ONE STORY, TWO ORDERS: TIME OVERRIDES TEXT IN LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

A story has two orders: the chronological order of its events and the text order in which they are mentioned. Chronological order can bias human memory for the order of presentation, but whether and how LLMs exhibit similar interference remains unclear. We independently manipulate text order and chronological order in controlled narratives across ten different LLMs. When the orders conflict, text-order accuracy drops by 63.3 percentage points on average, whereas chronological-order judgments are substantially more robust. Yet both ranks remain linearly decodable from pre-query hidden states, with nearly perfect text-order rank decoding at the final layer. Causal interventions trace part of this interference to content encoding chronological order in an unchanged carrier region before the question, with larger effects from values than keys. Inspired by this mechanistic understanding, we propose ReFrame (Reference-Frame Routing Adaptation), adapting only attention query and value projections through low-rank updates. The method raises mean text-order conflict accuracy to 99.9% on held-out original questions, improves chronological-order judgments, and extends these gains to additional question wordings. These findings expose a gap between what LLMs retain and what governs their answers. They identify selective use of preserved information as a distinct target for evaluating and improving reliability when chronological order and text order conflict.

1 INTRODUCTION

> *"Many years later, as he faced the firing squad, Colonel Aureliano Buendía was to remember that distant afternoon when his father took him to discover ice."*
>
> *One Hundred Years of Solitude* (García Márquez, 1970).

The opening of *One Hundred Years of Solitude* presents the firing squad before the childhood encounter with ice. The childhood event comes first in the character's life; the firing squad comes first on the page. Understanding the sentence requires keeping these two orders apart. "What happened first?" and "What was mentioned first?" are different questions, even when they concern the same events. We call these the *chronological order* and the *text order*, respectively (Figure 1). A language model should be able to reconstruct the first while remaining faithful to the second.

This distinction extends beyond literary flashbacks. A user may describe yesterday's meeting before last week's phone call; an incident report may reveal an earlier event only after describing its consequences. Reconstructing what happened is useful, but answering what was reported first requires preserving the history of the account itself. Human narrative memory illustrates the tension. Explicit chronological information can bias judgments about the order in which events were presented (Xu & Kwok, 2019), and causal and chronological relations can organize recall even when people are asked to follow the order of presentation (Antony et al., 2024). These findings motivate a question about language models: does constructing a timeline interfere with reporting the order in which the model encountered its events?

We identify a central gap in LLM narrative understanding: *selecting between text order and chronological order when they conflict*. Conflict occurs when one event is mentioned before another but happens after it, so the same pair requires opposite answers depending on the question. Existing studies establish sensitivity to input order and limitations in temporal judgments (Chen et al., 2024;

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.

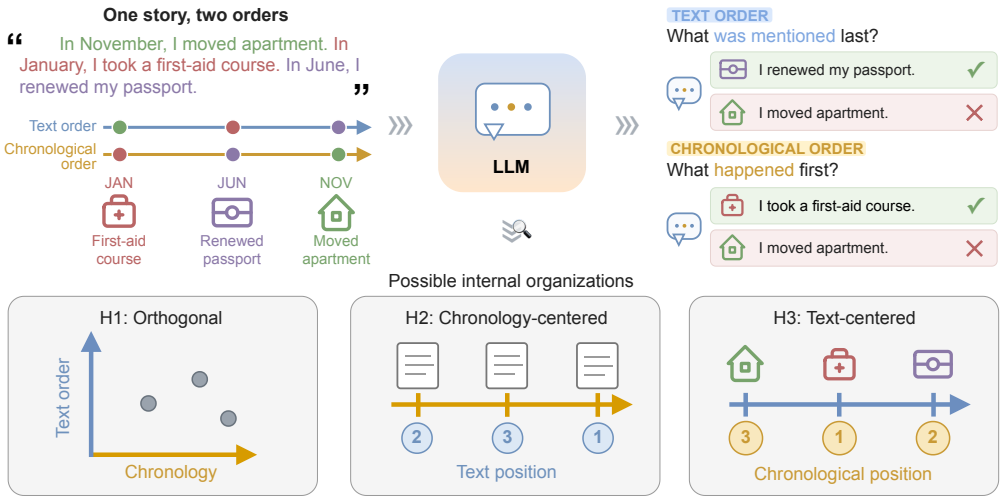

Figure 1: **One story, two orders.** Upper panels illustrate text-order and chronological-order questions about the same narrative. Lower panels contrast three possible organizations: separate order coordinates (H1), chronological-order-centered representations retaining text positions (H2), and text-order-centered representations retaining chronological positions (H3).

Fatemi et al., 2025); they leave open how a model resolves these two reporting targets within the same passage. To our knowledge, we provide the first controlled mechanistic study of this conflict, showing how chronological order can override text-order reporting even when text-order information remains available.

This failure is surprising because positional encodings explicitly supply Transformer LLMs with information about token positions (Vaswani et al., 2017). Text order therefore has a direct structural basis in the input sequence. At the same time, understanding the described world requires organizing events by when they happened, independently of where they appear in the passage. A model must support both the sequence in which information arrives and the timeline conveyed by its content. When these disagree, the reconstructed timeline can compete with the text order that the question asks the model to report. Positional information may remain available without controlling the answer. Figure 1 summarizes three hypotheses for how the orders coexist: separate, orthogonal order coordinates (H1); a representation organized primarily by chronological order that retains text positions (H2); or one organized primarily by text order that retains chronological positions (H3). These possibilities motivate testing both whether the two orders remain accessible and which one causally controls the answer.

We test this competition using stories in which sentence positions and explicit dates vary independently across ten different LLMs. Conflict reduces text-order accuracy by 63.3 percentage points on average, with a larger cost than for chronological-order judgments in every model. Yet linear probes recover both orders before the question, and final-layer text-order rank is nearly perfectly decodable. The model therefore retains text order even when chronological order controls its answer.

To explain this failure, we trace part of the interference to contextual states in an unchanged *carrier region* before the question. Matched state transfers and access interventions show how chronological-order content formed during passage processing influences text-order answers, with larger effects from values than keys. Inspired by these mechanistic findings, we propose ReFrame (Reference-Frame Routing Adaptation), which adapts the attention query and value projections through low-rank updates (Hu et al., 2022). Across ten models, ReFrame raises mean text-order conflict accuracy to 99.9% on held-out original questions, improves overall chronological-order accuracy, and transfers gains to additional question wordings. Behavioral readouts become more selective for the requested text order, linking the intervention to the conflict identified above.

Our contributions are threefold:

1. **Discovery of a reporting failure.** We explicitly separate the two orders as competing reporting targets and establish that chronological order systematically disrupts text-order answers despite preserved text-order information.
2. **Causal explanation.** Carrier transplantation and access interventions identify a pre-query pathway through which chronological-order content affects text-order answers, with a stronger contribution from values than keys.
3. **Mechanism-inspired solution.** We propose ReFrame based on these mechanistic insights and show improved accuracy and text-order readout selectivity across ten models.

2 Related work

Order sensitivity in language models. Prior work examines how the arrangement of a prompt affects downstream performance. Changing the order of demonstrations (Zhao et al., 2021; Lu et al., 2022), the location of evidence (Liu et al., 2024), or the sequence of premises (Chen et al., 2024) can substantially alter model accuracy. Mechanistic studies connect these effects to causal attention and positional encodings (Wang et al., 2025; Ok & Lee, 2026), while the Curse of Verbalization relates performance to alignment between presentation order and the order in which information must be used (Zhou et al., 2026). These studies treat input order as a factor affecting another task. Our question concerns *order itself as the information to be represented and reported*: how LLMs encode text order, and whether that representation controls the answer when chronological order provides a competing organization of the same events.

Temporal reasoning and order memory. Order-memory research directly examines how LLMs represent and retrieve text order. Pink et al. (2026) identify a one-dimensional temporal code and an attention head that reinstates it when models report where content appeared in a long input. Complementary work evaluates judgments about event chronology and temporal constraints (Qiu et al., 2024; Fatemi et al., 2025), including the internal mechanisms supporting temporal ordering (Yang et al., 2026). We investigate how these abilities interact when text order and chronological order conflict. Our finding is that text-order information remains highly decodable, yet chronological order strongly interferes with its behavioral report. The contribution is therefore to identify and causally trace interference in the use of preserved order information, beyond establishing the ability to represent either order in isolation.

Human representations of narrative time. Human narrative research provides a motivation for studying this competition. Readers organize events chronologically even when a story presents them through flashbacks (Claus & Kelter, 2006), and neural responses during nonlinear films track chronological relationships beyond proximity in the presented sequence (Grall et al., 2023). Causal and chronological relationships can also organize recall more strongly than presentation order, even when participants are asked to recall events in the order presented (Antony et al., 2024). When chronological positions are made explicit, they can bias memory for presentation order (Xu & Kwok, 2019). These findings motivate testing whether chronological order can influence text-order reports even when information about text order remains accessible.

3 Experiment setup

We use short, timestamped stories to separate interference between orders from differences in event content. Explicit dates specify chronological order without relying on familiar event scripts, while sentence positions specify text order. Varying these factors independently yields matched narratives in which the question determines which order should control the answer.

Task. Each passage contains four distinct events with explicit timestamps. Consider the following conflict example:

> On June 19, 2021, Mia attended a first-aid course.
> On January 8, 2021, Mia renewed a passport.
> On March 4, 2022, Mia bought a camera.
> On November 27, 2023, Mia moved into a new apartment.

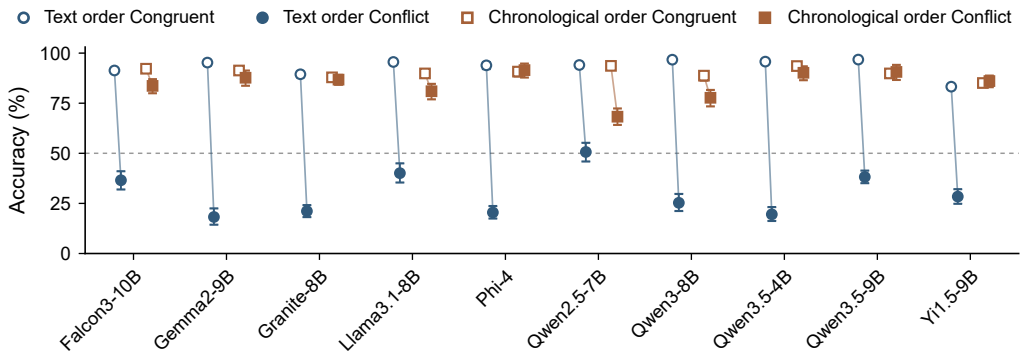

Figure 2: **Conflict penalizes text-order queries more than chronological-order queries.** Congruent and conflict accuracy for both query types across ten models.

The course is mentioned first, but the passport renewal happened first. For each event, sentence position $p(e)$ and chronological rank $\tau(e)$ lie in $\{1, 2, 3, 4\}$. A *text-order query* (P) asks which of two events is mentioned first; a *chronological-order query* (τ) asks which happened earlier. A pair is *congruent* when the orders agree and in *conflict* otherwise. Exchanging the queried timestamps changes chronological order while preserving text order; exchanging the complete sentences changes text order while preserving chronological order. Crossing these operations varies the two answers independently. We counterbalance option assignments. Appendix A gives the exact system instruction, questions, and option assignments. The passage contains no explicit text-position labels.

Models. We select ten different LLMs from seven families—Falcon, Gemma, Granite, Llama, Phi, Qwen, and Yi—to broaden coverage of model architectures and learned representations (Table 2). The panel includes full-attention models and Qwen3.5's hybrid full/linear-attention architecture. Models retain their native tokenizer, chat template, and output transformations. Pretrained weights remain frozen; Section 4.4 trains added low-rank parameters.

Evaluation. Behavioral evaluation uses 64 four-event worlds, all six event pairs, and balanced sentence and option orders. We report next-token A/B accuracy and the event-aligned margin

$$u_q(a, b) = \tfrac{1}{2}\big(m_{AB,q} - m_{BA,q}\big), \tag{1}$$

where $m_{AB,q}$ and $m_{BA,q}$ are A-minus-B logit margins under reversed event-to-option assignments. Pairing cancels a shared additive label preference. Unless noted otherwise, pointwise 95% intervals use 10,000 paired world-bootstrap draws after averaging effects within worlds. Appendix A details cohort differences, question wordings, and scoring.

4 Results

We first measure the behavioral asymmetry between the two orders, then use rank decoding to test whether text-order information remains available. Causal interventions examine how chronological-order information reaches text-order answers. Inspired by these mechanistic findings, we propose and evaluate Reference-Frame Routing Adaptation (ReFrame), implemented with low-rank query and value updates.

4.1 Behavior

When both orders agree, a correct answer cannot reveal which one the model used. Conflict makes that distinction observable: a response consistent with chronological order becomes an error for a text-order query. We therefore compare each task with its own congruent baseline to separate interference from baseline task difficulty. Conflict reduces text-order query accuracy by 43.4–77.0 percentage points, compared with -1.0–25.4 points for chronological-order queries (Figure 2). The text-order loss exceeds the chronological-order loss in all ten models. The consistent direction of this

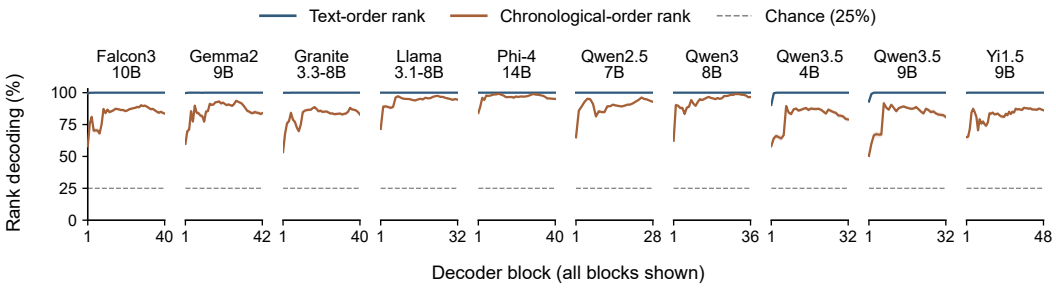

Figure 3: **Both orders remain linearly decodable before the question.** Text-order and chronological-order rank accuracy at every decoder block across ten models, using complete sentence-end states.

asymmetry identifies a preference between orders, beyond a general difficulty with conflicting inputs: chronological order disrupts text-order reporting more than text order disrupts chronological-order judgments. The text-order conflict cost occurs under all three original question wordings in every model (Appendix A.3, Table 4).

4.2 Rank decoding

To test whether this asymmetry reflects a loss of text-order information, we fit linear probes to pre-query sentence-end states at every decoder block across all ten models. Complete event descriptions are disjoint across the probe training, validation, and test splits. Probing before the question separates information retained from the passage from its subsequent use in answering. Text-order rank is readily decodable throughout the network, with approximately 100% accuracy at the final block (Figure 3). Chronological-order rank is also decodable. Without refitting, the final-block probes recover text-order rank with 93.3–100% accuracy on the original 64-world behavioral cohort, which is excluded from probe training and validation. This transfer connects information availability to the same materials on which the models fail to report text order reliably. The probe protocol is provided in Appendix B.

Poor reporting thus coexists with accessible text-order information. Linear decodability does not establish that the model uses this information appropriately (Elazar et al., 2021). The next question is how this accessible contextual information is organized and accessed during answering. We therefore use causal interventions to test how chronological order gains influence over text-order reports.

4.3 Causal interventions

Carrier transplantation. We hypothesize that constructing a timeline of the described events shapes the contextual states available during answering, allowing chronological order to compete with text order at readout. To test where this information becomes causally available, we focus on a short region after the final timestamp and before the question. These late positions can incorporate the preceding narrative while their own tokens remain unchanged across chronology manipulations. We call this region the *carrier*; it comprises the final sentence's post-timestamp text and separator.

We construct congruent and conflict runs differing only in the queried timestamps; text order, carrier tokens and positions, and the correct text-order answer remain fixed. Before presenting the question, we transplant the donor's carrier keys and values into the recipient at every full-attention layer, then continue with the recipient's question and options (Zhang & Nanda, 2024). A conflict donor tests *corruption* of a congruent recipient; a congruent donor tests *rescue* of a conflict recipient. No donor question or answer enters the recipient. Qwen3.5 recurrent states remain unchanged in this experiment.

Both directions affect all ten models (Figure 4). Corruption transfers 22.8–34.5% of the congruent–conflict margin difference, and rescue restores 20.1–34.2%. Rescue increases paired conflict accuracy by 19.3–33.7 points and individual-trial accuracy by 15.6–27.7 points. Effects exceed nonqueried-date sham transfers and persist when question-to-timestamp access is blocked; reconstruction and specificity controls appear in Appendix C.

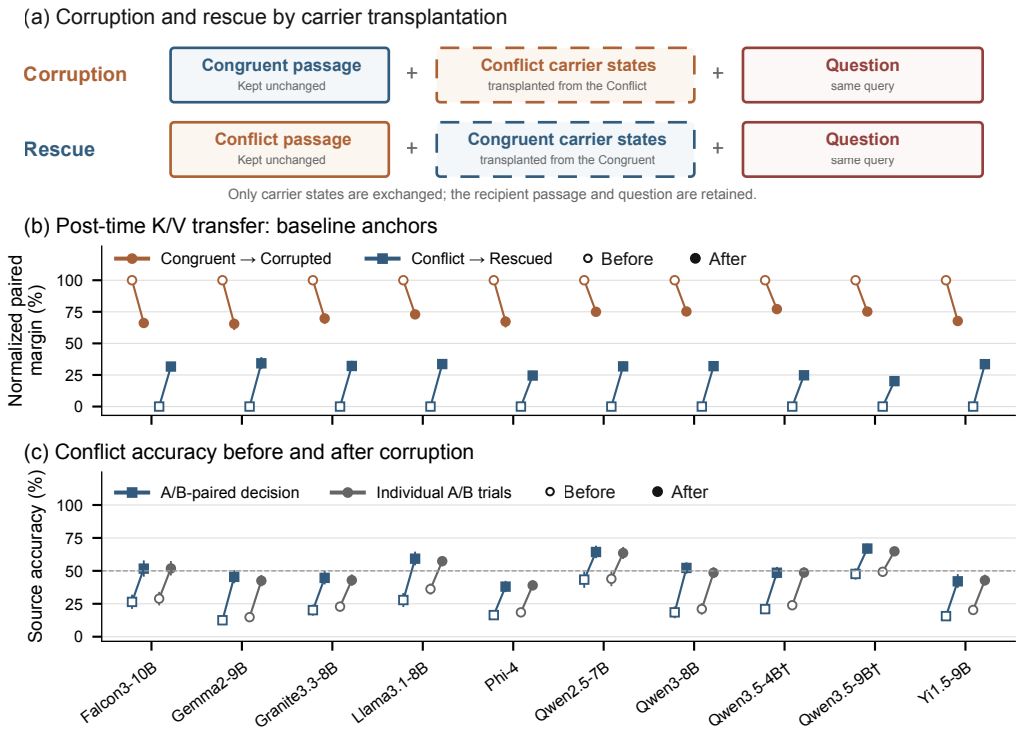

Figure 4: **Changing carrier states changes text-order judgments. (a)** Matched corruption and rescue transfers. **(b)** Paired margins before and after transfer, normalized to conflict (0) and congruent (100) baselines. **(c)** Conflict accuracy before and after rescue, for paired decisions and individual A/B trials. †: full-attention K/V only in Qwen3.5.

Identical carrier tokens acquire different contextual states under different chronologies, and exchanging those states changes how the model reports the unchanged text order. Chronological-order information has therefore been incorporated into carrier states before the question. Partial rescue supplies a timeline consistent with the text-order answer, but does not establish selective use of text order.

Together with rank decoding, these results suggest that both orders remain represented in contextual token states, where they can exert competing influences on answers. The carrier provides a causally effective route through which this information reaches the output. We next examine how the information is formed and whether its effect is carried primarily by the carrier's keys or values.

Access and value states. To distinguish direct timestamp access from chronological information already incorporated in the context, we block timestamp access during question processing, context processing, or both, using matched neutral-key controls (Figure 5). Among the eight nonrecurrent models, query-only blocking generally fails to suppress chronological-order influence. Context blocking reduces it in all eight and accounts for much of the reduction under joint blocking. It prevents non-timestamp tokens after the first timestamp from reading timestamp keys, covering more than the carrier alone. This pattern indicates that chronological information is incorporated into subsequent contextual states before the question, limiting the effect of restricting direct access at query time.

Extending the interventions to Qwen3.5 recurrent blocks reproduces the same pattern. Joint blocking nearly removes chronological-order gain for text-order queries but lowers chronological-order accuracy to 50% in both models. Removing the route can therefore also remove useful information. Implementation details, effect sizes, and controls are in Appendix C.4. The extension is necessary because full-attention masks alone leave recurrent paths open.

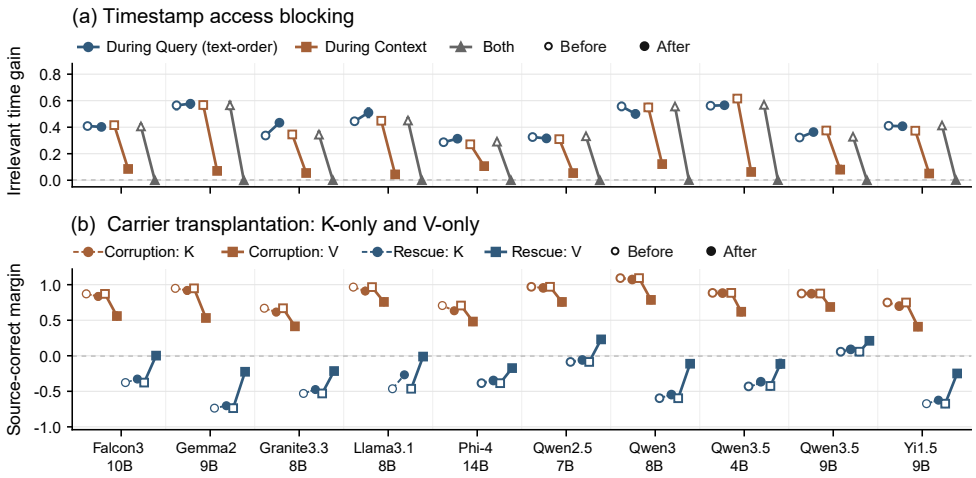

Figure 5: **Timestamp access and carrier values contribute to interference. (a)** Chronological-order gain for text-order queries under query, context, or joint timestamp-access blocking (reference SD per rank); open markers show matched neutral-key controls. **(b)** Text-order-correct margins before and after K-only or V-only corruption and rescue (baseline SD units). Qwen3.5 includes recurrent-path interventions (Appendix C.4).

We next separate carrier keys, which affect attention matching, from values, which supply retrieved content (Vaswani et al., 2017). This comparison tests whether the transferred effect depends more on attention matching or on the content made available through that attention. V-only transfer shifts text-order margins in both directions in every model and has substantially larger effects than K-only transfer, implicating the order-dependent content carried by values.

Together, these experiments show how chronological information incorporated into context states during passage processing can compete with text order at readout.

4.4 ReFrame improves selective use of order information

The preceding experiments reveal a gap between retaining text-order information and selecting it when chronological information competes at readout. Inspired by these mechanistic findings, we propose ReFrame (Reference-Frame Routing Adaptation) to improve selection of the order requested by the question while preserving useful chronological information. ReFrame applies low-rank updates to attention query and value projections (Hu et al., 2022), keeping pretrained weights frozen. The stronger carrier value effects motivate adapting the contextual content supplied by V, while Q updates can adjust question-dependent access to contextual states. Both operate during passage and question processing.

Training combines answer supervision with consistency under changes to the irrelevant order and limited capability-retention training: the consistency term penalizes irrelevant-order effects, while supervision requires sensitivity to the requested order. We apply the same recipe to all ten models, with three seeds per model. Appendix D.2 gives the training data, projection slices, parameter counts, objective, and validation-based selection rule.

Accuracy and transfer. Success on the original questions could reflect adaptation to a particular wording. We therefore change the question while holding the underlying order conflict fixed. Evaluation crosses a separate set of 64 held-out scenes with the original questions and four new question pairs that preserve the original task meaning. We hold passages, dates, and answer options fixed across wordings and average all four families equally.

Table 1 summarizes text-order conflict and overall chronological-order accuracy; Table 15 in Appendix D.2 reports the complete congruence breakdown. On original questions, ReFrame text-order conflict accuracy ranges from 99.5% to 100.0%, compared with 31.2–66.0% before training. Averag-

Table 1: **ReFrame on the ten-model panel.** Accuracy (%) on the same 64 held-out scenes. ReFrame averages three separately evaluated training seeds. Additional questions average four fixed question pairs absent from training, with events, dates, and options held constant. Text-order conflict uses incongruent cells; chronological order includes both congruence conditions. The full breakdown is in Table 15.

	Original questions				Additional questions			
	Text order Conflict		Chronological order All		Text order Conflict		Chronological order All	
Model	Native	ReFrame	Native	ReFrame	Native	ReFrame	Native	ReFrame
Falcon3-10B	52.7	99.5	82.2	94.7	53.9	97.9	79.0	94.2
Gemma-2-9B	39.5	100.0	89.5	94.6	42.4	100.0	87.3	94.8
Granite-3.3-8B	43.8	99.9	82.4	94.9	45.7	83.4	81.1	94.4
Llama-3.1-8B	46.5	100.0	72.1	94.4	50.5	98.8	68.9	94.3
Phi-4	31.2	100.0	95.1	98.5	28.4	99.5	94.5	97.7
Qwen2.5-7B	62.1	100.0	75.4	90.7	61.7	97.7	73.8	90.3
Qwen3-8B	34.4	100.0	86.7	95.6	30.9	98.4	85.1	95.1
Qwen3.5-4B	52.7	100.0	93.2	97.7	31.8	99.8	93.8	97.6
Qwen3.5-9B	66.0	100.0	90.6	97.9	56.7	100.0	90.0	97.6
Yi-1.5-9B	35.9	99.5	85.0	94.2	38.4	94.3	76.7	93.8

ing equally across the four additional question pairs, ReFrame text-order conflict accuracy ranges from 83.4% to 100.0%, improving over the native model in all ten models. Mean chronological-order accuracy across models changes from 85.2% to 95.3% on original questions and from 83.0% to 95.0% on additional questions. The paired design isolates wording changes from changes in passages or options. Variation across training seeds is reported in Appendix D.2.

Readout selectivity. H1 in Figure 1 provides a target for adaptation: the two orders should be separately usable when the question requires them. We therefore test whether low-rank updates disentangle their influences on answers. On the same original-question trials, independent sentence and date swaps give signed effects on the answer margin, $\beta_{q,p}$ and $\beta_{q,\tau}$. We express both behavioral readouts in shared coordinates:

$$\mathbf{v}_{\mathrm{P}} = (G_{\mathrm{P},p}, G_{\mathrm{P},\tau}), \qquad \mathbf{v}_{\tau} = (G_{\tau,p}, G_{\tau,\tau}), \qquad G_{q,r} = \frac{\overline{\beta}_{q,r}}{\sigma_{\text{native}}}. \tag{2}$$

The native-margin scale is fixed across axes, queries, and training states. Selective text-order and chronological-order readouts align with the positive horizontal and vertical axes, respectively. Moving closer to the target axis indicates less influence from the irrelevant order relative to the relevant one. We assess this alignment together with retained relevant influence, distinguishing improved relative selectivity from a reduction in absolute interference.

Figure 6 shows the shift from entangled influences toward separate readouts. The median angle between the two readouts increases from 30.7° to 87.7°, approaching 90° in all ten models. Median text-order axis deviation falls from 52.2° to 0.6°, improving in all ten; chronological-order deviation changes from 5.4° to 2.8°, improving in seven. Both tasks retain positive relevant influence in every model and seed. Mean absolute interference decreases in all ten text-order readouts; angular alignment can also improve through stronger relevant responses. Appendix D.2 reports estimation details, paired angle intervals, text-order influence magnitudes, and seed variation.

Figure 6 thus demonstrates the feasibility of more independent use of text order and chronological order at behavioral readout. ReFrame makes text-order responses more selective while retaining useful sensitivity to chronological order, approaching the functional goal of H1. This shows that an LLM can learn to use the same passage according to different ordering criteria, with the question guiding which order governs its answer.

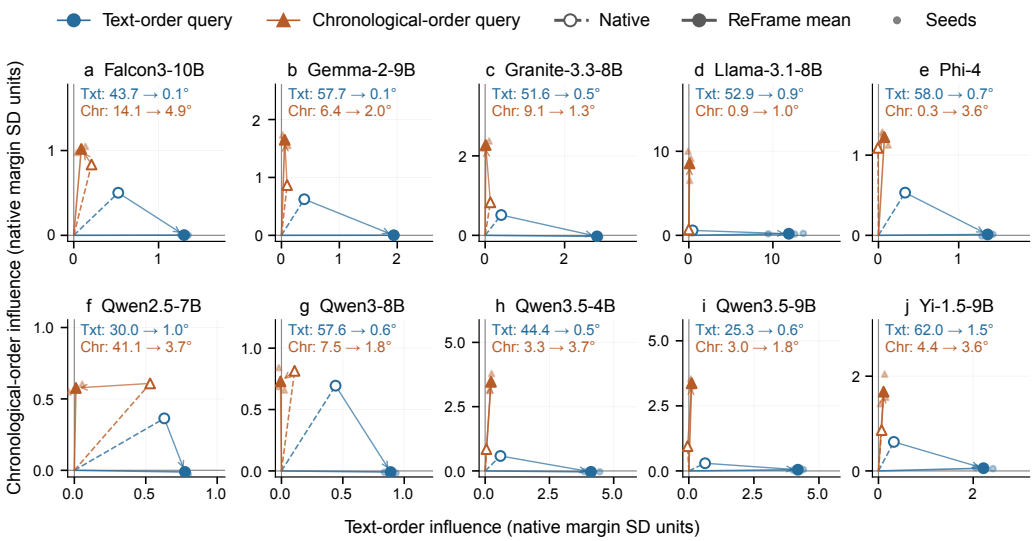

Figure 6: **ReFrame separates behavioral readouts toward their target axes.** Native and adapted influences on 64 held-out scenes with original questions. Horizontal and vertical axes correspond to text order and chronological order. Endpoints average three seeds; small markers show individual seeds. Annotations report target-axis deviations (Txt: text order; Chr: chronological order).

5 Conclusion

Understanding a narrative requires distinguishing the order of its telling from the order of its events. Across ten LLMs, chronological order strongly interferes with text-order reporting, exposing a gap between preserving an order and selecting it to answer a question.

Our analyses connect information availability, causal influence, and trainable selection. Rank decoding shows that both orders remain accessible before the question. Carrier transplantation and access interventions identify a contextual pathway through which chronological-order information contributes to text-order errors, with stronger effects from carrier values than keys.

Inspired by these mechanistic findings, we propose ReFrame (Reference-Frame Routing Adaptation), implemented through low-rank updates to attention query and value projections. This mechanism-inspired solution improves accuracy and question-wording transfer across the tested models. Its accuracy and influence changes support *functional disentanglement at readout*: the two orders become more independently usable while retaining useful sensitivity to both. ReFrame thus demonstrates that selective control over preserved information is a trainable component of reliable narrative understanding.

Limitations. The causal interventions motivate ReFrame, but its behavioral results do not directly establish changes to the same carrier pathway or to the geometry of hidden representations. We have not go beyond short, constructed English passages and constrained responses, the order conflict can be more complicated in longer narratives, extended contexts. ReFrame is evaluated for this specific order-conflict task, however broader transfer and effects on other tasks have not been systematically evaluated. The possible integration of this approach into pretraining or broader post-training remain unexplored.

AI USE STATEMENT

In this work, we used generative AI tools for **language polishing and editing of the manuscript** and **assistance with drafting and revising manuscript sections**. We have not used generative AI tools for **research ideation, experimental design, data collection, or result analysis**, and **synthetic data generation and mathematical and theoretical development (including formulating claims, proof assistance, and hypothesis refinement)** are not applicable to this work. Additionally, we used generative AI tools for **assistance with LaTeX formatting and reference organization**. We have reviewed all AI-assisted work. Specifically, all AI-assisted text was manually verified for factual accuracy against original sources, and any citations suggested by AI tools were independently confirmed through manual literature search. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

James Antony, Angelo Lozano, Pahul Dhoat, Janice Chen, and Kelly Bennion. Causal and chronological relationships predict memory organization for nonlinear narratives. *Journal of Cognitive Neuroscience*, 36:2368–2385, 2024. doi: 10.1162/jocn_a_02216.

Xinyun Chen, Ryan Andrew Chi, Xuezhi Wang, and Denny Zhou. Premise order matters in reasoning with large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 6596–6620. PMLR, 2024. URL https://proceedings.mlr.press/v235/chen24i.html.

Berry Claus and Stephanie Kelter. Comprehending narratives containing flashbacks: Evidence for temporally organized representations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 32(5):1031–1044, 2006. doi: 10.1037/0278-7393.32.5.1031.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021. doi: 10.1162/tacl_a_00359. URL https://aclanthology.org/2021.tacl-1.10/.

Bahare Fatemi, Seyed Mehran Kazemi, Anton Tsitsulin, Karishma Malkan, Jinyeong Yim, John Palowitch, Sungyong Seo, Jonathan Halcrow, and Bryan Perozzi. Test of time: A benchmark for evaluating llms on temporal reasoning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/eb7295a8bc613b375726659c2ecd6f14-Abstract-Conference.html.

Gabriel García Márquez. *One Hundred Years of Solitude*. Harper & Row, New York, 1970.

Clare Grall, Josefa Equita, and Emily S. Finn. Neural unscrambling of temporal information during a nonlinear narrative. *Cerebral Cortex*, 33:7001–7014, 2023. doi: 10.1093/cercor/bhad015.

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 *The Tenth International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2106.09685.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638. URL https://aclanthology.org/2024.tacl-1.9/.

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*, pp. 8086–8098, 2022. doi: 10.18653/v1/2022.acl-long.556. URL https://aclanthology.org/2022.acl-long.556/.

Hyunjong Ok and Jaeho Lee. Lost in the prompt order: Revealing the limitations of causal attention in language models. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 38566–38587, 2026. doi: 10.18653/v1/2026.findings-acl.1921. URL https://aclanthology.org/2026.findings-acl.1921/.

Mathis Pink, Vy Ai Vo, Qinyuan Wu, Jianing Mu, Javier Turek, Uri Hasson, Kenneth A. Norman, Sebastian Michelmann, Alexander Huth, and Mariya Toneva. Temporal context reinstatement drives episodic-like order memory in long-context language models. *arXiv preprint arXiv:2607.22575*, 2026. URL https://arxiv.org/abs/2607.22575.

Yifu Qiu, Zheng Zhao, Yftah Ziser, Anna Korhonen, Edoardo Ponti, and Shay B. Cohen. Are large language model temporally grounded? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 7064–7083, 2024. doi: 10.18653/v1/2024.naacl-long.391. URL https://aclanthology.org/2024.naacl-long.391/.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://arxiv.org/abs/1706.03762.

Ziqi Wang, Hanlin Zhang, Xiner Li, Kuan-Hao Huang, Chi Han, Shuiwang Ji, Sham M. Kakade, Hao Peng, and Heng Ji. Eliminating position bias of language models: A mechanistic approach. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=fvkElsJOsN.

Xinming Xu and Sze Chai Kwok. Temporal-order iconicity bias in narrative event understanding and memory. *Memory*, 27(8):1079–1090, 2019. doi: 10.1080/09658211.2019.1622734. URL https://doi.org/10.1080/09658211.2019.1622734.

Zhen Yang, Xinyue Zhang, Ping Jian, Chengzhi Li, Zhongbin Guo, Jiaping Feng, and Wenpeng Lu. Temporal token matters: Investigating and interpreting the consistency of temporal ordering in large language models. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 2577–2596, 2026. doi: 10.18653/v1/2026.findings-acl.123. URL https://aclanthology.org/2026.findings-acl.123/.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Hf17y6u9BC.

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

Yue Zhou, Henry Peng Zou, Barbara Di Eugenio, and Yang Zhang. The curse of verbalization: How presentation order constrains llm reasoning. In *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 4175–4185, 2026. URL https://aclanthology.org/2026.findings-eacl.218/.

A MATERIALS, MODELS, AND BEHAVIORAL EVALUATION

A.1 EXPERIMENTAL COHORTS

The paper combines complementary measurements, not a single pooled dataset. The behavioral panel in Figure 2 uses 64 four-event worlds and the original multi-wording design. Figures 4 and 5 reuse those worlds with a fully crossed manipulation of the queried pair. Figure 6 uses a separate 64-scene held-out cohort for the matched native–ReFrame comparison under original wording. Table 1 additionally crosses this same cohort with four additional question pairs. These questions are fixed before evaluation and absent from training; each scene retains identical events, dates, and options across wordings. ReFrame training uses 256 other scenes, with separate validation data (Appendix D.2). Figure 3 uses 256 generated worlds for training, 64 for validation, and 256 for matched testing. The 64 original worlds are also evaluated for text-order decoding without entering probe training or validation. The Qwen3.5 access extension uses worlds 0–15 and fixed pair 0, whereas its hybrid carrier-transfer extension retains all 64 worlds and six pairs (Appendix C.4).

Table 2: Models in the displayed family order. All layers are probed. The listed intervention scope describes the original experiments. Figure 5 additionally includes Qwen3.5 linear-block interventions (Appendix C.4).

Model	Blocks	Hidden width	Original scope
Falcon3-10B-Instruct	40	3072	All attention layers
Gemma-2-9B-IT	42	3584	All attention layers
Granite-3.3-8B-Instruct	40	4096	All attention layers
Llama-3.1-8B-Instruct	32	4096	All attention layers
Phi-4	40	5120	All attention layers
Qwen2.5-7B-Instruct	28	3584	All attention layers
Qwen3-8B	36	4096	All attention layers
Qwen3.5-4B	32	2560	Full-attention only
Qwen3.5-9B	32	4096	Full-attention only
Yi-1.5-9B-Chat	48	4096	All attention layers

All pretrained model parameters are frozen; the ReFrame study trains added low-rank parameters. We use the model's native tokenizer, chat template, and output-logit transformations. The model registry accompanying the paper records configuration hashes and intervention scope. The two Qwen3.5 models contain recurrent as well as full-attention blocks. Probes include every decoder block. Original token-indexed cache edits and attention masks affect only full-attention blocks; the Qwen3.5 extensions additionally intervene in linear blocks as specified in Appendix C.4. Native BF16 hidden values are retained losslessly for probing; the separate numerical protocol for interventions is described in Appendix C.5.

A.2 EXACT ENGLISH MATERIAL

The following is one unmodified world used by the timestamp experiments. Event sentences contain neither an explicit text position label nor a separate block stating relations between event IDs.

> **System instruction:** You are answering a controlled reading-comprehension test. Read carefully. Output exactly one letter: A or B. Do not explain.
>
> **Passage:** Read the passage below:
> On January 8, 2021, Mia attended a first-aid course.
> On June 19, 2021, Mia renewed a passport.
> On March 4, 2022, Mia bought a camera.
> On November 27, 2023, Mia moved into a new apartment.
>
> **Text-order question:** According only to the sequence in which the reports are written, which of these two events is mentioned first? Ignore the stated times or indices.
> A. Mia attended a first-aid course
> B. Mia renewed a passport
> Answer:
>
> **Chronological-order question:** According to the stated times, which event happened earlier? Ignore the order in which the reports are written.

For the chronological-order question, the passage, options, and answer prefix are unchanged. Swapping only "On January 8, 2021" and "On June 19, 2021" changes the chronological-order answer from A to B while leaving the text-order answer A. Swapping the two complete sentences instead changes only their text-order ranking. Crossing these operations produces four cells with text-order and chronological-order answers A/A, A/B, B/A, and B/B. Reversing the A/B option assignments supplies the corresponding matched orientations.

The 64 original worlds contain distinct person/event tuples but reuse 16 exact timestamp quadruples, four worlds per quadruple. Absolute dates, clock times, relative times, and other ordered time expressions vary across these sets. The decoding data randomize event-to-timestamp assignments and use disjoint complete event descriptions across train, validation, and test. Descriptions are composed from shared action, object, and location templates.

Trials and wordings. The original eight permutations are, in one-based event IDs, $(1234), (2341), (3412), (4123), (4321), (1432), (2143), (3214)$. Every event occupies every text position twice, and every pair appears in each relative text order four times. Figure 2 includes all six pairs, three text-order wordings, one chronological-order wording, and both option orientations: $64 \times 8 \times 6 \times 4 \times 2 = 24{,}576$ temporal trials per model.

The crossed-input experiments choose the base permutation by $(\text{world ID} + \text{pair ID}) \bmod 8$, then independently swap the queried sentences and their timestamps. Across both queries and A/B orientations, this gives $64 \times 6 \times 4 \times 2 \times 2 = 6{,}144$ ordinary trials per model. The rule is fixed by IDs, not outcomes. It balances permutations across the cohort, but not separately inside every timestamp set. The three text-order wordings used in the behavioral panel are evaluated separately in Table 4; only the direct wording is used in the crossed experiment.

The Direct wording is given in the example above. The other two text-order wordings are:

> **Beginning:** Looking only at placement in the passage, which event is mentioned closer to the beginning? Do not use when it happened or its assigned index.
>
> **Textual:** Which event comes earlier in the text itself? Answer by textual mention order, not chronological occurrence or assigned index.

Additional evaluation questions. The four question pairs used in Table 1 are:

1. **Text order:** According to the order of sentences in this passage, which of the two events is mentioned first? Ignore the dates.
   Chronological order: According to the dates in this passage, which of the two events happened first? Ignore sentence order.
2. **Text order:** Which of these two events appears first in the text? Answer using sentence order, not the event dates.
   Chronological order: Which of these two events occurred earlier? Answer using the event dates, not sentence order.
3. **Text order:** Considering only the order in which the events are described, which event is described first? Do not use the dates.
   Chronological order: Considering only the dates stated for the events, which event took place first? Do not use the order of description.
4. **Text order:** Which event is mentioned earlier in the passage? Compare the positions of the sentences and disregard the dates.
   Chronological order: Which event occurred earlier in time? Compare the dates and disregard the positions of the sentences.

A.3 Behavioral results and ability controls

Table 3 reports the two query-specific congruent baselines and their conflict scores. A lower conflict score alone is not an ability criterion. In the original design, a wording qualifies when congruent accuracy is at least 75% and semantic A/B-swap agreement at least 70%; a pooled text-order task requires at least two of three wordings to qualify. All ten selected models pass these temporal-task criteria.

Table 3: Behavioral panel underlying Figure 2. Congruent/conflict accuracies and losses are percentages or percentage points. The last column is the paired difference with a 95% world-bootstrap interval.

Model	Text-order C/F	Chronological order C/F	$\Delta_P - \Delta_\tau$ [95% CI]
Falcon3-10B	91.3/36.5	92.2/83.6	46.2 [38.6, 53.7]
Gemma2-9B	95.3/18.2	91.3/87.7	73.4 [65.4, 80.6]
Granite3.3-8B	89.4/21.1	87.9/86.8	67.2 [61.7, 72.3]
Llama3.1-8B	95.6/40.1	89.8/81.0	46.6 [38.2, 54.4]
Phi-4	93.9/20.5	90.8/91.5	74.2 [67.8, 79.9]
Qwen2.5-7B	94.1/50.7	93.6/68.3	18.0 [9.9, 26.4]
Qwen3-8B	96.7/25.3	88.7/77.7	60.4 [53.4, 66.9]
Qwen3.5-4B	95.7/19.5	93.5/90.2	72.8 [66.0, 78.9]
Qwen3.5-9B	96.8/38.2	89.8/90.6	59.3 [53.0, 65.1]
Yi1.5-9B	83.2/28.4	85.1/86.0	55.8 [49.7, 61.6]

Table 4: Text-order query conflict losses by wording in the 64-world behavioral panel (percentage points, 95% intervals).

Model	Direct	Beginning	Textual
Falcon3-10B	66.2 [61.5, 70.9]	53.1 [47.3, 59.0]	45.0 [40.2, 49.7]
Gemma2-9B	75.5 [70.0, 80.3]	77.7 [72.4, 82.8]	77.9 [72.5, 82.7]
Granite3.3-8B	63.1 [59.2, 66.8]	66.4 [61.8, 70.7]	75.4 [71.0, 79.3]
Llama3.1-8B	54.4 [49.1, 59.5]	46.7 [40.3, 53.1]	65.3 [60.1, 70.3]
Phi-4	72.4 [67.4, 76.7]	81.8 [77.3, 85.9]	66.0 [60.3, 71.6]
Qwen2.5-7B	44.5 [39.4, 49.6]	39.0 [34.1, 43.9]	46.7 [41.6, 51.6]
Qwen3-8B	72.5 [67.8, 76.6]	70.1 [65.2, 74.5]	71.6 [66.1, 76.7]
Qwen3.5-4B	69.0 [64.6, 73.0]	80.4 [75.7, 84.6]	79.2 [74.1, 83.8]
Qwen3.5-9B	43.4 [40.0, 46.7]	73.2 [68.6, 77.7]	59.1 [55.0, 63.1]
Yi1.5-9B	61.1 [57.0, 64.9]	50.9 [45.9, 55.8]	52.5 [48.5, 56.3]

Table 5: Behavioral checks on the crossed inputs used for Figures 4 and 5. Losses and paired contrasts are percentage points, with 95% intervals.

Model	Δ_P	Δ_τ	$\Delta_P - \Delta_\tau$
Falcon3-10B	65.4 [60.2, 70.5]	6.1 [2.7, 9.6]	59.2 [51.0, 67.2]
Gemma2-9B	79.8 [74.9, 84.2]	3.0 [0.1, 6.0]	76.8 [70.1, 83.1]
Granite3.3-8B	64.6 [60.1, 69.1]	-0.7 [-4.4, 3.2]	65.4 [58.8, 72.1]
Llama3.1-8B	52.9 [47.9, 57.6]	4.2 [0.5, 8.0]	48.7 [41.0, 56.1]
Phi-4	77.0 [72.9, 80.7]	-1.8 [-4.0, 0.4]	78.8 [73.6, 83.5]
Qwen2.5-7B	51.5 [46.2, 56.8]	23.1 [19.1, 27.1]	28.4 [19.8, 37.2]
Qwen3-8B	75.7 [70.7, 80.5]	11.6 [8.6, 14.6]	64.1 [56.8, 71.2]
Qwen3.5-4B	71.7 [66.8, 76.2]	1.8 [-0.3, 4.0]	70.0 [63.5, 76.1]
Qwen3.5-9B	46.3 [43.2, 49.4]	-3.1 [-5.5, -0.6]	49.3 [44.6, 53.9]
Yi1.5-9B	64.7 [60.0, 69.2]	-4.1 [-7.3, -1.0]	68.8 [63.0, 74.2]

The crossed-input results use the direct wording, their own inputs, and their own baseline scores. They establish that the later intervention cohort exhibits the same direction of behavioral asymmetry; they are not an independent replication on new worlds. Negative chronological-order query losses are retained. No model is excluded from the current mechanism figures because of a small effect, a null, or a control outcome.

A/B scoring. For each label, we retain the original exact token candidates and take their maximum native output logit. This is the archived label-scoring rule, not a sum over a label token set or unconstrained generation. Native output transforms are preserved, and the optimized readout is checked against the full native output layer for every model. Individual-trial ties use the original A/B rule; paired-margin ties receive accuracy 0.5. Equation 1 cancels a shared additive label bias but does not guarantee invariance to every option-dependent interaction. We therefore keep individual and paired accuracies separate.

B DECODING PROTOCOL

Features and splits. At each decoder block we extract the residual state at the final token of each event sentence, before the model's final normalization. The prefix excludes question and answer-option tokens. Both labels take four categorical values: the sentence's text position and its chronological rank among the world's four events. All eight permutations of a world remain in one split, giving 32 event observations per world. There are 8,192 training observations, 2,048 validation observations, and 8,192 matched-test observations. Each class is balanced in every split. All ten models and all 370 decoder blocks are included.

The 256/64/256 matched split shares the original 16 timestamp sets in balanced proportions. All complete event descriptions are disjoint across splits and from the original worlds, but vocabulary and compositional templates are shared. The original 64 worlds provide an additional text-order transfer test, without entering the probe training or validation sets. Their shorter event descriptions introduce a length and style shift as well as new event contexts.

Linear classifiers. Let H be the matrix of hidden states and Y the four-column one-hot target for one rank. We center features on the training mean, retain an intercept, and fit linear ridge scores by minimizing squared one-hot error plus $\alpha\|W\|_F^2$. Prediction is the class with the largest score. Text-order and chronological-order probes choose regularization independently using validation accuracy, then validation one-hot mean squared error, then stronger regularization as tie-breakers. The candidate set is

$$\{10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3\} \cup \{s\,10^k : k = -8, \ldots, 4\}, \qquad s = \|X_{\text{train}}\|_F^2/n, \tag{3}$$

plus an intercept-only model. Here n is the number of training observations and X the centered feature matrix. Neither layer selection nor penalty selection uses the test set. Solves use FP64 primal or dual eigensystems according to matrix size, with TF32 disabled. Independent direct solves agree to a maximum error of 4.22×10^{-13} in the retained checks; no selected finite penalty reaches the upper boundary of the expanded search.

On the original 64 worlds, final-block text-order decoding ranges from 93.3% to 100% without refitting. This test connects information availability to the original behavioral materials, whose descriptions introduce a length and style shift.

The reported probes use complete hidden states, including positional information; the results do not establish a position-independent code. All decoder blocks are shown in Figure 3. Conditional world-bootstrap intervals hold the fitted probes fixed and do not include variation from retraining.

C STATE TRANSFER, ACCESS MASKS, AND NUMERICAL CONTROLS

C.1 COUNTERFACTUAL ALIGNMENT AND INTERVENTION REGION

For each queried pair, donor and recipient differ only by exchange of that pair's in-sentence timestamps. Text order, event descriptions, query, options, carrier token IDs, absolute token positions, and total prefix length are matched. The carrier starts after the final timestamp token and ends before the question. It includes the remainder of the last event sentence and separator; it is not an independently inserted neutral memory buffer. Both prefixes are computed before the query.

For a K/V transfer, donor keys and values replace recipient entries at carrier positions at every full-attention layer. K-only and V-only variants replace only one cache tensor. The original Qwen3.5 transfers in Figure 4 leave recipient convolutional and recurrent states unchanged. This scope is the same for corruption and rescue. Figure 5b additionally includes recurrent carrier writes, described in Appendix C.4.

Let G and C be the model's mean text-order-correct paired margins under congruent and conflict baseline conditions. For the patched congruent and conflict margins N and D, respectively, we report

$$f_N = \frac{G - N}{G - C}, \qquad f_D = \frac{D - C}{G - C}. \tag{4}$$

Figure 4b shows $100(1 - f_N)$ and $100 f_D$ against baseline anchors 100 and 0. The anchors are a change of scale, not observations of 100% and 0% accuracy. Numerator and denominator are bootstrapped jointly; all denominator intervals exclude zero here. Values are not clipped to $[0, 100]$. These intervention fractions are not an additive or exhaustive decomposition of the total effect.

Table 6: Carrier K/V transfer on 64 worlds. Fractions of the baseline margin gap are percentages with 95% intervals. Accuracy columns show conflict baseline/rescued accuracy (percent); paired and single-trial readouts are separate.

Model	Corruption fraction	Rescue fraction	Paired acc.	Trial acc.
Falcon3-10B	33.9 [30.9, 37.3]	31.7 [28.7, 35.2]	26.4/51.7	28.9/51.9
Gemma2-9B	34.5 [30.3, 39.4]	34.2 [29.7, 39.1]	12.5/45.4	14.8/42.5
Granite3.3-8B	30.2 [26.5, 34.5]	32.1 [28.4, 36.2]	20.2/44.7	22.9/42.9
Llama3.1-8B	27.0 [23.6, 30.4]	33.6 [30.3, 37.3]	27.9/59.2	36.2/57.4
Phi-4	32.7 [28.6, 37.4]	24.5 [21.0, 28.3]	16.4/38.0	18.5/39.0
Qwen2.5-7B	25.0 [21.7, 28.7]	31.8 [28.4, 35.7]	43.4/64.3	43.9/63.5
Qwen3-8B	24.7 [22.6, 27.0]	32.0 [29.2, 34.9]	18.5/52.2	21.0/48.6
Qwen3.5-4B	22.8 [19.8, 26.4]	24.8 [22.0, 28.1]	21.0/48.6	23.9/48.7
Qwen3.5-9B	24.8 [21.6, 28.2]	20.1 [17.6, 22.8]	47.7/66.9	49.3/64.8
Yi1.5-9B	32.3 [29.4, 35.6]	33.6 [30.2, 37.4]	15.6/42.1	20.3/42.8

C.2 SPECIFICITY AND RECONSTRUCTION CONTROLS

Self-transfer replaces a cache with its own entries. Whole-prefix transfer replaces all prefix caches and, for Qwen3.5, the corresponding recurrent states; it should recover the donor continuation. The pre-carrier control transfers the full prefix before the suffix, not just timestamp keys. A sham donor exchanges the two nonqueried timestamps while keeping the queried pair fixed. This controls for a different timestamp change, although it need not produce zero contextual effects. Carrier-minus-sham corruption contrasts are positive in all ten models.

For the blocked-transfer control, question rows cannot access timestamp keys in the recipient cache. Carrier transfer still shifts the text-order margin in both directions in all models (Table 8). Since indirect routes remain, this result shows that the transferred suffix can influence the answer without the blocked question-to-timestamp edges; it does not prove that all other timestamp information has been eliminated.

"Rescue" is used only for text-order queries, where timestamp exchange preserves the correct answer. For chronological-order queries the same exchange changes the target itself. Those raw chronological-order query results are retained in the source tables but are not described as restoration of an unchanged answer.

Table 7: Change in text-order-correct paired margin under separate K and V transfers (baseline SD units, 95% intervals). Negative corruption and positive rescue changes are in the predicted directions.

Model	Direction	K only	V only
Falcon3-10B	Corruption	-0.04 [-0.04, -0.03]	-0.31 [-0.35, -0.28]
Falcon3-10B	Rescue	0.05 [0.04, 0.06]	0.38 [0.35, 0.41]
Gemma2-9B	Corruption	-0.03 [-0.04, -0.02]	-0.42 [-0.48, -0.36]
Gemma2-9B	Rescue	0.03 [0.03, 0.04]	0.51 [0.45, 0.57]
Granite3.3-8B	Corruption	-0.05 [-0.07, -0.04]	-0.25 [-0.29, -0.22]
Granite3.3-8B	Rescue	0.05 [0.05, 0.06]	0.32 [0.28, 0.35]
Llama3.1-8B	Corruption	-0.06 [-0.07, -0.04]	-0.21 [-0.24, -0.18]
Llama3.1-8B	Rescue	0.20 [0.17, 0.22]	0.45 [0.41, 0.50]
Phi-4	Corruption	-0.07 [-0.08, -0.06]	-0.23 [-0.25, -0.20]
Phi-4	Rescue	0.04 [0.03, 0.04]	0.21 [0.18, 0.24]
Qwen2.5-7B	Corruption	-0.02 [-0.02, -0.01]	-0.21 [-0.25, -0.18]
Qwen2.5-7B	Rescue	0.03 [0.02, 0.03]	0.32 [0.28, 0.36]
Qwen3-8B	Corruption	-0.02 [-0.02, -0.02]	-0.31 [-0.35, -0.27]
Qwen3-8B	Rescue	0.05 [0.04, 0.06]	0.49 [0.44, 0.53]
Qwen3.5-4B	Corruption	-0.01 [-0.02, -0.00]	-0.24 [-0.27, -0.22]
Qwen3.5-4B	Rescue	0.05 [0.04, 0.06]	0.29 [0.26, 0.32]
Qwen3.5-9B	Corruption	-0.00 [-0.01, 0.00]	-0.17 [-0.20, -0.15]
Qwen3.5-9B	Rescue	0.02 [0.02, 0.03]	0.14 [0.12, 0.16]
Yi1.5-9B	Corruption	-0.05 [-0.06, -0.04]	-0.34 [-0.37, -0.31]
Yi1.5-9B	Rescue	0.05 [0.04, 0.06]	0.43 [0.38, 0.47]

Table 8: Carrier specificity controls in baseline margin SD (95% intervals). Corruption is oriented as a positive reduction. Blocked conditions prevent the question from reading timestamp keys.

Model	Carrier − sham	Blocked corruption	Blocked rescue
Falcon3-10B	0.43 [0.39, 0.47]	0.51 [0.47, 0.56]	0.49 [0.45, 0.54]
Gemma2-9B	0.58 [0.51, 0.65]	0.65 [0.57, 0.72]	0.66 [0.58, 0.74]
Granite3.3-8B	0.36 [0.32, 0.41]	0.49 [0.43, 0.56]	0.46 [0.40, 0.52]
Llama3.1-8B	0.41 [0.36, 0.46]	0.53 [0.47, 0.59]	0.58 [0.52, 0.63]
Phi-4	0.36 [0.32, 0.39]	0.53 [0.47, 0.58]	0.51 [0.45, 0.56]
Qwen2.5-7B	0.26 [0.22, 0.30]	0.35 [0.30, 0.39]	0.39 [0.34, 0.44]
Qwen3-8B	0.41 [0.37, 0.46]	0.48 [0.43, 0.52]	0.54 [0.49, 0.59]
Qwen3.5-4B	0.32 [0.28, 0.36]	0.29 [0.25, 0.34]	0.36 [0.33, 0.40]
Qwen3.5-9B	0.22 [0.19, 0.24]	0.23 [0.20, 0.27]	0.26 [0.23, 0.29]
Yi1.5-9B	0.45 [0.40, 0.49]	0.59 [0.54, 0.64]	0.55 [0.50, 0.61]

C.3 Access masks and matched controls

The access experiment prepends a fixed neutral passage and computes its own baseline. Timestamp-key masks act at question rows, non-timestamp context rows after the first timestamp, or both. The context mask spans a broader part of the passage than the carrier used for transplantation. For every destination row, its neutral-key control removes the same number of causally available key positions. The two key sets need not have identical attention mass or content; the contrast tests the specified timestamp-access intervention against an equal-edge-count control.

In the original full-attention-only experiment, we compare chronological-order gain under each timestamp mask with the matched neutral mask. Query-only intervals are below zero for two models; context and joint intervals are below zero for nine. Qwen3.5-4B instead has a small positive context contrast. Qwen3.5-9B has negative but much smaller context and joint effects than the full-attention models (Table 9). In both hybrid models, masks leave recurrent pathways and propagation through unmasked timestamp-token rows intact. The Qwen3.5 points in Figure 5a instead show the row-scoped hybrid extension below; Table 9 retains the original conditions.

Table 9: Original full-attention-only masks on 64 worlds and six pairs: timestamp-key minus matched neutral-key change in chronological-order gain for text-order queries, in baseline SD per rank (95% intervals). Negative values indicate reduced influence. Updated Qwen3.5 hybrid results appear in Table 12.

Model	Query rows	Context rows	Both
Falcon3-10B	-0.007 [-0.016, 0.003]	-0.331 [-0.346, -0.315]	-0.406 [-0.424, -0.387]
Gemma2-9B	0.013 [0.006, 0.020]	-0.497 [-0.524, -0.467]	-0.566 [-0.594, -0.534]
Granite3.3-8B	0.097 [0.079, 0.114]	-0.291 [-0.313, -0.269]	-0.344 [-0.367, -0.319]
Llama3.1-8B	0.065 [0.046, 0.084]	-0.405 [-0.423, -0.385]	-0.450 [-0.472, -0.427]
Phi-4	0.026 [0.013, 0.040]	-0.165 [-0.179, -0.150]	-0.291 [-0.306, -0.275]
Qwen2.5-7B	-0.010 [-0.018, -0.002]	-0.256 [-0.278, -0.234]	-0.332 [-0.358, -0.307]
Qwen3-8B	-0.057 [-0.068, -0.046]	-0.428 [-0.452, -0.403]	-0.555 [-0.584, -0.525]
Qwen3.5-4B	0.006 [-0.005, 0.017]	0.008 [0.004, 0.011]	0.006 [-0.005, 0.016]
Qwen3.5-9B	-0.003 [-0.010, 0.004]	-0.006 [-0.008, -0.004]	-0.011 [-0.017, -0.004]
Yi1.5-9B	-0.003 [-0.013, 0.007]	-0.323 [-0.341, -0.305]	-0.412 [-0.434, -0.389]

Capability and positive-control results are reported separately for each query. The text-order positive control blocks query access to event-description keys; the chronological-order positive control blocks query access to timestamp keys, each against its matched neutral block. These tests have their own limitations when information is also available indirectly. No post-hoc threshold on these diagnostics removes a model from the original cohort. The extension has separate text-order and chronological-order accuracy checks.

Table 10: Access-experiment diagnostics for text-order queries. C/F and swap agreement are percentages; gains and positive-control changes use baseline SD per rank. Positive-control changes include 95% intervals. Models are not filtered by these outcomes.

Model	C/F	Swap	G_{rel}	G_{irr}	Control ΔG_{rel}
Falcon3-10B	95.0/28.7	87.8	0.170	0.418	-0.103 [-0.131, -0.075]
Gemma2-9B	94.1/17.6	87.6	0.103	0.567	-0.040 [-0.069, -0.012]
Granite3.3-8B	81.0/25.1	73.0	0.017	0.339	-0.006 [-0.032, 0.019]
Llama3.1-8B	91.7/36.4	68.9	0.190	0.442	-0.173 [-0.210, -0.138]
Phi-4	95.7/25.5	87.3	0.120	0.269	-0.115 [-0.131, -0.099]
Qwen2.5-7B	95.6/44.8	83.7	0.253	0.304	-0.188 [-0.216, -0.158]
Qwen3-8B	97.0/19.8	91.9	0.145	0.550	-0.128 [-0.155, -0.103]
Qwen3.5-4B	97.0/24.1	88.3	0.170	0.424	-0.089 [-0.113, -0.066]
Qwen3.5-9B	96.5/41.5	82.9	0.233	0.277	-0.211 [-0.229, -0.192]
Yi1.5-9B	86.5/24.5	75.1	0.078	0.371	-0.029 [-0.053, -0.006]

C.4 Hybrid access and carrier-write extensions

Row-scoped source exclusion. The Qwen3.5-4B and 9B extension uses worlds 0–15, covering the original 16 timestamp sets, and fixed event pair 0. Both text orders, timestamp assignments, query types, and A/B orientations give 256 inputs per model. A baseline and six interventions yield 1,792 primary scores per model. Every scope includes all eight full-attention and 24 linear-attention blocks. This is a smaller extension of the original materials, not a full-cohort replication.

Table 11: Access-experiment diagnostics for chronological-order queries. C/F and swap agreement are percentages; gains and positive-control changes use baseline SD per rank. Positive-control changes include 95% intervals. Models are not filtered by these outcomes.

Model	C/F	Swap	G_{rel}	G_{irr}	Control ΔG_{rel}
Falcon3-10B	90.1/84.2	82.9	0.581	0.045	-0.026 [-0.036, -0.017]
Gemma2-9B	92.4/88.3	91.0	0.566	0.000	-0.024 [-0.032, -0.016]
Granite3.3-8B	84.0/88.3	90.3	0.601	-0.058	-0.020 [-0.032, -0.008]
Llama3.1-8B	85.4/75.6	66.7	0.465	0.047	-0.060 [-0.073, -0.048]
Phi-4	91.7/94.1	94.0	0.712	-0.056	-0.143 [-0.175, -0.112]
Qwen2.5-7B	94.6/70.4	91.2	0.528	0.199	-0.034 [-0.044, -0.024]
Qwen3-8B	88.5/81.8	85.3	0.503	0.069	-0.073 [-0.091, -0.055]
Qwen3.5-4B	95.1/91.0	94.1	0.595	0.038	-0.040 [-0.051, -0.029]
Qwen3.5-9B	90.3/95.1	93.0	0.682	-0.043	-0.063 [-0.077, -0.049]
Yi1.5-9B	86.1/84.0	82.9	0.573	0.024	-0.133 [-0.152, -0.115]

Query rows run from `Question:` through the answer prefix and chat suffix. Context rows comprise non-timestamp positions between the first timestamp and the question; Both is their union. Full-attention masks retain the original edge definitions. In each linear block, native and source-excluded branches start from the same incoming hidden states. The latter zeros the selected source QKV projections before causal convolution and sets their write/erase gate $\beta = 0$ and log-decay $g = 0$, hence $\alpha = 1$. We run the native gated delta recurrence, including Q/K normalization, and select the excluded branch's core outputs only at the designated rows. Other rows retain the native branch outputs exactly for that layer input. Native gated normalization and output projection follow, and later layers receive the intervened hidden states normally. Computation uses complete uncached forwards, without installing an edited generation cache.

Thus Query-only and Context-only differ in linear readout rows as well as full-attention edges. Both uses the same row rule. Neutral-source controls match source-token counts, but not all positional or state distributions. All worlds, including timestamp swaps with different token boundaries, are retained.

Raw gains and their bootstrap bounds are divided by the original 64-world access baseline SD: 2.12409 for 4B and 3.52172 for 9B. These reference denominators are held fixed, so extension intervals condition on them. All scopes share 10,000 paired world-bootstrap draws, seed 20260924. Table 12 reports control-minus-intervention contrasts and balanced accuracy across congruent and conflict inputs. Comparisons against full-attention-only masks use the same 16 worlds and pair, avoiding a cohort-change interpretation.

Table 12: Qwen3.5 row-scoped full + linear source exclusion: 16 worlds and one fixed pair. Gain reduction is neutral-control minus timestamp intervention (reference SD per rank, 95% CI). Accuracy columns are A/B-paired percentages balanced over congruent and conflict inputs under the listed timestamp intervention; baseline is unmodified.

Model	Rows	Chronological-order gain removed	Text-order accuracy	Chronological-order accuracy
Qwen3.5-4B	Baseline	—	56.2	90.6
Qwen3.5-4B	Query	-0.004 [-0.070, 0.064]	65.6	85.9
Qwen3.5-4B	Context	0.554 [0.421, 0.682]	95.3	57.8
Qwen3.5-4B	Both	0.568 [0.434, 0.696]	96.9	50.0
Qwen3.5-9B	Baseline	—	65.6	93.8
Qwen3.5-9B	Query	-0.041 [-0.097, 0.013]	57.8	85.9
Qwen3.5-9B	Context	0.295 [0.229, 0.359]	96.9	70.3
Qwen3.5-9B	Both	0.328 [0.250, 0.403]	100.0	50.0

Context blocking reduces chronological-order gain for text-order queries in both models, whereas query-only contrasts include zero. Relative to full-attention-only masking on the same inputs, the additional context reduction is 0.577 [0.439, 0.711] SD per rank for 4B and 0.295 [0.230, 0.359] for 9B. Both leaves gains of 0.0007 and -0.0003, respectively, but balanced chronological-order query accuracy falls to 50%. Removing relevant chronological-order information is therefore part of the intervention's cost. The operation changes convolution, memory writing, erasure, and decay jointly; its effect is not an additive linear-path attribution or a selective repair.

Carrier K/V writes into recurrent states. The separate transfer extension retains all 64 worlds and all six pairs. Full-only conditions reproduce the original carrier K-only and V-only cache transfers in eight blocks. Linear-only conditions replace donor post-convolution carrier K or V in all 24 linear blocks and replay each prefix recurrence to its terminal state. Recipient Q, gates, the other K/V component, and noncarrier positions remain fixed; native normalization is retained. Replayed prefix outputs are discarded so deeper-layer donor and recipient activations remain native. Edited terminal states are installed at the question boundary. Hybrid conditions combine the two scopes; all retain recipient convolution caches.

This edits the carrier's state writes, rather than substituting the entire prefix memory. Linear K affects both writing and delta-rule correction and is not a strict analogue of a softmax addressing key. The three scopes can interact. Figure 5b shows the extended hybrid endpoints; Table 13 compares them with full-only transfers. All estimates reuse the original ordinary-forward SD and its recomputation in 10,000 world-bootstrap draws, seed 20260923. Self, nonqueried-timestamp sham, and whole-prefix controls remain separate. V-only hybrid rescue exceeds its matched sham by 0.317 [0.289, 0.347] SD in 4B and 0.148 [0.124, 0.174] in 9B.

Table 13: Hybrid carrier transfer on all 64 worlds and six pairs. Corruption is before minus after in congruent inputs; rescue is after minus before in conflict inputs. The last column is the paired additional rescue effect relative to full-attention-only transfer. Units are the original ordinary-forward baseline SD (95% CIs).

Model	Component	Corruption	Rescue	Hybrid − full
Qwen3.5-4B	K	0.003 [-0.005, 0.011]	0.050 [0.041, 0.061]	0.000 [-0.004, 0.004]
Qwen3.5-4B	V	0.266 [0.239, 0.295]	0.310 [0.278, 0.345]	0.020 [0.015, 0.024]
Qwen3.5-9B	K	0.003 [-0.003, 0.008]	0.033 [0.027, 0.038]	0.009 [0.006, 0.012]
Qwen3.5-9B	V	0.189 [0.165, 0.216]	0.153 [0.131, 0.177]	0.013 [0.010, 0.016]

Extension validation. Row-mask experiments use strict FP32 computation with BF16-stored weights and TF32 disabled. Each world's baseline is checked against frozen scores; world 0 additionally reproduces the six original full-only masks. Maximum margin discrepancies are 0.000597 and 0.003077 for 4B and 9B, below the predeclared 0.01 tolerance. Every edited block checks source zeroing, gate neutralization, and exact target/untargeted row selection. Independent tokenwise recurrence checks at fixed first, middle, and last linear blocks differ by at most 6.68×10^{-6}.

The carrier-write extension contains 73,728 condition scores per model and checks all original baseline/full-K/full-V scores. Maximum margin differences are 0.000870 and 0.001990; three individual-trial signs change across both models, but all reproduced A/B-paired decisions agree. Recurrence and attention use strict FP32; the qualified projection backend uses split-input TF32 with FP32 accumulation. Qualifications at fixed worlds 0 and 32, self identity, donor recovery, exact component checks, and independent recurrence checks are retained in the source audit. These numerical checks do not remove the scope and cohort limitations of either extension.

C.5 Numerical validation

The mechanism cohort contains 61,440 intervention scores and 6,144 ordinary full-forward scores per model; the access cohort contains 55,296 scores per model. Across ten different LLMs this is 1,228,800 condition scores, not that many independent samples. Attention, normalization, residuals, caches, and output readout use FP32 in the mechanism evaluation with separately qualified linear arithmetic. Each backend is compared with strict FP32 before its formal cohort. Access arithmetic is qualified separately on all nine masks. The acceptance tolerance remains 0.01 in logit margin.

Formal cells check self identity, whole-prefix donor recovery, and cached versus complete native forward agreement. Native output-layer transformations, including model-specific logit scaling, are retained. Per-model maximum errors and decision disagreements accompany the paper's source data.

D INPUT INFLUENCE AND ADAPTATION

D.1 ACCESS-EFFECT ESTIMATION

Signed finite differences. For fixed canonical events a, b, let $u_{st,q}$ be the paired margin when text-order swap s and timestamp swap t are each in $\{0, 1\}$. Let $d_p = p(b) - p(a)$ and $d_\tau = \tau(b) - \tau(a)$ in the unswapped cell. The per-rank gains are

$$J_{q \leftarrow p} = \frac{u_{00,q} + u_{01,q} - u_{10,q} - u_{11,q}}{4d_p}, \tag{5}$$

$$J_{q \leftarrow \tau} = \frac{u_{00,q} + u_{10,q} - u_{01,q} - u_{11,q}}{4d_\tau}. \tag{6}$$

Each contrast averages over the other factor and divides by the signed rank-distance change. Pair-level gains are averaged within each world before the cohort mean. For the access experiment in Figure 5, gains use the experiment's neutral-prefix baseline margin SD, shared across the query types and conditions within a model. The Qwen3.5 extension uses the reference scales specified in Appendix C.4.

Sampling unit. Behavioral and intervention intervals use 10,000 paired bootstrap draws of 64 worlds (seed 20260923), after averaging pairs, crossed directions, and option orientations within worlds. The same draws resample all conditions and components of each statistic. Intervals are pointwise and conditional on the timestamp sets used; model counts do not estimate population frequencies. Probe intervals resample their own held-out worlds with fitted classifiers fixed. ReFrame uncertainty is specified below.

D.2 MECHANISM-INSPIRED REFRAME ADAPTATION AND PAIRED READOUT GEOMETRY

Models and trainable parameters. We use the same training data and recipe with three seeds (11, 23, 47) for each of the ten models. All pretrained parameters are frozen. We add $\Delta W = (\alpha/r)BA$ to each Q and V projection, with $\alpha/r = 2$, no dropout, Kaiming-uniform A, and zero-initialized B. V has rank 4. Q ranks are distributed deterministically across layers to approximate the same total parameter budget as V; integer rounding is performed separately for each architectural group. Table 14 gives the resulting counts and selected epochs.

For Phi-4, only the Q and V row slices of its fused QKV projection are adapted. For Qwen3.5, we adapt Q/V slices in both full-attention and recurrent linear-attention blocks, excluding the extra query-gate rows. K, gates, output projections, MLPs, embeddings, and the output head remain frozen. The backbone uses BF16 and the low-rank parameters use FP32; models are not quantized. Answer scoring projects the final hidden state onto the native A/B output rows in FP32, followed by the model-specific output scaling or softcapping. Gemma retains its attention and output softcapping, and Granite retains its native output scaling. Qwen3.5 recurrent computation is checked against the native kernel. Native and trained comparisons share rendering, backend, and evaluation batches within a model. Before optimization, zero-initialized adapters are checked to reproduce native outputs exactly at fixed batching.

Training and selection. The task training set contains 256 four-event scenes and 4,096 trials: two query types, two text orders, two chronological orders, and two option orientations per scene. Half the scenes use the original wording; the others use four training-only paraphrase families. We also use 384 capability trials (64 cases per domain for facts, arithmetic, and logic, each in both option orientations). Let m be the A/B margin and $p = \mathrm{sigmoid}(m)$. The objective is binary answer cross-entropy plus 0.1 times the symmetric Bernoulli KL between paired trials that differ only in the irrelevant order, plus 0.1 times the native-to-adapted Bernoulli KL on capability trials. Correct-answer supervision also applies to the capability examples. The native capability probabilities are fixed teachers.

We use AdamW with learning rate 10^{-4}, weight decay 0.01, gradient clipping at 1, a 5% linear warm-up, and two epochs. A microbatch contains two paired units, with four-step gradient accumulation (304 optimizer updates total). Validation is performed every half epoch. It contains 64 original-wording scenes, 32 additional-wording scenes, and 32 capability cases per domain, totaling 1,728

Table 14: Trainable ReFrame parameters and validation-selected epochs. All runs allow the native epoch-zero state. Selection uses text-order conflict validation accuracy subject to retention constraints; no test accuracy enters the rule.

Model	Parameters	Seed 11	Seed 23	Seed 47
Falcon3-10B	1,306,624	2.0	2.0	2.0
Gemma-2-9B	1,890,816	1.5	2.0	2.0
Granite-3.3-8B	1,638,400	2.0	2.0	2.0
Llama-3.1-8B	1,310,720	1.5	1.5	2.0
Phi-4	2,048,000	2.0	2.0	2.0
Qwen2.5-7B	917,504	2.0	1.5	1.5
Qwen3-8B	1,474,560	1.5	2.0	2.0
Qwen3.5-4B	1,502,720	2.0	1.5	2.0
Qwen3.5-9B	1,900,544	1.5	2.0	2.0
Yi-1.5-9B	1,769,472	1.0	2.0	1.0

Table 15: **Complete ReFrame accuracy by query and congruence.** Accuracy (%) on the same 64 held-out scenes; each cell is native → ReFrame (three-seed mean). Additional questions average four fixed, unseen question pairs with events, dates, and options held constant.

Original questions

	Text-order query		Chronological-order query	
Model	Congruent	Conflict	Congruent	Conflict
Falcon3-10B	92.6 → 99.9	52.7 → 99.5	90.6 → 96.0	73.8 → 93.4
Gemma-2-9B	93.8 → 100.0	39.5 → 100.0	95.7 → 95.7	83.2 → 93.5
Granite-3.3-8B	92.6 → 99.9	43.8 → 99.9	88.3 → 94.0	76.6 → 95.8
Llama-3.1-8B	79.3 → 99.7	46.5 → 100.0	70.7 → 95.1	73.4 → 93.8
Phi-4	95.7 → 100.0	31.2 → 100.0	96.9 → 99.9	93.4 → 97.1
Qwen2.5-7B	97.3 → 100.0	62.1 → 100.0	95.7 → 91.9	55.1 → 89.5
Qwen3-8B	96.5 → 99.7	34.4 → 100.0	89.8 → 95.2	83.6 → 96.0
Qwen3.5-4B	97.7 → 100.0	52.7 → 100.0	96.1 → 99.2	90.2 → 96.1
Qwen3.5-9B	98.0 → 100.0	66.0 → 100.0	88.3 → 98.8	93.0 → 96.9
Yi-1.5-9B	91.4 → 100.0	35.9 → 99.5	88.3 → 96.4	81.6 → 92.1

Additional questions

	Text-order query		Chronological-order query	
Model	Congruent	Conflict	Congruent	Conflict
Falcon3-10B	95.3 → 100.0	53.9 → 97.9	86.9 → 96.4	71.1 → 91.9
Gemma-2-9B	94.3 → 100.0	42.4 → 100.0	94.3 → 96.0	80.3 → 93.6
Granite-3.3-8B	91.9 → 99.9	45.7 → 83.4	85.4 → 94.1	76.8 → 94.6
Llama-3.1-8B	91.9 → 99.3	50.5 → 98.8	71.6 → 95.8	66.2 → 92.8
Phi-4	95.7 → 100.0	28.4 → 99.5	96.9 → 99.8	92.2 → 95.5
Qwen2.5-7B	95.6 → 99.9	61.7 → 97.7	95.4 → 92.6	52.1 → 88.1
Qwen3-8B	95.3 → 99.3	30.9 → 98.4	91.1 → 96.4	79.1 → 93.8
Qwen3.5-4B	97.4 → 100.0	31.8 → 99.8	95.2 → 99.4	92.3 → 95.7
Qwen3.5-9B	95.2 → 100.0	56.7 → 100.0	86.0 → 99.1	93.9 → 96.1
Yi-1.5-9B	84.7 → 99.9	38.4 → 94.3	79.7 → 97.3	73.6 → 90.4

trials. Selection maximizes original-wording text-order conflict accuracy, breaking ties by cross-entropy over the validation trials. Relative to native validation accuracy, a candidate training state may lose at most two percentage points in chronological-order accuracy and text-order congruent accuracy, and five points in each capability domain. The untrained state is eligible. Additional-wording validation accuracy is recorded but is not an accuracy gate. No test result selects a training state or seed.

Each training seed is evaluated separately; reported ReFrame accuracies average the three seeds without ensembling logits. Table 15 gives the full query and congruence breakdown, Table 16 reports seed variation, and Table 17 reports chronological-order accuracy for each additional question pair. The exact questions appear in Appendix A.

Table 16: ReFrame accuracy (%, mean ± sample standard deviation across the three training seeds). These summarize training variability, not confidence intervals over scenes.

Model	Original Text-order conflict	Original Chronological order	Additional Text-order conflict	Additional Chronological order
Falcon3-10B	99.5 ± 0.6	94.7 ± 0.1	97.9 ± 0.9	94.2 ± 1.0
Gemma-2-9B	100.0 ± 0.0	94.6 ± 1.0	100.0 ± 0.1	94.8 ± 0.8
Granite-3.3-8B	99.9 ± 0.2	94.9 ± 0.9	83.4 ± 4.2	94.4 ± 0.9
Llama-3.1-8B	100.0 ± 0.0	94.4 ± 1.7	98.8 ± 0.4	94.3 ± 0.9
Phi-4	100.0 ± 0.0	98.5 ± 0.4	99.5 ± 0.3	97.7 ± 0.7
Qwen2.5-7B	100.0 ± 0.0	90.7 ± 1.0	97.7 ± 1.9	90.3 ± 1.0
Qwen3-8B	100.0 ± 0.0	95.6 ± 0.7	98.4 ± 0.4	95.1 ± 0.0
Qwen3.5-4B	100.0 ± 0.0	97.7 ± 0.2	99.8 ± 0.1	97.6 ± 0.3
Qwen3.5-9B	100.0 ± 0.0	97.9 ± 0.4	100.0 ± 0.1	97.6 ± 0.4
Yi-1.5-9B	99.5 ± 0.2	94.2 ± 0.9	94.3 ± 1.4	93.8 ± 0.2

Table 17: Each additional question pair: overall chronological-order accuracy (%). All four fixed families are retained. ReFrame averages three seeds.

	Pair 1		Pair 2		Pair 3		Pair 4	
Model	Native	ReFrame	Native	ReFrame	Native	ReFrame	Native	ReFrame
Falcon3-10B	78.5	93.9	78.5	93.9	78.5	94.5	80.5	94.4
Gemma-2-9B	85.2	94.7	88.7	94.8	86.5	94.7	88.9	95.0
Granite-3.3-8B	78.9	93.4	82.8	94.7	82.0	94.9	80.7	94.5
Llama-3.1-8B	74.0	94.3	60.4	95.5	76.6	93.2	64.6	94.3
Phi-4	92.0	97.7	95.3	98.5	93.8	95.7	97.1	98.8
Qwen2.5-7B	72.5	89.6	73.0	90.7	74.4	90.3	75.2	90.8
Qwen3-8B	85.2	95.2	81.6	94.0	87.1	95.9	86.5	95.1
Qwen3.5-4B	93.6	97.3	94.9	97.9	93.8	97.3	92.8	97.8
Qwen3.5-9B	90.2	97.1	91.0	98.2	88.3	97.7	90.4	97.5
Yi-1.5-9B	81.1	93.6	67.2	94.3	79.7	92.8	78.7	94.5

Exact factorial contrasts. For each scene, query, and order configuration, let $m^{(0)}$ and $m^{(1)}$ be the A/B margins under the two option orientations. We use the event-aligned margin $y = (m^{(0)} - m^{(1)})/2$. Let $s, t \in \{-1, +1\}$ indicate whether the canonical event is earlier in text order and chronological order. The four balanced cells have the exact decomposition

$$y_{s,t} = \beta_0 + \beta_p s + \beta_\tau t + \beta_{p\tau} st, \qquad \beta_p = \tfrac{1}{4} \sum_{s,t} s\, y_{s,t}, \quad \beta_\tau = \tfrac{1}{4} \sum_{s,t} t\, y_{s,t}. \tag{7}$$

These coefficients represent half the contrast from reversing the queried pair's relative order, averaged over the other factor. They use sign-coded factors and differ in units from the per-rank J coefficients used for access-effect estimation. No fitted probe or rotation of hidden states enters Figure 6.

For each model, σ_{native} is the population standard deviation of its native event-aligned margins pooled over the 64 scenes, both queries, and all four cells under original wording. We use this one denominator for both coordinates and every trained seed. Axes have equal scaling within each panel; plot limits can differ across models. This avoids interpreting raw logit scale as a cross-model difference or independently rescaling each trained vector.

Angles and retained influence. For each seed, we first average contrasts over scenes, then calculate

$$\delta_{\text{P}} = \operatorname{atan2}(|\overline{\beta}_{\text{P},\tau}|, \overline{\beta}_{\text{P},p}), \tag{8}$$

$$\delta_\tau = \operatorname{atan2}(|\overline{\beta}_{\tau,p}|, \overline{\beta}_{\tau,\tau}). \tag{9}$$

Both angles lie in $[0, 180°]$ and penalize relevant influence with the wrong sign. Zero-norm responses have undefined angles. We average the three seed-level angles for annotation and tabulation; filled Figure 6 endpoints average their coordinates. The small seed markers preserve the variation between them. The angle between the text-order and chronological-order vectors and its distance from 90° provide complementary diagnostics. We compute these for each seed before averaging across seeds within a model. On original questions, the median between-readout angle across models increases from 30.7° to 87.7°; adapted model means range from 84.9° to 91.0°. The mean absolute distance from 90° decreases in all ten models. Correct-axis alignment remains the primary criterion because perpendicular vectors can still answer the wrong questions.

Table 18: Text-order query deviation from its correct positive input axis (degrees). Brackets give pointwise 95% intervals from 2,000 paired scene-bootstrap draws, conditional on the selected training states. Negative ReFrame-minus-native changes indicate improved alignment.

Model	Native	ReFrame	Paired change
Falcon3-10B	43.7 [35.4, 52.2]	0.1 [0.0, 0.9]	-43.6 [-52.1, -35.1]
Gemma-2-9B	57.7 [47.7, 66.5]	0.1 [0.1, 0.3]	-57.5 [-66.3, -47.6]
Granite-3.3-8B	51.6 [42.0, 61.6]	0.5 [0.2, 0.9]	-51.1 [-61.0, -41.5]
Llama-3.1-8B	52.9 [45.8, 59.9]	0.9 [0.3, 1.7]	-51.9 [-58.9, -45.0]
Phi-4	58.0 [52.1, 64.2]	0.7 [0.5, 1.0]	-57.3 [-63.5, -51.4]
Qwen2.5-7B	30.0 [22.4, 39.3]	1.0 [0.7, 1.5]	-29.0 [-38.2, -21.4]
Qwen3-8B	57.6 [49.6, 64.6]	0.6 [0.2, 1.4]	-57.0 [-63.9, -49.1]
Qwen3.5-4B	44.4 [36.9, 52.0]	0.5 [0.3, 0.7]	-43.8 [-51.4, -36.4]
Qwen3.5-9B	25.3 [19.5, 32.1]	0.6 [0.4, 0.8]	-24.7 [-31.4, -19.1]
Yi-1.5-9B	62.0 [51.9, 71.7]	1.5 [0.8, 2.4]	-60.5 [-70.1, -50.7]

Table 19: Chronological-order query deviation from its correct positive input axis (degrees). Brackets give pointwise 95% intervals from 2,000 paired scene-bootstrap draws, conditional on the selected training states. Negative ReFrame-minus-native changes indicate improved alignment.

Model	Native	ReFrame	Paired change
Falcon3-10B	14.1 [8.3, 20.4]	4.9 [2.3, 8.4]	-9.2 [-13.6, -4.7]
Gemma-2-9B	6.4 [2.4, 11.1]	2.0 [0.8, 5.3]	-4.4 [-7.9, -0.7]
Granite-3.3-8B	9.1 [3.7, 15.1]	1.3 [0.9, 3.7]	-7.8 [-13.0, -0.9]
Llama-3.1-8B	0.9 [0.1, 9.3]	1.0 [0.5, 3.9]	+0.1 [-7.8, +2.3]
Phi-4	0.3 [0.0, 3.4]	3.6 [1.9, 5.9]	+3.3 [-0.7, +4.8]
Qwen2.5-7B	41.1 [31.5, 51.0]	3.7 [2.4, 6.9]	-37.4 [-45.6, -28.2]
Qwen3-8B	7.5 [3.1, 12.6]	1.8 [1.1, 3.7]	-5.6 [-10.4, -0.6]
Qwen3.5-4B	3.3 [0.4, 6.8]	3.7 [2.1, 5.8]	+0.4 [-2.6, +2.9]
Qwen3.5-9B	3.0 [0.2, 6.6]	1.8 [0.5, 3.9]	-1.1 [-5.6, +3.1]
Yi-1.5-9B	4.4 [0.3, 10.2]	3.6 [1.4, 7.3]	-0.8 [-4.7, +2.7]

We retain signed relevant and irrelevant gains, response norms, and mean absolute scene-level irrelevant gains. The last quantity detects interference that cancels in an aggregate signed coefficient. Likewise, a smaller angular deviation can reflect a larger relevant gain without reduced absolute interference; it is not by itself evidence of suppression. Tables 18 and 19 give the angle estimates, and Table 20 reports text-order influence magnitudes.

We use 2,000 paired bootstrap samples of scenes. A draw retains all queries, factorial cells, option orientations, wordings, and model states from each sampled scene. The native scale is recomputed in each draw and shared by the paired trained observations. Intervals are pointwise, unadjusted for multiplicity, and conditional on the three trained adapter states; they do not estimate uncertainty over a population of training seeds.

Held-out capability checks. We additionally evaluate facts, arithmetic, and logic on held-out examples (Table 21). All ten models match or exceed native accuracy in each tested domain when averaged over the three training seeds.

Table 20: Text-order query influence magnitudes on original questions, in common native-margin standard-deviation units. The final pair averages absolute scene-level contrasts and therefore does not hide cancellation between scenes.

	Relevant gain		Signed irrelevant gain		Mean |irrelevant gain|	
Model	Native	ReFrame	Native	ReFrame	Native	ReFrame
Falcon3-10B	0.52	1.30	0.50	0.00	0.50	0.04
Gemma-2-9B	0.39	1.94	0.62	0.00	0.66	0.02
Granite-3.3-8B	0.40	2.82	0.51	-0.03	0.54	0.07
Llama-3.1-8B	0.44	11.89	0.59	0.19	0.59	0.47
Phi-4	0.33	1.36	0.53	0.01	0.53	0.04
Qwen2.5-7B	0.63	0.77	0.36	-0.01	0.38	0.03
Qwen3-8B	0.44	0.89	0.69	-0.01	0.70	0.03
Qwen3.5-4B	0.59	4.12	0.58	-0.04	0.58	0.07
Qwen3.5-9B	0.62	4.19	0.29	0.05	0.30	0.06
Yi-1.5-9B	0.32	2.22	0.61	0.06	0.61	0.07

Table 21: Held-out capability checks (%; 64 cases per domain, two option orientations per case). These small constructed A/B tasks assess limited side effects; they are not a comprehensive capability benchmark. ReFrame averages three seeds.

	Facts		Arithmetic		Logic	
Model	Native	ReFrame	Native	ReFrame	Native	ReFrame
Falcon3-10B	100.0	100.0	56.2	66.9	100.0	100.0
Gemma-2-9B	100.0	100.0	55.5	99.7	100.0	100.0
Granite-3.3-8B	100.0	100.0	62.5	91.1	85.9	99.7
Llama-3.1-8B	100.0	100.0	52.3	53.1	100.0	100.0
Phi-4	100.0	100.0	68.8	88.5	100.0	100.0
Qwen2.5-7B	100.0	100.0	60.2	95.6	68.8	98.4
Qwen3-8B	100.0	100.0	60.9	95.1	99.2	99.7
Qwen3.5-4B	100.0	100.0	61.7	99.5	100.0	100.0
Qwen3.5-9B	100.0	100.0	60.9	100.0	100.0	100.0
Yi-1.5-9B	100.0	100.0	67.2	74.0	78.1	97.7