IDENTIFYING CHANGES IN CAUSAL ATTRIBUTION UNDER REASONING STEP REPLACEMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Language models can replace written reasoning steps with continuous hidden states. How does this change the contribution of the reasoning that remains visible? We measure how a change in the input affects the answer by intervening on the states stored for the prompt, written steps, and hidden steps. These parts interact, so assigning credit to one depends on how their joint effects are divided. We compare two reasoning modes using the same rule for assigning credit. For rules based on different orderings of the parts, we derive the exact range of possible changes in attribution. Unchanged interactions cancel; only changes in interactions make this comparison ambiguous. We apply the method to GPT-2 models trained on a synthetic policy task, holding each model's weights fixed across modes. Across three adaptation runs, replacing more steps increases the contribution of states at text positions, while accuracy and correct policy use decline. None of the runs supports an increase in the contribution of hidden-step states. In matched comparisons within two models, interventions on text inputs produce smaller attribution changes than interventions on the stored states. The results separate the influence carried by text-associated states from the model's ability to apply the policy correctly.

1 INTRODUCTION

Consider a model deciding whether a policy allows an entity to enter an archive. The correct decision depends on membership facts and which policies apply to the entity. If some written reasoning steps are replaced by continuous hidden states, how does the policy's influence on that decision change? The original prompt, the remaining written steps, and the hidden steps may all carry relevant information to the answer. Reading a shorter explanation does not reveal their contributions.

This question arises in methods that perform some reasoning without emitting words. Coconut feeds hidden states back into the model in place of initial written steps; CODI learns continuous reasoning through self-distillation (Hao et al., 2025; Shen et al., 2025). Even explicit chains of thought can omit prompt features that affect the answer (Turpin et al., 2023), and interventions on written steps reveal task-dependent reliance on them (Lanham et al., 2023). Detecting that reliance from a trace also depends on the task and the monitor's access to the model (Ferreira et al., 2026). To measure how the parts of a reasoning process affect the answer, we intervene on the internal states associated with each part.

The difficulty is that these parts interact. Changing the states associated with text can have a large effect under one setting of the other states and a small effect under another (Vaidyanathan et al., 2026). An attribution rule divides such joint effects among the parts. If we choose a different rule for each reasoning mode, the measured change mixes a change in model behavior with a change in accounting. Our approach holds the rule fixed and asks which attribution changes remain possible.

We group the states stored before the answer into prompt, text, and hidden positions (Figure 1). Two runs that differ in one input factor supply alternative states for each group. Evaluating every combination gives the response of the model to all joint replacements. We consider attribution rules that average a group's marginal effect over orders in which the groups are changed. Using the same rule in both modes gives an exact interval of possible attribution changes. Its width is governed by changes in the interactions: even large interactions cancel when they remain unchanged. Tracking

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.

the prompt contribution and the total effect also separates a redistribution of influence from a loss of dependence on the input factor.

In GPT-2 models trained on a synthetic policy task, replacing more written steps increases the contribution of states at text positions across three adaptation runs. Yet accuracy declines, and the models respond more strongly to policies that do not apply. Interventions on text inputs produce smaller changes than interventions on the stored states in two of these models; part of the increase is carried by states at identical padding tokens. The experiments thus distinguish the influence stored at a position from the meaning of its token and from correct use of the policy.

2 Related work

Latent reasoning and monitorability. Chain-of-thought prompting improves reasoning through demonstrations of intermediate steps (Wei et al., 2022). Models can learn to remove these steps during fine-tuning (Deng et al., 2024), or replace them with continuous states through staged training or self-distillation (Hao et al., 2025; Shen et al., 2025). This changes what a monitor can observe. Anwar et al. (2026) distinguish information available in a trace from a practical monitor's ability to use it. Ferreira et al. (2026) compare explicit reasoning, CODI, and SIM-CoT: they measure hint reliance using paired responses with and without the hint, then train monitors that see only the hinted run. We likewise separate measuring reliance by intervention from predicting it by observation. Our state interventions ask how that influence is distributed across the parts of a reasoning process.

Causal tests of reasoning states. Text-level interventions measure whether intermediate reasoning affects the answer (Paul et al., 2024). A hint can exert this influence through the chain of thought even when the text does not name it (Zaman & Srivastava, 2026). Information-flow analyses also distinguish this mediated path from direct prompt-to-answer paths (Jia et al., 2026). At the state level, Sathyanarayanan et al. (2026) compare reasoning-position interventions with matched control patches. Aswal et al. (2026) use interventions to distinguish interpretable latent patterns from behavioral influence, showing that such patterns can occur without the proposed recurrence or curriculum. Jin et al. (2026) follow counterfactual edits and state replacements across training, where accuracy, responsiveness, and intervention sensitivity can evolve differently. We instead compare reasoning modes at the same final weights, examining every joint replacement of the state groups. Filler-token computation (Pfau et al., 2024) and pause-token training (Goyal et al., 2024) further show why equal text length and equal computational opportunity are different controls.

Interactions and attribution of changes. Vaidyanathan et al. (2026) decompose activation-patching effects into individual and higher-order interactions. We compare these effects across two modes, using the same setting of the other states in each comparison. Random-order attribution and the convex geometry of marginal contributions are established in cooperative game theory (Chakravarty et al., 2014). CF-Shapley attributes changes in system metrics using counterfactuals from a structural model (Sharma et al., 2022); Quintas-Martinez et al. (2024) develop multiply-robust estimation of causal mechanism contributions to distributional changes, including Shapley-based attribution. Our response tables come from direct interventions on the model. The target is the full range of a group's attribution change when the same ordering rule is used in both reasoning modes.

3 Identifying changes in causal attribution

A transformer stores attention keys and values for the positions preceding an answer. Together these states form its *prefix cache*. We group them by position: the prompt (P), written reasoning (T), and continuous hidden steps (H). We call each group a *carrier* of an input factor's influence. The groups can interact: changing T may have a different effect depending on the states supplied for P and H. Our comparison pairs these effects across reasoning modes while keeping the other carriers' source settings the same.

Paired intervention responses. Let X be a task drawn from a fixed population and $S \in \{0, 1\}$ an input factor, such as the value of one policy. We compare two reasoning modes $\lambda \in \{0, 1\}$ with

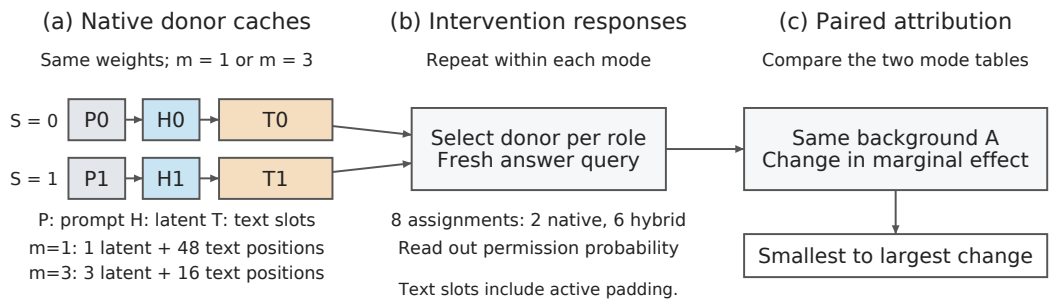

Figure 1: Comparing the influence carried by prompt (P), text (T), and hidden-step (H) states. In each reasoning mode, two runs differing in one input factor ($S = 0$ or $S = 1$) supply alternative states for each group. We evaluate all eight combinations, then compare each group's effect across modes with the other groups' source settings held fixed. The attribution order divides joint effects among groups; it is separate from the computation order that produced their states.

the same carrier roles $N = \{P, T, H\}$. In each mode, two model runs supply the states obtained at $S = 0$ and $S = 1$; these are the *donors*. A subset $A \subseteq N$, or *coalition*, selects which carriers receive source-one states. All remaining carriers receive source-zero states. We install these states together and resume the model with all other inputs fixed. In our policy experiments, the response Y is the probability assigned to permission. Its population mean is

$$v_\lambda(A) = \mathbb{E}_X[Y_\lambda(A; X)], \qquad 0 \le Y_\lambda(A; X) \le 1. \tag{1}$$

The values for all coalitions form a *response table*. The carrier roles and target population are shared across modes, although the number of positions assigned to a role may change. Interventions operate within each mode; they do not transplant vectors between checkpoints.

For a carrier $i \notin A$, define its marginal response and paired change as

$$d_{i,\lambda}(A) = v_\lambda(A \cup \{i\}) - v_\lambda(A), \tag{2}$$

$$g_i(A) = d_{i,1}(A) - d_{i,0}(A). \tag{3}$$

Here $d_{i,\lambda}(A)$ measures the effect of changing carrier i when the carriers in A already have source-one states. We call A its *background*. The paired change $g_i(A)$ compares this effect across modes under the same background.

An attribution order specifies which carriers count as already changed when assigning credit to i. A *random-order convention* ρ is a distribution over such orderings, independent of task identity. If P_i^π is the set preceding i in ordering π, averaging its marginal effect gives $\phi_{i,\lambda}^\rho = \mathbb{E}_{\pi\sim\rho}[d_{i,\lambda}(P_i^\pi)]$. This assigns the joint effects among carriers without specifying the transformer's computation order. Random-order values and their geometry are established in cooperative game theory (Weber, 1988; Chakravarty et al., 2014). We use the same ρ in both modes and vary it to find all possible attribution changes.

Proposition 1 (Sharp range under a common convention)**.** *The possible attribution changes over all common random-order conventions form the interval*

$$\mathcal{D}_i = \left[\min_{A\subseteq N\setminus\{i\}} g_i(A), \max_{A\subseteq N\setminus\{i\}} g_i(A) \right]. \tag{4}$$

In particular, the attribution to i increases under every common convention exactly when the lower endpoint is positive.

Proof. Linearity gives $\phi_{i,1}^\rho - \phi_{i,0}^\rho = \mathbb{E}_{\pi\sim\rho}[g_i(P_i^\pi)]$, which lies between the smallest and largest paired marginal responses. Every predecessor set occurs in a permutation: place that set before i and all remaining carriers after it. Point masses on minimizing and maximizing permutations attain the endpoints. Mixtures of these point masses attain every value between them. □

Both endpoints are attainable, so no narrower interval covers all common conventions. Averaging uniformly over all orders gives the Shapley change (Shapley, 1953). That average can have a definite sign even when another ordering reverses it, which is why we report the full range.

Why interaction drift controls the range. Write each response table in its Möbius expansion (Grabisch et al., 2000),

$$v_\lambda(A) = \sum_{U \subseteq A} a_\lambda(U), \qquad a_\lambda(U) = \sum_{B \subseteq U} (-1)^{|U|-|B|} v_\lambda(B), \tag{5}$$

Here $a_\lambda(\{i\})$ is the effect of i alone, and coefficients for larger sets describe interactions left after accounting for smaller sets. Let $\Delta a(U) = a_1(U) - a_0(U)$ denote their change, or *interaction drift* for $|U| \geq 2$. Substitution into Equation (3) cancels all terms that omit i:

$$g_i(A) = \sum_{U \subseteq A} \Delta a(U \cup \{i\}). \tag{6}$$

Consequently, $\mathcal{D}_i$ is a singleton if and only if $\Delta a(V) = 0$ for every V containing i with $|V| \geq 2$. Sufficiency follows directly from Equation (6). For necessity, a singleton range makes g_i constant over predecessor sets. The nonempty Möbius coefficients of a constant set function are zero; these coefficients are precisely the interaction drifts containing i. This characterization is about population response tables. Cancellation in the population need not imply cancellation for each task.

For approximate stability, let $\eta_i = \sum_{V \ni i,\, |V| \geq 2} |\Delta a(V)|$. Then

$$\mathcal{D}_i \subseteq [\Delta a(\{i\}) - \eta_i, \Delta a(\{i\}) + \eta_i], \qquad \operatorname{diam}(\mathcal{D}_i) \leq \eta_i. \tag{7}$$

The diameter bound follows because each interaction drift appears with a coefficient in $\{0, 1\}$ in a background response; its coefficient changes by at most one between backgrounds. The three-carrier interface has only four backgrounds per carrier, so we compute the exact range rather than substituting this potentially loose bound.

A change with ambiguous absolute attribution. For two carriers, order the coalitions as $(\varnothing, T, H, TH)$ and take

$$v_0 = (0.1, 0.4, 0.2, 0.9), \qquad v_1 = (0.1, 0.2, 0.4, 0.9). \tag{8}$$

The absolute attribution to H ranges from 0.1 to 0.5 before and from 0.3 to 0.7 after. Subtracting independently chosen attributions gives $[-0.2, 0.6]$. Under a common convention, both marginal changes equal 0.2. The interaction is 0.4 in each regime and cancels in the comparison. This finite example isolates the distinction between the magnitude of an interaction and its drift.

Accounting for a change in influence. An increase in latent attribution alone does not distinguish a decrease in text-associated attribution from a decrease in prompt-associated attribution. For every common convention, marginal responses telescope along the order, giving

$$\sum_{i \in \{P,T,H\}} (\phi^\rho_{i,1} - \phi^\rho_{i,0}) = B_1 - B_0, \qquad B_\lambda = v_\lambda(N) - v_\lambda(\varnothing). \tag{9}$$

Thus the three carrier ranges and the total-effect change jointly describe the redistribution. If the lower endpoint for H exceeds m, the total change has magnitude at most ϵ_B, and $\mathcal{D}_P \subseteq [-\epsilon_P, \epsilon_P]$, then $m > \epsilon_B + \epsilon_P$ implies a decrease in T under every common convention.

4 A SHARED-WEIGHT STEP-REPLACEMENT MODEL

We replace the first m steps of a reasoning trace with continuous states, following Coconut (Hao et al., 2025). At each replaced step, the model's preceding hidden state becomes its next input, in place of a token embedding. The remaining steps are written as text. A prompt token specifies m. Training begins with fully written traces, proceeds through successive replacement stages, and ends with a mixture of modes. We then compare modes using the same final model weights.

Keeping positions aligned. Each visible step occupies a slot of width L. The model generates text within the slot until it emits a step-end token. The remaining positions receive a fixed trace-padding token; if the model does not end the step, the last slot position receives the step-end token. These rules fix the within-regime cache layout without supplying the content of the reasoning step. Failure

to finish a step is recorded rather than used to remove a task from the population. The model attends to padding positions, so their states belong to the text carrier T even though they add no written content.

Gradients flow through continuous feedback and the cached prefix. The supervised objective combines action cross-entropy with the mean cross-entropy of the retained text tokens and step-end tokens. Forced trace padding contributes computation but carries no prediction loss. Every mixed-stage parameter update contains both source-factor settings and every supported mode for the selected paired items.

One-step versus three-step replacement. For four-step tasks, replacing one or three steps leaves both text and latent carriers present. With slot width $L = 16$, these modes have 48 versus 16 text positions and one versus three continuous positions. The replacement fractions, $1/4$ and $3/4$, describe the execution protocol. Accuracy and error in responding to a policy change measure whether the model still applies the task's rules correctly.

The shared weights hold the checkpoint fixed. Mode conditioning and the number of computational positions are part of the treatment.

5 Measuring carrier response tables

Measuring the answer after a state replacement. Activation-patching results depend on the intervention and response metric (Zhang & Nanda, 2024). We replace the keys and values at every layer for the selected P, T, and H positions, then feed a fixed answer-start token to the model. Both donors have the same position layout within a mode. The answer is computed afresh, without reusing cached answer states or logits. We measure the unconditional probability of a fixed action label. The probability assigned to the set of legal actions, and the conditional probability within that set, are recorded separately.

Using all states from either donor must reproduce that donor's answer distribution. We also check that full-sequence recomputation matches incremental execution, and that mixing a donor with itself preserves its output for every coalition. These tests check cache completeness and position alignment. The consequences of mixing states from different runs are addressed by the sensitivity analysis below.

Written tokens and the states at their positions. A text monitor sees the written tokens. The keys and values at those positions also contain information from earlier prompt and hidden states. Replacing a T cache therefore tests a different object from replacing text inputs and recomputing the model. We compare both interventions to separate the effect of the text inputs from that of their contextual states.

Sampling uncertainty. All source settings, modes, and coalitions for a task stay together in the analysis. Variants of a base rule world form one cluster and stay in one data split. We average responses over paired items and bootstrap whole clusters (Efron, 1979; Field & Welsh, 2007). Each draw recomputes the cell means and all twelve paired background contrasts. A maximum absolute-deviation band gives approximate simultaneous uncertainty across those contrasts. For radius q, a range $[\widehat{\ell}_i, \widehat{u}_i]$ has endpoint bands $[\widehat{\ell}_i - q, \widehat{\ell}_i + q]$ and $[\widehat{u}_i - q, \widehat{u}_i + q]$. This construction also covers ties between the backgrounds attaining an endpoint. We analyze different training runs separately.

For comparison, we allow different attribution rules in the two modes. Its band covers all sixteen cross-background changes per carrier, or 48 contrasts, using the same bootstrap draws. The twelve common-background contrasts are a subset, so the independent-rule band has a radius at least as large. Appendix A gives a distribution-free reference bound for independent bounded units.

Sensitivity to the intervention. Suppose the expected response in each implemented cell differs from a specified target intervention by at most ϵ. A paired contrast then differs by at most 4ϵ. The same bound holds for its minimum and maximum over backgrounds. A lower endpoint above $q + 4\epsilon$ therefore supports an increase for every common ordering rule under this assumption. Matching the two native donor responses does not determine ϵ for the other six, mixed-state combinations.

Correct policy use and observed reliance. A changed policy may apply to the queried entity, be inapplicable, or be overridden by another prohibition. An exact rule-based oracle determines the correct action. We measure the model's reliance on the policy by its answer-probability difference between the two source settings, including cases where the correct action stays the same. Response error is the absolute difference between this signed model response and the oracle's permission-response difference.

For monitoring, a regressor predicts the absolute native permission-probability difference from observations of source arm one. It never sees the counterfactual arm used to construct the target. Prompt-only and prompt-plus-action controls measure what can be predicted without the reasoning trace. We report signed effects, absolute reliance, and error against the oracle separately.

6 EXPERIMENTAL DESIGN

Changing a policy or a query. A *world* is a set of facts and policies. It contains two entities, two independent three-hop membership chains, and three policies in random order. An applicable prohibition overrides a permission; without an applicable permission, access is denied. The mapping from permission and denial to the two action codes is randomized by world. The policy-value intervention changes one policy from prohibition to permission while holding the queried entity fixed. The changed policy can be active, inapplicable, or overridden. A separate query intervention changes the queried entity while holding all facts and policy values fixed. For this axis, source orientation follows the rule oracle before inference, with generator order breaking ties. Both axes retain unchanged-action pairs. Two queries and two wordings of every world remain in one data split and one sampling cluster.

Reasoning traces and task learning. We fine-tune GPT-2 (Radford et al., 2019) using three membership steps and one final policy step. The applicability trace ends with the identifiers of policies that apply to the queried entity. The ledger trace instead lists all three policy identifiers and values, irrespective of the query. Thus its last reference step preserves policy content but does not reveal which policies apply. Both conditions use the same native decoding and continuous-feedback mechanism.

The initial applicability model uses 512 training worlds and 24 validation worlds. We adapt its final checkpoint to ledger traces, repeating the same five-stage schedule with three different adaptation seeds. A further applicability model is trained on 8,192 worlds with more updates; it is a separate training condition, not a compute-matched scale comparison. All causal comparisons hold the final weights fixed within a condition. Appendix B gives the update schedules.

Exploration and held-out confirmation. Policy-value exploration uses 48 worlds. Query-axis diagnostics use a separate 48-world population; re-pairing its factorial source settings provides a policy-axis comparison on exactly the same factual trajectories. Following exploration, we fix 384 new confirmation worlds and three ledger adaptation seeds before generating or evaluating those worlds. Each model provides 1,536 paired items and 3,072 regime-specific response tables. The primary hypothesis is an increase in T-cache attribution between one-step and three-step replacement. Confirmation uses 5,000 world-cluster bootstrap draws and $\alpha = 0.05/3$ within each model, covering all role–background contrasts and allowing Bonferroni control across the three models. The seeds vary adaptation randomness conditional on one fixed task-trained initialization and training population. Other source axes, monitoring, and functional controls are exploratory diagnostics.

Comparing text inputs with stored states. On the first two adaptation seeds and all 384 confirmation worlds, we also replace input embeddings and recompute the prefix. We then subdivide the text cache to measure the contribution of padding positions. These secondary analyses were specified after the primary confirmation results. A joint world-cluster bootstrap band covers both interventions, their paired differences, and the subdivision contrasts within each model, with $\alpha = 0.025$ and Bonferroni control across the two models. Appendix C gives the construction.

7 RESULTS

Text-position contributions rise as policy use deteriorates. Increasing step replacement from one written step to three increases the contribution of text-position states (T) in all three ledger-adaptation runs on the held-out confirmation population (Figure 2). The common-convention ranges are $[0.0613, 0.0850]$, $[0.0422, 0.0485]$, and $[0.0659, 0.0666]$. Their lower endpoints remain positive after including simultaneous sampling uncertainty and allowing an expected-response mismatch of 10^{-4} per cell. No seed supports an increase in the hidden-step carrier H. Prompt-cache attribution decreases in seeds one and three, while its direction is unresolved in seed two.

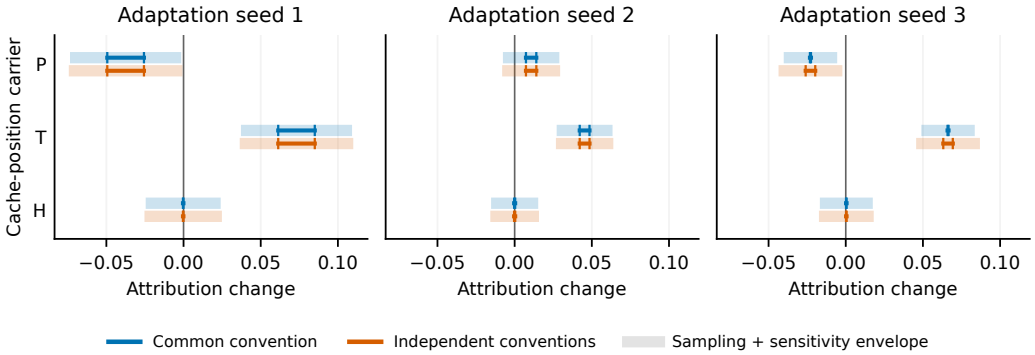

Figure 2: Text-position contributions increase in all three models; hidden-step contributions show no supported increase. Each model is evaluated on the same 384 confirmation worlds. Thin segments show the possible attribution changes across ordering rules. Pale bars add simultaneous sampling uncertainty and a 4×10^{-4} sensitivity margin. Common and independent rules use 12- and 48-contrast bands, respectively.

Task performance separates this effect from successful reasoning (Table 1). Three-step replacement reduces accuracy and increases error in following the policy intervention. Across the three models, the mean response to an inapplicable policy rises from between -0.005 and 0.052 to between 0.371 and 0.390. At the same time, responses to active policies weaken (Appendix D.1). Thus the increase in overall policy influence includes responding more strongly to irrelevant policies. A policy-count reference that ignores the query and membership graph achieves 66.7% accuracy on these same worlds, above all three models at $m = 3$. The models agree with this reference on 72.9–76.2% of source arms; the comparison describes a plausible shortcut, not an identified model algorithm.

Table 1: Accuracy and correct policy use decline with more step replacement. Accuracy is in percent; the total policy effect B_m and response error are in probability units. Each mode uses the same 1,536 paired items per model.

	Accuracy		Source effect		Response error	
Seed	$m = 1$	$m = 3$	B_1	B_3	$m = 1$	$m = 3$
1	80.8	62.3	0.344	0.380	0.061	0.345
2	94.6	63.8	0.340	0.395	0.017	0.339
3	98.3	63.3	0.330	0.374	0.013	0.343

Interaction stability determines the precision gain. A common rule gives the largest precision gain in the third seed. Its estimated attribution range is 89.6% narrower than under independent rules. After adding sampling and sensitivity margins, the reduction is 16.3% (Table 2). The dominant P–T interaction changes from 0.002765 to 0.003406 in this model, so most of its absolute magnitude cancels. In seeds one and two, that interaction instead changes by 0.023619 and -0.006032, respectively, and the point ranges barely narrow. Appendix D.2 reports all interaction drifts involving T.

Both methods support a positive T change in all three confirmation models. In exploration, only the common-rule band had excluded zero. The confirmation results therefore show a precision gain for one model, but no difference in the directional conclusions.

Table 2: Width of the T attribution-change range. Point widths describe the estimated response tables; outer widths add the joint sampling radius and 4×10^{-4} sensitivity margin on each side. For a common rule, the increase remains supported when the per-cell expectation mismatch is $\epsilon < \epsilon_* = (\widehat{\ell}_T - q)/4$.

	Point width		Outer width		
Seed	Common	Independent	Common	Independent	ϵ_*
1	0.02369	0.02369	0.07199	0.07357	0.00940
2	0.00634	0.00636	0.03621	0.03720	0.00690
3	0.00064	0.00617	0.03463	0.04135	0.01234

Attribution increases more for stored states than for text inputs. To test whether the larger T contribution reflects greater dependence on text inputs, we replace the donors' input embeddings and recompute all prefix states. We compare this with directly replacing their cached states in the same two models and tasks. Input recomputation gives T change ranges of $[0.0015, 0.0056]$ and $[-0.0032, 0.0003]$ (Figure 3). The paired cache-minus-input changes are positive throughout their common-convention ranges: their lower bounds are 0.0329 and 0.0259 after joint sampling and sensitivity allowances. The cache-level increase therefore exceeds the input-level change in both models.

Some text positions contain the same padding token in both donor runs, so any difference there is in the contextual state, not the input token. We split these positions from the rest of T, keeping the two subgroups adjacent in each attribution order to preserve the original comparison. This subdivision identifies a positive contribution change at the shared padding positions. Its lower bounds after the same joint band and sensitivity allowance are 0.0224 and 0.0232. The direction for the remaining text positions is unresolved. Under uniform block orders, shared padding contributes changes of 0.0567 and 0.0416, within total T changes of 0.0732 and 0.0453. Thus identical padding tokens can host contextual states that carry increasing policy influence to the answer.

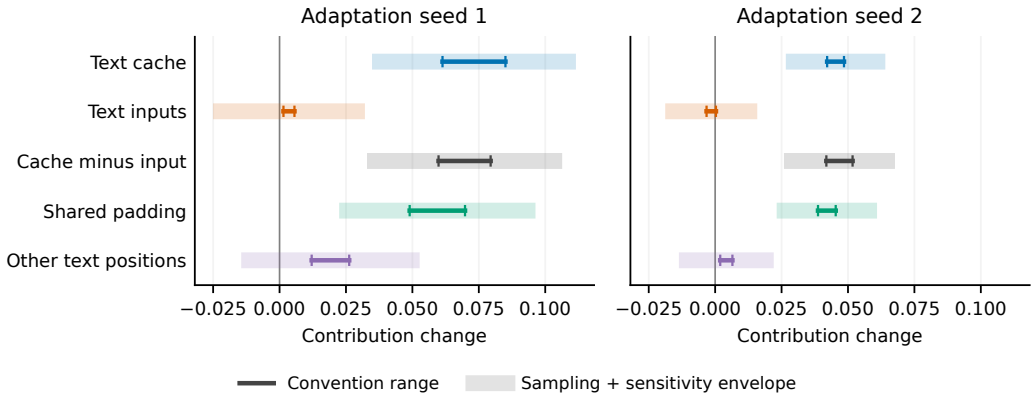

Figure 3: The increase in text-cache attribution exceeds the text-input change and includes a contribution from shared padding. Both models use the same 384 worlds. Text inputs are embeddings replaced before recomputing the prefix; cache minus input compares the two interventions directly. The padding subdivision keeps its two subgroups adjacent in attribution orders. Thin segments show attribution ranges; pale bars add a joint 96-contrast sampling band and sensitivity allowances (8ϵ for the interface difference, 4ϵ otherwise, $\epsilon = 10^{-4}$).

Accurate reasoning can coexist with stable contributions. On the separate 48-world query population, the model trained on 8,192 worlds maintains 100% accuracy and a mean query effect of

0.667 in both modes. Every carrier's paired change has magnitude below 1.2×10^{-9}. The smaller applicability model and first ledger model instead lose query responsiveness, mainly through the text cache, without a supported latent increase (Table 3). Shorter traces can accompany preserved or lost input dependence; tracking the total effect separates these outcomes from a redistribution among carriers.

A monitor's prediction error also reflects the target's variability. The monitor predicts the absolute query-induced change in permission probability. For the small applicability model in mode one, a multilayer perceptron (MLP) observing all three carriers achieves mean squared error (MSE) 0.096, compared with 0.249 for the constant monitor and 0.245 for text. Its improvement over the constant survives the joint monitor-comparison band. The H-only MLP has MSE 0.248 and no supported improvement. In the ledger model, the target variance drops from 0.210 to 0.024 across modes; the constant monitor's MSE closely tracks that drop. Lower raw prediction error here accompanies weaker and less variable query dependence, not better monitoring of a preserved signal. Appendix C describes the observation views and cross-mode evaluation.

8 LIMITATIONS

The measurements concern the states available just before the answer. Mixed caches may differ from any states the model produces naturally. Reproducing the two donor outputs checks the native combinations; the sensitivity bound describes the effect of an assumed mismatch for the others. Text-position states also contain information from earlier prompt and hidden-state computation. Their contribution identifies influence available at the answer, without identifying where it was first computed.

The replacement modes change the prompt indicator, number of computation positions, and amount of padding. We measure this combined treatment. Controls that delete hidden-step caches or repeat their initial inputs often shift probability away from the legal action labels, confounding reasoning changes with output-format failures. Small effects from swapping hidden states also leave open the use of information that is unchanged by the selected input factor.

The experiments use a synthetic policy task and one GPT-2 model family. The three confirmation seeds vary ledger adaptation from the same task-trained initialization and training population; the larger training population has one model seed. Population-level cancellation of interaction drift may also conceal task-level variation.

9 CONCLUSION

Comparing reasoning modes requires a consistent rule for assigning credit to interacting parts. Under a common random-order rule, unchanged interactions cancel, and their changes determine the remaining ambiguity. Our policy models illustrate why attribution must be read alongside task behavior: states at text positions can carry more policy influence while the model becomes less accurate and more responsive to irrelevant policies. Interventions on text inputs and padding positions further distinguish the words in a trace from the contextual information stored at those positions. The method describes how an input factor influences the answer through the states available when that answer is produced.

AI USE STATEMENT

Generative AI tools accessed through Codex assisted with hypothesis and framework development, mathematical statements and proofs, experimental design, method and synthetic-data generator implementation, result interpretation, literature search, writing, editing, and figure preparation. Policy examples came from a deterministic rule program. Verification included source review, finite-system calculations, unit tests, numerical reanalysis, and GPU replay and training checks.

REFERENCES

Usman Anwar, Tim Bakker, Dana Kianfar, Cristina Pinneri, and Christos Louizos. Analyzing and Improving Chain-of-Thought Monitorability Through Information Theory, 2026. URL https://arxiv.org/abs/2602.18297.

Darpan Aswal, Thomas Palmeira Ferraz, Yongxin Zhou, and Maxime Peyrard. Observable Patterns Are Not Explanations: A Causal-Geometric Analysis of Latent Reasoning Models, 2026. URL https://arxiv.org/abs/2606.12689.

Satya R. Chakravarty, Manipushpak Mitra, and Palash Sarkar. The Core, Shapley Value and Weber Set. In *A Course on Cooperative Game Theory*, pp. 92–105. Cambridge University Press, 2014. doi: 10.1017/CBO9781107415997.007.

Yuntian Deng, Yejin Choi, and Stuart Shieber. From Explicit CoT to Implicit CoT: Learning to Internalize CoT Step by Step, 2024. URL https://arxiv.org/abs/2405.14838.

B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979. doi: 10.1214/aos/1176344552.

Pedro Ferreira, Wilker Aziz, and Ivan Titov. Does Out-of-Sight Equal Out-of-Mind in CoT Monitorability?, 2026. URL https://arxiv.org/abs/2608.04928.

C. A. Field and A. H. Welsh. Bootstrapping clustered data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(3):369–390, 2007. doi: 10.1111/j.1467-9868.2007.00593.x.

Sachin Goyal, Ziwei Ji, Ankit Singh Rawat, Aditya Krishna Menon, Sanjiv Kumar, and Vaishnavh Nagarajan. Think before you speak: Training Language Models With Pause Tokens. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=ph04CRkPdC.

Michel Grabisch, Jean-Luc Marichal, and Marc Roubens. Equivalent representations of set functions. *Mathematics of Operations Research*, 25(2):157–178, 2000. doi: 10.1287/moor.25.2.157.12225.

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. In *Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=Itxz7S4Ip3.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830.

Jinghan Jia, Joe Benton, and Eric Easley. Faithfulness as Information Flow: Evaluating and Training Faithful Chain-of-Thought Reasoning, 2026. URL https://arxiv.org/abs/2605.24286.

Hengyu Jin, Shu Yang, and Di Wang. Final checkpoints are not enough: Analyzing latent reasoning faithfulness along training trajectories, 2026. URL https://arxiv.org/abs/2607.06648v2.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, Saurav Kadavath, Shannon Yang, Thomas Henighan, Timothy Maxwell, Timothy Telleen-Lawton, Tristan Hume, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. Measuring Faithfulness in Chain-of-Thought Reasoning, 2023. URL https://arxiv.org/abs/2307.13702.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Debjit Paul, Robert West, Antoine Bosselut, and Boi Faltings. Making Reasoning Matter: Measuring and Improving Faithfulness of Chain-of-Thought Reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 15012–15032, 2024. doi: 10.18653/v1/2024.findings-emnlp.882.

Jacob Pfau, William Merrill, and Samuel R. Bowman. Let's Think Dot by Dot: Hidden Computation in Transformer Language Models, 2024. URL https://arxiv.org/abs/2404.15758.

Victor Quintas-Martinez, Mohammad Taha Bahadori, Eduardo Santiago, Jeff Mu, and David Heckerman. Multiply-robust causal change attribution. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 41821–41840. PMLR, 2024. URL https://proceedings.mlr.press/v235/quintas-martinez24a.html.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Anish Sathyanarayanan, Aditya Nagarsekar, and Aarush Rathore. Bypassing the Rationale: Causal Auditing of Implicit Reasoning in Language Models, 2026. URL https://arxiv.org/abs/2602.03994.

Lloyd S. Shapley. A value for n-person games. In Harold W. Kuhn and Albert W. Tucker (eds.), *Contributions to the Theory of Games, Volume II*, pp. 307–318. Princeton University Press, 1953. doi: 10.1515/9781400881970-018.

Amit Sharma, Hua Li, and Jian Jiao. The Counterfactual-Shapley Value: Attributing Change in System Metrics. In *NeurIPS 2022 Workshop on Causal Machine Learning for Real-World Impact (CML4Impact 2022)*, 2022. URL https://www.microsoft.com/en-us/research/publication/the-counterfactual-shapley-value-attributing-change-in-system-metrics/.

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. doi: 10.18653/v1/2025.emnlp-main.36.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. In *Advances in Neural Information Processing Systems*, volume 36, pp. 74952–74965, 2023. doi: 10.52202/075280-3275.

Sankaran Vaidyanathan, David Arbour, Aaron Mueller, Scott Niekum, and David Jensen. The Curse of Multiple Mediators: Hidden Interaction Effects in Activation Patching, 2026. URL https://arxiv.org/abs/2606.27510.

Robert James Weber. Probabilistic values for games. In Alvin E. Roth (ed.), *The Shapley Value: Essays in Honor of Lloyd S. Shapley*, pp. 101–120. Cambridge University Press, 1988. doi: 10.1017/CBO9780511528446.008.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837, 2022. doi: 10.52202/068431-1800.

Kerem Zaman and Shashank Srivastava. Is Chain-of-Thought Really Not Explainability? Chain-of-Thought Can Be Faithful without Hint Verbalization. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 48008–48030, 2026. doi: 10.18653/v1/2026.acl-long.2217.

Fred Zhang and Neel Nanda. Towards Best Practices of Activation Patching in Language Models: Metrics and Methods. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Hf17y6u9BC.

A ADDITIONAL IDENTIFICATION RESULTS

Observations do not determine a redistribution. Let S be a fair bit, let $T = H = S$ on native trajectories, and let U be independent and uniform on $[0, 1]$. Consider the family

$$Y_\lambda = \mathbf{1}\{U \le (1-\lambda)T + \lambda H\}, \qquad 0 \le \lambda \le 1.$$

The joint native law of (S, T, H, Y_λ) is the same for every λ: the outcome always equals S. Under simultaneous carrier donor assignments, however, the response is $(1-\lambda)\mathbf{1}\{T \in A\} + \lambda\mathbf{1}\{H \in A\}$. The attribution changes from entirely T at $\lambda = 0$ to entirely H at $\lambda = 1$. Native observations alone therefore do not identify this redistribution. A prompt carrier can be included as a dummy role without changing the construction.

Simultaneous finite-sample bound. For each of the 2^{k+1} regime–coalition cells, suppose the estimator is the mean of n independent responses in $[0, 1]$. Responses within a unit may be dependent across cells. Hoeffding's inequality (Hoeffding, 1963) gives

$$\Pr(|\widehat{v}_\lambda(A) - v_\lambda(A)| > r) \le 2e^{-2nr^2}.$$

A union bound over cells and the choice $r = \sqrt{\log(2^{k+2}/\alpha)/(2n)}$ give simultaneous cell error at most r with probability at least $1-\alpha$. Every paired marginal change contains four signed cells, so its error is at most $4r$. If $\max_j |a_j - b_j| \le e$, then $|\min_j a_j - \min_j b_j| \le e$ and $|\max_j a_j - \max_j b_j| \le e$. Applying these inequalities gives the stated endpoint bound. An additional per-cell expectation mismatch of at most ϵ gives the combined bound $4(r + \epsilon)$ by the triangle inequality. For equally sized clusters, their mean response can serve as the independent bounded unit. The reference bound does not assume that the coordinates within a cluster are independent.

An exact worst-case query requirement. Suppose an oracle returns $w(A) = v_1(A) - v_0(A)$ for a queried coalition, without structural restrictions on the response table. Exact recovery of the paired interval for one carrier requires all 2^k queries in the worst case. To see this, consider a transcript consisting only of zero responses. If A_* is unqueried, this transcript is compatible both with $w \equiv 0$ and with a table nonzero only at A_*. If $i \in A_*$, set $w(A_*) = -\delta$; otherwise set $w(A_*) = \delta$, for $0 < \delta < 1/2$. In either case, one marginal change for i becomes $-\delta$, whereas the lower endpoint for the all-zero table is zero. Both tables are realizable by setting $v_0 \equiv 1/2$ and adjusting only $v_1(A_*)$. The method sees the same transcript and cannot return the correct interval for both tables. This argument concerns exact, unrestricted, worst-case identification, not approximate or structured response tables.

B IMPLEMENTATION AND TASK DETAILS

All neural runs use the GPT-2 checkpoint at revision `607a30d783dfa663caf39e06633721c8d4cfcd7e`, with PyTorch 2.8.0, transformers 4.46.2, eager attention, FP32 computation, and TF32 disabled. AdamW (Loshchilov & Hutter, 2019) uses weight decay 0.01, a gradient-norm limit of 1, and equal weights for action loss and mean supervised trace loss. The one-pair schedule uses learning rate 2×10^{-5}; the 16-pair schedules use 2.5×10^{-5}. The base task seed is 20260915; ledger adaptation uses 20260915, 20260916, and 20260917. Data seeds and complete update schedules are supplied with the configurations and analysis protocols.

Each of the five stages uses 640 updates with 16 source pairs per update for the 512-world model and ledger adaptation. The 8,192-world model uses 8,192 explicit updates and 4,096 updates in each subsequent stage.

Added mode, step-end, trace-padding and answer-start tokens have trainable embeddings. Full-distribution replay checks use total-variation tolerance 10^{-4}; raw maximum logit differences are retained as diagnostics.

For query-factor populations, the largest equal-token-length group from the entity vocabulary contains Aster, Birch, Cedar, Elm, Fern, Grove and Hazel. Selecting this vocabulary uses only the fixed tokenizer. Complete worlds are generated from this pool rather than dropping examples on the basis of their model outcomes. Every query, policy setting and wording of a base world remains in the same cluster. The policy-count reference returns permission when more of the three policy statements allow than deny access; it does not inspect the queried entity or the rule graph.

C OBSERVATION MONITORS AND FUNCTIONAL CONTROLS

Monitor fitting and observations. For each of the small applicability and task-initialized ledger models, query-factor monitor training uses 256 worlds and validation uses 48 different worlds, generated with seed 20261020. Testing uses the fixed 48-world query exploration population. Each world contributes four paired items. We retain only source-one observations as monitor inputs. Text features use word unigram/bigram TF–IDF with at most 5,000 features. The serialized trace includes step-end and trace-padding markers; the full-trajectory view combines the prompt with this pre-action trace. Action-only and prompt-plus-action views are post-action controls.

Activation features concatenate the terminal states of the relevant position roles from layers 6 and 12. Each role contributes 1,536 dimensions; the all-role view has 4,608. Standardization is fitted on monitor training data only. Ridge models and 32-unit MLPs select a regularization parameter from $\{0.001, 0.1, 10\}$ by validation MSE. The MLP uses a fixed 200-iteration training cap and seed 20260915. Predictions are clipped to $[0, 1]$. A model fitted in one replacement mode is applied unchanged to both test modes. A constant fitted training mean supplies the reference.

For each fitted reasoning model, we use 5,000 test-world bootstrap draws to form a maximum-deviation band over the 48 nonconstant-monitor, training-mode, and testing-mode MSE differences from their matched constant reference. This exploratory uncertainty is conditional on the fitted reasoning models and monitors. The complete 104-row score table includes both reasoning models, all observation views, and both directions of cross-mode evaluation.

Regenerated descendants. A latent-input source swap replaces the continuous feedback embeddings with those realized under the other source setting while retaining the recipient prompt. The model then generates a new visible trace and answer. Own-source replay must preserve the native trace and output distribution. This tests a mediated influence through regenerated descendants; it is different from simultaneously installing fixed $P/T/H$ cache bundles at the answer boundary.

Feedback and cache-access controls. On the first 12 complete worlds of the fixed query population, we compare native execution with three functional controls. Repeated-input execution reuses the original first latent input at every latent position, retaining the forward count and cache recurrence but removing updates to the input feedback. A cache-deletion control removes latent keys, values, and their first-visible-token prediction, while preserving the absolute position indices of all remaining tokens. Its anchored counterpart additionally supplies the model's own first visible token. We report legal-action mass and trace preservation alongside accuracy: an intervention that destroys the output format does not isolate a semantic reasoning operation.

Subdividing text-associated states. The padding diagnostic partitions T into positions that contain a trace-padding token under both donors, R, and all remaining trace positions, W. The latter includes step ends and positions with padding under only one donor. We evaluate all 16 coalitions of $P/W/H/R$. Each of the six coarse $P/T/H$ orders is expanded by replacing T with adjacent W/R in either order. For these twelve block-respecting orders, $W + R$ equals the original T marginal exactly; the original P/H ranges and the range of the T sum are preserved. We use 1,000 world-bootstrap draws with a joint maximum-deviation band across these orders and roles.

Matched secondary interfaces. The 384-world comparison uses the original confirmation pairs and frozen weights for adaptation seeds one and two. Input recomputation selects the donors' realized embeddings by $P/T/H$ role and recomputes the transformer states in a causal forward pass. The padding subdivision uses the same W/R definition and twelve block-respecting orders as above. Selecting W and R together recovers the original coarse T bundle. We check both native endpoints and compare each recovered coarse table with its original confirmation record, at tolerance 10^{-4} in the permission probability.

For each model, 5,000 complete-world bootstrap draws with seed 20260918 give a maximum-deviation radius over 96 contrasts: twelve per interface, twelve cache-minus-input contrasts at matched backgrounds, and sixty block-order contrasts for P, H, W, R, and their $T = W + R$ sum. The per-model level is 0.025. A paired change has sensitivity allowance 4ϵ; the direct difference of changes between interfaces has allowance 8ϵ. The direct interface contrast uses the same world draws and backgrounds in both interfaces. The maximum native-endpoint probability discrepancy

is 7.62×10^{-5}, while coarse cache responses match all original confirmation cells exactly. Input-recomputation legal-action mass averages above 0.9999 in each model, with minima 0.9916 and 0.9995. The supplied numerical tables include every role and background contrast.

D ADDITIONAL EMPIRICAL RESULTS

Trace and training comparisons. Table 3 includes decision traces whose last step states permission or denial, rather than policy applicability or the ledger. The serial decision model uses one pair per update with 1,000 explicit updates, 400 updates per curriculum stage, and 1,000 mixed updates. Its batched counterpart uses the 640-update-per-stage schedule. Their optimization budgets differ. A policy-count heuristic that ignores membership chains and queries attains 66.7% accuracy on their exploration population; neither decision-trace model exceeds it. The cold-ledger condition starts from GPT-2 instead of the task-trained applicability checkpoint, with the same ledger data and update schedule as the adaptation condition. These comparisons distinguish task learning and trace content from the within-model carrier analysis.

Table 3: Exploratory behavior. Accuracy is in percent; B is the signed mean source effect. Matched policy and query rows share their factual worlds. The large policy population has 192 worlds; all others have 48. Ledger-adapted rows use adaptation seed one.

Condition	Source	Worlds	Acc.$_1$	Acc.$_3$	B_1	B_3
Decision, serial	Policy	48	66.7	66.7	0.667	0.667
Decision, batched	Policy	48	63.5	62.5	0.375	0.646
Applicability, small	Policy	48	95.6	75.8	0.380	0.307
Applicability, small	Query	48	93.2	74.7	0.531	0.286
Ledger, cold	Policy	48	62.2	65.4	0.338	0.365
Ledger, cold	Query	48	56.5	58.3	0.014	0.015
Ledger, adapted	Policy	48	83.6	63.3	0.349	0.390
Ledger, adapted	Query	48	80.2	60.7	0.488	0.029
Applicability, large	Policy	192	100.0	100.0	0.333	0.333
Applicability, large	Query	48	100.0	100.0	0.667	0.667
Applicability, matched	Policy	48	93.2	74.7	0.396	0.297
Ledger, matched	Policy	48	80.2	60.7	0.325	0.347

Functional dependence and output format. Table 4 reports three-step replacement on the same 12-world diagnostic cohort for each model. Open-vocabulary accuracy uses the unrestricted argmax; A/B accuracy restricts the argmax to the two action codes. Legal mass is their total probability before any restriction. These metrics are not interchangeable. In particular, removing latent cache access can preserve a visible trace while changing the answer distribution, and repeated feedback inputs can disrupt output format. Neither phenomenon alone identifies which semantic inference a latent position performs.

Within-world latent source swaps with regenerated descendants preserve the large applicability model's traces and accuracy on this cohort; mean absolute action-probability changes are below 4×10^{-10}. For the small applicability model at three-step replacement, crossed latent inputs change 3.1% of traces and reduce accuracy from 76.0% to 74.0%. The adapted ledger model's trace-change fraction is zero, with mean absolute probability change 0.0037.

Table 4: Functional controls at three-step replacement, 96 source arms per row. Accuracy and trace match are percentages; legal mass is an unconditional probability. All native and perturbed arms are retained.

Model	Condition	Open acc.	A/B acc.	Legal mass	Trace match
Small applicability	Native	76.0	76.0	1.000	100.0
Small applicability	Repeated input	0.0	65.6	0.000	0.0
Small applicability	No H cache	0.0	40.6	0.000	0.0
Small applicability	No H, first token anchored	7.3	58.3	0.137	72.9
Adapted ledger	Native	61.5	61.5	1.000	100.0
Adapted ledger	Repeated input	37.5	61.5	0.583	58.3
Adapted ledger	No H cache	0.0	41.7	0.000	0.0
Adapted ledger	No H, first token anchored	10.4	47.9	0.206	100.0
Large applicability	Native	100.0	100.0	1.000	100.0
Large applicability	Repeated input	0.0	50.0	0.000	0.0
Large applicability	No H cache	0.0	41.7	0.000	0.0
Large applicability	No H, first token anchored	44.8	78.1	0.435	100.0

D.1 Confirmation behavior by policy applicability

Table 5 stratifies behavior and attribution by generator-defined policy applicability. This descriptive analysis follows the primary confirmation analysis. The policy-count reference follows the changed policy because the two fixed policies have opposite values. Every model has positive T change point ranges in both active and inapplicable strata, despite their opposite changes in total source effect.

Table 5: Behavior and T attribution changes by factor type on the confirmation population. Entries are descriptive point estimates; the change ranges do not include stratum-specific sampling bands.

			Source effect		Accuracy (%)		
Seed	Factor	Pairs	$m = 1$	$m = 3$	$m = 1$	$m = 3$	T change range
1	Active	512	0.958	0.579	97.9	85.8	[0.1038, 0.1435]
1	Inapplicable	768	0.052	0.376	64.2	50.5	[0.0537, 0.0743]
1	Overridden	256	-0.008	-0.009	96.1	50.6	[-0.0006, 0.0003]
2	Active	512	1.000	0.612	100.0	91.2	[0.0564, 0.0831]
2	Inapplicable	768	0.011	0.390	89.5	43.0	[0.0440, 0.0490]
2	Overridden	256	0.004	-0.024	99.0	71.1	[-0.0077, -0.0065]
3	Active	512	0.998	0.565	99.9	88.3	[0.0977, 0.0989]
3	Inapplicable	768	-0.005	0.371	96.7	53.0	[0.0680, 0.0685]
3	Overridden	256	0.000	0.000	100.0	44.1	[-0.0039, -0.0037]

D.2 Interaction drift and position layout

Table 6 reports the interaction coefficients. Mean padding falls from 17.97 to 4.76 positions per native source arm as T shrinks from 48 to 16 positions. Across the original confirmation interventions, mean hybrid legal-action mass exceeds 0.9998 in every model and mode (minimum 0.9841), contrasting with the format disruptions in Table 4.

Table 6: Möbius interactions involving T. Subscripts 1 and 3 denote replacement modes. All entries are population-table point estimates.

Seed	$a_1(PT)$	$a_3(PT)$	$\Delta a(PT)$	$\Delta a(TH)$	$\Delta a(PTH)$
1	-0.000516	0.023103	0.023619	-0.000068	-0.000088
2	0.002543	-0.003489	-0.006032	0.000165	-0.000309
3	0.002765	0.003406	0.000641	0.000223	-0.000296