BOUNDARY-RESOLVED ANSWER RELAYS IN LATENT REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

How does answer information pass from continuous latent reasoning to discrete generation? We test whether the contextual cache of a naturally generated answer prefix relays information from native latent states. We introduce **Boundary-Resolved Answer Transport and Sensitivity** (**BATS**), a protocol that crosses latent and prefix caches, recomputes the terminal reader, and measures donor-answer transfer separately from destructive sensitivity. In CODI–GPT-2, native latent values repeatedly write answer information into the contextual values of "The answer is," which terminal attention heads combine to produce an answer. On 64 new arithmetic pairs, the fixed setup yields 61 donor-answer transfers and 64 recipient recoveries under unrestricted greedy generation. All three prefix positions are required by the native-patching criterion, but redistributing their mean value difference at matched readout dose enables single-position control. In two prospectively selected Llama-based checkpoints, the same bidirectional test supports relay sufficiency in SIM-CoT, whereas Zen-E falls below the sufficiency threshold. A separate fresh-input study shows that Zen-E's patched prefix can recover the original donor answer with latent support from different-answer problems, although latent content still influences the output. Complementary boundary studies show persistent latent-answer transport and transport–sensitivity dissociation. Together, these findings support an operator-defined latent-to-prefix relay and show why answer-transport claims must specify the intervention, upstream state, and decoding boundary.

1 INTRODUCTION

Two reasoning trajectories can generate the identical prefix "The answer is:" and continue with different answers. Coconut (Hao et al., 2024) and CODI (Shen et al., 2025) perform intermediate reasoning through continuous hidden states before returning to token generation. Even fixed prefix words acquire contextual cache states that depend on this preceding computation. Understanding the transition therefore requires tracing how latent information reaches the answer through these generated states. We ask whether the prefix acts as an answer relay, which positions participate, and what upstream support its readout requires.

Research gap. Interchange intervention, causal tracing, and path patching test how changes to internal states affect model outputs (Geiger et al., 2022; Vig et al., 2020; Meng et al., 2022; Goldowsky-Dill et al., 2023). SCIT (Ding et al., 2026) uses donor interchange, K/V component tests, semantic controls, and corruption to identify which cache segments affect the answer. Other studies examine the effects of individual latent steps and when prompt caches remain useful (Li et al., 2026; Kumar et al., 2026). Separate changes to latent states and prefix states can produce the same donor answer. This alone does not show whether the model uses them independently or whether the latent states affect the answer by changing the prefix states. Producing a donor answer after replacing states also differs from losing the original answer after corruption. Both tests depend on how the states are changed, how outputs are measured, and which other states are present (Zhang & Nanda, 2024; Vaidyanathan et al., 2026).

Proposed method. We introduce Boundary-Resolved Answer Transport and Sensitivity (BATS), a protocol that crosses clean and donor-latent-patched recipient states at naturally reached answer boundaries (Figure 1b). Both parent paths generate the same visible prefix; its contextual cache records the clean or patched computation. We exchange these prefix caches across latent backgrounds

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.

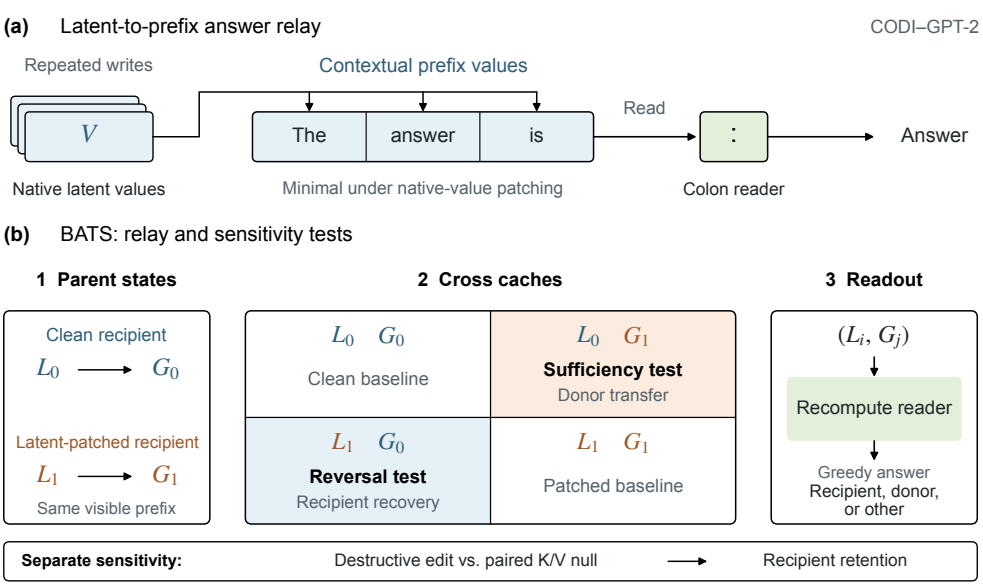


Figure 1: **A latent-to-prefix relay and our BATS protocol.** (a) Native-latent values repeatedly write contextual prefix values for the terminal colon reader in CODI–GPT-2. Arrows summarize the tested relation; coalition minimality is relative to native-value patching. (b) We cross clean and donor-latent-patched recipient states to test forward donor transfer and reverse recipient recovery, recomputing the reader in every cell. The lower strip separately measures destructive sensitivity through recipient retention and paired nulls. Blue/orange indicate clean/donor-induced states; green marks the reader. The recipient prompt stays fixed in the four-cell relay tests. Boundaries must be naturally reached and tokenizer-aligned.

and recompute the terminal reader in all four combinations. Forward transfer tests prefix sufficiency, while reverse restoration tests whether the clean prefix recovers the recipient answer despite the latent patch. We measure unrestricted greedy answers and separately evaluate destructive retention against identity replay and paired algebraic nulls.

In CODI–GPT-2, we find that latent values repeatedly affect the prefix states used by the terminal reader (Figure 1a). Across two prospectively selected Llama-based checkpoints, SIM-CoT (Wei et al., 2026) passes the relay test, whereas Zen-E's prefix falls below the transfer threshold when the original latent states are kept.

Contributions. Our contributions are threefold:

- We introduce BATS, a four-condition test that checks whether a changed prefix transfers a donor answer and whether restoring the prefix recovers the original answer. Testing both original and patched latent states shows how these effects depend on the latent states. Separate corruption tests measure answer loss.
- We show that latent values repeatedly affect prefix states, which fixed attention heads then use to generate the answer. On 64 new arithmetic pairs, the same heads and intervention rules give 61/64 donor transfers and 64/64 recipient recoveries under greedy generation. Native-value patches require three prefix positions, while using their mean value difference at matched dose allows single-position control.
- We show that useful latent support does not require a matching final answer. On fresh inputs, latent states from problems whose answers differ from the original donor's raise Zen-E's donor-answer recovery from 17/64 to 42/64 in travel and from 20/64 to 43/64 in production.

2 TESTING ANSWER TRANSPORT, RELAYS, AND DESTRUCTIVE SENSITIVITY

2.1 SEGMENTS AND ANSWER BOUNDARIES

For each model we define four non-overlapping segment classes: (i) the *raw prompt*, containing task tokens only; (ii) BOT, the shared delimiter or beginning-of-thought position; (iii) the *native latent tail*, containing model-generated latent states before answer decoding; and (iv) the *generated prefix*, containing contextual cache states created while processing answer-prefix tokens. The distinction is semantic as well as positional: generated-prefix values belong to neither the raw prompt nor the latent tail, and BOT is treated as a separate segment.

We index tokenizer-specific boundaries t_j after replaying exactly j specified prefix tokens. For GPT-2, "The answer is:" contains four specified tokens. We define t_0 immediately before "The" and t_1, t_2, t_3, t_4 after processing each successive prefix token. For the Llama study, the specified prefix is "The answer is:" followed by a trailing-space token, so terminal t_5 follows five tokens. Source and recipient boundaries are located independently after tokenization and replayed with identity checks; equal character strings are never assumed to imply equal cache widths.

2.2 SEGMENT-LEVEL CAUSAL INTERCHANGE

Let R be a clean recipient example, S a clean source (donor) example with a different target answer, and X_u a segment at intervention boundary u. The interchange intervention

$$R[X_u \leftarrow S] \tag{1}$$

replaces only the recipient's prespecified segment with the corresponding source segment, preserving all recipient positions outside X_u. The donor-sufficiency rate is

$$\mathrm{Transfer}(X, u) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}[y(R_i[X_u \leftarrow S_i]) = y(S_i)] \,. \tag{2}$$

Here, y denotes the parsed generated answer, and N is the number of qualified recipient–donor pairs in the evaluated panel. We require distinct-answer pairs for the headline statistic. Same-answer, partial-variable, random-source, and wrong-location controls test whether an intervention transports donor semantics rather than merely perturbing the recipient or inserting a generic answer direction.

We measure transport by unrestricted greedy generation on valid, parseable clean examples. We report sequence-level candidate rankings separately and never substitute them for greedy counts.

2.3 TESTING THE RELAY WITH BIDIRECTIONAL OVERWRITES

We test a prespecified set of jointly patched prefix positions (a coalition) with four $L \times G$ combinations. Clean and donor-latent-patched recipient paths generate L_0G_0 and L_1G_1, respectively. Thus G_1 is written in the patched recipient trajectory; it is not copied directly from a separate donor trajectory. The latent background L includes its recomputed latent-end delimiter (EOT) where present. L_0G_1 tests coalition sufficiency in the clean background; L_1G_0 tests reversal by restoring the clean prefix in the patched background. The exchanged history G excludes the final prefix token (the reader). Each cell recomputes that token using both segments, then decodes greedily. A coalition satisfies *relay sufficiency* when it passes the clean and total-path baselines, forward donor transfer, and reverse recipient recovery criteria, including the study's residual-donor limits. This bidirectional criterion concerns answer outcomes under the specified overwrite; continuous upstream effects may remain. This design leaves latent–prefix interactions and alternative relay sets unresolved (Vaidyanathan et al., 2026).

2.4 DESTRUCTIVE RETENTION TESTS AND NULL INTERVENTIONS

Relay tests establish donor transmission. We separately measure *retention*, the fraction of recipient answers preserved after a destructive operator D:

$$\mathrm{Retention}(D(X), u) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}[y(R_i[D(X_u)]) = y(R_i)] \,. \tag{3}$$

Low retention indicates destructive sensitivity when identity and paired-null controls preserve the answer; its dependence interpretation applies only to the tested segment–operator pair. Gaussian corruption probes sensitivity to perturbed state, at the risk of moving activations off distribution. Our strongest prompt-dependent result therefore uses *multiset-preserving* binding tests: key-only, value-only, and independent key/value permutations preserve the respective tensor-value multisets while disrupting which value each key retrieves. Applying one permutation jointly to bound (K, V) pairs leaves the evaluated attention read algebraically unchanged, which makes it a natural implementation and symmetry null. Since the null leaves the attention sum intact, it checks that our intervention machinery behaves as intended rather than matching the severity of a destructive test. The two intervention families are also logically independent: Appendix E constructs response tables in which qualified donor transfer and destructive retention vary independently, with clean, null, and semantic-control outcomes held fixed.

Scope of the comparison. Rates describe the fixed inputs, checkpoint, and operator. A later observation may follow an earlier cache edit, and requires the edited branch to reach the observation boundary naturally. Appendix E gives the conditional estimands. A *matched endpoint* shares units, intervention timing, and observation boundary with a paired algebraic null; span width and perturbation severity can still differ across segments (Grant et al., 2026).

Inference and reporting rules. A qualified donor-specific interchange supports answer transport by X, while recipient loss under a within-segment destructive operator, read against identity replay and its paired algebraic null, supports dependence evidence for that segment–operator pair. We report concordant evidence when both analyses implicate the same segment, and a *transport–dependence dissociation* when they implicate different segment–operator pairs. We distinguish three inconclusive statuses: qualification failure; non-estimability when the required natural-path contrast is unavailable; and indeterminacy when valid observations do not resolve the classification under the prespecified thresholds. These statuses are distinct from a measured null effect.

3 Experimental Design

Table 1 summarizes localization, new-input confirmation, and boundary studies. The main mechanism evidence comes from the 64-pair CODI–GPT-2 confirmation; the earlier panels determine its heads and intervention rules. Checkpoint identities and protocols appear in the appendix; exact inputs are provided in the executable supplement.

Tasks and units. Balls computes remaining red balls plus blue balls. For example, one confirmation recipient starts with 26 red and 7 blue balls and gives away 16 red balls (answer 17); its donor starts with 34 red balls, leaving the other quantities fixed (answer 25). Both naturally emit "The answer is:" before their answers. We test whether the donor-patched prefix cache can change the recipient's answer from 17 to 25. Travel and production compute duration times rate plus an offset; equal groups computes group count times group size. Each six-role matrix contains a recipient, a distinct-answer donor, two single-variable controls, a same-answer donor with different factors, and a random-source control. We analyze each checkpoint–schema stratum separately. Counts describe fixed, rule-selected inputs; full dose curves use the initial GPT-2 panel, while the 64-pair confirmation tests the fixed native and unit-dose conditions.

Qualification and fixed comparisons. Qualification requires correct, parseable clean answers, a naturally generated common prefix, and faithful model and cache replay (Appendix H). Later-boundary contrasts additionally require the edited branch to reach the boundary naturally. The public-checkpoint comparison fixes both checkpoints, the five-token prefix, four $L \times G$ cells, and six-slot latent-value intervention before evaluation. Its winner and competitor thresholds are 39/48 and 12/48. We additionally require the donor-transfer rate to decrease by at least 0.50 (50 percentage points) from L_1G_1 to L_1G_0. The fresh content-control study follows same-panel diagnostics but fixes its own inputs and analysis before observing new-input model outcomes (Appendix C). Qualification failures and unavailable contrasts are reported in Appendix G.

Table 1: **Study design and independent units.** GPT-2 mechanism panels use Balls arithmetic pairs; other panels use six-role matrices per schema (travel, production, equal groups). Parentheses give totals across the three schemas. Zen-E and SIM-CoT share each panel. Roles, boundaries, and repeated interventions are dependent; shared matrices and runtime checks do not enlarge the independent sample.

Study	Checkpoint	Input panel	Experimental stage
Fixed relay mechanism	CODI–GPT-2	8 + 8 pairs	Discovery + validation
		64 new pairs	**Fixed-rule confirmation**
		16 new pairs	Earlier whole-segment test
Targeted cross-check	CODI–Llama-1B	16 new / schema	Endpoint fixed after screening
Prefix sufficiency	Zen-E, SIM-CoT	48 / schema (144)	Checkpoints fixed in advance
Latent-content support	Zen-E, SIM-CoT	64 new / schema (192)	Protocol fixed before selection
Transport persistence	Coconut–GPT-2	16 / schema (48)	Separate boundary study
Transport vs. sensitivity	CODI–Llama-1B	48 / schema (144)	16/schema discovery; held out

4 Results: Relay Mechanism and Boundary Cases

4.1 A latent-to-prefix relay with a fixed writer and reader

Does the localized relay continue to operate when its heads and intervention rules are fixed and its inputs change? On 64 new CODI–GPT-2 arithmetic pairs, patching the native values of "The," "answer," and "is" in eight fixed reader heads transfers the donor answer on 61/64 unrestricted greedy continuations; restoring their clean values recovers the recipient answer on 64/64. All pairs qualify, with none replaced. The eight reader heads, prefix positions, and thresholds were fixed after localization on eight discovery and eight validation pairs. The criterion requires forward donor and reverse recipient rates of at least 0.8, and a reverse donor rate of at most 0.25. The full coalition passes, with zero reverse donor answers; sequence scoring gives 62/64 forward and 64/64 reverse (Table 8).

Repeated writing from upstream latent states. To distinguish repeated writing from autoregressive copying, we independently vary the upstream state U and the previously generated prefix P, with 0 denoting clean and 1 patched. At each tested position, we extract the newly written value difference δv and repeat it at all three positions in a common clean terminal background. This readout isolates the new writer signal while holding the number of contributing positions fixed; previously patched prefix slots cannot supply extra dose. For "answer" and "is," the upstream-only condition U_1P_0 produces 62/64 and 58/64 greedy donor answers, whereas prior-prefix-only U_0P_1 produces 0/64 at either position. Projection onto the full writer difference Δv, measured by $\beta = \langle \delta v, \Delta v \rangle / \|\Delta v\|^2$, is 0.945 and 0.916 for upstream-only writing, versus 0.043 and 0.067 for prior-prefix-only writing. Most of the donor-aligned writer projection is therefore recovered from the upstream state, even when the preceding prefix is clean. Prior prefix states alone do not reproduce that signal, ruling out a purely autoregressive prefix-copying explanation under this readout. Retaining the older five-slot latent tail while restoring the final-latent state and later history preserves essentially the full writer projection. These projection and common-readout results do not imply full-vector equality or describe an unmodified downstream trajectory.

Reading the contextual prefix. Swapping the prefix's attention-weighted value contributions in the eight fixed layer-10/11 heads yields 61/64 greedy donor transfers and 64/64 recipient recoveries; each wrong-span control (prompt-tail including BOT, latent-tail, EOT, or colon-self) gives 0/64 and 1/64, respectively. Together, the writer and reader tests support repeated latent-to-prefix transmission followed by terminal attention readout. Appendix A.1 gives every factorial cell and both answer endpoints.

Table 2: **Fixed mechanism, new inputs.** Unrestricted greedy outcomes on 64 separate confirmation pairs, all qualified. Heads and intervention rules remain fixed. Ranges cover three dependent subsets of the same 64 pairs. Native patches exchange corresponding values; matched-dose patches redistribute the mean three-position difference. Only the full native coalition meets the bidirectional criterion; each matched-dose singleton meets it. Reverse donor counts are zero for these passing conditions. Appendix Table 8 gives every subset and both endpoints.

Prefix-value intervention	Forward donor	Reverse recipient
Native singletons	0–1/64	9–13/64
Native pairs	33–41/64	59–61/64
Native all three	61/64	64/64
Matched-dose singletons	61–62/64	64/64

A separate disjoint 16-pair evaluation locates this relay between whole segments: late latent values transfer 15/16 donor answers before the prefix, and bidirectional prefix overwrites give 15/16 forward transfers and 16/16 recipient recoveries at the terminal reader. Books also passes the whole-segment relay criteria; Coins fails random-source specificity despite the behavioral pattern. The full boundary progression, template controls, and qualification failures appear in Appendix A and Table 13.

4.2 Coalition minimality depends on the intervention

Why must all three positions be patched together? At native amplitude, every proper subset fails the same bidirectional criterion on the 64 new pairs: singletons transfer at most 1/64 donor answers, and pairs transfer 33–41/64 (Table 2). This identifies an inclusion-minimal coalition for the tested native-value patches. To test whether this minimum depends on how content and readout dose are distributed, we pool the three value differences and redistribute them across fewer positions.

For each pair and reader head (ℓ, h), let $G = \{\text{The}, \text{answer}, \text{is}\}$ and define

$$M_{\ell h} = \frac{1}{3}\sum_{s\in G}(V^1_{\ell hs} - V^0_{\ell hs}), \qquad A_C = \sum_{s\in C}\alpha_{\ell hs}, \qquad a_C = a\frac{A_G}{A_C}. \tag{4}$$

For a subset $C \subseteq G$, forward patches add $a_C M_{\ell h}$ to its clean values; reverse patches subtract it from patched values, using the corresponding parent path's colon attention weights α. Here, V^0 and V^1 are the corresponding prefix values from the clean and latent-patched recipient paths. At unit amplitude $a = 1$, every singleton passes: forward donor counts are 61–62/64 and reverse recipient counts are 64/64. The prescribed attention-weighted contribution is equalized by construction; successful answer transfer and recovery are empirical outcomes. Because M pools all three positions, this establishes control by redistributed content at matched dose, rather than semantic interchangeability of the native positions. The original dose-response analysis remains limited to the eight-pair validation panel.

4.3 Relay sufficiency differs across checkpoints

Does the same bidirectional criterion support relay sufficiency in other checkpoints? A targeted CODI–Llama-1B cross-check reproduces the relay on fresh matrices: forward transfer is 16/16 for travel and 14/16 for production, with 16/16 recipient recoveries in both. This motivates a separate comparison in which Zen-E and SIM-CoT are selected before either checkpoint is evaluated in the study. Both pass 32/32 qualification cases and complete the shared 48-matrix-per-schema panel once.

On the prospectively fixed checkpoint panel, SIM-CoT meets the operational criterion for relay sufficiency in both primary schemas (Table 3A). Donor-answer counts in $L_1G_1/L_0G_1/L_1G_0$ are 48/41/1 for travel and 48/42/0 for production; reverse overwrite restores the recipient on 47/48 and 48/48 matrices, respectively. Zen-E passes the baseline, total-path, and reverse criteria, but its donor-patched prefix alone transfers only 11/48 travel and 17/48 production answers, below the 39/48 sufficiency threshold. This pattern indicates that the patched prefix is not sufficient in the clean latent background under the tested overwrite criterion. Both checkpoints preserve every equal-groups recipient in all four cells. Under the same overwrite criterion, SIM-CoT exposes a sufficient prefix relay, while Zen-E's prefix falls below the sufficiency threshold without additional upstream support.

Table 3: **Relay sufficiency across checkpoints and the role of latent content.** Panel A gives recipient (R) : donor (D) answer counts out of 48; residuals are other parseable answers. L_0G_0/L_1G_1 are clean/latent-patched baselines; L_0G_1 tests prefix sufficiency, and L_1G_0 tests reversal by restoring the clean prefix. The prespecified criteria require a winner rate of at least 39/48, a competitor cap of 12/48, and a donor-transfer rate decrease of at least 0.50 (50 percentage points) from L_1G_1 to L_1G_0. Panel B reports original-donor hits on new inputs, with the original donor-patched prefix G_1 and recipient delimiter E fixed. Same/different refers to the latent source's answer relative to the original donor. Both Zen-E paired tests pass $\alpha = 0.025$; SIM-CoT production is nonsignificant (n.s.). Equal groups preserves every recipient in both panels. Samples are not pooled across panels or arms; Appendix Tables 10–12 give uncertainty and complete controls.

A. Prospective checkpoint test: R : D counts ($n = 48$ per schema)

Checkpoint	Schema	L_0G_0	L_1G_1	L_0G_1	L_1G_0	Status
Zen-E	Travel	48 : 0	0 : 47	33 : 11	48 : 0	prefix insufficient alone
Zen-E	Production	48 : 0	0 : 48	22 : 17	47 : 1	prefix insufficient alone
Zen-E	Equal	48 : 0	48 : 0	48 : 0	48 : 0	control passed
SIM-CoT	Travel	48 : 0	0 : 48	3 : 41	47 : 1	relay sufficiency
SIM-CoT	Production	48 : 0	0 : 48	2 : 42	48 : 0	relay sufficiency
SIM-CoT	Equal	48 : 0	48 : 0	48 : 0	48 : 0	control passed

B. Separate fresh-input test: 64 matrices per schema, fixed G_1 and clean E

Checkpoint	Schema	Recipient L	Different L	Same L	Restored : lost	Status
Zen-E	Travel	17/64	42/64	64/64	27 : 2	primary pass
Zen-E	Production	20/64	43/64	64/64	25 : 2	primary pass
SIM-CoT	Travel	54/64	61/64	64/64	7 : 0	secondary
SIM-CoT	Production	59/64	61/64	64/64	5 : 3	secondary, n.s.

4.4 What upstream support does prefix readout require?

Zen-E's patched prefix falls below the sufficiency threshold in a clean latent background. We therefore ask whether useful latent support must come from a problem whose final answer matches the original donor's. On the original 48-matrix panel, restoring donor latent values with G_1 fixed and EOT clean yields 47/48 and 48/48 donor hits in travel and production, respectively. Restoring only EOT yields 11/48 and 18/48. These same-panel diagnostics motivate the separate fresh-input test of latent content (Appendix C). The fresh test fixes the original donor-patched prefix and keeps the delimiter clean, while supplying latent values from the recipient, a same-answer source, or a different-answer source. Here, "same" and "different" refer to the original donor's final answer.

On 64 new matrices per schema, different-answer latent values raise Zen-E's original-donor hits from 17/64 to 42/64 for travel and from 20/64 to 43/64 for production (Table 3B). The paired restoration/loss counts are 27/2 and 25/2. Both prespecified one-sided tests pass $\alpha = 0.025$ ($p < 3 \times 10^{-6}$); net gains are 39.1 and 35.9 percentage points, with descriptive paired-bootstrap 95% intervals of [25.0, 53.1] and [21.9, 50.0] points for these selected panels. Thus the supporting latent source need not have the donor's final answer.

Latent content nevertheless influences generation. Same-answer sources preserve 64/64 original-donor hits in both tasks, while different-answer sources elicit their own answers on 5/64 and 10/64. The secondary SIM-CoT production comparison shows five restorations and three losses, without a statistically significant net recovery. Both checkpoints show donor-margin gains, including SIM-CoT despite its operational relay sufficiency; these gains need not preserve the greedy winner (Appendix Table 12). The evidence separates final-answer matching from useful upstream support, but does not isolate an answer-independent latent representation from shared computation or other content effects.

4.5 Boundary cases and limits of transport evidence

Latent transport can persist through a prefix. An official GSM8K-trained Coconut–GPT-2 checkpoint (Hao et al., 2024; Cobbe et al., 2021) passes 32/32 qualification cases. Native-latent K/V interchange transfers 15/16 travel, 16/16 equal-groups, and 16/16 production donor answers at both t_0 and t_1; raw-prompt and BOT transfer is 0/16 throughout, with semantic-role controls supporting

donor specificity. Every branch naturally emits the complete one-token pre-answer prefix "###," and the answer begins immediately afterward. These matched results show that transport persists across this checkpoint's entire prefix, arguing against a universal one-step handoff. Its paired binding tests leave dependence indeterminate under the fixed thresholds (Appendix D).

Transport and destructive sensitivity can implicate different segments. The CODI–Llama-1B boundary study asks whether the segment that transports a donor answer is also the segment whose unbinding removes the recipient answer. After a 16-matrix-per-schema discovery screen, it fixes 48 held-out matrices per schema, observation boundaries t_0, t_5, K/V-unbinding operators, and paired algebraic nulls. All cache edits occur before EOT. At t_0, raw/BOT/latent-KV/latent-V transfer is $0/0/45/47$ for travel and $3/0/46/47$ for production, each out of 48. Raw key-only, value-only, and independent unbinding each retain 0/48 recipients, whereas the corresponding latent interventions retain 44/48, 48/48, and 48/48; paired nulls pass. Value-only and independent unbinding keep every output parseable. These two operators distinguish latent donor transport from raw-prompt destructive sensitivity; key-only results are confounded by mostly unparseable outputs. Raw spans cover 24–36 positions and latent spans six, so the contrast compares these segment–operator pairs without equalizing perturbation dose.

Unreached boundaries are not null results. The study's joint rule requires qualified transport and retention at both boundaries. Although donor-interchange trajectories reach t_5, raw destructive trajectories do so naturally on 0/48 matrices in every schema. Later-boundary retention is therefore not estimable, and the joint rule remains unresolved despite the informative secondary t_0 result. Equal groups illustrates a different limit: its raw transport is 48/48 and raw retention is $12/48, 0/48, 0/48$, so tightening the 12/48 cap to 11/48 changes the label from concordant to indeterminate (Table 16).

Supporting analyses retain the same distinctions. A separate held-out suite reproduces latent transport and raw-prompt sensitivity (Appendix D); a post hoc boundary-local intervention reuses existing units (Appendix H). Position-count controls, forced-prefix experiments, and observational comparisons appear in Appendices H, G, and F. Forced prefixes can alter transport, while attention, probes, cosine similarity, and K/V magnitude can disagree with interchange.

5 Related Work

Latent-reasoning representations. Coconut recycles hidden states as continuous thoughts, CODI self-distills explicit CoT into latent states, and KaVa distills compressed teacher K/V trajectories into a latent-reasoning student (Hao et al., 2024; Shen et al., 2025; Kuzina et al., 2026). These methods construct or supervise latent computation, but leave open how answer information is transmitted when latent, prompt, and generated states coexist during decoding. SIM-CoT (Wei et al., 2026) adds step-level supervision through a training-time auxiliary decoder. Other work separates storage from computation, characterizes latent dynamics, or intervenes without retraining (Goyal et al., 2025; Duraipandian et al., 2026; Chang et al., 2026). We study whether, and under what upstream conditions, native latent states transmit answer information through the cache of a naturally generated prefix.

Time-resolved causal analysis of reasoning. Step-wise interventions test necessity and influence propagation across latent steps (Li et al., 2026); prompt-cache splicing sweeps layers and decode time (Kumar et al., 2026); and visible-CoT patching reveals answer bypass or recovery (Sathyanarayanan et al., 2026; Mehrafarin et al., 2026). Their intervention units are latent steps, prompt spans, or visible-token states rather than the specific shared comparison studied here among raw prompt, BOT, native latent, and generated-prefix cache segments. SCIT is closest to our setting: its donor interchange, K/V component tests, semantic controls, and corruption already compare counterfactual transfer with sensitivity at evaluated boundaries (Ding et al., 2026). BATS uses these ingredients to test a specific transmission mechanism: donor identity passes through naturally generated prefix states, and forward and reverse overwrites separate prefix sufficiency from the surviving latent effect. SCIT identifies answer-carrying segments, whereas BATS tests whether native latent states transmit donor-specific information through generated-prefix caches and what upstream support that readout requires. Position interventions show that the three-position minimum depends on how value content and readout dose are distributed. The empirical additions are the resolved writer–prefix–reader

relation, its confirmation with fixed rules on new inputs, and the checkpoint and latent-content contrasts.

Causal localization and intervention validity. Interchange intervention, causal mediation and tracing, activation patching, and path patching provide the methodological basis for causal localization (Geiger et al., 2022; Vig et al., 2020; Meng et al., 2022; Goldowsky-Dill et al., 2023; Wu et al., 2024). Corruption methods and evaluation metrics can change patching-based localization (Zhang & Nanda, 2024). Recent work shows that patched representations can leave the model's natural distribution and that interactions among mediators can hide or inflate componentwise effects (Grant et al., 2026; Vaidyanathan et al., 2026). We therefore interpret patching as evidence about a specified counterfactual operator, not automatic proof of natural necessity. Replay and locality checks, semantic specificity, natural-path eligibility, identity controls, and algebraic nulls delimit the reported finite-set response profile.

6 Discussion and Limitations

The fixed-mechanism tests show that a shared visible prefix can serve as a contextual relay: upstream latent states repeatedly write answer information into the prefix's contextual values, and a terminal reader combines those values to produce an answer. Three positions are minimal for native patches, while a single position suffices under redistributed content and dose. These interventions change both factors, so they do not isolate dose as the sole explanation. Minimality characterizes the intervention family rather than a fixed number of storage sites.

The checkpoint contrast concerns the upstream state in which prefix readout succeeds. SIM-CoT passes prefix sufficiency in a clean latent background, yet additional latent support still raises donor margins. Zen-E falls below that sufficiency threshold, but latent values from different-answer problems improve readout of its patched prefix. These findings distinguish final-answer matching from useful support; they do not identify an answer-independent latent component.

Scope and remaining questions. We study templated arithmetic, GPT-2- and Llama-1B-scale latent reasoners, and short fixed prefixes. The 64-pair mechanism confirmation tests new inputs within one checkpoint and arithmetic family. Extension to open-ended reasoning, variable-length latent trajectories, or larger instruction-tuned models remains untested. The fresh content controls reuse number and template families and share an additive subproblem. Separating shared computation from magnitude and answer-specific content would require varying those factors independently while keeping the patched prefix fixed. Neither checkpoint study estimates prevalence or controlled training-seed effects; shared matrices are not pooled as independent evidence.

Transport and destructive sensitivity remain specific to the tested boundary and operator. Raw and latent spans differ in width and semantic extent; the matched-footprint grid equates position counts without equating perturbation severity. In the same-prefix study, the equal-groups label lies at its decision threshold, and the exploratory boundary-local branch reuses existing units. When edited trajectories never reach a later boundary, retention cannot be estimated there; this is an eligibility limit, not evidence of an absent route. The experiments establish transport along the tested segments. Whether the route is unique, naturally necessary, or causally determined by task structure remains open.

7 Conclusion

Across the tested settings, a fixed visible prefix can serve as an operator-defined relay between latent reasoning and answer generation. In CODI–GPT-2, 64 new pairs support repeated writing by native latent values and readout by fixed terminal heads. Three positions are jointly required under native-value patching, whereas redistributing their mean difference at matched dose enables single-position control. The same bidirectional test separates relay sufficiency in SIM-CoT from upstream-dependent readout in Zen-E. Fresh-input controls further show that useful upstream support need not match the donor's final answer, although its content still matters. Together with persistent transport and transport–sensitivity dissociation, these results show that an answer-relay claim is inseparable from its intervention, background state, and decoding boundary.

REPRODUCIBILITY STATEMENT.

The appendix gives boundaries, panel splits, and the 64-pair fixed-mechanism protocol. The executable supplement provides original checkpoint inputs, weight revisions, replay code, per-cell references, and GPT-2 mechanism routines. The separate confirmation record supplies the fixed inputs, per-condition outputs, evaluated runtime, GPU replay command, and CPU table reconstruction. The content-control extension includes the fixed fresh inputs, selection exclusions, protocol, complete outputs, and CPU reconstruction of Tables 3B and 10–12. The outcome archive checks counts and reporting logic across the wider study. Model replay and outcome reconstruction are separate; neither adds independent evidence.

AI USE STATEMENT

We used OpenAI Codex to help develop the conceptual framework, formulate mathematical claims, refine hypotheses, design and implement experiments, interpret results, and draft and edit the manuscript. Codex also assisted with literature search, citation checking, code and artifact auditing, automated tests, LaTeX builds, and PDF inspection. The authors take responsibility for the final text, claims, code, figures, and artifacts.

REFERENCES

Shuochen Chang, Tong Bai, Xiaofeng Zhang, Qianli Ma, Qingyang Liu, Zhaohe Liao, Yibo Miao, and Li Niu. Unlocking the black box of latent reasoning: An interpretability-guided approach to intervention. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 34019–34032, 2026.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Yi Ding, Lijun Huang, and Menglin Yang. SCIT: Testing causal cache carriers in latent chain-of-thought models. *arXiv preprint arXiv:2608.27265*, 2026.

Sabari Iyyappan Duraipandian, Shreya Sanjay Boyane, Manju Nagesh, Jerome Francis, Archana Vaidheeswaran, and Kevin Zhu. Interpreting latent CoT reasoning as dynamical systems. *arXiv preprint arXiv:2607.09698*, 2026.

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

Nicholas Goldowsky-Dill, Chris MacLeod, Lucas Sato, and Aryaman Arora. Localizing model behavior with path patching. *arXiv preprint arXiv:2304.05969*, 2023.

Sayam Goyal, Brad Peters, María Emilia Granda, Akshath Vijayakumar Narmadha, Dharunish Yugeswardeenoo, Callum Stuart McDougall, Sean O'Brien, Ashwinee Panda, Kevin Zhu, and Cole Blondin. Scratchpad thinking: Alternation between storage and computation in latent reasoning models. In *Mechanistic Interpretability Workshop at NeurIPS 2025*, 2025.

Satchel Grant, Simon Jerome Han, Alexa Tartaglini, and Christopher Potts. Addressing divergent representations from causal interventions on neural networks. In *International Conference on Learning Representations*, volume 2026, pp. 11330–11356, 2026.

Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. Training large language models to reason in a continuous latent space. *arXiv preprint arXiv:2412.06769*, 2024.

Vinayshekhar Bannihatti Kumar, Manoj Ghuhan Arivazhagan, Disha Makhija, and Rashmi Gangadharaiah. Probing the prompt KV cache: Where it becomes dispensable. *arXiv preprint arXiv:2605.30574*, 2026.

Anna Kuzina, Maciej Pioro, and Babak Ehteshami Bejnordi. KaVa: Latent reasoning via compressed KV-cache distillation. In *International Conference on Learning Representations*, volume 2026, pp. 129875–129902, 2026.

Zirui Li, Xuefeng Bai, Kehai Chen, Yizhi Li, Jian Yang, Chenghua Lin, and Min Zhang. Dynamics within latent chain-of-thought: An empirical study of causal structure. *arXiv preprint arXiv:2602.08783*, 2026.

Houman Mehrafarin, Amit Parekh, and Ioannis Konstas. When chain-of-thought fails, the solution hides in the hidden states. *arXiv preprint arXiv:2604.23351*, 2026.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. *Advances in neural information processing systems*, 35:17359–17372, 2022.

Anish Sathyanarayanan, Aditya Nagarsekar, and Aarush Rathore. Bypassing the rationale: Causal auditing of implicit reasoning in language models. *arXiv preprint arXiv:2602.03994*, 2026.

Zhenyi Shen, Hanqi Yan, Linhai Zhang, Zhanghao Hu, Yali Du, and Yulan He. CODI: Compressing chain-of-thought into continuous space via self-distillation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 677–693, 2025.

Sankaran Vaidyanathan, David Arbour, Aaron Mueller, Scott Niekum, and David Jensen. The curse of multiple mediators: Hidden interaction effects in activation patching. *arXiv preprint arXiv:2606.27510*, 2026.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Simas Sakenis, Jason Huang, Yaron Singer, and Stuart Shieber. Causal mediation analysis for interpreting neural NLP: The case of gender bias. *arXiv preprint arXiv:2004.12265*, 2020.

Xilin Wei, Xiaoran Liu, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Jiaqi Wang, Xipeng Qiu, and Dahua Lin. Sim-cot: Supervised implicit chain-of-thought. In *International Conference on Learning Representations*, volume 2026, pp. 56721–56742, 2026.

Zhengxuan Wu, Atticus Geiger, Aryaman Arora, Jing Huang, Zheng Wang, Noah Goodman, Christopher D Manning, and Christopher Potts. pyvene: A library for understanding and improving PyTorch models via interventions. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: System Demonstrations)*, pp. 158–165, 2024.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations*, volume 2024, pp. 1651–1678, 2024.

A Mechanism Experiments and Quantitative Readouts

Panels, selection, and endpoints. The mechanism model uses the public CODI-GPT-2 weights, with six latent iterations and a 768-dimensional backbone. This is the original mechanism checkpoint studied by SCIT, not an additional replication arm. The initial mechanism panel contains 16 arithmetic counterfactual pairs: indices 0–7 are discovery and 8–15 validation. Reader heads are selected on discovery pairs and fixed for position, dose, and writer analyses. These initial analyses reuse the same eight validation pairs. A separate 64-pair panel confirms the fixed mechanism tests without a new discovery split (Appendix A.1). The earlier whole-segment held-out confirmation uses 16 new pairs with 96 distinct recipient/control prompts, no prompt overlap with the initial panel, and no new head, position, or dose selection. In that confirmation the source latent intervention replaces values at five pre-final latent positions in layers 8–11. Prefix experiments exchange contextual K/V at generated-history positions, then recompute the shared colon token. The public-checkpoint tests instead replace values at all six native latent positions in all layers and exchange the four-token "The answer is:" history before recomputing its trailing-space reader. They test the relay between whole segments, not the exact GPT-2 prefix positions or heads.

Boundary progression in the held-out confirmation. For each history length j, the four cells share a terminal answer endpoint: exchange the first j generated-prefix cache positions and replay the remaining specified prefix to the colon. Both parent paths must naturally generate the common prefix. Let m_{ab} be the sequence log-probability of the donor answer minus that of the recipient in cell L_aG_b. Report total $m_{11} - m_{00}$, forward $m_{01} - m_{00}$, residual direct $m_{10} - m_{00}$, and interaction $m_{11} - m_{10} - m_{01} + m_{00}$ separately. These are finite-panel, operator-defined contrasts. Candidate sequence winners and unrestricted greedy answers are different endpoints.

Table 4: **The relay accumulates across the prefix.** All entries use the 16 held-out confirmation pairs and sequence-level scoring. In each row the clean/total baselines have donor counts 0/16 and 16/16, with mean margins $m_{00} = -15.285$ and $m_{11} = 12.781$. After exchanging three prefix positions, unrestricted greedy donor counts in $L_1G_1/L_0G_1/L_1G_0$ are 15/16, 15/16, and 0/16; reverse recipient recovery is 16/16. Thus the winner criterion supports relay sufficiency, not zero residual numerical effect or perfect greedy transfer.

Exchanged history	**Donor wins / 16**		**Mean answer margin**	
	Forward	Reverse	m_{01}	m_{10}
"The"	0/16	11/16	-7.655	3.328
"The answer"	8/16	2/16	1.900	-6.784
"The answer is"	16/16	0/16	11.856	-14.763

Native-amplitude positions and direct reader contributions. The frozen GPT-2 reader nodes are L10H2 and L11H0/H3/H4/H6/H8/H10/H11 (zero-based indices). On the eight discovery pairs, we rank all 48 heads in layers 8–11 by the minimum of four mean normalized answer-margin effects: forward and reverse prefix-value patching, and forward and reverse colon-reader output patching before the output projection. Ties are resolved by layer and head index. We evaluate cumulative top-K sets for $k \in \{1, 2, 4, 8, 12, 16, 24, 32, 48\}$ and select the first set meeting all four directional criteria; $k = 8$ passes discovery and the separate eight-pair validation. This is the smallest passing set on that grid, not an exhaustively minimal head subset. Forward intervention replaces selected prefix values in a clean background; reverse intervention restores their clean values in the patched background. The native-amplitude criterion requires at least 0.8 forward donor and reverse recipient rates, and at most 0.25 reverse donor rate, on each discovery/validation split. Table 5 reports every nonempty position subset on the initial validation panel. For direct-contribution tests we modify the selected head's scaled dot-product attention (SDPA) value contribution before head concatenation and output projection. Swapping the three prefix contributions in the eight frozen heads gives forward donor 8/8 and reverse recipient 8/8. Applying the same tests separately to colon-self, the pre-latent prompt tail, latent tail, or EOT gives forward donor 0/8 and reverse donor 8/8 for each group. The prompt-tail control comprises the final two raw-prompt positions and BOT; it is not a raw-prompt-only intervention. Attention-only interchange also gives forward donor 0/8 and reverse donor 8/8. These are controls for the terminal reader within the selected head set, not a unique minimal full circuit.

Table 5: **Initial position and dose analysis.** All entries are sequence-level endpoints on the same eight validation pairs in the mechanism panel, with frozen layer-10/11 reader heads. Native patches use the corresponding values from the latent-patched recipient path. Dose-matched patches instead add/remove the mean three-position value difference, scaled to equalize attention-weighted dose. Each "any" row contains three dependent position choices, not 24 independent pairs. The unchanged three-position native intervention is inclusion-minimal under the original criterion; the amplitude intervention tests a different operator.

Prefix-value positions / intervention	Forward donor	Reverse recipient
"The" / native amplitude	0/8	3/8
"answer" / native amplitude	1/8	4/8
"is" / native amplitude	1/8	4/8
"The + answer" / native amplitude	3/8	7/8
"The + is" / native amplitude	3/8	7/8
"answer + is" / native amplitude	4/8	8/8
All three / native amplitude	8/8	8/8
Any singleton / matched dose	8/8 each	8/8 each
Any pair / matched dose	8/8 each	8/8 each

Dose construction and quantitative comparison. Section 4.2 defines the mean-difference intervention and its attention-matched scale (Eq. 4). The initial panel also uses fixed-per-slot $a_C = a$ and fixed-unweighted-total $a_C = 3a/|C|$ schedules. With attention matching at unit amplitude, every singleton and every pair passes on each of the same eight validation pairs (three dependent position choices per cardinality).

The initial-panel grid crosses all seven nonempty subsets, amplitudes $0, .25, .5, .75, 1, 1.25, 1.5$, layer groups L10, L11, and L10+L11, the three dose schedules, and both directions. Normalize each margin effect by the clean-to-full-patch margin in its direction. Piecewise-linear interpolation of the discovery three-position amplitude curve predicts validation effects from the corresponding effective dose; the fits below use the full L10+L11 reader set.

Table 6: **A dose model predicts the constructed interventions.** The schedules evaluate dependent interventions on eight validation pairs. Dose matching uses the attention sum by construction; the fits assess the resulting answer-margin response within this operator family. Attention-matched and unweighted fits are close because native attention is approximately uniform, so the data do not uniquely identify either weighting rule or a linear native circuit.

Dose schedule	**Forward**		**Reverse**	
	MAE	R^2	MAE	R^2
Fixed per slot + attention mass	0.0400	0.9720	0.0397	0.9725
Fixed unweighted total	0.0457	0.9783	0.0425	0.9788
Attention matched	0.0432	0.9795	0.0394	0.9802

Separating writing from copying. The U/P factorization, common three-copy readout, and projection coefficient β are defined in Section 4.1. The initial panel also reports residual-vector diagnostics alongside the projection and answer readout. A second factorization splits the older five-position tail T from the derived final-latent state F, restoring all other later history to clean. This tests whether the final-latent state alone mediates writing. The reference differences are family-specific: $\Delta v_{UP} = v(U_1P_1) - v(U_0P_0)$ and $\Delta v_{TF} = v(T_1F_1) - v(T_0F_0)$, with β computed relative to the corresponding reference. In the T/F factorization, the delimiter and preceding prefix history are restored to clean even in T_1F_1, so its reference vector and answer readout need not match U_1P_1. Projections are computed per pair by concatenating the selected value-head coordinates, then averaged across pairs.

Table 7: **The older latent tail supplies the writer.** Means and sequence-source readouts use the same eight validation pairs, value-only interventions, and the common three-copy readout. The first test leaves a small prior-prefix influence; the second retains essentially the full writer projection after restoring the final-latent and later-history states to clean. Both are conditional mechanism analyses on the earlier panel.

Writer test	**Hybrid**	**"answer"**		**"is"**	
		β	Donor / 8	β	Donor / 8
Upstream / prior prefix	U_0P_1	0.01822	0/8	0.03391	0/8
	U_1P_0	0.97655	8/8	0.95737	8/8
Older tail / final latent	T_0F_1	-0.00007	0/8	-0.00005	0/8
	T_1F_0	1.00003	8/8	1.00003	8/8

A.1 Fixed-mechanism confirmation on new inputs

Selection and scope. The checkpoint, six latent iterations, five pre-final value slots in layers 8–11, four-token prefix, eight reader heads, native position subsets, unit-dose rule, and writer operators were fixed before evaluating new inputs. This locally specified confirmation adds 64 independent Balls pairs; it is not an externally registered or checkpoint-blind study. The generator seed is 20260906

and source-control seed is 40. Remaining red balls and blue balls range from 2 to 18. We excluded 226,758 historical prompt strings, including candidate pools and constructed controls. Before model inference, ranges 19–27 and 28–36 for the number of red balls given away supplied only one and zero eligible pairs. A model-free availability inventory selected 10–18, with 1,808 unused canonical prompts out of 2,601; this range lies between the initial and earlier confirmation ranges. The first 64 candidates meeting semantic-role, equal-token-width, and prompt-disjointness constraints were taken from 998 candidates. Their 384 actual recipient/control prompts are mutually distinct and absent from the recorded historical inputs. The range adjustment used no model outcomes; the panel was then fixed.

Qualification and estimand. For each pair, the recipient and five source roles must give the correct clean answer and naturally generate the common prefix; the oracle-patched parent must also reach that prefix naturally. At least 60/64 complete pairs must qualify for confirmation. All 64 did: all 384 clean-role checks passed, and no pair was excluded, replaced, or filtered on transfer success or positive margins. A pair is the inference unit. The 60 dependent conditions per pair comprise four whole-prefix relay cells, 14 native-subset cells, ten direct-reader/control cells, 16 zero/unit-dose cells, and 16 writer cells. Sequence scores and unrestricted greedy continuations are recorded separately for all 3,840 conditions.

Fixed criteria and results. We retain the 0.80 forward-donor/reverse-recipient floors and 0.25 reverse-donor ceiling. The three-position coalition is inclusion-minimal within the fixed head set when it passes and every proper subset fails this same bidirectional criterion; this is an operational panel result, not proof of natural necessity or a population-wide minimal circuit. Writer interpretations additionally require upstream/tail-only mean $\beta \geq 0.75$, prior-prefix/final-only mean $\beta \leq 0.25$, the corresponding donor/recipient winner pattern, and at least 0.90 natural factor-parent and next-token agreement. All these tests pass on both endpoints; natural agreement is 1.0 throughout. The whole-prefix four cells give greedy donor counts $0, 63, 61, 0$ in $L_0G_0, L_1G_1, L_0G_1, L_1G_0$, respectively; the clean and reverse cells recover 64/64 recipients. Sequence counts are $0, 63, 62, 0$. The two writer factorizations use only oracle value interventions at "answer" and "is" with the original common three-copy background. This confirmation does not refit the full dose curve, test matched-dose pairs, or repeat the earlier key-only and full semantic-control grids. The new colon-self control covers both frozen reader layers; the initial panel tested layer 11 only. Of the three greedy failures under the full native coalition, one retains the recipient answer and two produce other numeric answers. One of the latter still ranks the donor first under candidate-sequence scoring, explaining the 62/64 sequence versus 61/64 greedy transfer counts. All three cases remain in the reported denominator.

Table 8: **Complete position and reader outcomes on the 64-pair confirmation panel.** Every entry is a count out of 64. F_D is forward donor transfer; R_R is reverse recipient recovery; R_D is donor retention after reverse overwrite. The bidirectional criterion requires $F_D/64, R_R/64 \geq 0.8$ and $R_D/64 \leq 0.25$. The 16 rows are dependent conditions, not additional pairs. Native proper subsets all fail the same criterion. Wrong-span controls retain the donor on 61/64 greedy reverse runs and recover the recipient on 1/64; they are not zero-effect claims.

Intervention / positions	Sequence			Greedy		
	F_D	R_R	R_D	F_D	R_R	R_D
Native values						
The	1	11	47	1	9	41
answer	1	12	45	1	10	39
is	0	15	42	0	13	33
The + answer	42	63	0	33	61	0
The + is	45	61	1	39	60	1
answer + is	47	61	1	41	59	1
The + answer + is	62	64	0	61	64	0
Attention-matched mean difference, unit dose						
The	62	64	0	62	64	0
answer	62	64	0	61	64	0
is	62	64	0	61	64	0
The + answer + is	62	64	0	61	64	0
Direct reader contributions						
Three prefix positions	62	64	0	61	64	0
Prompt tail + BOT	0	1	62	0	1	61
Latent tail	0	1	62	0	1	61
EOT	0	1	62	0	1	61
Colon self	0	1	62	0	1	61

Table 9: **Fixed writer factorizations on 64 new pairs.** S_D and G_D are sequence and greedy donor-answer counts out of 64. Mean projections and answer counts use value-only oracle interventions and the common three-copy readout. Baseline and crossed cells are all shown. U_1P_0 preserves most of the writer projection; U_0P_1 has a small nonzero influence but no donor winners. T_1F_0 preserves essentially the full projection after restoring the final-latent state and later history. All factor-parent prefixes and tested next tokens are natural. This readout tests the newly written vector in a shared background, not the unmodified downstream trajectory.

Writer factors	Cell	"answer"			"is"		
		β	S_D	G_D	β	S_D	G_D
Upstream / prefix	U_0P_0	0.000000	0	0	0.000000	0	0
	U_1P_1	1.000000	62	61	1.000000	62	61
	U_0P_1	0.043032	0	0	0.066737	0	0
	U_1P_0	0.945198	62	62	0.915761	60	58
Tail / final latent	T_0F_0	0.000000	0	0	0.000000	0	0
	T_1F_1	1.000000	62	62	1.000000	60	58
	T_0F_1	0.000004	0	0	0.000007	0	0
	T_1F_0	0.999996	62	62	0.999987	60	58

Uncertainty and execution. Descriptive 95% Wilson intervals are 87.1–98.4% for 61/64, 89.3–99.1% for 62/64, and 94.3–100% for 64/64. They use a binomial working model for pair-level outcomes. The inputs are rule-selected, so these are descriptive model-based intervals, not probability-sampling guarantees or simultaneous confidence bounds for all mechanism criteria. Two historical pairs reproduce 464 comparable recorded fields before fresh inference. The formal run uses one H100 for 9m17s including startup (506.50s in the measured model phase), with peak allocated/reserved CUDA memory 0.442/0.467 GiB and no training. PyTorch 2.5.1, Transformers 5.7.0 and PEFT 0.19.1 retain the evaluated FP16 runtime. Exact replay errors are zero; the maximum manual-attention reconstruction error is 0.005187, below the 0.01 tolerance specified in the fresh driver. CPU reconstruction from per-pair outputs reproduces all counts and verdicts. This reconstruction checks reported measurements; it is separate from executing the model interventions.

B Public Checkpoints and Execution Semantics

Checkpoint identities. The two prospectively fixed arms are `zen-E/CODI-llama3.2-1b-Instruct` and `internlm/SIM_COT-LLaMA3-CODI-1B`; the links identify the evaluated revisions. Both use the Llama-3.2-1B-Instruct backbone, six latent iterations, LoRA rank 128/alpha 32, and a hidden-to-2048-to-hidden GELU projection followed by layer normalization. The recorded inference environment uses Python 3.10, PyTorch 2.5.1, Transformers 4.46.2, PEFT 0.15.2, Safetensors 0.7.0, and eager attention. The base model and projection use BF16; PEFT's LoRA parameters retain FP32. Casting the entire model to BF16 changes the loaded adapter values and does not reproduce the recorded execution. Loading requires an exact match to the parameters used by the answer-generation model. Zen-E's four auxiliary dynamic-classifier tensors and SIM-CoT's auxiliary decoder subtree are excluded because they are not constructed or called by the answer-generation model; remaining missing or extra keys invalidate loading.

Boundary-study checkpoint names. CODI–Llama-1B in the targeted cross-check and transport–sensitivity studies denotes `connordilgren/llama32-1b-gsm8k-codi`, using `pytorch_model.bin`. It is a separate released checkpoint from the Zen-E arm above; the targeted cross-check and the boundary studies reuse this same checkpoint. Coconut–GPT-2 denotes `connordilgren/gpt2-gsm8k-coconut`, using `checkpoint_33`. The links specify the evaluated revisions. CODI–GPT-2, used for the detailed mechanism, is identified in Appendix A.

Input and latent boundaries. Tokenize the question with the base tokenizer's special-token handling and no added chat template; append BOT. If the encoded question has width n, its cache span is $[0, n)$, BOT is $[n, n+1)$, six executed latent inputs occupy $[n+1, n+7)$, and EOT occupies $[n+7, n+8)$. Each latent input is the projected last-layer hidden state from the preceding execution. Cache spans refer to inputs actually processed by the model, not placeholders counted from an output string. Recipient and donor are separately tokenized and must meet the prespecified equal-width constraint. Original prompt and BOT remain recipient-owned in the latent-value intervention.

Four-cell replay. Starting from the cache immediately before EOT, form L_0 from the recipient and L_1 by replacing only the six native-latent values with donor values in every layer. Recompute EOT separately for each branch and greedily follow the shared prefix, requiring token IDs $[791, 4320, 374, 25, 220]$. Save the contextual K/V after the first four prefix tokens and before the trailing-space reader; this four-token span is G. Construct L_0G_1 and L_1G_0 by copying only that span between the two branch caches. Recompute the trailing-space token once in every cell, including identity cells, before decoding up to 16 answer tokens. Do not transplant the already-computed reader cache or reuse its logits: either would obscure which background computes the final answer. EOS and the native turn-end token stop generation; the latent-only vocabulary item is excluded from answer-time argmax, matching the native runtime. A parent that does not generate the prespecified prefix contributes a non-estimable path, not a teacher-forced replacement or a resampled example. Record generated tokens/text, parsed recipient/source/other answers, and natural-path eligibility separately.

Targeted independent-checkpoint relay cross-check. After the sufficiency screen identified a natural five-token prefix and latent transport in travel and production, a separate protocol freezes the

four factorial cells, an all-layer six-slot native-latent value swap, fresh ascending model-blind data, and fixed 13/16 winner and 4/16 competitor thresholds before relay inference. The 48 matrices have zero exact-prompt or candidate overlap with the preceding prespecified suites. All 96 clean recipient/donor generations are correct and naturally prefixed; every one of the 192 four-cell outcomes is parseable and parent-path natural. Travel donor counts across $L_0G_0, L_1G_1, L_0G_1, L_1G_0$ are $0, 16, 16, 0$ of 16, and production counts are $0, 14, 14, 0$; reverse overwrite restores the recipient on 16/16 in both strata. Equal groups retains the recipient on 16/16 and transfers the source on 0/16 in every cell. All 96 cache-patch checks and exact identity replays pass, as does an independent result reconstruction. The checkpoint is trained separately from the main GPT-2 relay checkpoint and was selected after the earlier transport screen, so it establishes the relay on a second checkpoint while leaving checkpoint success rate, a shared minimal coalition, and training-seed replication for a checkpoint-blind sample.

Executable supplement and scope. The execution package contains the original 32 qualification prompts and 144 matrices, preserving question text, answers, order, and schema membership under anonymous matrix identifiers. It includes checkpoint revisions, the four-cell runtime, and recorded per-cell tokens and outcomes. Replaying both public arms matches all 1,152 intervention outputs token for token, with 32/32 native qualification sequences matching per arm. All factorial outputs are parseable and naturally supported. Identity, locality, load, row-key, and classification checks pass. These replays use the existing matrices. The separate GPT-2 environment supplies the earlier 8/8 discovery/validation panel, the existing 16-pair held-out relay panel, the position, dose, and writer routines, and CPU reconstruction of the initial mechanism Tables 4–7. The separate 64-pair confirmation record reconstructs Tables 2, 8 and 9 from its per-pair outputs and includes the evaluated runtime for replaying all 60 conditions on each fixed pair. Model weights and the pinned upstream CODI module are obtained from their public providers; the package includes the evaluated source patch and acquisition commands. The content-control extension includes both same-panel diagnostics and the separate 192-matrix fresh panel, its local protocol, historical selection exclusions, and all 3,072 recorded cell outputs with complete-answer token scores. Its CPU entry point rebuilds Tables 3B and 10–12; model replay reuses the original cache operators. The earlier outcome archive remains a count and reporting-logic audit for the wider study. Secondary studies outside these execution paths are not claimed as end-to-end reproductions. The fresh panel tests new inputs rather than reproducing the prespecified 48-matrix counts.

C Latent Content Controls and Fresh-Input Prediction

Intervention and development evidence. The original four-cell background bundles the native latent values with the latent-end delimiter's downstream K/V state. We first separate latent V (L), delimiter K/V (E), and generated-prefix K/V (G) in an eight-cell diagnostic on the original 48 matrices per schema and checkpoint. If the prompt has n tokens, these spans are $[n+1, n+7)$, $[n+7, n+8)$, and $[n+8, n+12)$; the final prefix token at $n+12$ is recomputed as the reader. The delimiter is the model's latent-end token, not a chat turn terminator. With the original donor-patched prefix G_1 fixed, restoring donor latent V alone in a clean E background gives 47/48 travel and 48/48 production donor hits in Zen-E, versus 11/48 and 17/48 with recipient latent. Restoring only E gives 11/48 and 18/48: E has one observed answer effect but does not explain most of the deficit. Adding E to donor L/G changes no generated tokens on either checkpoint's 144 matrices.

A second diagnostic fixes E clean and crosses G_0/G_1 with recipient, original-donor, same-answer, and different-answer latent V. All other cache state remains recipient-side. Each additional source must generate its correct answer naturally, and its latent-patched recipient parent must reach the natural prefix; G_1 always comes from the *original* donor-patched parent. With G_1 fixed, same-answer latent values reproduce the original-donor latent condition's output token for token on all 144 matrices per arm. Different-answer latent values give 33/48 and 39/48 original-donor hits in Zen-E, restoring 22 and losing zero in each task. These two diagnostics reuse the original panel, add no independent units, and motivate the following prediction; they are not retroactively confirmatory.

New inputs and prespecified test. Before generating new data or observing its model outcomes, we locally fixed 64 matrices per schema, both known checkpoints, all eight content-control cells, and the analysis below. This is a prospectively fixed local protocol, not an externally registered or checkpoint-blind study. The original arithmetic generator scans indices 200000–265535 in ascending

order and selects the first eligible matrices. Eligibility depends only on semantic validity, prompt overlap, role-wise numeric-configuration overlap, and equal within-matrix token width under both tokenizers; it never depends on model answers. We exclude 4,568 historical prompt strings and 757 six-role numeric configurations, including renamed versions of those configurations. The selected 192 matrices contain 1,152 distinct prompts. Names, number families, and templates are reused; all roles share the additive constant in travel and production. After exclusion, travel uses template 0 on 63 matrices and template 2 on one, production uses template 2 throughout, and equal groups uses templates 0/1/2/3 on 13/13/24/14 matrices. This tests new combinations of inputs within the selected distribution, not generalization to unseen templates.

The primary comparison is original-donor recovery with different-answer versus recipient latent, at fixed G_1 and clean E. For each Zen-E task, let b count newly restored donor answers and c lost donor answers. Each of u non-comparable matrices is conservatively counted as an additional loss. The fixed one-sided paired test uses

$$p = 2^{-(b+c+u)} \sum_{k=b}^{b+c+u} \binom{b+c+u}{k}, \qquad \widehat{\Delta} = (b-c-u)/64.$$

Both Zen-E tasks must have positive gain and $p \leq 0.025$, with Bonferroni control over the two primary tests. SIM-CoT is secondary and equal groups is the negative control. At the planning probabilities of 0.35 restoration and 0.05 loss, the exact test has 0.9801 power per task at $n = 64$; these are planning assumptions. Descriptive percentile intervals use 10,000 paired matrix bootstraps with seed 40. This is model-based paired uncertainty for the selected input family, not probability-sampling inference over checkpoints or tasks; cells, schemas, and reuse across checkpoints are not pooled.

Execution and results. Historical-input smoke checks precede the one full new-input run; no new-input smoke, replacement, or sample expansion occurs. All 64 old qualification sequences match their native references. Every matrix and additional source qualifies, all 3,072 cell outputs are available and parseable, and all complete-answer scores are finite. CPU tests cover cache composition, independent score reconstruction, selection, and conservative paired inference. Base/projection BF16 and LoRA FP32 match the validated runtime. Table 10 reports the primary and secondary paired contrasts; Table 11 retains all output categories.

Table 10: **Fresh-input rescue with the original donor-patched prefix fixed and E clean.** Counts are original-donor hits, not the different source's own answer. Both Zen-E primary tests pass the fixed 0.025 threshold. SIM-CoT is secondary; its production interval spans zero. Intervals are unadjusted descriptive paired bootstraps. Every task has 64 comparable matrices and zero qualification failures.

Arm	Task	Original-donor hits		Restored/ lost	Gain (pp)	95% interval (pp)	p
		Recipient L	Different L	b/c			
Zen-E	Travel	17/64	42/64	27/2	39.1	[25.0, 53.1]	8.12×10^{-7}
	Production	20/64	43/64	25/2	35.9	[21.9, 50.0]	2.82×10^{-6}
SIM-CoT	Travel	54/64	61/64	7/0	10.9	[4.7, 18.8]	0.0078
	Production	59/64	61/64	5/3	3.1	[-4.7, 10.9]	0.3633

Table 11: **Complete fresh-input content controls.** Each entry is recipient / original donor / different source / other, out of 64; unparseable and unavailable counts are zero in every cell. E is always clean. Equal groups, omitted from the columns, is 64/0/0/0 in every cell of both checkpoints (1,024 dependent outputs). With G_1 fixed, same-answer latent values reproduce the original-donor latent condition's output token for token on all 192 matrices per checkpoint.

Latent source	**Prefix**	**Zen-E**		**SIM-CoT**	
		Travel	Production	Travel	Production
Recipient	G_0	64/0/0/0	64/0/0/0	64/0/0/0	64/0/0/0
	G_1	37/17/0/10	38/20/0/6	7/54/0/3	2/59/0/3
Original donor	G_0	64/0/0/0	63/0/0/1	61/0/0/3	63/0/0/1
	G_1	0/64/0/0	0/64/0/0	0/64/0/0	0/64/0/0
Same-answer	G_0	64/0/0/0	63/0/0/1	61/0/0/3	63/0/0/1
	G_1	0/64/0/0	0/64/0/0	0/64/0/0	0/64/0/0
Different-answer	G_0	63/0/0/1	61/0/0/3	61/0/1/2	63/0/0/1
	G_1	3/42/5/14	2/43/10/9	0/61/3/0	0/61/1/2

Table 12: **Continuous effects coexist with greedy prefix sufficiency.** Complete-answer log-probability margins are $m = \log p(a_D) - \log p(a_R)$ and $q = \log p(a_M) - \log p(a_D)$, where M is the different-answer source. Entries are mean changes from recipient latent at the same G_1, in natural-log units. "Positive" counts positive different-answer Δm. Both checkpoints show donor support and increasing relative support for the different source's answer.

Arm	**Task**	**Donor margin Δm**		**Positive different Δm**	**Different-source margin Δq**
		Same-answer L	Different L		
Zen-E	Travel	11.4087	5.8096	64/64	4.3018
	Production	13.0500	6.7295	64/64	5.3511
SIM-CoT	Travel	10.0479	5.2791	63/64	5.1960
	Production	8.3588	4.6816	63/64	4.7179

Interpretation and limits. The new inputs confirm Zen-E's positive net rescue prediction, not monotonic rescue: each task loses two baseline donor hits, all to other numeric answers. Same-answer controls preserve the original rescue, while different-answer sources are weaker and sometimes redirect generation to their own answers. G_0 also has exceptions, including one different-donor answer in SIM-CoT travel. Thus upstream support persists without final-answer matching, while content influences the output; these controls neither isolate an answer-independent latent component nor separate norm, shared subcomputation, and answer content. As a descriptive readout, each Zen-E task has 64/64 matrices with donor-versus-recipient margin ordered same-answer latent $>$ different-answer latent $>$ recipient latent. Nevertheless, two donor hits per task are lost to other answers; this two-answer margin does not determine unrestricted generation. The same-answer source changes factorization but preserves the product; the different-answer source still shares the task and additive subproblem. Table 12 further shows that the original relay sufficiency label for SIM-CoT does not imply zero latent contribution. This follow-up tests new inputs on known checkpoints; it neither discovers a new checkpoint nor confirms GPT-2's exact heads or coalition.

D WHOLE-SEGMENT AND BOUNDARY STUDIES

Table 13: Whole-segment and boundary-study counts. Counts summarize fixed input panels separately by schema and unit type. Fixed-mechanism outcomes appear in Tables 8–9; fresh content controls appear in Tables 10–12.

Experiment	Observed result	Conclusion / scope
Balls	$15/16$ latent greedy at t_0; $16/16$ prefix greedy at t_4; four-cell greedy L_1G_1/L_0G_1 both $15/16$	prospective relay mechanism study
Books	$15/16$ latent; $16/16$ prefix; semantic relay-coalition criterion met	whole-segment relay replication
Coins	$16/16$ latent; $16/16$ prefix; random-source specificity $11/16$	behavioral replication only
Natural travel / production	latent 15–$16/16$ / $16/16$; raw/BOT $0/16$	prospective held-out confirmation
Natural equal groups	raw $16/16$; prespecified and full latent $0/16$, five roles	prospective held-out sufficiency
Equal-groups onset analysis	raw key/value retention $8/16, 5/16$; full prefix $0/16$; latent all $16/16$	prospective follow-up; limited scope
Latent-route binding analysis	latent/raw source $16/16, 0/16$; raw/latent unbinding retention 0–$1/16, 15$–$16/16$	prospective latent transport with raw-unbinding dependence
GSM8K-trained GPT-2 Coconut	qualification $32/32$; latent-KV donor travel/equal/production $15/16, 16/16, 16/16$ at t_0, t_1; raw/BOT $0/16$ each; descriptive total $47/48$	public native-latent transport; dependence indeterminate
GSM8K-trained Llama-1B CODI	qualification $32/32$; held-out clean $96/96$ per schema; travel latent-KV/V $16/16$, production latent-V $15/16$, equal raw $16/16$, unchanged t_0–t_5	prospective same-prefix boundary condition; sufficiency only
Llama-1B targeted relay cross-check	clean/natural recipient and donor $16/16$ per schema; travel $L_1G_1/L_0G_1/L_1G_0$ source $16/16, 16/16, 0/16$; production $14/16, 14/16, 0/16$; equal source $0/16$ in all cells	independently verified relay sufficiency on two latent-carried strata; prompt-carried negative control; checkpoint selection is targeted
Zen-E prospective checkpoint test	qualification $32/32$; travel $L_1G_1/L_0G_1/L_1G_0$ source $47/48, 11/48, 0/48$; production $48/48, 17/48, 1/48$; equal source $0/48$ in every cell	prefix insufficient alone on both primary schemas; exact negative-control pass; no checkpoint or intervention search
SIM-CoT prospective checkpoint test	qualification $32/32$; travel $L_1G_1/L_0G_1/L_1G_0$ source $48/48, 41/48, 1/48$; production $48/48, 42/48, 0/48$; equal source $0/48$ in every cell	relay sufficiency on both primary schemas; exact negative-control pass; relay replicated in this checkpoint

Continued on the next page

Table 13 (continued)

Experiment	Observed result	Conclusion / scope
Matched endpoint + boundary-local analysis	Endpoint donor: travel latent-KV/V $45/48, 47/48$, production $46/48, 47/48$, equal raw $48/48$. Boundary-local t_5 raw/latent K/V/I: travel/production $0/48, 0/48, 0/48$ vs. $48/48, 48/48, 48/48$, equal $16/48, 0/48, 0/48$ vs. $48/48, 48/48, 48/48$	Joint endpoint unresolved; schema-level dissociation; exploratory analysis adds no independent matrices; equal/joint indeterminate

Coconut binding tests. Latent dependence remains indeterminate in this checkpoint. At t_0, paired K/V nulls retain all 48 distinct units, while independent-unbinding retention varies across the three 16-unit strata and meets neither prespecified classification; the totals are 13 retained answers after raw unbinding and 31 after latent unbinding.

Practical and variable-length use. A staged application can first enforce clean qualification, then scan coarse segments for donor transport, run destructive/null contrasts only on qualified candidates and competitors, and reserve layer/head localization for supported cells. The evaluated checkpoints expose fixed latent interfaces; for variable-length rollouts, boundaries should instead be indexed by each trajectory's actual stopping point, with source and recipient states aligned by the final k executed latent positions or a prespecified task event. Carrier conclusions should then be reported separately for groups with the same stopping length, and unequal or unsupported alignments should remain non-estimable rather than being filled by a fixed global suffix.

E Inference Scope and Decision Sensitivity

Relay classifications concern the specified crossed-cache operator and answer criterion. Relay sufficiency can coexist with continuous upstream effects, as the SIM-CoT content controls demonstrate (Appendix C). Transport and destructive retention are reported separately, with natural-path eligibility determining which contrasts can be estimated. Table 15 summarizes the scope of each study.

1. Fix units, strata, intervention and observation boundaries, operators, semantic controls, and primary/joint decision criteria before the corresponding evaluation.
2. Require correct, parseable clean outputs and faithful cache/identity replay. Track natural-path eligibility for each edited branch; unsupported contrasts are non-estimable.
3. Measure unrestricted donor transfer for candidate segments, with the protocol's same-answer, partial-variable, random-source, and wrong-location controls.
4. For a proposed relay, construct clean and latent-patched parents, cross their L and G caches, and recompute the reader in every cell. Report relay sufficiency only when both baseline and crossed-cell criteria pass.
5. Separately measure recipient retention under the selected destructive operators, alongside identity replay and paired algebraic nulls.
6. Report relay classification and transport/retention profiles separately. Report concordance when qualified transport and destructive/null evidence implicate the same segment; otherwise distinguish dissociation, indeterminacy, and non-estimability. Combine counts only within fixed homogeneous strata and retain each study's prospective, discovery, or post hoc status.

Algorithm 1: BATS relay testing and separate transport/retention analysis

Operator-relative estimands and identification. Fix checkpoint θ, generator g, finite set Q, intervention boundary u, and operator bundle O, including reader replay. The observation boundary t may follow the cache edit at u. With natural-path eligibility $E_O(u, t)$, the transport and retention

estimands are

$$\tau^{\mathrm{car}}_{X,u\to t}(\theta, g, Q, O) = \mathbb{E}_Q[\mathbf{1}\{y(R[X_u \leftarrow_O S]) = y(S)\} \mid E_O(u,t) = 1], \tag{5}$$

$$\tau^{\mathrm{ret}}_{X,u\to t}(\theta, g, Q, O) = \mathbb{E}_Q[\mathbf{1}\{y(R[D_O(X_u)]) = y(R)\} \mid E_O(u,t) = 1]. \tag{6}$$

Reported rates are exact finite-set proportions over Q, conditional on (θ, g, Q, O). Identification assumes operator support on the qualified units, consistency of boundary replay and operator semantics (checked by identity replay), and no interference across independent units; roles and repeated boundaries within a unit may interact, so we treat them as dependent. Later observations condition on the edited branch reaching t naturally; they need not involve a new intervention at t. The contrast is not estimable without comparable eligible trajectories.

Response-table observation (non-identification without cross-intervention restrictions). Fix a checkpoint, generator, boundary, specified operator bundle, and common eligible finite set. Assume replay and intervention consistency, boundary-local replacement, no interference across independent units, distinct intervention states, and no structural restriction linking donor-patched to destructively perturbed potential outputs. In the unrestricted class of deterministic acyclic boundary-response systems, qualified donor-transfer and destructive-retention rates are variation independent: either rate can vary without constraining the other, even with clean, identity, paired-null, and semantic-control outcomes fixed. Consequently, either intervention family alone leaves the other unidentified; observing both identifies only their specified finite-set response profile.

Proof. For one unit, let recipient, donor, and a valid third output be r, s, w, with $r \neq s$ and $w \neq r$, and hold every clean, identity, paired-null, and control output at the same specified value in all systems. For the donor-patched and destructive states respectively, define four deterministic response tables with outputs

$$(r, r), \qquad (s, r), \qquad (r, w), \qquad (s, w).$$

These tables agree on all fixed control outcomes but realize all four combinations of transfer and recipient retention. Each is implementable by an acyclic lookup function with a distinct input for each intervention state; additional control states can be assigned identical outputs. Choosing the four tables independently across units yields any attainable pair of finite-set counts. A different operator can reach an unconstrained fifth state, so the observed profile cannot identify an operator-invariant route. □

The observation requires common eligibility. If a destructive branch never reaches the boundary naturally, its retention rate is undefined. If measured rates leave a classification indeterminate, the other intervention family does not resolve it. The construction is deliberately unrestricted, and its role is to rule out the logical implications transport ⇒ destructive dependence and destructive dependence ⇒ transport in the absence of linking assumptions. How the two relate under particular transformer structure, and what would positively identify a natural route, are separate questions.

Table 14: **Why isolated intervention readings are insufficient.** Each row shows how omitting one requirement changes the interpretation of the same six observed endpoints; no experiment is rerun and no missing outcome is imputed. Donor interchange does not measure destructive retention, destructive unbinding lacks donor identity, and the paired algebraic null checks implementation symmetry rather than severity.

Omission	**Effect on endpoint interpretation**	**Unsupported conclusion prevented**
Patching alone (omit destructive/null evidence)	All six endpoints retain a unique donor-transport winner; the t_0 disagreement between transport and dependence can no longer be identified.	Prevents equating answer transport with operator-relative dependence evidence.
Natural-path criterion	Using off-path outcomes would wrongly support a t_5 classification for travel and production; equal groups remains indeterminate.	Prevents treating off-path destructive outcomes as naturally reachable.
Paired K/V algebraic null	The six endpoint labels do not change, but the implementation symmetry check is removed.	Detects violations of the intended permutation symmetry, indexing errors, or locality bugs; it does not match destructive severity or by itself identify binding-specific dependence.

Study-design status. Relay rules and the matched-endpoint units, endpoints, and decision criteria were specified before inspecting their corresponding outcomes. The t_0 decomposition is secondary under the same prespecified criteria. The boundary-local t_5 analysis was designed after the endpoint results and reuses the same strata; it is therefore post hoc, adds no independent units, and does not resolve the primary joint endpoint.

Table 15: **Evidence hierarchy for the boundary-resolved claims.** Repeated boundaries, factorial cells, and reuse across checkpoints are dependent diagnostics; checkpoint evidence is reported per fixed arm rather than pooled.

Empirical target	Independent variation available	Supported scope / remaining gap
Resolved latent-to-prefix relay	Initial GPT-2 localization (8 discovery + 8 validation), separate fixed-mechanism confirmation (64 new pairs), and a screened CODI–Llama-1B four-cell cross-check	Detailed mechanism confirmed on new GPT-2 inputs; the second checkpoint tests the relay between whole segments, not shared heads or a minimal circuit.
Prospective checkpoint replication	Two prespecified public checkpoints; the same 48 model-blind matrices per schema and four dependent cells	SIM-CoT supports relay sufficiency; Zen-E's prefix is insufficient alone. This establishes existence beyond the screened checkpoint, not prevalence, additional independent task units beyond the shared 48, or seed-controlled generalization.
Persistence across prefix onset	Three separate 16-matrix strata in an official Coconut–GPT-2 checkpoint, evaluated at t_0 and after its complete one-token prefix	Refutes a universal immediate-handoff rule in this setting; the checkpoint offers a single-token prefix, so it yields transport evidence with an indeterminate dependence profile.
Same-prefix transport–dependence endpoint	Three disjoint 48-matrix strata within one public CODI–Llama-1B checkpoint	Secondary t_0 dissociation for travel and production; the joint t_0/t_5 rule remains unresolved and has no independent-checkpoint replication.

Table 16: **Exhaustive integer grids for transport and dependence count thresholds.** For each row, only the indicated integer cap/floor pair varies over $0 \le c < h \le 48$; observed counts, natural support, and every other criterion remain fixed. Here c is the low-side cap, h the high-side floor, and tr marks transport. At $(12, 39)$, equal groups has zero cap margin, which reflects decision-rule sensitivity rather than sampling or effect-size uncertainty, and no pair resolves t_5. Structural footprint, trajectory support, and eligible output diagnostics are reported in Table 18.

Analysis	Exact label-preserving region	Threshold pairs / 1176	Cap/floor margin
Transport: all schemas, roles, endpoints	$3 \le c_{\mathrm{tr}} < h_{\mathrm{tr}} \le 44$	861	$9/5$
Secondary t_0: travel/production dependence	$0 \le c < h \le 44$	990	$12/5$
Secondary t_0: equal-groups dependence	$12 \le c < h \le 48$	666	$0/9$
Joint t_0, t_5 endpoint	unresolved for all $0 \le c < h \le 48$	1176	n/a ($t_5 : 0/48$)

F Comparison with Observational Baselines

Models, inputs, and boundaries. Table 17 reuses the earlier 16 CODI–GPT-2 Balls pairs (32 recipient/donor trajectories) and a separate relation-chain panel of 16 2×2 matrices (64 prompts, grouped by matrix). The relation-chain checkpoint (V7.1, seed 40) uses Llama-3.2-1B-Instruct with six latent steps, evaluated on two-step person–badge–room lookup with eight room-name answers. GPT-2 is evaluated before and after its natural "The answer is:" prefix. The relation-chain model normally answers directly; its second boundary follows an imposed copy of that prefix. Intervention-based reference labels were fixed before these observational comparisons.

Features and ranking. Probe features concatenate all layers' K and V after mean-pooling within each segment. Four outer folds keep each pair or matrix intact; PCA (at most 16 components) and scaling are fit on training groups only, with ridge strength selected by grouped inner validation from $\{0.01, 0.1, 1, 10, 100\}$. The GPT-2 probe predicts the numeric answer and ranks segments by lowest held-out mean absolute error normalized by the training-answer standard deviation; the relation-chain probe predicts a one-hot target over eight room names and ranks by highest held-out accuracy. Other scores rank by highest attention mass per token (averaged over heads, layers, and inputs), K/V RMS, or counterfactual cosine distance. Cosine uses the same pooled vectors: recipient/donor distance for GPT-2, and the mean of all six prompt-pair distances within each relation-chain matrix.

Table 17: **Observational attribution versus causal intervention.** Parentheses give the prespecified BATS attribution; bold cells match. R/B/L/G denote raw prompt/BOT/native latent/generated prefix; RMS denotes root-mean-square magnitude. Boundary agreement counts matching segment rankings across these four conditions, not correct answers across inputs. Only the relation-chain "Forced" condition uses an imposed prefix; the other conditions follow natural trajectories.

Attribution score	**CODI–GPT-2**		**Relation chain**		**Boundary agreement**
	t_0 (L)	t_4 (G)	t_0 (L)	Forced (R)	
Nested ridge probe	R	**G**	**L**	L	2/4
Attention per token	R	**G**	**L**	**R**	3/4
K/V RMS	R	R	B	B	0/4
Counterfactual cosine	**L**	L	**L**	L	2/4

No score matches the intervention-based reference at all four boundaries. These are dependent comparisons within two existing panels, not four independent studies or new relay confirmations.

G Controls, Exclusions, and Indeterminate Results

The following observations constrain the interpretation:

- **Operator and mechanism scope.** A colon alone and the suffix "Answer:" preserve the relation-chain latent route; matched width and absolute position are insufficient explanations.
- Multiple reordered forced prefixes reconfigure routing priority, so canonical grammatical order is not necessary; the full lexical/compositional trigger rule remains unresolved.
- Generated history does not acquire donor specificity in the relation-chain forced-prefix setting, ruling it out as the terminal donor carrier there.
- Paired (K, V) permutation preserves attention outputs and is used only as a null; key-only, value-only, and independent unbinding provide the destructive binding tests.
- Proposed single-edge and portable-direction explanations fail their prespecified discovery or donor-specificity criteria, so we leave both mechanisms open.
- **Qualification and unavailable cells.** The earlier 48-pair GPT-2 range extension misses its prespecified clean-validity criterion: the lowest accuracy across the six clean roles is 43/48, below the required 47/48. Although its intervention criteria pass, its counts remain

descriptive, and no examples are resampled. This failed panel is separate from the later 64-pair confirmation.

- An earlier relation-chain screen finds no naturally emitted standard prefix among its eligible candidates, so forced-prefix evidence is never substituted for a natural relay. The targeted Llama cross-check instead uses a qualified checkpoint with its own naturally emitted five-token prefix.
- **Natural-prefix and endpoint estimability.** The public GSM8K-trained Coconut checkpoint gives a second native-latent transport result. Its complete natural shared prefix is the single token "###", with no generated-prefix relay coalition or later boundary before answer content, so a full relay and a joint assessment of transport and dependence are not testable on this checkpoint.
- Its multiset-preserving binding analysis is indeterminate: paired K/V nulls are exact 48/48, but raw and latent destructive retention falls between the prespecified concordance and dissociation thresholds and differs by schema. Donor transport is therefore not relabeled as concordant transport and dependence.
- The six-boundary natural-prefix sufficiency screen rules out prefix identity as a universal routing rule in one public CODI–Llama-1B checkpoint: the identical natural five-token prefix preserves latent donor sufficiency for travel and production but prompt donor sufficiency for equal groups from t_0 through t_5. The screen measures sufficiency only; destructive binding tests and paired nulls are supplied by the separate endpoint study (Appendix H).
- Separate prospective onset and latent-route analyses support one narrow operator-specific pattern: raw unbinding is destructive while latent donor transport persists. Full-prefix prompt dependence, uniform latent-binding dependence, and equation-level causation lie beyond this pattern.

H Implementation and Data Integrity

Runtime qualification. Before inference, each condition freezes examples, answer-width constraints, segment boundaries, operators, and decision criteria, and must pass clean generation, parsing, native-wrapper equivalence, latent-step replay, finite-logit, boundary, identity-replay, and independent-cache checks; failure withholds carrier attribution.

All recipient/source boundaries are independently tokenized and replayed. The relation-chain data audit checks online matrix generation, tokenizer-level equal width, semantic role alignment, and train/evaluation overlap. Unrestricted greedy generation separates clean validity, parse validity, recipient answer, donor answer, and other answers or unparseable outputs. The code audit reproduces winner and role conclusions after correcting ambiguous diagnostic field names; the affected names were not used to define the headline flip.

Known-route calibration. We calibrate the response-profile rule on three associative K/V systems whose routes are hidden from the inference procedure, each constructed to read exactly one of raw prompt, native latent, or generated prefix over 48 independent units. The inference function receives observations without route labels and fixes all three predictions before the expected routes are revealed; all three constructed routes are classified correctly. For each system's constructed route, identity and semantic-donor outcomes are 48/48, key-only, value-only, and independent K/V unbinding each retain 0/48 recipient answers, and paired K/V permutation retains 48/48. This checks route-blind prediction and classification logic on controlled associative systems, with a transformer-native three-route calibration left open. Separately, on a fixed prompt-only transformer graph, BATS recovers raw prompt on 64/64 relation-chain cells and 16/16 matrices, while BOT and masked latent transfer 0/64. The engineered transformer latent-only and generated-prefix-only variants fail clean-generation qualification despite exact mask/no-mask API equivalence; they therefore cannot calibrate transformer-native routes.

Observational-baseline analysis. Probe folds keep complete counterfactual units together and fit preprocessing only on training groups. All four boundary rankings in Appendix F are recomputed with the same segment labels used by the intervention analysis.

Six-boundary natural-prefix sufficiency screen. The four checkpoint identities are fixed by repository revision and model/tokenizer identity. Data are generated before inference: 48 qualification and 48 held-out matrices have zero prompt overlap with each other and with historical suites, and every prespecified tokenizer-width bucket meets its width criterion. The trajectory exporter validates semantic spans against actual cache width and aborts evaluation if any K/V snapshot shares tensor storage with another snapshot; the independent verification program separately reconstructs boundaries, answers, row keys, natural-prefix flags, classifications, and all reported counts. All qualified rows follow natural replay paths rather than conditional replay, and all six independent result checks pass.

Prespecified transport–retention evaluation. All cache edits occur before EOT. The t_0, t_5 labels index observations after that same early edit, with t_5 eligible only when the edited branch itself generates the complete prefix; no new patch is applied at t_5. Using the same official CODI–Llama-1B checkpoint, this evaluation analyzes 48 held-out six-role matrices per schema, generated before inference and disjoint from the prior suites. Every schema–role cell is 48/48 correct, parseable, and naturally prefixed. The joint endpoint remains unresolved because t_5 dependence is not estimable. At both endpoints, every sufficiency condition is naturally eligible on 48/48 matrices; transfer rates ordered as raw prompt, BOT, latent K/V, and latent V are $0/48, 0/48, 45/48, 47/48$ for travel, $3/48, 0/48, 46/48, 47/48$ for production, and $48/48, 0/48, 0/48, 0/48$ for equal groups. Each of four additional donor roles meets the same-route criterion under the prespecified carrier $\geq 39/48$ and competitor $\leq 12/48$ thresholds.

In the t_0 schema-level analysis, reported as secondary under prespecified criteria, raw versus latent K/V/I unbinding retains $0/48, 0/48, 0/48$ versus $44/48, 48/48, 48/48$ recipients for travel and production, and $12/48, 0/48, 0/48$ versus $48/48, 48/48, 48/48$ for equal groups. Identity and paired K/V nulls retain 48/48 throughout. These results show latent donor transport and recipient-answer loss under raw-prompt unbinding for travel and production and a threshold-bound raw-prompt concordance classification for equal groups. At t_5, raw destructive conditions have 0/48 natural-path eligibility in every schema, whereas identity, paired nulls, and latent destructive conditions have 48/48; conditional raw outcomes are excluded, leaving destructive retention non-estimable and the joint endpoint unresolved.

Independent verification reconstructs the clean, sufficiency, destructive-retention, and permutation-control counts and their complete condition grid. The independent unit is one matrix (48 per schema); roles, conditions, rows, and repeated boundaries are dependent within it.

Post hoc boundary-local t_5 intervention analysis. On the three fixed 48-matrix strata (0 new units; no pooled denominator), this post hoc follow-up forks the intact cache after prefix tokens p_0–p_3, applies one prespecified intervention, and replays p_4 for fresh logits. Every schema passes the clean, natural-prefix, reconstruction, identity, and paired-null checks on 48/48 matrices. Raw versus latent K/V/I unbinding retains $0/48, 0/48, 0/48$ versus $48/48, 48/48, 48/48$ answers for travel and production, meeting the boundary-local response-profile criterion; equal groups yields $16/48, 0/48, 0/48$ versus $48/48, 48/48, 48/48$, exceeding the prespecified 12/48 maximum and leaving its evidence indeterminate. Independent verification reconstructs the complete condition grid and counts. Designed after the endpoint results, this analysis is exploratory, and checkpoint/seed and schema/task confounding remain unresolved within it.

Post-result matched-operator robustness grid. A secondary analysis applies four masks and doses 3–6 to standard Coconut and CODI at t_0 and each checkpoint's native terminal boundary, using 48 matrices per schema. Unlike natural-path conditional retention, this grid counts invalid or prefix-mismatched outputs as zero in a fixed denominator. For each matrix, segment s, and operator o, average retention across four masks at each dose k to obtain $\bar{r}_{s,o,k}$, then compute the normalized dose-area score $A_{s,o} = (\bar{r}_{s,o,3} + 2\bar{r}_{s,o,4} + 2\bar{r}_{s,o,5} + \bar{r}_{s,o,6})/6$. The contrasts are $\Delta_{\text{ind}} = A_{\text{latent,ind}} - A_{\text{raw,ind}}$ and $U_{\text{raw}} - U_{\text{latent}}$, where $U_s = A_{s,\text{null}} - A_{s,\text{ind}}$. Positive values indicate greater raw-prompt damage under independent unbinding, before or after null correction, respectively.

CODI passes all identity and paired-null criteria. At its native terminal boundary, both contrasts are zero on all 144 matrices. At t_0, mean Δ_{ind} is 0.0208, 0.0269, and 0.0365 for equal groups, production, and travel, with positive/negative/tied matrix counts 10/0/38, 19/5/24, and 13/4/31; null correction

gives the same values here, not a general algebraic identity. All six Coconut boundary–schema cells fail the identity and paired-null criteria, so their zero contrasts are uninterpretable. No cross-schema pooling, intervals, significance tests, or dose/mask selection are performed. This post-result evidence leaves the primary verdict unchanged: CODI shows only a weak local t_0 pattern, without a stable cross-checkpoint or cross-boundary effect.

Table 18: **Operator construction, trajectory support, and eligible output impact.** Span entries are width (number of matrices); all Panel B entries are counts out of 48, and K/V/I denotes key-only, value-only, and independent unbinding. Eligible, Parseable, Retained, and Δ_{null} triplets follow K/V/I order, and Null reports eligibility/retention. For each operator, Δ_{null} is the paired-null retained count minus the destructive retained count, out of 48 matrices. Dashes mark non-estimable t_5 raw contrasts, not zero retention; identity retains 48/48 throughout. The local follow-up† reuses the same units, and the 1,152 check rows are not independent samples. Raw spans (24–36) differ from latent spans (6), so the table establishes locality, trajectory support, and output diagnostics rather than equal intervention severity or circuit localization.

A. Specified intervention spans and implementation checks

Schema	Raw-span width (number of matrices)	Latent span
Equal groups	$24(11), 25(1), 29(11), 30(13), 31(11), 32(1)$	$6(48)$
Production	$30(37), 31(9), 34(2)$	$6(48)$
Travel	$32(2), 33(1), 34(38), 36(7)$	$6(48)$

Shared footprint: matched endpoint and boundary-local follow-up†. *Implementation checks:* specified permutation, shape preservation, and outside-segment exactness each pass $1152/1152$ at the matched endpoint. The boundary-local follow-up passes the same checks plus record mapping and outside-segment exactness, each $1152/1152$.

B. Natural-path eligibility and output diagnostics

Schema	Seg.	Eligible	Parseable	Retained	Null	Δ_{null}
Matched endpoint: t_0						
Travel	raw	48/48/48	1/48/48	0/0/0	48/48	48/48/48
Production	raw	48/48/48	0/48/48	0/0/0	48/48	48/48/48
Equal groups	raw	48/48/48	41/48/48	12/0/0	48/48	36/48/48
Travel, production	latent	48/48/48	48/48/48	44/48/48	48/48	4/0/0
Equal groups	latent	48/48/48	48/48/48	48/48/48	48/48	0/0/0
Matched endpoint: t_5						
All three	raw	0/0/0	—	—	48/48	—
Boundary-local follow-up: $t_5^\dagger$						
Travel, production	raw	48/48/48	48/48/48	0/0/0	48/48	48/48/48
Equal groups	raw	48/48/48	48/48/48	16/0/0	48/48	32/48/48
All three	latent	48/48/48	48/48/48	48/48/48	48/48	0/0/0