From Hidden-State Decoding to Selective Temporal Repair

Anonymous authors
Paper under double-blind review

Abstract

When can temporal information recovered from a language model's hidden states improve its answers? We study Qwen2.5-3B, Llama-3.2-3B, and Gemma-2-2B through four interfaces: local decoding, frozen transfer, graph composition, and selective replacement. On 1,082 held-out OmniTemp event pairs, frozen readouts exceed a lexical control fitted on the same labels by 12.07–16.58 Macro-F1 points, with positive document-bootstrap intervals. A separate coupled mechanism combines per-model adapters, a shared event-timeline decoder, source-fitted gates, and cross-model agreement. On 706 TDG edges, it improves all three native outputs by 3.85–11.14 points, with positive gain intervals and more helpful than harmful edits. The broader evaluation identifies the scope of these gains. Local readouts fail the predeclared strict-order criterion; only Gemma retains a positive lexical-control margin under MATRES transfer; and readout-driven graph solving reaches 31.7–33.3% accuracy against a 100% gold-edge oracle. TDG's gains coexist with low overlap recall, while later, separately fitted variants fail joint transfer criteria on MEANTIME and AQUAINT. These results establish benefits from a shared selective mechanism on the TDG population and motivate evaluating coverage, class recall, and harmful edits alongside decoding accuracy.

1 Introduction

A hidden-state readout can recover a temporal relation that a language model answers incorrectly. The practical question is what to do with that signal. Deploying every decoded relation can introduce errors under a new annotation distribution, make a graph inconsistent, or overwrite a correct native answer. Temporal repair therefore requires evidence about the interface that uses a representation, as well as evidence that the representation is informative.

Temporal reasoning makes this problem concrete. Local decisions classify an event pair as before, after, equal, or vague. A downstream graph query asks whether an event has zero, one, or multiple feasible positions among consistent orderings. A readout may improve average pairwise classification while omitting a decisive edge or adding a contradictory one. For selective replacement, the corresponding failure is an intervention that changes a correct answer into an incorrect one.

We examine two complementary systems across four evaluation interfaces in three frozen 2–3B language models. Per-model logistic readouts test whether hidden states support local classification, frozen cross-corpus transfer, and graph composition. A separate coupled mechanism tests selective replacement: model-specific adapters feed a shared event-timeline decoder, and source-fitted gates combine its candidates with cross-model agreement. Rejected edits retain the native answer. The shared decoder couples the learned relation representation; the acceptance rules determine which proposed changes reach the output.

The experiments reveal useful signal and a restricted opportunity to act on it. All three readouts outperform an equal-fit-label lexical control on held-out OmniTemp documents. The coupled mechanism also improves all three outputs on the TDG three-class stress test, with positive document-bootstrap gain intervals and more helpful than harmful edits. These gains coexist with low overlap recall and nonzero harm. Frozen readout transfer and graph composition are less successful, and later, separately fitted repair variants fail joint criteria on MEANTIME and AQUAINT. Because the

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.

systems and target populations differ, these tests characterize their respective deployment outcomes rather than isolate a single cause of transfer failure.

We contribute (1) a coupled selective-repair mechanism that produces simultaneous net gains for three model families on TDG; (2) a four-interface empirical evaluation linking local decoding to transfer, graph answers, and the benefit and harm of replacement; and (3) a reporting contract that makes fitting labels, evaluation units, class recall, coverage, and harmful edits explicit. These experiments were frozen separately and organized retrospectively around the deployment question. Their original endpoints and the later adaptive source-only search remain part of the reported evidence.

2 Background and research questions

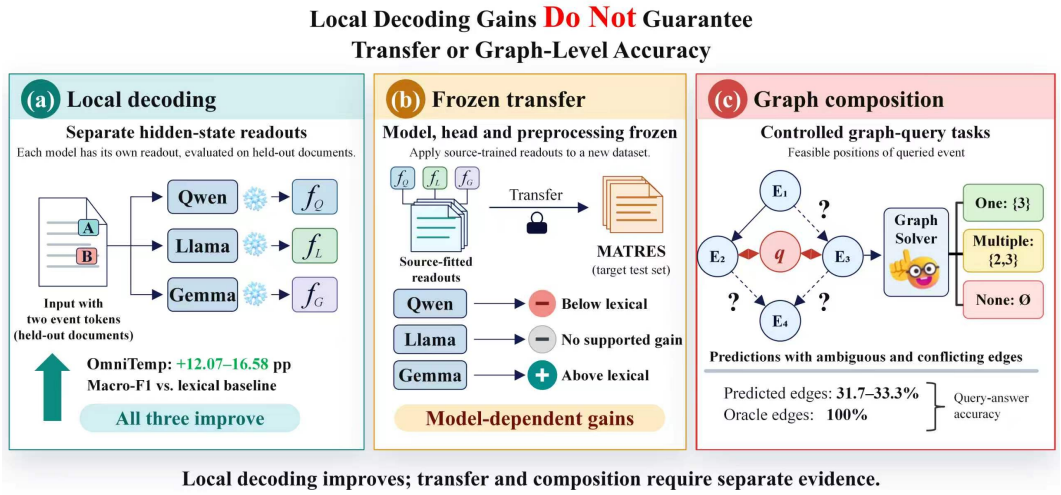


Figure 1: Three interfaces for a frozen temporal readout. (a) Readouts improve local four-class decoding over a lexical baseline fitted on the same labels. (b) Frozen transfer to MATRES yields model-dependent outcomes. (c) Composing a subset of MATRES contexts into graph queries falls far below a gold-edge oracle. The panels use different evaluation units and metrics; (b) and (c) share underlying evidence and are not independent replications.

Temporal relations. Temporal relation extraction connects linguistic evidence to event ordering. MATRES emphasizes a multi-axis annotation scheme and event start-point relations (Ning et al., 2018). OmniTemp provides complete annotation of targeted event pairs and accompanies a global temporal graph generation approach (Eirew et al., 2025). Shared label names can have different annotation distributions and semantics, especially for simultaneity and unresolved relations. We use official annotations and state each cross-corpus label mapping explicitly.

Probing versus use. Supervised probes measure what a specified decoder can recover from representations under a specified training protocol. They do not establish that the original model uses that information when answering. Probe control tasks are therefore important for interpreting apparent accessibility (Hewitt & Liang, 2019). Our lexical and rule comparisons address specific shallow alternatives, not every possible shortcut. Likewise, frozen cross-corpus evaluation tests one learned interface, not all information that any future decoder could recover.

Four questions. Does a frozen readout improve local classification beyond native and shallow controls (Q1)? Does its advantage survive frozen transfer (Q2)? Do its relations improve exact graph-level answers (Q3)? Can a coupled selective mechanism improve native outputs while accounting for coverage, class recall, and harm (Q4)? Success at one interface does not waive the next test.

3 Related work

Temporal relations and global structure. TimeML introduced a markup language for events, time expressions, and their links (Pustejovsky et al., 2003), and TempEval-3 operationalized temporal extraction as a shared task over this representation (UzZaman et al., 2013). Sparse corpora motivated dense event-ordering annotation and multi-pass extraction: TimeBank-Dense labels local event links (Cassidy et al., 2014), and a sieve architecture propagates those links into global orderings (Chambers et al., 2014). MATRES instead adopts a multi-axis start-point scheme (Ning et al., 2018), later supplemented by an improved neural baseline (Ning et al., 2019); parallel multilingual corpora such as MEANTIME extend TimeML-style event and time annotation across languages (Minard et al., 2016). Structured temporal extraction goes beyond independent pair classification. Ning et al. (2017) combine pairwise scores with global consistency; Han et al. (2019) combine neural scoring with structured prediction; and Wang et al. (2020) encode temporal and subevent constraints in a joint learning objective. These approaches motivate constraint-aware inference, but do not guarantee improvement when a frozen readout supplies noisy hard constraints. Crowdsourced temporal dependency graphs provide a related but distinct representation in which temporal links form tree-like dependencies (Yao et al., 2020); we use this corpus only for a separately declared transfer stress test. Recent dense timeline annotation re-examines the same TimeBank documents rather than providing an independent text source (Rogers et al., 2024), while OmniTemp studies global zero-shot temporal graph generation (Eirew et al., 2025). TRACIE concerns implicit-event reasoning (Zhou et al., 2021), whereas METRE models genuinely admissible ambiguity (Hu et al., 2024). These differing targets reinforce the need to state label semantics and downstream tasks rather than treating all temporal scores as interchangeable. Large language models have since made zero-shot temporal prompting a concrete baseline, with consistency-focused global inference showing that pairwise gains need not translate into globally valid graphs (Yuan et al., 2023; Eirew et al., 2025). Our controlled graph study asks the complementary question of whether a frozen hidden-state readout can supply edges to an exact solver.

Probing interpretation. Diagnostic classifiers were introduced as post-hoc probes of intermediate layers (Alain & Bengio, 2017); sentence- and token-level probes subsequently showed that apparent encodings depend on the control task and probe complexity (Conneau et al., 2018; Belinkov et al., 2017; Tenney et al., 2019). Control-task probes caution that probe accuracy can reflect easy surface shortcuts rather than task-specific content (Hewitt & Liang, 2019); minimum-description-length probing similarly separates recoverable information from the cost of the decoder (Voita & Titov, 2020), and information-theoretic probing frames the measured quantity as information recoverable by a specified decoder under its training protocol (Pimentel et al., 2020). Both views support our distinction between what a supervised head recovers on held-out documents and what the underlying model natively uses or what a downstream system can safely deploy. A readout that improves over lexical and native controls under document-disjoint fitting is evidence for a usable interface, not a causal claim about the model's internal computation or about transfer. We complement aggregate Macro-F1 with class recall, reverse-order outcomes, and grouped intervals.

Selective deployment and safety. Selective prediction trades coverage for risk, with a risk–coverage characterization of abstention, including selective prediction for deep networks (Geifman & El-Yaniv, 2017; El-Yaniv & Wiener, 2010); domain-shift question answering shows that this trade-off must be calibrated out of distribution rather than assumed from in-domain confidence (Kamath et al., 2020). In our selective-replacement setting, an abstention retains the native answer rather than withholding an answer. We therefore distinguish an edit's benefit from its harm: a previously correct answer may become wrong even when aggregate Macro-F1 improves. Our coupled repair test measures this trade-off and coverage; it does not provide a risk guarantee.

4 Experimental framework

4.1 Models, labels, and frozen readouts

We study Qwen2.5-3B-Instruct, Llama-3.2-3B-Instruct, and Gemma-2-2B-it. Model bodies remain frozen. The natural-relation readout predicts {before, after, equal, vague}. Let $h^{\rightarrow}$ and $h^{\leftarrow}$ be the states at relative layer depth 0.75 for forward- and reverse-direction questions about the same marked

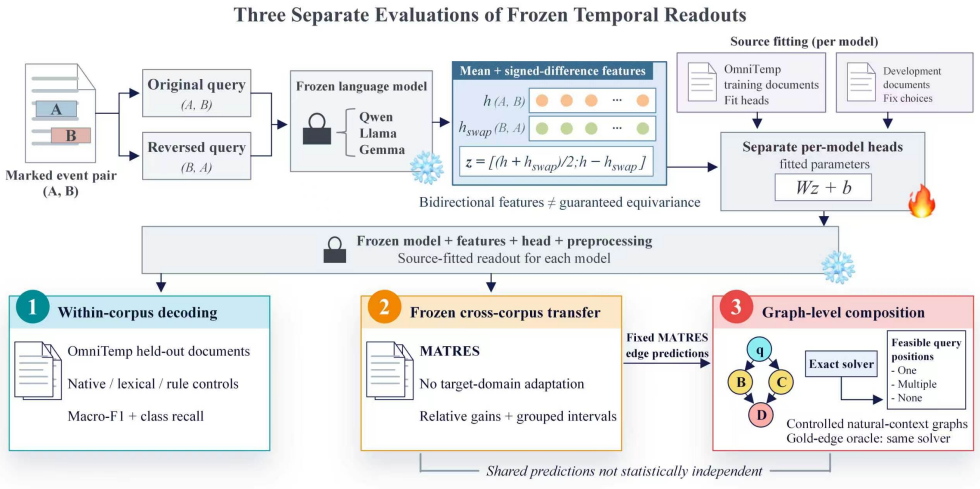


Figure 2: Frozen readout pipeline and its three separately frozen evaluations. A marked event pair is queried in both directions; each frozen model produces a bidirectional mean-plus-signed-difference feature that feeds a per-model fitted head. The same frozen readout is then evaluated on three distinct populations and endpoints—within-corpus decoding on held-out OmniTemp, frozen cross-corpus transfer to MATRES, and graph-level composition—with transfer and composition sharing the same edge predictions and therefore not statistically independent. The later selective-repair system (Section 5.5) is a separate mechanism and is not part of this readout pipeline.

Table 1: Distinct evaluation populations and interfaces. Counts are evaluated items/documents, except graph composition (graphs/matched triplets). The later coupled repair has a different fitting budget and label mapping; its result is not a replication of the first three tests.

Evaluation	Population	Frozen interface	Endpoint
OmniTemp	1,082 / 10	Per-model logistic readout	Four-class F1
MATRES	837 / 20	Same readout, no adaptation	Four-class F1
Composition	60 / 20	Readout + exact solver	Query answer
TDG	706 / 50	Coupled timeline + gate	Three-class F1
MEANTIME	526 / 112	Decision-preserving variant	Three-class F1

document and event pair. Only the first two answer options reverse their event arguments; document markers are unchanged. The frozen representation is

$$z(x, A, B) = \left[\frac{h^{\rightarrow} + h^{\leftarrow}}{2}; \; h^{\rightarrow} - h^{\leftarrow} \right]. \tag{1}$$

A class-balanced multinomial logistic head is fitted separately for each model. The frozen regularization values are $C = 0.01$ for Qwen and Gemma and $C = 0.1$ for Llama. The workflow is shared, but parameters are not. Reversed questions provide representational contrast, not an exact prediction-equivariance constraint.

The fitting set contains 1,887 event pairs from 16 OmniTemp training documents. Four separate development documents (514 pairs) determine regularization and are not added to fitting. The ten held-out documents contain 1,082 pairs: 419 before, 431 after, 60 equal, and 172 vague. This experiment does *not* use a 148-label budget; that budget belongs to later exploratory studies.

4.2 Controls and statistical units

Native decoding uses a canonical four-option next-token decision. The text-order rule predicts before because event A is the first textual mention in this evaluation. A frozen local-connector rule adds limited local linguistic evidence. The lexical control is TF-IDF unigram/bigram logistic regression,

with $C = 0.001$, fitted on the same training examples. Equal fit budget refers to the same labeled examples, not identical feature capacity or compute.

A supplementary four-shot native control uses one source demonstration per class, excluding every query's held-out documents. On the separate 280-pair repair-development set it changes Qwen/Llama/Gemma Macro-F1 by $-7.95/-3.46/+1.79$ points; only the last interval crosses zero. This bounded prompt check does not replace the main-study baseline or test optimized reasoning prompts (Appendix G).

Four-class performance is measured by Macro-F1 $= \frac{1}{4}\sum_c \mathrm{F1}_c$, together with per-class recall. Strict before/after balanced accuracy is a separate endpoint. Intervals are obtained using 20,000 paired bootstrap draws over documents: each sampled document contributes all its event pairs, and both systems are scored on the same draw. We report the archived percentile intervals; these are not simultaneous family-wise guarantees over all comparisons.

The preregistered primary gate requires strict-order gains of at least five percentage points over each shallow control, positive interval lower bounds, reverse-order recall of at least 50%, and a positive lower bound against native. The separate secondary gate requires four-class Macro-F1 gains of at least five points over each shallow control with positive lower bounds, plus at least 20% recall for both equal and vague. Passing it does not override a failed primary gate.

4.3 Cross-corpus and composition tests

Frozen transfer. The readouts, feature construction, prompts, regularization, class weights, and seeds are transferred unchanged to 837 MATRES Platinum pairs from 20 news documents. Official labels comprise 424 before, 269 after, 31 equal, and 113 vague. Event positions are matched against official tuples; six stale preprocessing labels are replaced with official annotations. The recorded overlap audit finds no exact or seven-token-shingle match against the ten OmniTemp held-out documents. This limited check does not establish absence of pretraining contamination.

Controlled composition. We construct 20 matched graph triplets, totaling 60 graphs, from 510 distinct strict-order contexts selected deterministically from MATRES Platinum, independently of model predictions. Thirty graphs have eight nodes and thirty have nine. The answer classes are unique, multiple, and no valid ordering, with 20 examples each. These are controlled compositions of natural passages, *not* intact document graphs or an independent new-corpus replication of transfer. New graphs are checked against 3,780 earlier graphs under raw and rooted transitive-closure isomorphism. Frozen MATRES relation predictions supply the edges without retraining; native graph answers are obtained directly, without the solver.

For a predicted graph $\hat{G}$ and query event q, the solver enumerates the feasible positions of q across valid orderings, and determines whether this set has cardinality zero, one, or greater than one. A unique query position does not imply a unique complete ordering. Vague predictions omit a constraint. Under the distinct-position semantics of this construction, equal predictions introduce mutual precedence and hence inconsistency. This is an explicit task convention, not a general definition of temporal equality. The oracle supplies gold edges to the same solver. Uncertainty is grouped by matched triplet, not by individual edge.

5 Results

5.1 Useful local information is recoverable

Table 2 gives a clear affirmative answer to Q1 at the level of four-class Macro-F1. Each model exceeds the strongest shallow control, with positive document-bootstrap intervals. Text order and the local rule achieve only 13.96% and 14.18%, respectively. The result therefore survives these specific tests of mention-order and shallow lexical explanations. It does not establish freedom from all lexical or dataset-specific cues.

The gain over native decoding is also substantial: 15.87 points for Qwen, 22.26 for Llama, and 30.56 for Gemma (differences computed before rounding). Native output performance therefore understates the information accessible to this supervised readout. However, the readout receives

Table 2: OmniTemp four-class Macro-F1 (%) on ten held-out documents. The lexical head and hidden-state heads use the same fitting labels; the two rules require no fitted head. Intervals are 95% paired document-bootstrap CIs for gain over lexical. All exclude zero, but no readout passes the strict-order primary gate.

Group	Method	Macro-F1	95% CI of gain
Shallow controls	Text order	13.96	—
	Local rule	14.18	—
	Lexical (equal budget)	30.32	—
Native decoding	Qwen2.5-3B	26.52	—
	Llama-3.2-3B	20.33	—
	Gemma-2-2B	16.33	—
Frozen readout	Qwen2.5-3B	42.39	[+2.41,+22.57]
	Llama-3.2-3B	42.60	[+5.74,+19.99]
	Gemma-2-2B	46.89	[+9.26,+27.74]

Table 3: OmniTemp's positive average scores coexist with incomplete class recovery. Strict BA is before/after balanced accuracy; reverse recall concerns after against textual mention order. Values are percentages.

Model	Strict BA	Reverse	Equal	Vague	Primary	Secondary
Qwen	54.14	44.32	43.33	27.33	Fail	Pass
Llama	56.61	46.87	11.67	48.26	Fail	Fail
Gemma	58.26	48.49	33.33	36.63	Fail	Pass

task labels and learns a new decision interface, so these comparisons are not evidence that native inference already performs the same computation internally. The corresponding gain-with-interval plot is provided in Figure 6 (Appendix).

The preregistered primary claim nevertheless fails for all three models (Table 3). Reverse-order recall remains below 50%, and strict-order comparisons against the literal rules do not establish all required gains. Llama also fails the secondary gate because equal recall is 11.67%. Thus, the positive finding is shared aggregate improvement on held-out documents, not complete or uniformly safe repair.

5.2 Frozen transfer yields different outcomes across models

The same interface no longer produces a shared advantage on MATRES (Table 4). Qwen falls below both native and lexical decoding. Llama also falls below native decoding and does not establish an advantage over the lexical control. Gemma retains a positive lexical-control margin, but its class profile remains uneven.

In before, after, equal, and vague order, the recall vectors are (0.00, 6.32, 9.68, 80.53) for Qwen, (14.15, 36.43, 61.29, 0.00) for Llama, and (68.87, 18.22, 22.58, 7.08) for Gemma. These are not interchangeable failure modes: one readout largely loses concrete-order recall, another fails to recover vague, and the strongest aggregate readout still has very low vague recall (Appendix Figure 7). All three fail the full class-protection requirements.

This result establishes a limitation of the *frozen trained readout*, not the absence of transferable information in every hidden layer. Changes in corpus composition, annotation, language patterns, and calibration are plausible contributors. The present experiment does not isolate their causal effects. Its operational conclusion is narrower and sufficient: source-corpus probe gains are not a reliable certificate for deploying the unchanged readout on this target corpus.

5.3 The decoded relations do not improve graph answers

The graph test answers Q3 negatively for the evaluated pipeline. Readout-plus-solver accuracy is 31.7–33.3%, compared with 36.7% for each of the shallow graph controls. No model establishes the required positive grouped interval, and all fail the frozen 60% accuracy and 50% worst-class-recall

Table 4: MATRES frozen transfer, four-class Macro-F1 (%) on the same frozen split. No target-corpus adaptation is performed. Readout intervals are 95% paired document-bootstrap CIs of the gain over the equal-budget lexical control (18.23%). Only Gemma keeps a positive margin.

Group	Method	Macro-F1	95% CI of gain
Shallow control: lexical (equal budget)		18.23	—
Native decoding	Qwen2.5-3B	16.98	—
	Llama-3.2-3B	23.03	—
	Gemma-2-2B	19.02	—
Frozen readout	Qwen2.5-3B	9.50	[−11.72,−5.87]
	Llama-3.2-3B	16.99	[−5.20,+2.10]
	Gemma-2-2B	27.14	[+3.50,+13.48]

Table 5: Controlled natural-context graph composition (60 graphs). Native denotes direct graph answers; readout and lexical predictions feed the solver. No evaluated non-oracle arm exceeds 36.7% accuracy, whereas gold edges yield 100%. Strict coverage / orientation are edge-level diagnostics.

Model	Native	Readout+solver	Lexical+solver	Gold-edge oracle	Strict cov.	Strict orient.
Qwen	33.3	31.7	36.7	100.0	4.9	44.0
Llama	33.3	33.3	36.7	100.0	40.6	53.6
Gemma	31.7	31.7	36.7	100.0	68.8	72.6

gates. Graph-answer recalls for unique/multiple/none are 0/25/70 for Qwen, 0/0/100 for Llama, and 0/0/95 for Gemma. We stress that these numbers inherit the construction's label semantics: vague pairs omit a constraint, while equal is converted to mutual precedence and hence inconsistency. This is an explicit task convention for the distinct-position query, not a general definition of temporal equality, and it differs from treatments that model vague as multiple simultaneously admissible relations (Hu et al., 2024). Results should therefore be read as composition behavior under that declared conversion.

Exact predicted-graph recovery is zero for every model despite differing strict-edge coverage, while the gold-edge oracle reaches 100% answer accuracy. Appendix E examines omission and addition errors and shows the corresponding figure.

5.4 Retrospective robustness and error attribution

Deleting each OmniTemp document in turn preserves every model's gain over native and lexical decoding; the MATRES point-estimate pattern also survives each deletion. In the graph experiment, isolated oracle corrections implicate vague omissions for Qwen and equal-to-mutual-precedence conversions for Llama/Gemma. These retrospective diagnostics do not define a deployable repair or independent confirmation. Appendix J gives all sensitivity ranges and isolated interventions.

5.5 Selective coupling improves all three TDG outputs

The preceding tests use independent logistic readouts. For Q4, we instead couple the models through a shared event-timeline decoder. Frozen event-token features at approximately two-thirds depth feed per-model adapters and a common decoder of event center, width, and uncertainty. Its candidate relation is used selectively, not substituted unconditionally. If n_m is model m's native label, c_m its candidate, and g_m its source-fitted acceptance decision, the output is

$$\hat{y}_m = \begin{cases} c_m, & g_m = 1 \ \wedge \ \exists j \neq m : \ c_j = c_m, \\ n_m, & \text{otherwise.} \end{cases} \tag{2}$$

The common decoder and agreement rule couple the outputs; adapters and gates remain model-specific. Agreement is a deployment condition, not a guarantee that the common answer is correct. The head and gate each use 148 source labels with 16 shared IDs—280 distinct fitting labels, excluding earlier exploratory access. This design is distinct from the logistic readout and exact graph solver.

Table 6: Two separately frozen coupled systems on two target populations. Three-class Macro-F1 (%), gain in points, and 95% document-bootstrap interval. Help/harm count wrong-to-correct / correct-to-wrong edits; coverage counts changed answers. Rows across blocks are not a same-system cross-corpus ablation.

Model	Native	Repair	Gain [95% CI]	Help/harm	Coverage
TDG: original coupled mechanism, 706 pairs / 50 documents					
Qwen	21.29	32.43	+11.14 [7.46,15.20]	82/40	31.73%
Llama	20.63	27.32	+6.70 [3.46,9.56]	48/22	15.44%
Gemma	14.37	18.21	+3.85 [0.98,5.97]	32/2	14.02%
MEANTIME: retained V15 variant, 526 pairs / 112 documents					
Qwen	28.68	32.83	+4.15 [1.12,7.45]	31/15	12.36%
Llama	26.29	26.50	+0.21 [−0.88,1.30]	2/8	5.51%
Gemma	14.74	26.60	+11.86 [8.66,15.18]	47/8	15.40%

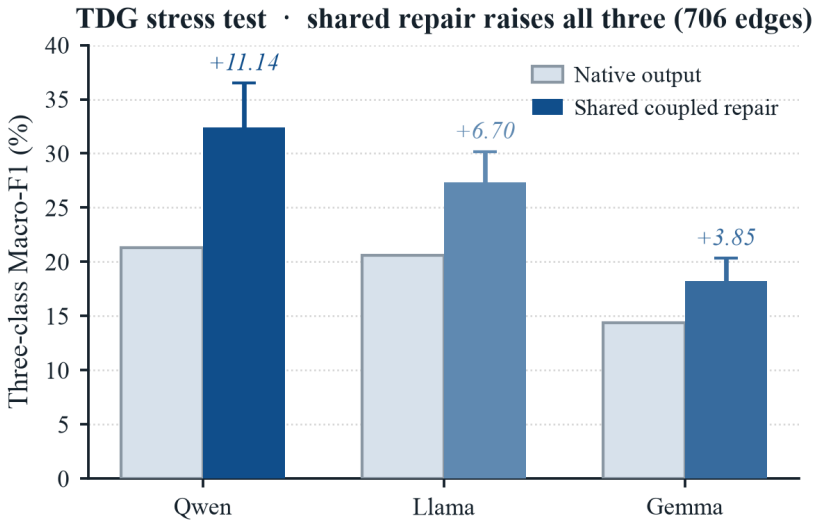


Figure 3: The coupled mechanism improves all three native outputs on 706 TDG edges. Bars show three-class Macro-F1; whiskers give 95% document-bootstrap gain intervals (all lower bounds >0). The supported gold classes are before, after, and overlap. Low overlap recall and nonzero harm limit the claim to aggregate benefit on this population (Table 6).

The frozen stress test covers all 706 explicit event–event edges from 50 TDG test documents (Yao et al., 2020). Its gold classes are before, after, and overlap; overlap is operationally mapped to equal, although interval overlap does not imply equal start times. Vague outputs remain errors. No TDG labels select fitted parameters, and all predictions are sealed before scoring. Three-class Macro-F1 and paired document-bootstrap intervals (10,000 draws) are reported in Table 6; full protocol and limitations appear in Appendix A.

Every TDG output passes that test's original net-benefit criteria: positive gain and interval lower bound, help exceeding harm, coverage of at least 10%, positive recall in every supported class, and no single document dominating positive document net gains. Nonetheless, overlap recall is only 6.01%, 6.01%, and 4.50%; Gemma's after recall is 4.29%. Harm per edited edge is 17.86%, 20.18%, and 2.02%. The coupled outputs support net transfer improvement, not three independent safe policies, recovery of the absent vague class, or reliable graph composition. Later source-only attempts to satisfy stricter four-class criteria are disclosed in Appendix B, without inheriting the TDG result.

A new frozen test of the decision-preserving variant. V15 augments candidate fitting with a zero-margin penalty when an originally correct native decision loses its top rank. It is retained after source development despite one unmet recall floor, then refitted and frozen before MEANTIME

inference. All target predictions are sealed before scoring; no target labels select the variant or thresholds. On 526 compatible explicit pairs, Qwen and Gemma gain, whereas Llama's interval crosses zero and its harmful edits exceed helpful ones (Table 6). Equal recall is only 4.79/1.06/3.19%; the joint target criterion fails. MEANTIME's compatible population has no vague gold labels. This result bounds the retained variant, not the earlier TDG system; neither average source gains nor the 9.72% source vague recall established transfer stability (Appendix C).

6 Implications for selective repair

The deployment question connects the four interfaces. Local decoding establishes that a trained head can recover useful relations. Transfer tests whether that head remains useful on a new population; composition tests the effect of its errors on a different answer space. Selective replacement introduces a further decision: whether a proposed change improves on the native answer. The TDG result shows that a shared mechanism can make beneficial edits across three model families, while its low overlap recall shows how aggregate improvement can leave a relation class largely unrecovered.

The later variants clarify the remaining challenge. MEANTIME does not establish a Llama gain, and the V18-derived AQUAINT deployment also fails joint transfer (Appendix D). These outcomes concern the interaction among the candidate, the native answer, and the acceptance rule. They do not identify an intrinsic defect in one model family. A useful repair evaluation must expose that interaction through coverage, class recall, and harmful edits, in addition to average accuracy.

7 Limitations and reproducibility

The local and transfer studies have only ten and twenty held-out documents; the graph task is a constructed diagnostic. The native interface uses one fixed option verbalizer, not an optimized chain-of-thought or self-consistency baseline; lexical controls are not randomized-label selectivity tests. Only 2–3B models are studied. The framing is retrospective, and repeated source-label inspection makes development intervals unadjusted for adaptive search. TDG and MEANTIME lack vague gold and do not prove four-class safety; a later V18-derived AQUAINT test also fails joint transfer (Appendix D). Hashes and audits establish computational traceability, not scientific certification or a completed public release. Appendix F details these limits.

8 Conclusion

Temporal information in frozen hidden states can support both stronger local decoding and beneficial selective edits. Our shared timeline mechanism improves all three outputs on TDG, while the transfer and composition tests show why these gains require a precise deployment scope. Evaluating the proposed edit, its effect on the native answer, and its behavior across classes is essential to turning readable information into reliable repair.

AI use statement

Generative AI tools were used to propose and refine hypotheses and experimental designs, implement and debug methods, construct controlled graph instances, preprocess data, orchestrate experiments, and analyze and interpret results. They also assisted literature and dataset search, scientific-figure preparation, manuscript drafting and revision, reference formatting, and readability checks. Corpus relation labels were retained rather than relabeled to improve results; controlled graph-answer labels were computed from the declared graph semantics. Recorded protocols, prediction hashes, and rerun diagnostics support computational checks but do not replace author review or independent scientific validation. The authors retain responsibility for the final text, claims, code, and artifacts.

References

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *International Conference on Learning Representations*, 2017. URL `https://openreview.net/forum?id=HJ4-rAVtl`.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. What do neural machine translation models learn about morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 861–872, 2017. doi: 10.18653/v1/P17-1080. URL https://aclanthology.org/P17-1080/.

Taylor Cassidy, Bill McDowell, Nathanael Chambers, and Steven Bethard. An annotation framework for dense event ordering. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 501–506, 2014. doi: 10.3115/v1/P14-2082. URL https://aclanthology.org/P14-2082/.

Nathanael Chambers, Taylor Cassidy, Bill McDowell, and Steven Bethard. Dense event ordering with a multi-pass architecture. *Transactions of the Association for Computational Linguistics*, 2: 273–284, 2014. doi: 10.1162/tacl_a_00182. URL https://aclanthology.org/Q14-1022/.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. What you can cram into a single $&!#*$ vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2126–2136, 2018. doi: 10.18653/v1/P18-1198. URL https://aclanthology.org/P18-1198/.

Alon Eirew, Kfir Bar, and Ido Dagan. Beyond pairwise: Global zero-shot temporal graph generation. In *Proceedings of EMNLP*, pp. 31440–31458, 2025. doi: 10.18653/v1/2025.emnlp-main.1601. URL https://aclanthology.org/2025.emnlp-main.1601/.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(53):1605–1641, 2010. URL https://www.jmlr.org/papers/v11/el-yaniv10a.html.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, pp. 4878–4887, 2017. URL https://arxiv.org/abs/1705.08500.

Rujun Han, I-Hung Hsu, Mu Yang, Aram Galstyan, Ralph Weischedel, and Nanyun Peng. Deep structured neural network for event temporal relation extraction. In *Proceedings of CoNLL*, 2019. doi: 10.18653/v1/K19-1062. URL https://aclanthology.org/K19-1062/.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of EMNLP-IJCNLP*, pp. 2733–2743, 2019. URL https://aclanthology.org/D19-1275/.

Yutong Hu, Quzhe Huang, and Yansong Feng. Only one relation possible? modeling the ambiguity in temporal relation extraction. In *Proceedings of NLPCC 2024*, pp. 444–456, 2024. doi: 10.1007/978-981-97-9431-7_34. URL https://doi.org/10.1007/978-981-97-9431-7_34.

Amita Kamath, Robin Jia, and Percy Liang. Selective question answering under domain shift. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5684–5696, 2020. doi: 10.18653/v1/2020.acl-main.503. URL https://aclanthology.org/2020.acl-main.503/.

Anne-Lyse Minard, Manuela Speranza, Ruben Urizar, Begoña Altuna, Marieke van Erp, Anneleen Schoen, and Chantal van Son. MEANTIME, the NewsReader multilingual event and time corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 4417–4422, 2016. URL https://aclanthology.org/L16-1699/.

Qiang Ning, Zhili Feng, and Dan Roth. A structured learning approach to temporal relation extraction. In *Proceedings of EMNLP*, pp. 1027–1037, 2017. doi: 10.18653/v1/D17-1108. URL https://aclanthology.org/D17-1108/.

Qiang Ning, Hao Wu, and Dan Roth. A multi-axis annotation scheme for event temporal relations. In *Proceedings of ACL*, pp. 1318–1328, 2018. doi: 10.18653/v1/P18-1122. URL https://aclanthology.org/P18-1122/.

Qiang Ning, Sanjay Subramanian, and Dan Roth. An improved neural baseline for temporal relation extraction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 6203–6209, 2019. doi: 10.18653/v1/D19-1642. URL https://aclanthology.org/D19-1642/.

Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. Information-theoretic probing for linguistic structure. In *Proceedings of ACL*, pp. 4609–4622, 2020. URL https://aclanthology.org/2020.acl-main.420/.

James Pustejovsky, José Castaño, Robert Ingria, Roser Saurí, Robert Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir Radev. Timeml: Robust specification of event and temporal expressions in text. In *New Directions in Question Answering*, pp. 28–34, 2003. URL https://dl.acm.org/doi/10.5555/1225753.1225774.

Anna Rogers, Marzena Karpinska, Ankita Gupta, Vladislav Lialin, Gregory Smelkov, and Anna Rumshisky. NarrativeTime: Dense temporal annotation on a timeline. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 12053–12073, 2024. URL https://aclanthology.org/2024.lrec-main.1054/.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4593–4601, 2019. doi: 10.18653/v1/P19-1452. URL https://aclanthology.org/P19-1452/.

Naushad UzZaman, Hector Llorens, Leon Derczynski, James Allen, Marc Verhagen, and James Pustejovsky. SemEval-2013 task 1: TempEval-3: Evaluating time expressions, events, and temporal relations. In *Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, volume 2, pp. 1–9, 2013. URL https://aclanthology.org/S13-2001/.

Elena Voita and Ivan Titov. Information-theoretic probing with minimum description length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 183–196, 2020. doi: 10.18653/v1/2020.emnlp-main.14. URL https://aclanthology.org/2020.emnlp-main.14/.

Haoyu Wang, Muhao Chen, Hongming Zhang, and Dan Roth. Joint constrained learning for event-event relation extraction. In *Proceedings of EMNLP*, pp. 696–706, 2020. doi: 10.18653/v1/2020.emnlp-main.51. URL https://aclanthology.org/2020.emnlp-main.51/.

Jiarui Yao, Haoling Qiu, Bonan Min, and Nianwen Xue. Annotating temporal dependency graphs via crowdsourcing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5368–5380, 2020. doi: 10.18653/v1/2020.emnlp-main.432. URL https://aclanthology.org/2020.emnlp-main.432/.

Chenhan Yuan, Qianqian Xie, and Sophia Ananiadou. Zero-shot temporal relation extraction with ChatGPT. In *Proceedings of the 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pp. 92–102, 2023. doi: 10.18653/v1/2023.bionlp-1.7. URL https://aclanthology.org/2023.bionlp-1.7/.

Ben Zhou, Kyle Richardson, Qiang Ning, Tushar Khot, Ashish Sabharwal, and Dan Roth. Temporal reasoning on implicit events from distant supervision. In *Proceedings of NAACL-HLT*, pp. 1361–1371, 2021. doi: 10.18653/v1/2021.naacl-main.107. URL https://aclanthology.org/2021.naacl-main.107/.