizhao Zhang, Rongrong Ji, and Yue Gao. Hypergraph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 3558–3565, 2019. doi: 10.1609/aaai.v33i01.33013558. URL https://doi.org/10.1609/aaai.v33i01.33013558.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. doi: 10.1038/s42256-020-00257-z. URL https://www.nature.com/articles/s42256-020-00257-z.

Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1263–1272. PMLR, 2017. URL https://proceedings.mlr.press/v70/gilmer17a.html.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

Peter Hase, Mona Diab, Asli Celikyilmaz, Xian Li, Zornitsa Kozareva, Veselin Stoyanov, Mohit Bansal, and Srinivasan Iyer. Methods for measuring, updating, and visualizing factual beliefs in language models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 2714–2731. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.eacl-main.199. URL https://aclanthology.org/2023.eacl-main.199/.

Peter Hase, Thomas Hofweber, Xiang Zhou, Elias Stengel-Eskin, and Mohit Bansal. Fundamental problems with model editing: How should rational belief revision work in LLMs? *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=LRf19n5Ly3.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Nora Kassner, Oyvind Tafjord, Hinrich Schütze, and Peter Clark. BeliefBank: Adding memory to a pre-trained language model for a systematic notion of belief. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 8849–8861. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.697. URL https://aclanthology.org/2021.emnlp-main.697/.

Daniel Keysers, Nathanael Schärli, Nathan Scales, Hylke Buisman, Daniel Furrer, Sergii Kashubin, Nikola Momchev, Danila Sinopalnikov, Lukasz Stafiniak, Tibor Tihon, Dmitry Tsarkov, Xiao Wang, Marc van Zee, and Olivier Bousquet. Measuring compositional generalization: A comprehensive method on realistic data. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/1912.09713.

Najoung Kim and Tal Linzen. COGS: A compositional generalization challenge based on semantic interpretation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9087–9105. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.731. URL https://aclanthology.org/2020.emnlp-main.731/.

Brenden Lake and Marco Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2873–2882. PMLR, 2018. URL https://proceedings.mlr.press/v80/lake18a.html.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8086–8098. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-long.556. URL https://aclanthology.org/2022.acl-long.556/.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17359–17372, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/6f1d43d5a82a37e89b0665b33bf3a182-Abstract-Conference.html.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.07229.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11048–11064. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.emnlp-main.759. URL https://aclanthology.org/2022.emnlp-main.759/.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D. Manning. Fast model editing at scale. In *International Conference on Learning Representations*, 2022a. URL https://arxiv.org/abs/2110.11309.

Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D. Manning, and Chelsea Finn. Memory-based model editing at scale. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 15817–15831. PMLR, 2022b. URL https://proceedings.mlr.press/v162/mitchell22a.html.

Terufumi Morishita, Gaku Morio, Atsuki Yamaguchi, and Yasuhiro Sogawa. Learning deductive reasoning from synthetic corpus based on formal logic. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 25254–25274. PMLR, 2023. URL https://proceedings.mlr.press/v202/morishita23a.html.

Terufumi Morishita, Gaku Morio, Atsuki Yamaguchi, and Yasuhiro Sogawa. Enhancing reasoning capabilities of LLMs via principled synthetic logic corpus. In *Advances in Neural Information Processing Systems*, volume 37, pp. 73572–73604, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/8678da90126aa58326b2fc0254b33a8c-Abstract-Conference.html.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In Dan Jurafsky, Joyce Chai, Natalie Schluter,

and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4902–4912, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.442. URL https://aclanthology.org/2020.acl-main.442/.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.11324.

Oyvind Tafjord, Bhavana Dalvi, and Peter Clark. ProofWriter: Generating implications, proofs, and abductive statements over natural language. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 3621–3634. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.findings-acl.317. URL https://aclanthology.org/2021.findings-acl.317/.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems*, volume 36, pp. 74952–74965, 2023. doi: 10.52202/075280-3275. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/ed3fea9033a80fea1376299fa7863f4a-Abstract-Conference.html.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 7534–7550. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.609. URL https://aclanthology.org/2020.emnlp-main.609/.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2203.11171.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837, 2022. doi: 10.52202/068431-1800. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

Bryan Wilie, Samuel Cahyawijaya, Etsuko Ishii, Junxian He, and Pascale Fung. Belief revision: The adaptability of large language models reasoning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 10480–10496. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.586. URL https://aclanthology.org/2024.emnlp-main.586/.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12697–12706. PMLR, 2021. URL https://proceedings.mlr.press/v139/zhao21c.html.

Zexuan Zhong, Zhengxuan Wu, Christopher D. Manning, Christopher Potts, and Danqi Chen. MQuAKE: Assessing knowledge editing in language models via multi-hop questions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 15686–15702. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.971. URL https://aclanthology.org/2023.emnlp-main.971/.

A EVALUATION PROVENANCE, CHRONOLOGY, AND STATISTICS

Train–test isolation. The primary tests contain 374 pairs from 186 sources (EntailmentBank), 629/400 (FLD), and 179/179 (SciFact). Source components link shared identifiers across examples and versions. Training and test have zero overlap in registered source components, normalized evidence–claim endpoints, and pair identifiers. The source-disjoint FLD extension test comprises three mutually disjoint blocks, each also disjoint from the primary test and training (Table 4). This audit establishes separation of registered sources and normalized endpoints; linguistic templates and pretraining provenance lie outside its scope.

Fixed main evaluation and targeted follow-ups. The initial protocol fixed the complete matched-$R = 2$ structure comparison, the True $R = 2, 3, 4, 5$ sweep, and six training epochs before training began. All 54 main trained configurations report epoch 6, with no best-epoch selection, and the complete sweep is reported. Follow-up propagation and readout studies selected EB $R = 3$ and FLD $R = 4$ after observing the initial sweep, then fixed those configurations before repetitions and extension evaluation. Diagnostics on the primary test use the same underlying endpoints; recombinations introduce new pairings rather than new sources.

Extension protocols. Extension A froze its 800-source admission, training controls, and prefix-by-residual analysis on September 20; B froze its separate 800-source admission and readout/source-freezing protocol on September 21. Their first evaluations replicate the corresponding fixed comparisons on new source-disjoint cohorts. The later crossed attribution study also evaluates B. Extension C is the prospective replication of that crossed study: four checkpoints, all eight cells, eligibility, and the family of 16 primary contrasts were fixed before any predictions on the additional sources on September 25. It uses all 36 eligible claims from 28 sources; its evaluation denominator is this mechanism cohort. An audit against 582 historical data files found zero admitted source or normalized-endpoint overlap. The 200-source pilot of A is a subset of A and is counted once.

Table 4: **FLD test provenance and mechanism cohorts.** The primary test and all three extension blocks are disjoint from training and one another in registered source components and normalized evidence–claim endpoints. Pairs and sources describe each constructed block; M and partial/full cells give claims/sources. Block C evaluates only its 36-claim/28-source partial/full mechanism cohort, so its evaluation denominator is that mechanism cohort.

Test cohort	Original pairs	Sources	M: n/s	Partial/full M: n/s
Primary	629	400	135/98	133/97
Extension A	1277	800	247/189	246/189
Extension B	1288	800	250/194	242/189
Extension C	207	127	37/29	36/28
Extension inventory	2772	1727	534/412	524/406

Extension A supplies the prefix-by-residual controls; B supplies endpoint-factorization, source-freezing and crossed attribution; C supplies the prospective crossed-attribution replication. Blocks A, B and C use separate frozen admission and analysis protocols. The 200-source pilot is a subset of A and is counted once. Each cohort reports the effect defined by its intervention protocol.

Aggregation. For a claim-level score z_{spc} in source component $s \in \mathcal{S}$, pair $p \in \mathcal{P}_s$, and evaluation scope C_{sp}, we compute

$$\text{Score} = \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} \frac{1}{|\mathcal{P}_s|} \sum_{p \in \mathcal{P}_s} \frac{1}{|C_{sp}|} \sum_{c \in C_{sp}} z_{spc}. \tag{5}$$

Only pairs and sources with a nonempty scope enter its denominator. Source-macro and micro counts are retained separately. An empty multihop scope is undefined, not zero. Diagonal endpoint accuracy first averages the two predictions for the same semantic endpoint within a claim, then follows the same aggregation.

State-specific repair summaries (Table 18) average diagonal endpoint observations within original pairs, then sources. Readout-interface comparisons (Table 22) first average repeated views or directions within each claim. These conventions agree on the all-partial cohort, where each claim contributes one eligible endpoint; other states can include both endpoints of a claim.

Uncertainty and denominators. We report each block under its actual intervention protocol and source denominator. The extension inventory has 2,772 pairs from 1,727 sources; each protocol uses its own evaluation cohort. In particular, A's 246 eligible claims from 189 sources and B's 242/189 are different cohorts. Ordinary fixed-checkpoint intervals use 10,000 paired-source bootstrap resamples. Crossed attribution uses a family of eight contrasts per checkpoint/cohort (four partial-target and four full-target contrasts); extension C uses 100,000 resamples and approximate two-sided Bonferroni intervals for its predeclared family of 16 partial-target contrasts. Training-seed SDs and conditional source intervals quantify different variation and are reported separately.

B Implementation details

Optimization. All trained state-prediction conditions use completion-token cross-entropy with no auxiliary loss, Adam with decoupled weight decay (AdamW; moment decays $\beta_1 = 0.9$, $\beta_2 = 0.999$, numerical stabilizer $\epsilon = 10^{-8}$, weight decay zero), learning rate 8×10^{-5}, warmup ratio 0.03, and gradient norm clipping at 1.0. Training lasts six epochs with effective batch size 16 pairs and microbatch size one. LoRA targets query, key, value, attention output, gate, up, and down projections; its rank is $r_{\mathrm{LoRA}} = 64$, $\alpha = 32$ (scaling $\alpha / r_{\mathrm{LoRA}} = 0.5$), and dropout is 0.05. Graph rank is 16 and graph learning rate is 8×10^{-5}. Models use bfloat16, scaled dot-product attention, non-reentrant gradient checkpointing, and no quantization. The sequence cap is 16,384 tokens; overlength examples fail admission rather than being silently truncated.

LoRA plus the graph branch trains 174.78M parameters for Qwen and 167.97M for Llama (2.13%/2.09% of the respective frozen base parameter counts); the graph branch has 196,640 parameters. Ranks and learning rates are fixed across the propagation/readout comparison.

Selection and decoding. The main sweep evaluates the test set after each epoch; every condition reports fixed epoch 6 without checkpoint selection. The remote repetitions and follow-up training controls evaluate only after epoch 6. The seed-44 propagation repetitions use RTX 3090 GPUs rather than A800 GPUs, with the same model revision, data, training hyperparameters, and decoding; hardware-dependent numerical differences can affect greedy decoding. Greedy decoding uses one beam and no sampling, with a per-input generation budget determined by the output grammar. Paired-task evaluation batches contain four pairs. The paired task generates freely over the full native vocabulary. The output grammar assigns before/after state codes to each claim ID, and invalid pairs are recorded explicitly. Gold targets enter supervised training and post-generation evaluation, not inference inputs.

View-local evidence representations and controlled readout. The auxiliary sequence for each view lists its evidence before all query claims; its hidden states come from a different forward context than the joint-pair decoder, with shared LM and LoRA parameters. For hidden dimension D, let h_j^{view} be the final-normalized hidden vector at position j and T_n the token positions of node n. Evidence node e pools T_e; claim node c uses its final token:

$$z_e^{(0)} = \frac{1}{|T_e|} \sum_{j \in T_e} h_j^{\mathrm{view}}, \qquad z_c^{(0)} = h_{\mathrm{last}(T_c)}^{\mathrm{view}}. \tag{6}$$

The output grammar identifies the claim and endpoint at each state-code position; `before_view` maps that endpoint to its evidence view. The corresponding residual modifies the final-normalized main hidden state used to predict the next token, immediately before the vocabulary head. Other positions receive zero residual. Head-local injection keeps propagated contributions outside the transformer stream and KV cache, so they can remain fixed while decoder context changes. Generated states remain in the joint prefix, coupling endpoint judgments despite view-local graph branches.

Training objective and graph initialization. For batch B, completion-token positions Y_i, target tokens $a_{i,j}$, and visible groups $\mathcal{G}_{i,v}$ for example i and view v, the training objective is

$$\mathcal{L}_{\mathrm{CE}} = -\frac{1}{|B|}\sum_{i\in B}\frac{1}{|Y_i|}\sum_{j\in Y_i}\log p_\theta(a_{i,j} \mid x_i, \mathcal{G}_{i,0}, \mathcal{G}_{i,1}, a_{i,<j}). \tag{7}$$

Training uses teacher forcing, averaging completion-token CE within each pair and then across pairs; only LoRA and graph parameters are trained, while the base model and native vocabulary head remain frozen. The completion includes claim identifiers, separators, state codes, and the end-of-sequence (EOS) token; prompt tokens are masked. Graph inputs are omitted for NoGraph. The shared projections have dimensions $W_\downarrow \in \mathbb{R}^{16\times 2D}$ and $W_\uparrow \in \mathbb{R}^{D\times 16}$. The two polarity vectors $s_{p_g} \in \mathbb{R}^{16}$ are learned and initialized to ones, so negative polarity is not a fixed sign inversion. The up-projection is zero-initialized.

Frozen retrieval encoder. Similarity uses `Alibaba-NLP/gte-Qwen2-1.5B-instruct`, revision `a9af15a6372d7d6b25e9fb07c2ccb9e1fe645644`. Claims receive the prefix `Instruct: Retrieve evidence relevant to the given claim.`, followed by a newline and `Query: `; evidence receives no prefix. We take the final non-padding token vector, normalize it in L_2, rank evidence by cosine similarity, and break ties by evidence ID. Encoding uses bfloat16 without text truncation, with a 16,384-token admission cap. The encoder is frozen, with no supervised fitting or threshold calibration. Each claim/view budget equals its number of unique visible gold evidence ancestors; a zero budget yields no retrieval group.

Checkpoints and inference reproducibility. Artifacts bind model identity, configuration, data, code, epoch, and predictions. Attention and key–value caches are isolated per input; batch composition does not provide access to other examples. Bfloat16 greedy predictions can nevertheless vary across batch shapes or hardware. Diagnostic original pairs are therefore freshly decoded in matched quartet batches, and fixed-batch repeat checks accompany the local interventions. Comparisons use those matched outputs rather than archived predictions from another execution context.

Source-specific construction. EntailmentBank uses proof leaves as evidence and grouped proof steps for intermediate claims and the hypothesis. Its test has 187 proof-leaf deletions and 187 redundancy controls (duplicate a leaf and its dependent groups, then delete the original). FLD uses supplied facts/rules and admitted forward/refutation proofs, considers used/unused evidence deletions, and admits quantifier-free instances only when SAT checks agree with full support/opposition at both endpoints; partial versus none remains proof-coverage dependent. SciFact uses abstract sentences and rationale sets, with 23 sentence deletions and 156 rationale-set deletions. All labels are recomputed after intervention; temporal direction is assigned independently. Targets use registered $\mathcal{H}$; propagation receives $\mathcal{G}_v$, which omits entire groups with missing source identifiers.

Representational scope. Codes 0, 1, 2, 3, and 6 denote none/none, none/partial, none/full, partial/none, and full/none, respectively. States 4, 5, 7, and 8 represent simultaneous nonzero support and opposition. The present datasets instantiate the five single-polarity states listed above, which define the empirical scope. Registered coverage also differs from unconstrained logical entailment: the partial/none distinction depends on the admitted proof or rationale structure. This is why the visible-graph recovery audit accompanies the endpoint diagnostic.

C PROPAGATION CONTROLS AND DEPTH

C.1 MATCHED PROPAGATION CONTROLS

Table 5 combines all matched propagation seeds and the additional FLD reference arms.

Table 5: Qwen propagation repetitions, fixed epoch 6. Top: all matched dynamic/StaticSource M scores and paired-source 95% intervals (pp), at EB $R = 3$ (33 sources) and FLD $R = 4$ (98 M sources). Mean gains are 34.34 (SD 4.63) and 9.01 (SD 3.83); SD is across training runs, not sources. Bottom: all FLD reference and propagation runs, A/P/M source-macro percentages on the common primary test.

Data	Seed	Dynamic M	StaticSource M	Gain [95% interval]
EB	42	96.97	57.58	+39.39 [24.24,54.55]
EB	43	93.94	63.64	+30.30 [12.12,48.48]
EB	44	93.94	60.61	+33.33 [18.18,48.48]
FLD	42	95.41	84.18	+11.22 [5.10,17.86]
FLD	43	92.86	81.63	+11.22 [4.59,17.86]
FLD	44	87.24	82.65	+4.59 [−2.55,11.73]

Condition	Seed 42 A/P/M	Seed 43 A/P/M	Seed 44 A/P/M
NoGraph	92.03/94.55/89.80	87.62/95.23/82.65	87.75/95.20/77.04
Similarity	90.70/97.76/69.90	92.03/97.67/75.00	89.89/96.78/73.47
True	96.45/98.93/95.41	96.25/98.39/92.86	94.78/98.98/87.24
StaticSource	93.78/97.45/84.18	93.11/97.76/81.63	93.71/98.03/82.65

Separately trained StaticSource retains target updates, trainable parameters, and gradient flow, matching dataset, seed, graph, and rounds. All six dynamic-minus-static M estimates are positive; the FLD seed-44 source interval includes zero. Training-seed SDs and fixed-checkpoint source intervals measure different uncertainty.

Fixed-checkpoint source freezing. On extension B, source freezing at inference preserves the trained weights, graph topology, target updates, and readout interface, but supplies initial source states at each propagation round. Qwen $R = 4$ M-Joint drops from 79.38% to 37.37%; Llama drops from 81.70% to 41.24% (250 claims/194 sources). This intervention complements the separately trained StaticSource comparison. It differs from zeroing the graph residual at the vocabulary head, which tests the immediate contribution of the entire residual rather than intermediate-source updating.

C.2 DIAGNOSING THE SHALLOW FLD DEFICIT

Registered intervention depth. The depth field records affected claims with $d \geq 2$; missing entries must not be treated as depth zero. EntailmentBank test has 33 multihop claims, all at $d = 2$. FLD has 98 at $d = 2$ and 37 at $d = 3$. SciFact has no multihop scope. Table 6 reports every observed stratum for the seven reported conditions, separately from neural rounds R.

Depth-level effects are non-monotonic: Llama FLD $d = 3$ scores 70.27% at $R = 4$, below NoGraph's 75.68%, but 81.08% at $R = 5$. Each R is separately trained. Sources can span depths, so aggregate M must be reconstructed from claim records rather than averaged over strata. The complete sweep remains in Appendix I.

Table 6: Joint accuracy by registered intervention-propagation depth d, on the evaluation partition. The n/s column gives claims / source components. Scores are source-macro percentages; d is a data annotation and R is the model's message-passing round count. All observed multihop depths are retained.

		NoGraph	True				Random	Similarity
d	n/s	R=0	R=2	R=3	R=4	R=5	R=2	R=2
Qwen / EntailmentBank								
2	33 / 33	66.67	72.73	96.97	96.97	96.97	45.45	42.42
Llama / EntailmentBank								
2	33 / 33	78.79	84.85	96.97	96.97	96.97	39.39	39.39
Qwen / FLDx2-formal								
2	98 / 98	89.80	85.71	91.84	96.94	95.92	71.43	69.39
3	37 / 37	78.38	67.57	78.38	86.49	91.89	54.05	59.46
Llama / FLDx2-formal								
2	98 / 98	82.65	72.45	85.71	95.92	97.96	63.27	60.20
3	37 / 37	75.68	59.46	67.57	70.27	81.08	48.65	54.05

Each depth is aggregated as claim means within pairs, pair means within source components, and an equal-source mean. Sources may occur at multiple depths; these rows are not independent cohorts and their means should not be averaged to reconstruct the overall M score. SciFact has no registered multihop cohort. NoChange scores and all micro counts are retained in the accompanying depth audit.

Table 7 distinguishes missed changes from spurious changes. Relative to NoGraph, True $R = 2$ increases M-NoChange by 4.59/7.65 points (Qwen/Llama), while decreasing P-SpuriousChange by 2.55/2.00 points and improving affected non-M accuracy. Its 20 Qwen and 38 Llama unchanged multihop prediction records all assign full coverage at both endpoints. Among claims correct under NoGraph but incorrect under $R = 2$, 15/19 and 22/25 become full/full; conversely, $R = 2$ corrects 11 and nine formerly incorrect claims. The deficit therefore concentrates on multihop partial-coverage judgments. Full/full assigns complete coverage to both temporal endpoints, collapsing the required change in either direction.

Table 7: FLD shallow-propagation error profile (seed 42; source-macro percentages). M: 135 claims/98 sources; affected non-M: 249/249; preserved: 705/384. M-NoChange and M-WrongChange partition M errors here; all four conditions have valid outputs. The models are separately trained, so these are behavioral comparisons.

Model	Condition	M-Joint	M-NoChange	M-Wrong	A non-M	P-Spurious
Qwen	NoGraph	89.80	7.65	2.55	91.16	3.29
Qwen	True $R = 2$	83.16	12.24	4.59	95.18	0.73
Llama	NoGraph	82.65	17.35	0.00	89.96	2.22
Llama	True $R = 2$	71.94	25.00	3.06	97.59	0.23

The depth-3 stratum has larger percentage losses, but depth-2 also declines (Table 6). Preserving each claim's weight in the full M score, the depth-2/depth-3 contributions to the overall difference are $-4.59/-2.04$ points for Qwen and $-7.65/-3.06$ for Llama. These contributions sum to the overall loss; separately normalized depth scores do not. Supporting and opposing partial/full transitions both lose accuracy. Together, these strata localize the shortfall to missed partial-coverage changes across both polarities and depths.

Rounds and selection. Registered depth d is an intervention path; R counts learned updates, each using all visible groups and contextual initial states. Selecting R for a new setting requires a separate tuning split and joint revision/preservation criteria. The actual selection chronology is in Appendix A.

Separating residual utility from source-update dependence at $R = 2$. On the full primary FLD M cohort (135 claims/98 sources), matched inference gives normal/source-frozen/residual-zero M-Joint of 83.67/82.14/41.33% for Qwen and 71.94/76.53/30.10% for Llama. Normal-minus-zero is +42.35/+41.84 points, with family-of-four intervals [25.51,58.16]/[25.51,57.14]. Normal-minus-frozen is +1.53/-4.59, with intervals [-2.55,7.14]/[-11.22,1.02]. The residual benefit is large, whereas both source-update intervals span zero. This fixed-checkpoint contrast separates residual utility from source-update dependence; the trained $R = 2$ versus NoGraph comparison measures the effect of the full training configuration. Qwen's matched normal replay differs from its historical score by 0.51 points; only matched batches enter intervention contrasts. Source freezing measures the aggregate effect of updated source activations.

D ENDPOINT RECOMBINATION AND READOUT LOCALIZATION

Construction and recovery. Reversal swaps before/after assignment and targets without exchanging evidence views. Diagonals duplicate one evidence view E_v and its visible graph $\mathcal{G}_v$; labels defined by $(E_v, \mathcal{H})$ and original source weights are retained. Analysis follows the original multihop cohort although diagonal targets become preserved. Partial diagonals pair familiar endpoints in an unseen preservation composition. Replay on registered $\mathcal{H}$ reproduces all 270 endpoint labels for 135 original M claims. Replay on filtered $\mathcal{G}_v$ recovers 121 of 133 partial targets (89 sources); twelve receive none (11 sources, overlapping). Admission uses visible-graph replay agreement with registered labels, not neural correctness. The scanned raw/canonicalized FLD train, test, and diagnostic inputs contain zero exact input–label conflicts.

Table 8: Endpoint recombination and position localization (source-macro %). Each cell P/F reports partial/full endpoints of the same claims. O/R score the endpoint in the original/reversed pair; D1/D2 are diagonal positions; DJ requires both diagonal predictions correct. The endpoint mean is $(\mathrm{D1} + \mathrm{D2})/2$. Recoverability is selected by graph–label agreement, not model predictions.

Model	Condition	O (P/F)	R (P/F)	D1 (P/F)	D2 (P/F)	DJ (P/F)
Full cohort: 133 claims / 97 sources						
Qwen	NoGraph R0	89.18/96.91	86.60/96.91	42.27/94.33	7.73/92.78	0.00/88.66
Qwen	Similarity R2	70.10/94.85	66.49/95.88	47.94/95.88	7.22/97.42	5.15/95.36
Qwen	True R4	94.33/98.45	94.33/97.94	82.47/98.97	14.95/98.97	11.34/98.97
Qwen	StaticSource R4	85.05/96.39	82.47/95.88	60.82/93.81	6.19/96.91	1.03/93.81
Llama	NoGraph R0	82.47/98.97	81.44/98.97	26.29/91.75	14.95/94.33	0.00/87.63
Llama	True R4	91.75/97.42	91.75/97.42	85.57/98.97	13.40/98.45	8.25/98.45
Graph-recoverable cohort: 121 claims / 89 sources						
Qwen	NoGraph R0	–	–	46.63/93.82	8.43/92.13	0.00/87.64
Qwen	Similarity R2	–	–	51.12/95.51	5.62/97.19	3.37/94.94
Qwen	True R4	–	–	88.20/98.88	15.73/98.88	11.24/98.88
Qwen	StaticSource R4	–	–	67.42/94.38	6.74/97.75	1.12/94.38
Llama	NoGraph R0	–	–	29.21/92.13	16.29/93.82	0.00/87.64
Llama	True R4	–	–	86.52/98.88	10.11/98.31	4.49/98.31

Table 9: An exact-input copy baseline on the post-hoc graph-recoverable FLDx2-formal cohort (121 claims, 89 sources). Joint and Exact copy score partial-diagonal pairs; Full copy scores corresponding full diagonals. Rates are source-macro percentages. Copy replaces the second state with the first only when visible evidence and supplied graph inputs match exactly; malformed pairs remain failures. Its partial Joint equals first-position accuracy by construction: the second output reuses the first judgment.

Model	Condition	Joint	Exact copy	Full copy
Qwen	NoGraph	0.00	46.63	93.82
Qwen	Similarity	3.37	51.12	95.51
Qwen	True R4	11.24	88.20	98.88
Qwen	StaticSource	1.12	67.42	94.38
Llama	True R4	4.49	86.52	98.88
Llama	NoGraph	0.00	29.21	92.13

D.1 Endpoint capability probe and local prefix interventions

Single-endpoint capability probe. The same Qwen/Llama True $R = 4$ checkpoints receive one evidence view, its visible graph, and the original ordered claim list, predicting one state per claim with greedy native-vocabulary decoding. Claims within an endpoint remain autoregressive. Invalid completions count as errors; Table 12 separately reports validity and routed-token accuracy. This fixed-checkpoint prompt/grammar probe is distinct from the trained endpoint-factorized interface in Section 6.7. Endpoints correct independently and in their original pair still fail on the diagonal: Joint is 14.06%/5.00% on Qwen's 74-claim/64-source and Llama's 65/60 subsets. These prediction-selected subsets furnish concrete cases of pairing-induced failure.

Table 10: Partial endpoint accuracy (source-macro %) on graph-recoverable FLDx2-formal (121 claims/89 sources). True $R = 4$.

	Diagonal		Independent	First-state prefix		
Model	Pos. 1	Pos. 2	Strict	Free	Partial	Full
Qwen	88.20	15.73	67.98	16.29	12.92	82.58
Llama	86.52	10.11	59.55	10.11	5.06	72.47

Endpoint scores, not Joint. Independent changes prompt/grammar; invalid outputs count as errors. Prefix columns score the second state with preceding outputs fixed, changing only the current first-state token to observed (Free), partial, or full. Partial/Full denote coverage states.

Local single-token intervention and replay. At each second-state position, replay fixes the paired input and all earlier generated tokens, replacing only the current claim's first code with partial or full on its active polarity axis. Full diagonals are controls. Observed/partial/full/duplicate-observed prefixes share a four-item batch; duplicate shams agree, and all target positions are traceable. All 388 quartet completions per model match the preceding diagnostic. Qwen recoverable free-prefix accuracy is 16.29%, versus 15.73% in cached generation; Llama is 10.11% in both. Intervention contrasts therefore use matched replay, unrestricted vocabulary argmax, and full-vocabulary probabilities. Label-assisted scores measure conditional second-state accuracy; free-running Joint scores both generated states. Paired-source intervals use 10,000 draws conditional on fixed checkpoints.

Table 11: Paired source contrasts on graph-recoverable partial endpoints (percentage points; 95% percentile bootstrap intervals, 10,000 resamples). Checkpoints and sources are fixed.

Contrast	Qwen	Llama
Independent − diagonal position 2	+52.25 [+40.45, +63.48]	+49.44 [+37.64, +60.67]
Correct partial prefix − matched free	-3.37 [-7.30, -0.56]	-5.06 [-10.11, -0.56]
Counterfactual full − partial prefix	+69.66 [+60.67, +78.09]	+67.42 [+58.43, +76.40]
Gold probability: full − partial prefix	+66.07 [+58.18, +73.74]	+63.92 [+56.17, +71.48]

Intervals quantify source-sampling variation across the same 89 sources conditional on the fixed checkpoints. The independent contrast changes the prompt/grammar; prefix contrasts are local single-token interventions within the original paired grammar.

Complete decoding controls. Tables 12–13 retain both the full and graph-recoverable cohorts, both target states, completion validity, and full-vocabulary probabilities for the endpoint and prefix diagnostics.

Table 12: Endpoint decoding controls (source-macro %). All scores use the two existing True $R = 4$ checkpoints without training. Full cohort: 133 claims/97 sources; graph-recoverable: 121 claims/89 sources.

Model	State	Original	Diag. 1	Diag. 2	Indep.	Valid	Routed
Full cohort							
Qwen	Partial	94.33	82.47	14.95	63.92	96.91	64.95
Qwen	Full	98.45	98.97	98.97	88.14	100.00	88.14
Llama	Partial	91.75	85.57	13.40	59.79	91.75	65.46
Llama	Full	97.42	98.97	98.45	80.41	90.72	89.69
Graph-recoverable							
Qwen	Partial	94.94	88.20	15.73	67.98	96.63	69.10
Qwen	Full	98.31	98.88	98.88	88.20	100.00	88.20
Llama	Partial	91.01	86.52	10.11	59.55	91.01	65.17
Llama	Full	97.19	98.88	98.31	79.78	89.89	89.89

Original measures the selected endpoint in its original pairing, not pairwise Joint. Indep. requires a valid complete single-view output; Valid is its format-validity rate. Routed scores the native full-vocabulary argmax at the claim's routed state position, irrespective of completion validity. Full-state controls also degrade under the independent prompt, which changes both the visible views and the trained output grammar. All diagnostic conditions share the FLD primary test cohort.

Table 13: Prefix intervention controls. Accuracy and gold-state probability (%) use the full vocabulary. The intervention keeps the free-running prefix fixed except for the current first-state token; every selected claim is covered.

		Second-state accuracy			Gold-state probability		
Model	State	Free	Partial	Full	Free	Partial	Full
Full cohort							
Qwen	Partial	15.46	14.95	81.96	17.32	17.07	80.88
Qwen	Full	98.97	100.00	100.00	98.34	99.94	99.37
Llama	Partial	13.40	10.82	74.23	12.97	11.03	72.02
Llama	Full	98.45	100.00	99.48	97.41	99.99	98.44
Graph-recoverable							
Qwen	Partial	16.29	12.92	82.58	17.82	15.07	81.13
Qwen	Full	98.88	100.00	100.00	98.22	99.93	99.35
Llama	Partial	10.11	5.06	72.47	10.26	5.98	69.90
Llama	Full	98.31	100.00	99.44	97.30	99.99	98.43

Free is matched prefix replay, the appropriate numerical control for forced-prefix comparisons; Qwen's original diagonal second-state scores differ slightly from replay (Table 12). Partial/Full specify forced coverage with the endpoint's original polarity. For a partial endpoint, Full is deliberately incorrect. These label-assisted interventions are diagnostic, not deployment scores.

Stable residual, changing decoder. The residual is bitwise identical across pairings (maximum difference zero), consistent with its view-local construction. Main hidden-state cosine similarity is 0.949/0.733; partial-minus-full margins fall from 13.78/16.73 to 0.73/1.61. Fixed residual margins of $-5.90/-9.52$ favor full; the decoder-margin shift reverses the combined preference. Individual branch projections use float32 and combined logits model precision; rounding can affect exact additivity. Observational decomposition is complemented by the removal intervention below.

Training-only transition reference. The empirical modal successor of a first partial code is full (215/215 for code 1; 512/512 for code 3). Full is also modal after full (575/888 for code 2; 2,220/2,860 for code 6). A first-code-only conditional-mode predictor emits full after either prefix, contrary to the observed partial recovery after a full prefix. Restricting counts to affected claims would require gold affectedness at inference. The relevant readout depends on the broader context, as the crossed interventions show.

D.2 Local residual intervention

We cross each observed/partial/full prefix with the learned residual versus zero at the current vocabulary head, holding the main hidden vector fixed within each prefix. Earlier outputs are not regenerated. NoGraph identity checks agree exactly because its residual is zero. Missing routes count as errors. Table 14 retains both target states and all graph-model outcomes; repeated NoGraph cells are omitted.

Table 14: Local prefix-by-residual readout controls (source-macro accuracy, %). Primary has 133 partial/full target claims from 97 sources, extension A 246/189, and EB 33/33. These are fixed-prefix next-token judgments, not free-running Joint accuracy.

Model/arm	Cohort	Target	Observed prefix		Partial prefix		Full prefix	
			With	Zero	With	Zero	With	Zero
Qwen True R4	Primary	Partial	15.46	37.11	14.95	34.54	81.96	94.33
Qwen True R4	Primary	Full	98.97	0.00	100.00	37.63	100.00	0.00
Qwen True R4	Extension A	Partial	15.87	26.46	29.37	18.52	74.60	90.21
Qwen True R4	Extension A	Full	88.62	0.00	100.00	8.73	100.00	0.00
Qwen True R2	EB	Partial	27.27	100.00	27.27	100.00	18.18	100.00
Qwen True R2	EB	Full	100.00	6.06	100.00	0.00	100.00	6.06
Llama True R4	Primary	Partial	13.40	49.48	10.82	44.85	74.23	96.91
Llama True R4	Primary	Full	98.45	0.00	100.00	42.27	99.48	0.00
Llama True R4	Extension A	Partial	11.38	46.30	11.11	45.77	67.99	97.88
Llama True R4	Extension A	Full	91.53	0.00	100.00	25.66	100.00	0.00
Llama True R2	EB	Partial	9.09	100.00	6.06	100.00	33.33	100.00
Llama True R2	EB	Full	100.00	0.00	100.00	0.00	100.00	0.00

On primary FLD, removal improves observed-prefix partial accuracy (15.46 to 37.11% Qwen; 13.40 to 49.48% Llama) but destroys full accuracy (98.97/98.45% to zero). Extension A repeats the direction; EB also gains partial accuracy while losing full accuracy. The residual's contribution depends on state and prefix. Unconditional removal trades partial recovery for full-state failure. Qwen EB $R = 2$ has a negative local token-effect estimate, further locating the dissociation by configuration.

E Crossed attribution and interface robustness

Separating token, companion, and earlier history. For a fixed target endpoint, we cross three binary factors: companion evidence $C \in \{F, P\}$, the current claim's preceding state code $T \in \{F, P\}$, and the remaining generated prefix $H \in \{H_F, H_P\}$. Here H_F and H_P are the natural histories generated with full and partial companions, not gold labels assigned to a history. We replay each history across both companions and force only the current preceding state code. Target evidence, visible graph $\mathcal{G}_v$, output position, and graph residual remain fixed. The full-vocabulary argmax scores the second state. Replays with unchanged inputs and prefixes provide sham controls; full targets provide state controls. These local conditional judgments are distinct from free-running Joint accuracy.

Table 15 reports all four partial-target token contrasts per checkpoint/cohort. All 32 FLD contrasts are positive (19.31–92.27 points), as are all eight EB $R = 3$ contrasts. Each has a positive lower bound under a family-of-eight correction applied separately within each checkpoint/cohort. All bound result-file hashes pass verification and all selected FLD targets have routed outcomes.

Table 15: Crossed attribution on partial targets. Each entry is the full-minus-partial preceding-token effect on second-state accuracy (percentage points), holding target evidence, graph residual, companion C, and remaining history H fixed. Remaining histories H_F/H_P are naturally generated under full/partial companions; C_F/C_P independently set the companion view. All entries have positive family-of-eight Bonferroni interval lower bounds (eight includes four full-target controls within each checkpoint/cohort); this is not joint correction across rows. Counts are eligible claims/sources.

Dataset	Model/seed	n	H_F, C_F	H_F, C_P	H_P, C_F	H_P, C_P
Extension B	Qwen 42	242/189	79.63	75.13	53.97	47.88
Primary	Qwen 42	133/97	81.44	82.99	67.01	67.53
Extension B	Llama 42	242/189	91.53	91.27	67.99	62.96
Primary	Llama 42	133/97	92.27	91.75	63.92	63.92
Extension B	Qwen 43	242/189	77.25	75.93	54.76	48.94
Primary	Qwen 43	133/97	80.41	80.41	54.64	47.94
Extension B	Qwen 44	242/189	69.58	62.70	42.59	19.31
Primary	Qwen 44	133/97	78.87	79.38	60.82	54.64
EB	Qwen 42	33/33	69.70	54.55	75.76	36.36
EB	Llama 42	33/33	36.36	30.30	42.42	36.36

Additional contextual paths. Companion evidence and earlier history contribute alongside the strong token effect. On extension B, Qwen seed 44 with H_P has a companion contrast $C_F - C_P$ of $+11.11$ points $[6.35, 16.14]$ under T_F, versus -12.17 $[-17.20, -7.41]$ under T_P. Their interaction is $+23.28$ $[16.40, 30.69]$. With C_P, T_F fixed, changing H_P to H_F yields $+30.95$ $[24.87, 37.30]$. These are secondary pointwise 95% source intervals. Across FLD conditions, companion contrasts range from -12.17 to $+11.11$ points and history contrasts from -15.98 to $+30.95$. Naturally generated first codes can also remain unchanged while second-state correctness changes: five Qwen claims and three Llama claims provide such counterexamples, spanning seven distinct sources in their union. Thus generated state, companion evidence, and earlier history jointly condition the readout.

Prospective replication. Extension C keeps the four FLD checkpoints, eligibility rule, crossed protocol, and statistical family fixed before any predictions on the additional sources. All 16 primary point estimates are positive; 15 simultaneous intervals exclude zero (Table 16). The Qwen-44 H_P, C_P estimate is $+14.29$ points, ordinary 95% interval $[-7.14, 35.71]$, simultaneous interval $[-17.86, 46.43]$. All 16 conditions and the original cohort enter the analysis. Full-target secondary effects are zero in 14 of 16 cells and -3.57 points in two Qwen-42 cells; these values describe observed fixed-checkpoint effects. All sham checks pass, and no primary routed response is missing.

Table 16: Prospective replication on 36 partial targets from 28 sources. Entries give the full-minus-partial preceding-token effect on second-state accuracy in percentage points, with approximate family-of-16 Bonferroni intervals from 100,000 paired-source bootstrap draws. All 16 estimates are positive; 15 intervals exclude zero. The final Qwen-44 interval spans zero, so 15 of 16 contrasts meet the predeclared simultaneous threshold. The checkpoints share the same 28 sources.

Checkpoint	History	Companion	Effect [simultaneous 95% interval]
Qwen 42	F	F	+69.64 [44.64, 91.07]
	F	P	+67.86 [42.86, 89.29]
	P	F	+50.00 [25.00, 75.00]
	P	P	+39.29 [8.93, 67.86]
Llama 42	F	F	+85.71 [64.29, 100.00]
	F	P	+92.86 [77.23, 100.00]
	P	F	+58.93 [28.57, 85.71]
	P	P	+53.57 [23.21, 80.36]
Qwen 43	F	F	+85.71 [64.29, 100.00]
	F	P	+75.00 [50.00, 94.64]
	P	F	+58.93 [33.93, 82.14]
	P	P	+41.07 [10.71, 69.64]
Qwen 44	F	F	+75.00 [50.00, 94.64]
	F	P	+66.07 [41.07, 89.29]
	P	F	+39.29 [16.07, 64.29]
	P	P	+14.29 [−17.86, 46.43]

Configuration boundary on EntailmentBank. EB has 33 multihop targets from 33 sources. Both models have positive crossed token effects at $R = 3$ (Table 15). At $R = 2$, however, the local full-minus-partial token effect is -9.09 points $[-27.27, 9.09]$ for Qwen and $+27.27$ $[12.12, 42.42]$ for Llama. The corresponding fixed-checkpoint dynamic-minus-frozen-source M effects are $+6.06$ $[-9.09, 21.21]$ and $+3.03$ $[-15.15, 21.21]$; at $R = 3$ their point estimates are $+69.70$ and $+81.82$. These are separately trained R configurations. These results establish how the dissociation varies across neural round settings.

Training-consistent interface changes. Table 17 changes output order or code identity during both training and evaluation. Reversed order emits the after-state first and the before-state second; endpoint routing and supervision change consistently. The Qwen code permutation maps c to $(c + 4) \bmod 9$ in instructions and targets and applies the inverse map in scoring. All runs start from the Base model and train for six epochs and 1,206 optimizer updates. Partial-diagonal failure persists with either change, while original multihop and full-diagonal accuracy remain high. These controls establish that partial-diagonal failure survives both digit reassignment and temporal-order reversal; the crossed experiments separately quantify token-intervention effects. Endpoint-factorized training retains both evidence views but separates generation histories and KV caches. This comparison measures the joint effect of separate histories, grammar, and objective.

Table 17: Training-consistent interface controls (seed 42, True $R = 4$, six epochs and 1,206 updates). Reversing endpoint output order and consistently permuting digits preserve low partial-preservation accuracy despite high original multihop accuracy. Permutation maps c to $(c+4) \bmod 9$ throughout instruction, supervision, and inverse scoring. Partial diagonal cohorts: 276 claims/174 sources and 543/361, respectively; original M: 135/98 and 250/194. These controls establish persistence of the failure; token interventions under permuted encoding are not evaluated.

		Primary			Extension B		
Model	Interface	M	Partial	Full	M	Partial	Full
Qwen	Joint	94.39	9.77	99.48	79.38	5.17	95.06
Qwen	Reversed output order	94.39	8.72	98.94	81.19	7.29	94.56
Qwen	Consistent code permutation	92.86	11.97	99.02	80.67	6.93	94.30
Qwen	Endpoint-factorized	90.31	46.74	99.02	72.42	26.13	94.83
Llama	Joint	91.84	10.25	99.39	81.70	6.28	92.24
Llama	Reversed output order	94.39	6.70	99.01	82.73	3.19	93.84
Llama	Endpoint-factorized	91.33	49.43	99.09	75.26	34.86	93.94

F TRAINING-DISTRIBUTION AND REPAIR CONTROLS

Matched evaluation and scope. The FLD training controls start from Base and train six epochs/1,206 steps. Partial10/Full10 replace the same 321 of 3,209 pairs; Partial1/Full1 replace the same 32. The controls match replaced rows and optimizer updates; state marginals, token counts, and compute vary with the replacement content. Independent is a trained single-view interface, distinct from the fixed-checkpoint probe in Appendix D.1; it changes the available evidence context and output grammar. Table 18 uses matched quartet replays on the source-disjoint primary test; these can differ from canonical batch scores in Table 5.

Table 18: Qwen FLD training controls. All entries are percentages. Partial and code 0 (no coverage) are source-macro diagonal Joint scores over 276 endpoint observations/174 sources and 410/257, respectively; Balanced weights the five observed states equally. Both scopes include endpoints beyond the multihop cohort.

Graph	Training	A	P	M	Partial	Code 0	Balanced
NoGraph	Original replay	91.63	94.77	88.78	0.00	92.41	56.00
NoGraph	Partial10	84.94	94.04	77.55	69.64	6.23	53.68
NoGraph	Full10	86.14	95.63	77.04	0.00	89.11	57.20
NoGraph	Independent	56.56	87.29	61.73	49.14	84.44	74.09
True $R = 4$	Original replay	95.78	99.06	94.39	9.77	96.89	62.24
True $R = 4$	Partial10	91.03	98.79	84.69	79.12	7.20	69.98
True $R = 4$	Full10	91.77	98.96	81.63	9.39	98.05	61.95
True $R = 4$	Partial1	95.72	98.61	90.82	44.35	95.72	76.78
True $R = 4$	Full1	95.92	98.98	91.33	7.38	97.67	61.62
True $R = 4$	Independent	70.15	94.12	81.63	64.46	88.33	82.35

Separating total regression from composition-specific errors. Partial10 raises partial preservation by 69.35 points but lowers A/M by 4.75/9.69. Full10 also lowers A/M by 4.02/12.76. Partial10-minus-Full10 is -0.74 A [-3.82,2.21] and +3.06 M [-2.04,8.67]. Both replacement controls incur A/M regression; their direct contrasts span zero. The state-specific analysis below identifies the distinctive Partial10 cost (Table 19).

Table 19: Matched replacement contrasts, percentage points with pointwise paired-source 95% intervals at seed 42. Identical evaluation batches and source weights are used in each contrast. Partial/Full replacements match rows and updates, but not state marginals or text budgets.

Graph	Metric	Partial10 − Full10	Full10 − Original
True $R = 4$	A	-0.74 [-3.82, 2.21]	-4.02 [-6.76, -1.34]
	M	$+3.06$ [-2.04, 8.67]	-12.76 [-19.39, -6.63]
	Code 0	-90.86 [-94.16, -87.35]	$+1.17$ [-0.78, 3.31]
NoGraph	A	-1.20 [-5.49, 3.15]	-5.49 [-9.10, -2.14]
	M	$+0.51$ [-7.14, 8.16]	-11.73 [-20.41, -3.06]
	Code 0	-82.88 [-88.13, -77.43]	-3.31 [-6.23, -0.58]

The state-specific error flow. Partial10 emits partial/partial for 389 of 410 no-coverage endpoint observations (218 code-1, 171 code-3 pairs), leaving 21 correct versus 399 originally and 404 under Full10. These micro counts differ from source-macro scores. Code-0 training endpoints actually increase from 1,942 to 1,984, and code-0-preserving claims from 715 to 757. Both ten-percent controls reduce affected claims from 1,928 to 1,387, but only Partial10 adds 499 partial-preserving claims. The collapse is state redistribution, not merely removal of no-coverage labels.

Dosage separates improvement from collapse. Partial1 reaches 44.35% partial Joint, 95.72% Code 0, and 76.78% balanced accuracy. Its partial gain over Full1 is +36.97 points [30.27,43.68]. Relative to original, A changes by -0.07 [-2.34,2.34], M by -3.57 [-8.16,0.51], and Code 0 by -1.17 [-3.50,0.97]. Lower-dose augmentation therefore improves the preservation–revision tradeoff, with the measured residual costs reported above. Intervals use 10,000 paired-source draws, fixed seed 42, without family correction or training-seed uncertainty.

A depth-associated boundary within EntailmentBank. All-endpoint partial-diagonal accuracy is 92.47%/91.67% (Qwen/Llama), masking the multihop deficit in Table 20. Depth is assigned from the original revision; the 188 non-M observations/186 sources and 33 M/33 sources can overlap in source membership. These strata measure the association with registered depth at fixed checkpoints.

Table 20: EB partial diagonal Joint with explicit micro denominators. Non-M denotes targets outside the original multihop affected subset. These micro values are separate from the all-endpoint source-macro averages.

Model	Condition	Non-M partial	M partial
Qwen	NoGraph	3/188 (1.60%)	0/33 (0.00%)
Qwen	True $R = 2$	185/188 (98.40%)	8/33 (24.24%)
Llama	NoGraph	3/188 (1.60%)	1/33 (3.03%)
Llama	True $R = 2$	188/188 (100.00%)	2/33 (6.06%)

The failure extends to non-identical inputs. One companion view receives a duplicate existing fact under a new identifier, without changing proof groups or target coverage. We test both directions on the same 133 partial targets/97 sources (Table 21). A four-row batch contains the diagonal, both directions, and a duplicate diagonal; duplicate predictions agree. Each contrast uses its matched batch. The perturbation tests robustness to redundant facts.

Table 21: FLD partial preservation under redundant-fact perturbations. Source-macro Joint (%); identical 133-target/97-source cohort in all columns. Directions 1 and 2 reverse the before/after assignment of the perturbed pair.

Model	Condition	Matched diagonal	Direction 1	Direction 2
Qwen	NoGraph original	0.00	0.00	1.03
Qwen	NoGraph Partial10	86.08	4.64	10.31
Qwen	NoGraph Full10	0.00	0.00	0.00
Qwen	True original	11.34	11.34	12.89
Qwen	True Partial10	93.30	56.19	58.25
Qwen	True Full10	4.64	4.12	4.12
Llama	NoGraph Partial10	90.21	64.43	61.86
Llama	NoGraph Full10	0.00	0.00	0.00

Partial10 improves preservation across both non-identical directions, extending the benefit beyond exact duplication.

Factorized readout: gains beyond exact copying. Endpoint-factorized training retains both evidence views but separates generation histories and KV caches; Independent instead uses a single view. The comparison measures the joint effect of separate histories, grammar, and objective at matched epochs and updates; token and compute budgets vary by interface. Table 22 separates graph-recoverable 121/89 from all-partial 276/174 cohorts. All-partial diagonal Joint averages 45.27%, versus Exact-View Copy 44.64% (difference 0.64 points [−2.14,3.54]). On non-identical inputs, factorization improves 11.45% to 45.23% (+33.78 [28.30,39.29]), where copying is inactive. These intervals resample shared sources jointly across fixed seeds and average seed effects, preserving dependence across seed–source cells. Original A/P/M losses remain explicit. Non-identical scores here use all partial endpoints, unlike the restricted cohort in Table 21.

Table 22: Qwen FLD readout across all three seeds on the same primary test. Source-macro Joint (%). J/F: joint/endpoint-factorized; means and sample SDs describe training runs. Recoverable: 121 claims/89 sources. All-partial: 276 endpoints/174 sources; non-identical averages both perturbation directions. A/P/M: 384/705/135 claims and 249/384/98 sources, with matched quartet replays.

Metric	Joint mean ± SD	Factorized mean ± SD	Seed 42 J/F	Seed 43 J/F	Seed 44 J/F
Graph-recoverable cohort					
Original	92.32±3.19	89.70±2.88	94.94/90.45	93.26/92.13	88.76/86.52
Partial diagonal	15.36±7.14	90.64±1.72	11.24/92.13	23.60/91.01	11.24/88.76
Full diagonal	97.94±0.86	98.31±0.00	98.88/98.31	97.75/98.31	97.19/98.31
Original task and all-partial endpoints					
A	95.74±0.74	93.82±0.48	95.78/93.51	96.45/94.38	94.98/93.57
P	98.81±0.37	98.13±0.16	99.06/97.95	98.39/98.27	98.98/98.16
M	92.18±3.40	89.46±2.41	94.39/90.31	93.88/91.33	88.27/86.73
Partial diagonal	11.97±3.90	45.27±1.78	9.77/46.74	16.48/45.79	9.67/43.30
Non-identical	11.45±3.70	45.23±1.26	9.53/46.07	15.71/45.83	9.10/43.77

G STRUCTURE ROBUSTNESS

G.1 DETERMINISTIC VISIBLE-GRAPH REFERENCE

Table 23: Deterministic replay of view-visible $\mathcal{G}_v$ filtered from registered gold $\mathcal{H}$. Source-macro evaluation percentages. This baseline iterates to a fixed point; neural propagation uses a fixed round budget R.

Dataset	A	P	M
EntailmentBank	8.60	100.00	96.97
FLDx2-formal	55.29	100.00	90.31
SciFact	87.15	100.00	N/A

No training or text encoder. Groups omitted from $\mathcal{G}_v$ remain absent; targets retain the scope of $\mathcal{H}$. This reference measures recovery from the supplied structural information.

This evaluator replays only $\mathcal{G}_v$ and is scored against targets defined over $\mathcal{H}$. Claim coverage starts at none and visible evidence at full. A visible group contributes full if all premises are full, partial if some coverage remains but is incomplete, and none otherwise; target coverage takes the maximum separately by polarity. Claim premises inherit positive coverage, as in the registered contract. Updates continue to a fixed point, independently for each view. The algorithm does not receive gold states, the opposite view's evidence, propagation-depth annotations, or hidden incomplete edges. Target/audit-field replacement, sidecar metadata replacement, and opposite-view replacement leave its predictions unchanged; its replay agrees with the reference coverage implementation given the same visible groups on all evaluated views.

G.2 PROPAGATION AND READOUT UNDER PREDICTED GRAPHS

Learned structure and supervision. A separate Qwen3-8B-Base graph predictor learns registered premise groups and polarities from training sources (three epochs, learning rate 10^{-4}, effective batch size 16, seed 42). The graph-prediction pipeline uses two source-disjoint cross-fit folds; its predictor fitted to all training sources supplies the source-disjoint test graphs. Thus "cross-fitted predicted graphs" describes the pipeline: test sources never enter predictor fitting. Its test prompt contains supplied evidence and claim identifiers/text, without state labels, gold evidence counts, or before/after roles. Parsing enforces identifier validity and removes malformed groups, self-loops, and duplicates; cycles and semantic graph errors remain possible. Gold graphs supervise predictor training, but are not inputs to the predicted-graph inference arm. Evidence views and claim segmentation remain supplied.

Matched fixed-checkpoint protocol. We reuse the main gold-trained Qwen $R = 4$, seed-42, epoch-6 consumer, replacing view-visible gold groups $\mathcal{G}_v$ with predicted groups. Both graph arms use the same 133 partial/full multihop claims from 97 sources, selected from annotations independently of predictions, rather than the full 135-claim/98-source M scope. Each original view's graph follows it under reversal and duplication. Original, reversed, and diagonal inputs use matched quartet batches on the same device. Source freezing retains the checkpoint and target updates. Prefix interventions change only the current claim's first state code while retaining earlier generated tokens and the target evidence. All 133 targets in each arm are traceable; duplicate-prefix shams agree. Conditional vocabulary argmax is distinct from free-running Joint. Numerical replay differences are recorded without outcome filtering.

Table 24: Predicted-graph robustness of the fixed gold-trained Qwen $R = 4$ checkpoint (seed 42). Same 133 claims/97 sources and matched quartet batches throughout. Source-macro percentages over the full partial/full cohort.

Outcome	Gold graph	Predicted graph
Original Joint	94.33	82.99
Source-frozen original Joint	52.06	50.52
Dynamic minus frozen (pp)	42.27	32.47
Partial diagonal: position 1	82.47	75.26
Partial diagonal: position 2	14.95	13.92
Partial diagonal: Joint	11.34	12.37
Full diagonal: Joint	98.97	92.27
Partial target: force partial	14.95	14.95
Partial target: force full	81.96	70.62
Full target: force partial	100.00	99.48
Full target: force full	100.00	93.81

Attenuation without removal of the dissociation. Replacing gold with predicted graphs reduces original Joint from 94.33% to 82.99%, but the dynamic-minus-frozen-source gain remains 32.47 points (versus 42.27 with gold). Partial-diagonal Joint remains low at 12.37%, with first/second accuracy 75.26%/13.92%; the full diagonal reaches 92.27%. For partial targets, the full-minus-partial preceding-token effect is 55.67 points [46.39,64.43], compared with 67.01 [58.25,75.26] under gold graphs. These are pointwise 95% paired-source intervals from 10,000 resamples, conditional on the fixed checkpoint. The full-target effect also changes: forcing full rather than partial reduces full accuracy by 5.67 points $[-10.82, -1.55]$ with predicted graphs, versus zero with gold. The dissociation survives graph errors, with attenuated effects and a measurable full-state cost. This matched graph swap isolates inference-time structure quality within the gold-trained Qwen $R = 4$ checkpoint on the primary cohort.

H Scope and limitations

The experiments establish the propagation/readout dissociation in two base models under the specified interface and five observed single-polarity coverage states. EB contributes 33 multihop sources; SciFact evaluates single-hop revision, with preservation measured on 13 claims/11 sources. Main $\mathcal{G}_v$ graphs are filtered from oracle $\mathcal{H}$; retrieval counts also use oracle structure. The predicted-graph comparison evaluates inference-time graph errors at a fixed gold-trained consumer, retaining supplied views and claims and gold-graph predictor supervision. Primary diagnostics reuse test sources, and follow-up rounds were selected after the initial sweep; extension C provides the pre-specified prospective crossed test. Endpoint-factorized training measures the joint effect of separate histories, grammar, and objective, with original-task costs reported alongside preservation gains.

I COMPLETE RESULT MATRICES AND BEHAVIOR SUMMARY

The complete main matrix retains every registered condition, including unfavorable results and untrained raw-base references; scores do not determine inclusion. Follow-up configuration selection is disclosed in Appendix A. Bold marks the highest unrounded A/P/M among trained conditions in each model/dataset block, including exact ties; diagnostic columns are not ranked. Diagnostic, intervention, and robustness tables use no numerical bolding. Tables 12–13 in Appendix D.1 retain the complete endpoint decoding and forced-prefix controls.

Table 25: Complete Qwen matrix (30 rows), source-macro percentages. Raw Base is untrained; remaining rows use CE. M-NC: multihop NoChange; Invalid: malformed-pair count.

Data	Condition	R	A	P	M	M-NC	Invalid
EB	Raw Base	0	0.00	0.00	0.00	96.97	88
EB	NoGraph	0	92.47	98.43	66.67	30.30	0
EB	All-evidence res.	2	93.28	98.72	63.64	36.36	0
EB	Shuffled	2	93.28	98.01	69.70	27.27	0
EB	Random	2	95.16	99.78	45.45	51.52	0
EB	Similarity	2	94.89	99.85	42.42	54.55	0
EB	True	2	97.58	99.78	72.73	27.27	0
EB	True	3	**99.73**	99.98	**96.97**	3.03	0
EB	True	4	**99.73**	**100.00**	**96.97**	3.03	0
EB	True	5	**99.73**	**100.00**	**96.97**	3.03	0
FLD	Raw Base	0	0.00	35.55	0.00	98.98	35
FLD	NoGraph	0	92.03	94.55	89.80	7.65	0
FLD	All-evidence res.	2	88.35	96.26	78.57	20.41	0
FLD	Shuffled	2	88.49	94.82	78.57	16.84	0
FLD	Random	2	90.23	97.47	70.92	22.45	0
FLD	Similarity	2	90.70	97.76	69.90	25.00	0
FLD	True	2	92.84	98.73	83.16	12.24	0
FLD	True	3	94.51	98.25	90.31	6.12	0
FLD	True	4	**96.45**	**98.93**	**95.41**	3.06	0
FLD	True	5	96.32	98.42	94.90	1.02	0
SciFact	Raw Base	0	0.00	45.45	N/A	N/A	0
SciFact	NoGraph	0	42.74	18.18	N/A	N/A	0
SciFact	All-evidence res.	2	37.71	18.18	N/A	N/A	0
SciFact	Shuffled	2	39.11	27.27	N/A	N/A	0
SciFact	Random	2	68.44	90.91	N/A	N/A	0
SciFact	Similarity	2	69.55	**100.00**	N/A	N/A	0
SciFact	True	2	**71.23**	**100.00**	N/A	N/A	0
SciFact	True	3	68.44	**100.00**	N/A	N/A	0
SciFact	True	4	70.39	**100.00**	N/A	N/A	0
SciFact	True	5	70.67	90.91	N/A	N/A	0

Table 26: Complete Llama matrix (30 rows), source-macro percentages. Raw Base is untrained; remaining rows use CE. M-NC: multihop NoChange; Invalid: malformed-pair count.

Data	Condition	R	A	P	M	M-NC	Invalid
EB	Raw Base	0	0.00	0.00	0.00	0.00	374
EB	NoGraph	0	93.82	98.55	78.79	21.21	0
EB	All-evidence res.	2	91.40	98.27	63.64	36.36	0
EB	Shuffled	2	93.82	98.06	78.79	21.21	0
EB	Random	2	94.62	99.88	39.39	60.61	0
EB	Similarity	2	94.62	99.85	39.39	60.61	0
EB	True	2	98.66	99.78	84.85	15.15	0
EB	True	3	**99.73**	99.94	**96.97**	3.03	0
EB	True	4	**99.73**	**99.97**	**96.97**	3.03	0
EB	True	5	**99.73**	**99.97**	**96.97**	3.03	0
FLD	Raw Base	0	0.00	0.00	0.00	0.00	629
FLD	NoGraph	0	90.03	95.18	82.65	17.35	0
FLD	All-evidence res.	2	90.56	94.40	86.22	11.73	0
FLD	Shuffled	2	86.41	96.22	76.53	22.45	0
FLD	Random	2	88.76	98.24	65.31	26.02	0
FLD	Similarity	2	87.15	97.72	62.24	29.08	0
FLD	True	2	91.90	98.62	71.94	25.00	0
FLD	True	3	93.44	98.31	84.69	13.78	0
FLD	True	4	94.51	98.66	91.84	5.61	0
FLD	True	5	**96.39**	**99.20**	**94.90**	3.57	0
SciFact	Raw Base	0	0.00	0.00	N/A	N/A	179
SciFact	NoGraph	0	38.27	31.82	N/A	N/A	0
SciFact	All-evidence res.	2	34.92	13.64	N/A	N/A	0
SciFact	Shuffled	2	38.27	18.18	N/A	N/A	0
SciFact	Random	2	65.92	81.82	N/A	N/A	0
SciFact	Similarity	2	65.08	81.82	N/A	N/A	0
SciFact	True	2	65.92	81.82	N/A	N/A	0
SciFact	True	3	**66.48**	81.82	N/A	N/A	0
SciFact	True	4	65.27	**90.91**	N/A	N/A	0
SciFact	True	5	65.64	72.73	N/A	N/A	0

I.1 Behavior summary

All CE predictions in the main paired-task matrix parse successfully. A-DeltaMatch (equal to A-Joint here), A-NoChange, and A-WrongChange sum to 100%, as do P-StayRate and P-SpuriousChange. StayRate includes unchanged wrong states; malformed raw-base outputs do not count as preservation.

Table 27: True ($R = 2$) minus NoGraph ($R = 0$), source-macro behavior changes (pp). DM: DeltaMatch; NC: NoChange; WC: WrongChange; Stay: StayRate; Spur: SpuriousChange. Arrows indicate preferred directions.

Dataset	A-DM ↑	A-NC ↓	A-WC ↓	P-Stay ↑	P-Spur ↓
Qwen3-8B-Base					
EntailmentBank	+5.11	-4.84	-0.27	+1.35	-1.35
FLDx2-formal	+0.80	-2.07	+1.27	+2.55	-2.55
SciFact	+28.49	-0.56	-27.93	+81.82	-81.82
Llama-3.1-8B					
EntailmentBank	+4.84	-4.84	0.00	+1.23	-1.23
FLDx2-formal	+1.87	-2.14	+0.27	+2.00	-2.00
SciFact	+27.65	-3.35	-24.30	+59.09	-59.09