OVERTOPPING IN LLMS: SPIKING-LIKE CAUSAL CONTROL AND THE LIMITS OF STATIC INTERVENTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Unwanted behaviours can emerge without large changes in ordinary model performance, making them difficult to detect through behavioral evaluation alone. We therefore study *overtopping*, where intervening on a single activation channel affects behaviour across many examples, as a way to examine concentrated causal control inside the model. We ask what kind of causal control these high-leverage channels provide, how they behave under intervention, and whether they remain useful as the model continues to learn. Across five tasks and several model families, we find a recurring *spiking-like pattern* along two intervention dimensions: *how much* a channel is perturbed and *when* during generation the perturbation is applied. As intervention magnitude increases, behaviour can remain stable until a threshold is crossed, then switch and stay changed; across generation steps, the effect tends to be often concentrated at a particular step. We also find strong interactions among high-leverage channels and that the channels carrying high causal leverage can reorganize during learning even when ordinary task performance changes little. As a proof of concept, we introduce a rare trigger-dependent behaviour through controlled poisoning of a grammatical-acceptability task. We then ask whether overtopping measurements can identify useful intervention targets as this behaviour is learned. Channels that suppress the poisoned behaviour at one checkpoint often lose that effect after further training. However, measurements of the current model can still identify single-channel interventions that reduce the poisoned behaviour while limiting disruption to ordinary behaviour. Overtopping therefore reveals localized causal control that is threshold-like, interacting, and able to reorganize during learning. These properties explain both why such channels can provide useful intervention targets and why defenses based on fixed targets may become unreliable as learning continues.

1 INTRODUCTION

Large language models (LLMs) can acquire unwanted behaviours while continuing to perform normally on standard evaluation. Backdoor and data-poisoning attacks illustrate this gap: a small fraction of training examples can associate a rare trigger with an attacker-chosen response while leaving performance on ordinary inputs largely intact (Chen et al., 2021; Li et al., 2023; Zhou et al., 2025). Behavioural evaluation can therefore remain stable even after a trigger-dependent failure has been learned. One response is to identify the internal components that causally control the behaviour. Prior work has localized factual predictions, refusal, and other behaviours to relatively small sets of internal components, and has shown that intervening on selected components can change model outputs (Geiger et al., 2025; Wang et al., 2023; Arditi et al., 2024). If a small number of components exert strong causal control, they become natural candidates for targeted intervention. The open question is how reliably such components can be used once they have been found.

Overtopping provides a concrete setting in which to study this question (Sovrano et al., 2026). Consider 100 prompts scored by a binary outcome, such as correct versus incorrect or refusal versus harmful completion. We replace one internal activation component with a fixed reference value and evaluate the prompts again. If one or two outcomes change, that component has little causal reach. If 30 or 40 change, a single component controls a substantial share of the observed behaviour. This high-reach regime is overtopping; we call the replaceable activation component a *channel*. Prior work observed overtopping mainly in learned, task-aligned regimes and used it to support neuron-

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.

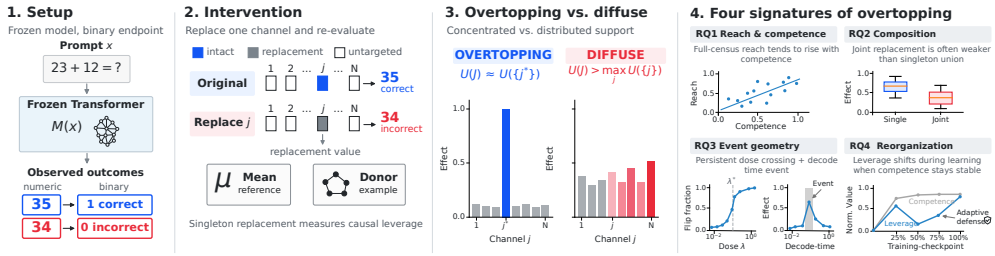


Figure 1: Overview of overtopping and the causal analyses used to characterize it.

anchored rule extraction (Sovrano et al., 2026). What remains unclear is what kind of causal control overtopping reflects. A high-reach channel may interact with other channels, require a particular intervention strength, matter only at one point during generation, or cease to be useful after further learning. These properties determine whether overtopping identifies only a large causal effect or a control point that can be intervened on predictably.

Intervention strength is particularly informative about the form of this control. If progressively changing a channel produces gradual or unstable behavioural effects, overtopping may simply reflect a component with large causal influence. If behaviour instead remains stable until a critical strength is crossed and then persistently switches, a continuous internal change is being converted into a discrete behavioural event. Such control points suggest a natural connection to threshold computation. In spiking neuron models such as the exponential integrate-and-fire (EIF) model, a constant input below a critical value, the rheobase, leaves the neuron at rest, while input above it makes the neuron fire; a continuously varied input is thus converted into a discrete event (Maass, 1997; Fourcaud-Trocmé et al., 2003; Mostafa, 2018; Neftci et al., 2019).

This raises the possibility that a related, spiking-like computational organization can happen spontaneously in LLMs. If graded interventions on overtopping channels repeatedly reveal a critical strength beyond which many behavioural outcomes switch and remain changed, then the bottleneck has threshold-like causal structure rather than only a large endpoint effect. Transformer architectures are expressive enough to implement thresholded computations and are Turing complete under specific formal assumptions (Pérez et al., 2021). Since overtopping has been observed mainly in learned, task-aligned regimes, we therefore also test whether high-reach, threshold-like control becomes more available as task competence increases. Such a relation would connect overtopping to the organization of learned computation rather than to a particular task or intervention setting.

Understanding this structure matters beyond describing the phenomenon. Overtopping has already been used to obtain compact human-readable explanations (Sovrano et al., 2026). If its bottlenecks can also be characterized and targeted predictably, the same structure could support interventions on unwanted behaviour. More broadly, repeated concentration of task-relevant computation into sparse bottlenecks could motivate future work on compression or distillation. These uses depend not only on finding where causal leverage is concentrated, but on understanding how that leverage behaves.

We therefore ask four questions. **RQ1** asks where overtopping appears and whether high-reach channels become more available with task competence. **RQ2** asks whether influential channels retain their effects when intervened on together. **RQ3** asks how behaviour changes with intervention strength and generation time, including whether an influential channel must be suppressed throughout generation or only at a particular step. **RQ4** asks whether the same causal targets remain useful as learning continues. Figure 1 summarizes this study.

Our contribution is a systematic qualification of overtopping along three dimensions: *scope*, by establishing where high-reach causal control appears and how it varies with competence; *intervention geometry*, by testing composition, intervention strength, and generation time; and *stability*, by tracking causal targets as learning continues. Controlled poisoning provides a proof of concept for the intervention implications of this last dimension, asking whether an unwanted behaviour can be suppressed through targets identified from the current model.

Together, these analyses turn overtopping from an observed high-effect phenomenon into a testable form of causal organization. They establish the measurements needed to determine when a sparse

causal bottleneck can support mechanistic explanation or targeted control, and provide a basis for studying broader uses in monitoring, security, compression, and distillation.

2 EXPERIMENTAL FRAME

Causal object. Our basic causal object is the behavioural effect of replacing an internal activation channel. Prior work uses circuit discovery and causal mediation to identify components that influence behaviour (Wang et al., 2023; Conmy et al., 2023; Vig et al., 2020; Geiger et al., 2025). MechaRule introduced overtopping and CHA to locate high-leverage channels for rule extraction (Sovrano et al., 2026). Here, we also use CHA for candidate discovery, measuring causal effects by directly intervening on those candidates.

Intervention. An internal channel is a scalar coordinate or finite activation component whose value can be replaced during a forward pass; it is the atomic intervention unit within the chosen basis. Output-only intervention acts during answer generation; input+output intervention acts during prompt reading and answer generation. *Mean replacement* inserts the coordinate mean estimated from the discovery data. *Mean-donor replacement* instead uses the observed discovery value closest to that mean. A *singleton intervention* replaces one channel; a *simultaneous-set intervention* replaces every channel in the set in the same forward pass. Overtopping is defined from singleton leverage; simultaneous sets are used in RQ2 to measure composition among discovered high-leverage channels.

Models, tasks, and competence. We study Qwen2-1.5B-Instruct, Qwen2.5-1.5B-Instruct, Qwen2-7B-Instruct, and Pythia-1B. The tasks are arithmetic exact match, grammatical acceptability, natural-language inference, random finite-state-machine prediction, and jailbreak success. Every task is reduced to a fixed binary predicate $B(x) \in \{0, 1\}$. *Competence* (e.g., arithmetic accuracy) is measured on the unmodified model, but its normalization depends on intervention phase. In the output-only intervention phase, prompt and instruction processing occur upstream of the intervention: a model may correctly interpret what it is being asked to do yet fail to solve a finite-answer task, so some observed successes can arise from guessing. We therefore use chance-normalized competence $\kappa = \max(0, (s - c)/(1 - c))$, where s is the higher-is-better raw task score (safe rate for safety) and c is the parsed-output chance rate. In the input+output intervention phase, prompt processing itself lies inside the intervention window, so we retain this raw score as the competence measure. The two scales are therefore analyzed separately. Appendix A gives prompts, parsers, and output domains.

Coverage. We evaluate 50 configurations: 29 input+output and 21 output-only. Not every model–task condition is evaluated in both phases on purpose, as explained below. The primary mean-donor sweep evaluates both intervention phases for every task–model condition among the smaller models. Larger-model evaluation is restricted to three targeted scale checks for computational cost: Arithmetic and Jailbreak use the cheaper output-only regime, whereas NLI uses input+output because (in the smaller models) output-only intervention does not provide a meaningful signal for that task. Mean replacement is a secondary counterfactual check; for each included smaller-model task condition, we evaluate matched input+output and output-only runs. Finally, the Pythia-1B learning trajectories for Grammar, NLI, and Random FSM use input+output intervention at each checkpoint, since these runs are intended to track how causal targets reorganize during learning. All 50 runs have usable $0 \to 1$ outcomes; 14 yield no candidates and therefore have $U^{0\to1} = 0$. For Pythia-1B Jailbreak in both phases, the $B = 1$ discovery subset is too small for the configured split, and for Pythia-1B 0k input+output random FSM it is empty. These three directions are therefore excluded only from $1 \to 0$ inference. Intervention phases are analyzed separately because input+output replacement affects prompt processing as well as generation.

Discovery, held-out evaluation, and directions. For each condition, candidate discovery and causal evaluation use disjoint example sets. On the discovery split, EAP-IG ranks channels (Sovrano et al., 2026) and CHA recursively refines intervention groups to one frozen candidate set J. Candidate identities, ranking, replacement values, and the discovery procedure are fixed before held-out evaluation. The same frozen J is then evaluated on held-out examples from both source states when both cohorts exist. The $0 \to 1$ direction conditions on examples with natural state $B = 0$ and counts interventions that change the endpoint to 1; $1 \to 0$ is defined analogously. For finite-answer tasks these are corrective and disruptive changes; for safety they increase and decrease jailbreak success. RQ1 uses conditions as analysis units, RQ2 analyzes intervention regimes separately, and RQ3 summarizes at the run/baseline/direction level after candidates and examples are aggregated within each condition.

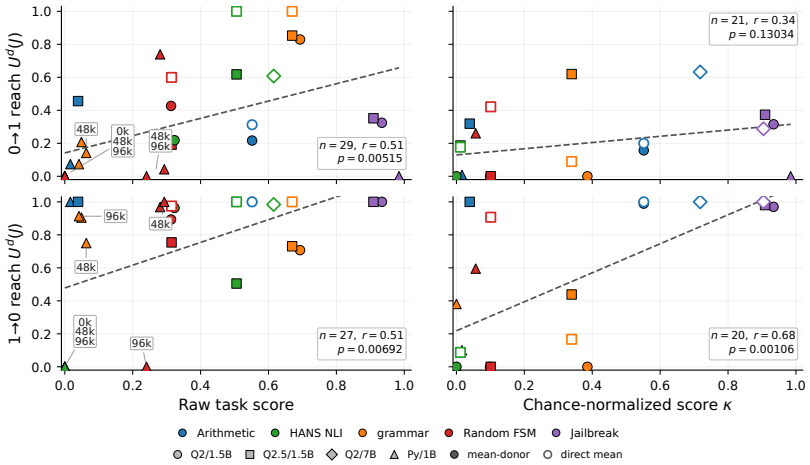


Figure 2: Competence versus directional singleton-union reach. Rows are flip directions and columns are intervention phases; dashed lines show within-phase fits; Py=Pythia, Q2=Qwen2, Q2.5=Qwen2.5.

3 RQ1: Competence, Causal Reach, and Direction

Preliminaries. For direction $d \in \{0 \to 1, 1 \to 0\}$, let $F_j^d(x)$ indicate that replacing candidate j changes the binary endpoint in direction d. *Singleton-union reach* is $U^d(J) = \Pr(\cup_{j \in J} F_j^d = 1)$ over direction-eligible examples. It measures how much of the held-out population is reachable by at least one discovered singleton. Because overtopping is a single-channel phenomenon, we do not interpret large $U^d(J)$ by itself as concentrated control; Appendix Table 2 also reports the strongest singleton reach $s_1^d = \max_j \Pr(F_j^d = 1)$.

Methods. We relate competence to $U^{0\to1}$ and $U^{1\to0}$ separately for input+output and output-only intervention. Zero-candidate conditions contribute zero reach; runs without defined held-out reach are not included in the fits. Figure 2 shows every analyzed condition and its within-phase linear association. Appendix B reports the 50-run analysis, including held-out source-state cohort sizes and counts of candidates with at least 5% direction-conditioned singleton reach, plus a sensitivity analysis restricted to final model states, removing intermediate checkpoints and repeated replacement baselines.

Results. Figure 2 shows positive competence associations for directional reach. For $0 \to 1$, input+output gives $n = 29$, $r = 0.506$, $p = .00515$, while output-only gives $n = 21$, $r = 0.341$, $p = .130$. For $1 \to 0$, input+output gives $n = 27$, $r = 0.507$, $p = .00692$, and output-only gives $n = 20$, $r = 0.676$, $p = .00106$. The abovementioned sensitivity analysis gives positive correlations in all four cases, but strong evidence remains only for output-only $1 \to 0$ (Appendix B).

Discussion. Breaking a successful computation is comparatively easy: a $1 \to 0$ intervention can disrupt any component on which that computation depends. Creating success is more demanding because enough of the computation must already be in place for one local replacement to complete or redirect it. In the full analysis, higher competence is associated with greater $0 \to 1$ causal reach, consistent with more competent conditions exposing more examples reachable by a localized causal change. RQ1 thus identifies the access points. The next question is whether those points behave like independent controls when several are used together, which motivates RQ2.

4 RQ2: Composition Under Simultaneous Intervention

Preliminaries. Singleton reach and simultaneous intervention are distinct estimands. Let $S_J(x) = \mathbf{1}[\bigcup_{j \in J} F_j(x) = 1]$ and $U(J) = \Pr(S_J = 1)$. Let $F_J(x) = 1$ when replacing the full frozen candidate set in one forward pass changes the endpoint, with $E(J) = \Pr(F_J = 1)$. The composition gap is $\Delta_{\text{comp}} = E(J) - U(J) = \Pr(\text{coalition-only}) - \Pr(\text{suppressed})$ on the shared complete-case population. CHA group interventions are part of candidate discovery: they refine a search down to the

frozen set J. Overtopping itself is evaluated afterward with singleton interventions at the candidate's structural locus, and controls are matched at the same layer or attention head. Discovery by group testing therefore does not imply that intervening on all members of J simultaneously must preserve their singleton effects.

Methods. We compare singleton-union and full-set flip masks on the same held-out examples for every nonempty candidate set with compatible simultaneous-set output. Each example falls into one of four mutually exclusive categories: *preserved*, flipped by at least one singleton and by the full set; *suppressed*, flipped by a singleton but not by the full set; *coalition-only*, flipped only by the full set; or *unaffected*. Hence $\Delta_{\text{comp}} = \Pr(\text{coalition-only}) - \Pr(\text{suppressed})$. We analyze the two intervention regimes separately because mean-donor copies an observed reference activation, whereas mean replacement inserts an average activation. Structurally matched noncandidate sets provide the specificity comparison in Appendix C.

Results. Among 25 evaluable mean-donor sets, 22 have $E(J) < U(J)$, two have $E(J) > U(J)$, and one is tied; the median composition gap is -0.146. For mean replacement, 10 of 11 evaluable sets have negative gaps and one has a positive gap, with median -0.166. Because $\Delta_{\text{comp}} = E(J) - U(J)$ contains $-U(J)$ algebraically, we do not interpret reach–gap correlations; the composition result rests on the predominance and example-level decomposition of negative gaps. Figure 3 shows the gap distribution by intervention direction and overall. Appendix Figure 8 shows the example-level source of the gap, and Appendix Table 3 summarizes the two intervention regimes, and Table 4 gives the evaluable mean-donor conditions.

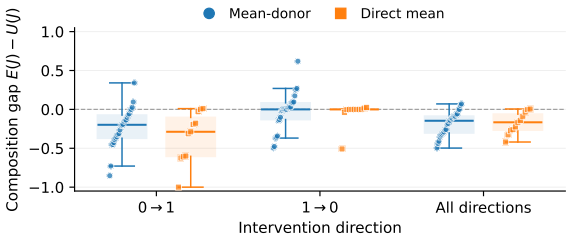


Figure 3: Composition gap by intervention direction. Negative values indicate suppression of singleton effects under joint replacement.

Discussion. RQ1 might suggest that several high-leverage channels can simply be stacked for stronger control. RQ2 shows why that inference fails. A channel can flip an example alone and lose the effect when other high-leverage channels move with it, while coalition-only examples show the converse. Appendix Figure 8 makes both interactions visible. For intervention design, the consequence is concrete: selecting the top channels individually is not enough to predict a combined edit or defense. Candidate coalitions have to be tested jointly, and the same conclusion holds for both mean-donor and mean counterfactuals. RQ3 next asks whether an individual access point concentrates its leverage around an event in dose and time.

5 RQ3: Dose Thresholds and Temporal Event Localization

Preliminaries. For a fixed candidate j, replacement strength is $h_j(\lambda, x) = (1-\lambda)h_j(x) + \lambda\widetilde{h}_j$, with $\lambda \in \{0, 0.1, \ldots, 1\}$. The *known-flip support* S_j^+ contains held-out source-state examples already known to change outcome under full replacement; S_j^- contains same-channel source-state examples that do not change. A *single persistent crossing* changes the binary endpoint exactly once and preserves the changed state at every higher sampled dose. Full-dose susceptibility guarantees a crossing when both endpoints reproduce, but it does not determine where the transition occurs or whether the path reverses. In the EIF model, rheobase is the minimum constant drive at which the stable subthreshold state disappears and spiking begins (Fourcaud-Trocmé et al., 2003); RQ3 asks whether intervention-defined causal leverage has an analogous persistent threshold geometry and whether decoding-time ablations localize a corresponding event.

Methods. We measure individual crossing paths because an aggregate rise in flip rate can hide reversals. Testing S_j^- alongside S_j^+ asks whether submaximal replacement also destabilizes examples unaffected at full-dose. A binary decision can switch even when its underlying score varies smoothly; to examine the response beneath that decision, the dose-0 and dose-1 completions define a shared token prefix and the first token where they differ. Fixing this prefix compares token competition in

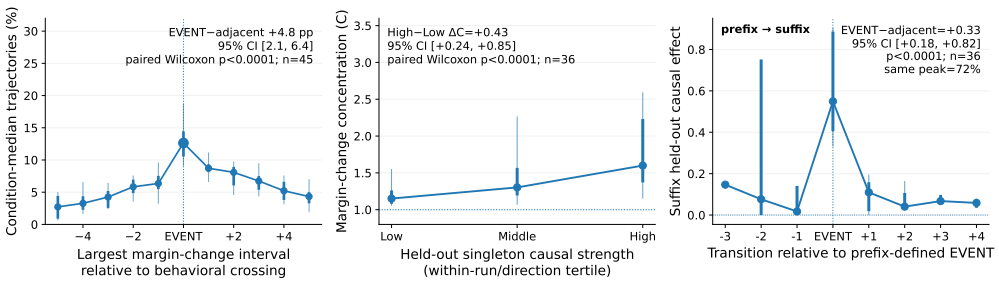


Figure 4: Dose and time localize corresponding causal events: *(a)* flips coincide with peak margin change, *(b)* stronger channels sharpen it, and *(c)* temporal sweeps recover the same decode step.

a shared context, before the continuations diverge. The *divergence-token margin* $m(\lambda)$ is the logit of the dose-1 token minus that of the dose-0 token. For endpoint-reproduced full-grid trajectories, let $\Delta m_k = m(\lambda_{k+1}) - m(\lambda_k)$. EVENT denotes the dose interval ending at the first persistent behavioural crossing. Aligning each trajectory at EVENT accommodates different thresholds and tests whether the strongest continuous change coincides with the behavioural transition. We compare peak-$|\Delta m_k|$ frequency at EVENT with adjacent offsets. Margin-change concentration is $C = (K-1)\max_k |\Delta m_k| / \sum_k |\Delta m_k|$ for $K = 11$ doses; equal changes at every interval give $C = 1$. Tertiles formed within each run/direction condition ask whether stronger channels concentrate their response more sharply while comparing channels in the same model and task.

A dose threshold does not reveal when replacement matters during generation. In a separate experiment, we hold $\lambda = 1$ and vary the temporal support of the intervention. A prefix schedule replaces the channel for the first t decode transitions, while a separately-generated complementary suffix schedule replaces it for the final t transitions. The two sweeps therefore use the same intervention strength and active-transition count but place the intervention at opposite ends of decoding. Output-only keeps prompt processing clean; input+output applies replacement during prompt processing for both schedules, holding that prompt-phase intervention fixed while decode-time support changes. The transition with the largest positive incremental gain in one sweep defines the temporal EVENT, and the complementary sweep is aligned to that event for validation, so event selection and evaluation use different temporal trajectories.

Results. Candidates exceed same-locus controls in singleton causal strength by a paired median 0.204 across 48 run/direction conditions (condition bootstrap 95% CI $[0.158, 0.240]$; one-sided paired $p = 1.39 \times 10^{-9}$; Appendix Figure 10). Across 48 known-flip conditions, the median single-persistent-crossing rate is 0.989, while the median same-channel non-flip stability rate is 0.989 (Appendix Table 6). Across individual channels, the median first persistent dose is $\lambda = 0.50$; the median of condition summaries is 0.588. Thus, on known-flip support, the typical channel reaches its persistent behavioural transition after about half of the displacement toward the full replacement, rather than requiring the endpoint intervention. Figure 4a shows peak margin-change localization at EVENT, with a median EVENT-minus-adjacent difference of 4.80 percentage points (condition bootstrap 95% CI $[2.14, 6.40]$; one-sided paired $p = 7.04 \times 10^{-7}$; $n = 45$ conditions). Stronger channels concentrate more of their total margin change into one interval: High minus Low $\Delta C = 0.431$ (95% CI $[0.243, 0.847]$; $p = 2.53 \times 10^{-8}$; $n = 36$; panel b).

The temporal sweep gives a cross-sweep localization result. Across 36 conditions with both schedules, a prefix-defined EVENT predicts excess effect in the suffix sweep of $+0.326$ (95% CI $[0.179, 0.824]$; one-sided paired $p = 1.72 \times 10^{-5}$; Figure 4c). Reversing discovery and validation gives $+0.440$ (95% CI $[0.381, 0.744]$; $p = 3.68 \times 10^{-6}$). The two sweeps select exactly the same peak transition in 72% of paired conditions and peaks within one transition in 83%. The temporal evidence therefore comes from reciprocal event localization, not from a general claim that the effect is front-loaded. Appendix Figure 16 and Appendix D.1 give the phase-resolved results.

Discussion. RQ3 asks whether a discovered access point concentrates its influence around a behavioural event. Full-dose susceptibility guarantees some crossing, but not a persistent one or its location. The individual-channel median EVENT at $\lambda = 0.50$, together with peak continuous margin change at the same crossing, shows that the binary transition and its underlying score geometry align.

Stronger channels also compress more margin change into one dose interval. The temporal experiment strengthens this interpretation because it changes a different axis: prefix and suffix schedules, selected and validated separately, localize the same narrow decode region. Dose asks how much intervention is needed; time asks when it has greatest leverage. Agreement across these two perturbations is what makes the spiking analogy informative: the evidence is for a localized threshold event in causal response, not simply for a discrete output crossing. RQ4 next asks how these control points change during learning. Appendix D.2 formalizes the threshold correspondence to EIF rheobase crossing.

6 RQ4: Learning-Time Causal Organization

Preliminaries. Mechanistic studies of grokking and induction heads show that identifiable computations can emerge or reorganize on timescales that aggregate performance does not reveal (Power et al., 2022; Nanda et al., 2023; Olsson et al., 2022). RQ4 asks two related questions: whether the channels carrying causal leverage remain stable during learning, and whether tracking this leverage can help intervene on a learned backdoor. We study these questions in two settings. First, natural Pythia checkpoints show how competence and singleton-union reach change during pretraining. Second, we introduce a backdoor during Grammar fine-tuning by poisoning training examples: a rare trigger is associated with a fixed target response (Li et al., 2023; Fenoglio et al., 2026). We compare the poisoned model with a clean model fine-tuned under the same schedule but without the trigger–target association. This gives us a known hidden behaviour that ordinary Grammar accuracy may not reveal, allowing us to test whether causal measurements can track and help control it during learning.

Methods. For the Pythia analysis, we measure competence and $U(J)$ at multiple checkpoints. Candidates are rediscovered independently at each checkpoint, so this analysis measures how checkpoint-local causal access changes during pretraining rather than tracking fixed channels over time. For Grammar poisoning, we evaluate clean and poisoned models at 0, 10, 25, 50, 75, and 100% of fine-tuning. We track ordinary Grammar accuracy, trigger-conditioned conversion to the planted target response, and the reach $U(J)$ of a fixed union of channels discovered across checkpoints. We then follow these same channel identities over training to measure how their causal roles change, using the clean model to control for changes due to ordinary fine-tuning.

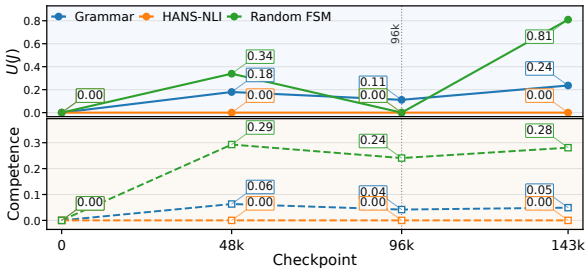


Figure 5: Natural checkpoint trajectories: $U(J)$ can change while competence changes little.

We define two channel-selection rules without using trigger annotations or attack outcomes during selection. A useful defense should suppress the trigger-induced target response while limiting damage to ordinary Grammar predictions. The first, *previous-checkpoint* rule, selects up to three overtopping channels at one checkpoint and reuses them at the next, directly testing whether intervention targets remain useful as training proceeds. The second, *checkpoint-aligned clean-reference* rule, reselects channels at each checkpoint. Among channels in the fixed union discovered at or before the current checkpoint, it ranks those whose intervention disrupts ordinary Grammar correctness more in the poisoned model than in the clean model. Selection is restricted to channels whose intervention causes at most τ ordinary-prediction damage in the poisoned model, and retains up to three channels. In both analyses, selected channels are evaluated individually and we report their mean singleton defense leverage. Figure 6 shows the clean-reference budget curve and the operating point $\tau = 0.30$.

Results. Figure 5 shows that reach and competence can change independently. In input+output random FSM, reach rises from 0.339 at $48\,000$ training steps to 0.810 at $143\,000$, while competence changes only from 0.293 to 0.281. For Grammar, reach is 0, 0.180, 0.111, and 0.236 across the displayed checkpoints despite consistently low competence.

The poisoning experiment shows the same decoupling more directly. Across the three training runs, ordinary Grammar accuracy remains similar between clean and poisoned models, while the backdoor is rapidly learned. Median attack success rate rises to 0.573 at 10%, 0.944 at 25%, and 1.000 by

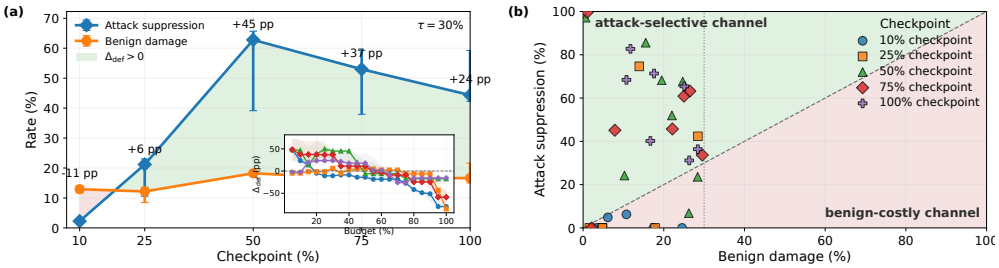


Figure 6: Clean-reference Grammar defense across three training runs. *(a)* Median suppression and benign damage at budget (τ) 30%; whiskers show IQR, inset shows budgeted leverage. *(b)* Channels with the largest poisoned-versus-clean causal excess under τ, across checkpoints.

50% of fine-tuning, while remaining near zero in the clean model. At the same time, the reach of the fixed channel union evolves differently in the clean and poisoned models (Appendix Table 8). Thus, ordinary accuracy gives little indication that the backdoor is being acquired or that the underlying causal organization is changing.

This reorganization makes static intervention targets unreliable. Channels selected at the previous checkpoint have negative median defense leverage at every subsequent checkpoint, ranging from -17.8 to -59.4 percentage points across the three runs (Appendix Table 9). In contrast, reselecting channels at the current checkpoint performs substantially better. At $\tau = 0.30$, median clean-reference leverage is -10.7 percentage points at 10% of training, followed by $+6.2$, $+44.9$, $+36.5$, and $+23.5$ points at 25, 50, 75, and 100%, respectively (Figure 6). At 10%, only two runs contain a channel satisfying the budget; from 25% onward, all three runs contribute and median leverage is positive.

Discussion. RQ4 shows that learning can change where causal leverage resides without a comparable change in ordinary task performance. Natural checkpoints show changes in checkpoint-local reach, while the fixed-union poisoning analysis directly tracks changes in channel role. As a result, a channel that is useful for suppressing the backdoor at one checkpoint may become ineffective or even counterproductive after further training. Freezing one such channel throughout training would therefore not constitute a durable defense in this setting: the model can redistribute causal leverage to other coordinates. The clean-reference analysis shows that checkpoint-aligned rescoring can identify useful singleton targets in this controlled trajectory. From 25% of fine-tuning onward, it produces positive defense leverage under budget τ, but selection and evaluation occur at the same checkpoint, so this does not by itself establish prospective transfer. Nonetheless, this result also shows that current causal measurements can in principle identify useful targets as the backdoor develops. These results concern singleton interventions. They do not imply that the highest-ranked channels can simply be combined into a multi-channel defense: RQ2 shows that simultaneous interventions can interact non-additively and must be validated jointly. Appendix E reports the full learning trajectory and defense comparison.

7 Final Discussion & Related Work

Cross-RQ synthesis. The four questions progressively narrow the picture. RQ1 shows that localized causal access is common enough to study systematically and that corrective reach tends to rise with competence. RQ2 shows that these access points are not independent knobs: joint intervention can suppress singleton effects or create coalition-only effects. RQ3 shows that one access point often has a persistent dose threshold and a reproducible high-leverage event in decoding time. RQ4 shows that the coordinates carrying this leverage can redistribute during learning while ordinary competence changes little. Together, overtopping is best understood as localized causal access embedded in a distributed, interacting, and changing computation.

What the four RQs permit, and rule out, for poisoning defense. A fixed single-channel freeze or blacklist is not robust across checkpoints because RQ4 shows that leverage redistributes and previous targets can reverse sign. RQ2 shows that a top-k list cannot be treated as additive, while RQ1 shows that generation-only auditing can miss access exposed during prompt processing. Behaviour-only

monitoring is also insufficient here because ordinary Grammar accuracy stays close while the trigger-dependent behaviour is acquired. The supported alternative is adaptive: recompute leverage, compare the current model with a matched clean reference under a benign-damage budget, and validate selected coalitions jointly. RQ3 suggests a further control axis, event-localized inference-time intervention, which remains to be tested directly together with training-time prevention and coalition deployment.

Localization alone does not give composition. Circuit discovery, sparse representations, and compact steering directions show that small internal sets can carry large behavioural influence (Wang et al., 2023; Conmy et al., 2023; Cunningham et al., 2024; Arditi et al., 2024). RQ1 fits that picture, but RQ2 shows why localization is incomplete: singleton leverage does not determine joint effect. The strong directional difference suggests that the behavioural role of a localized component depends on the surrounding intervention context. This supports MechaRule's regime-conditioned account of overtopping (Sovrano et al., 2026) and is compatible with the "bag of heuristics" account of arithmetic, where behaviour is produced by combinations of localized heuristic neurons (Nikankin et al., 2025). It also parallels causal mediation results in which sparse mediators can amplify or repress one another's effects (Vig et al., 2020).

Threshold geometry adds a response-level view. Graded steering and confidence studies show that continuous internal changes can produce threshold-like choices (Turner et al., 2023; Rimsky et al., 2024; Kumaran et al., 2026), while discontinuous metrics can make smooth scaling trends appear emergent (Schaeffer et al., 2023). RQ3 instead measures the response of an identified causal channel across dose and decoding time. A smooth latent trajectory can cross a decision boundary this way; the spiking connection is the persistent event geometry, not a claim of biological identity.

Learning changes the meaning of a causal target. Grokking and induction-head studies show that computations can form and reorganize during training (Power et al., 2022; Nanda et al., 2023; Olsson et al., 2022). RQ4 adds that intervention leverage itself moves without comparable competence change, so a channel useful at one checkpoint need not remain a durable control target. Together with RQ2, this links developmental reorganization to the need for remapping and joint validation.

8 Limitations

Construct validity. Overtopping depends on basis, replacement counterfactual, phase, and endpoint. The Grammar defense experiment operationalizes defense as suppression after poisoning is learned, not prevention during fine-tuning. Nonetheless, we separate phases and replacement rules, freeze candidates before held-out evaluation, use structurally matched controls (Appendix C), pair known-flip dose sweeps with same-channel non-flip controls (Appendix D), and evaluate defense through prospective transfer and clean-reference suppression together with ordinary-behaviour damage.

Internal validity. Task, model, phase, and replacement axes are not fully crossed, and the 11-dose RQ3 grid limits threshold resolution. However, RQ1 analyzes the two phases separately, a reduced sensitivity analysis removes intermediate Pythia checkpoints and repeated replacement observations (Appendix B), and RQ3 resamples run/direction conditions rather than nested examples or channels.

External validity. The sampled architectures, training procedures, and channel bases do not cover all language models. Nonetheless, five behaviours, several model families and sizes, two phases, and both natural and controlled learning trajectories broaden the tested scope.

9 Conclusion

Overtopping reveals a recurring spiking-like causal response geometry in LLMs: causal leverage is direction-dependent, singleton effects need not compose, and susceptible examples often undergo a single persistent transition under graded intervention, aligned with a localized high-leverage event during decoding. At the same time, the coordinates carrying this leverage can reorganize during learning even when ordinary behaviour changes little. This makes static intervention targets unreliable in the settings we study and motivates adaptive causal control based on current-model measurements and jointly validated interventions. Future work should test whether circuit topology, redundancy, response curvature, or training dynamics predict where causal access moves, linking mechanistic description to editing, monitoring, and training objectives before failures become visible in behaviour.

AI USE DISCLOSURE

Generative AI tools are the subject of the experiments; they were also used for proofreading and editing the paper. The authors reviewed the AI-generated text and take responsibility for the submission.

REPRODUCIBILITY STATEMENT

The appendices specify the task predicates, prompt/parser conventions, intervention definitions, held-out protocol, statistical units, diagnostics, and formal threshold correspondence. Code and exact experimental configuration are available at https://anonymous.4open.science/r/overtopping-phenomenology-770F.

REFERENCES

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-4322.

Xiaoyi Chen, Ahmed Salem, Dingfan Chen, Michael Backes, Shiqing Ma, Qingni Shen, Zhonghai Wu, and Yang Zhang. BadNL: Backdoor Attacks against NLP Models with Semantic-preserving Improvements. In *Proceedings of the 37th Annual Computer Security Applications Conference*, ACSAC '21, pages 554–569, New York, NY, USA, dec 2021. Association for Computing Machinery. ISBN 978-1-4503-8579-4. doi: 10.1145/3485832.3485837.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems*, volume 36, pages 16318–16352, 2023. doi: 10.52202/075280-0719.

Hoagy Cunningham, Aidan Ewart, Logan Riggs Smith, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations*, 2024. doi: 10.48550/arXiv.2309.08600.

Dario Fenoglio, Martin Gjoreski, Edoardo Gabrielli, and Marc Langheinrich. Security and Privacy in Federated Learning: A Survey and Unified Taxonomy of Threat Models, Attacks, and Defenses. *SSRN Electronic Journal*, (6771005), May 2026. doi: 10.2139/ssrn.6771005.

Nicolas Fourcaud-Trocmé, David Hansel, Carl van Vreeswijk, and Nicolas Brunel. How spike generation mechanisms determine the neuronal response to fluctuating inputs. *Journal of Neuroscience*, 23(37):11628–11640, 2003. doi: 10.1523/JNEUROSCI.23-37-11628.2003.

Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah D. Goodman, Christopher Potts, and Thomas Icard. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, 26(83):1–64, 2025. doi: 10.48550/arXiv.2301.04709. URL https://www.jmlr.org/papers/v26/23-0058.html.

Dharshan Kumaran, Nathaniel Daw, Simon Osindero, Petar Veličković, and Viorica Patraucean. Causal evidence that language models use confidence to drive behaviour. *Nature Machine Intelligence*, 2026. doi: 10.1038/s42256-026-01293-x.

Yudong Li, Shigeng Zhang, Weiping Wang, and Hong Song. Backdoor Attacks to Deep Learning Models and Countermeasures: A Survey. *IEEE Open Journal of the Computer Society*, 4:134–146, 2023. ISSN 2644-1268. doi: 10.1109/OJCS.2023.3267221.

Wolfgang Maass. Networks of spiking neurons: The third generation of neural network models. *Neural Networks*, 10(9):1659–1671, 1997. doi: 10.1016/S0893-6080(97)00011-7.

Hesham Mostafa. Supervised learning based on temporal coding in spiking neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 29(7):3227–3235, 2018. doi: 10.1109/TNNLS.2017.2726060.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. In *International Conference on Learning Representations*, 2023. doi: 10.48550/arXiv.2301.05217.

Emre O. Neftci, Hesham Mostafa, and Friedemann Zenke. Surrogate gradient learning in spiking neural networks: Bringing the power of gradient-based optimization to spiking neural networks. *IEEE Signal Processing Magazine*, 36(6):51–63, 2019. doi: 10.1109/MSP.2019.2931595.

Yaniv Nikankin, Anja Reusch, Aaron Mueller, and Yonatan Belinkov. Arithmetic without algorithms: Language models solve math with a bag of heuristics. In *International Conference on Learning Representations*, 2025. doi: 10.48550/arXiv.2410.21272.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads. *Transformer Circuits Thread*, 2022. doi: 10.48550/arXiv.2209.11895. URL https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html.

Jorge Pérez, Pablo Barceló, and Javier Marinkovic. Attention is turing-complete. *Journal of Machine Learning Research*, 22(75):1–35, 2021. URL https://www.jmlr.org/papers/v22/20-302.html.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv*, 2022. doi: 10.48550/arXiv.2201.02177.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering Llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.828.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-2425.

Francesco Sovrano, Gabriele Dominici, and Marc Langheinrich. Neuron-Anchored Rule Extraction for Large Language Models via Contrastive Hierarchical Ablation. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '26, pages 4369–4380. Association for Computing Machinery, 2026. ISBN 9798400722592. doi: 10.1145/3770855.3818091.

Alexander Matt Turner, Lisa Thiergart, David Udell, Gavin Leech, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv*, 2023. doi: 10.48550/arXiv.2308.10248.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. volume 33, pages 12388–12401, 2020. URL https://papers.neurips.cc/paper_files/paper/2020/hash/92650b2e92217715fe312e6fa7b90d82-Abstract.html.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: A circuit for indirect object identification in GPT-2 small. In *International Conference on Learning Representations*, 2023. doi: 10.48550/arXiv.2211.00593. URL https://openreview.net/forum?id=NpsVSN6o4ul.

Xiangyu Zhou, Yao Qiang, Saleh Zare Zade, Mohammad Amin Roshani, Prashant Khanduri, Douglas Zytko, and Dongxiao Zhu. Learning to Poison Large Language Models for Downstream Manipulation. *arXiv*, (arXiv:2402.13459), May 2025. doi: 10.48550/arXiv.2402.13459.

APPENDIX ROADMAP

Section	Contents
A	Common task, prompt, parser, intervention, and held-out conventions.
B	RQ1 full directional reach/density results and phase/size sensitivity plots.
C	RQ2 per-condition composition decomposition, coalition boundary cases, and matched-set specificity.
D	RQ3 graded-response geometry, complementary temporal ablations, formal EIF rheobase correspondence, and secondary diagnostics.
E	RQ4 natural checkpoints and the controlled Grammar same-channel analysis.

A COMMON TASK, PROMPT, AND INTERVENTION PROTOCOL

Each task is reduced to a binary behaviour predicate used consistently before and after intervention. The protocol fixes the model checkpoint, tokenizer, prompt template, parser, admissible-output set, task slice, intervention phase, and replacement baseline. Table 1 summarizes the task interfaces.

Each evaluation example consists of a prompt x, a target answer or target label y, a raw model continuation $g_\theta(x)$, a parser π that maps the raw continuation to a normalized answer, and a binary predicate $B(x) = \mathbf{1}\{\pi(g_\theta(x)) = y\}$. For jailbreaking, $B(x) = 1$ when the continuation is classified as a successful jailbreak. A singleton or group intervention is counted as a flip when the same parser and predicate assign a different binary value after replacement.

Task family	Prompt/output interface	Parsed output domain	Binary predicate	Chance
Arithmetic	Expression prompts request the result of an arithmetic expression and a direct numerical continuation.	Normalized numerical answer.	Exact match between parsed answer and target numerical answer.	0
Grammar acceptability	A sentence is presented and the model gives a short acceptability judgment.	Boolean acceptability label.	Correct binary acceptability judgment.	0.5
NLI	A premise and hypothesis are presented with an instruction to output only E or N.	Entailment/non-entailment label.	Correct binary entailment judgment.	0.5
Random FSM	A deterministic binary finite-state machine, start state, transitions, and input bitstring are listed; the model gives the final state.	Standalone generated state label.	Exact match with simulator-computed final state.	mean over sampled K
Safety	A safety prompt is answered by the model and scored by the evaluator.	Successful harmful completion versus non-jailbreak.	Jailbreak predicate.	n/a

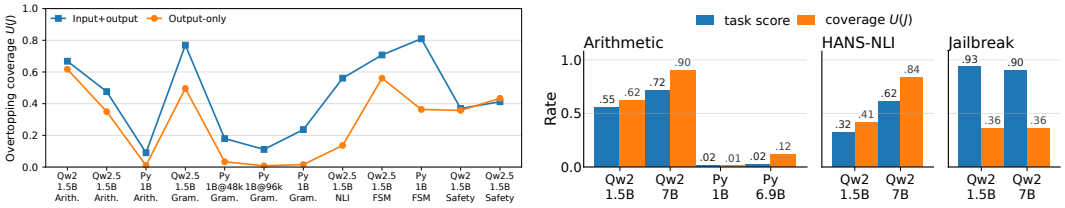
Table 1: Task interfaces, binary predicates, and chance baselines.

Arithmetic uses bare expression prompts; Grammar asks for a short acceptability judgment; NLI requests an E/N decision; random FSM supplies a freshly sampled transition table, start state, and input bitstring; safety prompts are evaluated by a fixed binary harmful-completion classifier. The same textual prompt and parser are used before and after intervention. Random FSM is a constructed prompt-following diagnostic: the transition table changes by example, so the model cannot solve it by memorizing one dataset-level input–output map. In low-competence conditions, intervention coverage is interpreted as causal control over the prompted decision process or output schema rather than as evidence of task mastery.

B RQ1 Supplement: Directional Reach and Sensitivity

Table 2 gives the 50-run analysis. For $b \in \{0, 1\}$, let $n_b = \#\{x : B(x) = b\}$ denote the held-out source-state cohort size. For direction d, let $N^d_\tau = \#\{j \in J : \Pr(F^d_j = 1) \geq \tau\}$, so $N^d_{.05}$ counts candidates whose direction-conditioned singleton reach is at least 0.05. Fourteen rows have $|J| = 0$. For the two Pythia-1B Jailbreak phases and Pythia-1B 0k input+output random FSM, the $0 \to 1$ search completed with no retained candidates and its zero reach is included in RQ1. Their displayed $1 \to 0$ zeros are not used in $1 \to 0$ inference because the required $B = 1$ discovery subset is insufficient or empty.

Figure 7 isolates two design sensitivities that help interpret the primary association. The phase comparison shows how changing the intervention locus from output-only to input+output can shift reach within matched conditions. The size comparison gives the available same-family model contrasts. These comparisons are descriptive because the experimental matrix is not a complete factorial crossing of phase and model size.



(a) Matched intervention-phase comparison.

(b) Available same-family size comparisons.

Figure 7: RQ1 sensitivity to intervention phase and available within-family size comparisons.

Cohort size and concentration of reach. The additional diagnostics qualify the directional-reach estimates in two ways. First, the source-state cohorts can be strongly imbalanced: several large $1! \to !0$ reaches are estimated from comparatively small $B = 1$ cohorts, so n_0/n_1 provides the relevant denominator for interpreting those proportions. Second, $N^d_{.05}$ distinguishes reach concentrated in a few strong singleton channels from reach distributed across many candidates. This distinction is substantial across conditions. For example, Jailbreak has relatively few $0! \to !1$ candidates above 5% reach despite nonzero union reach, whereas its $1! \to !0$ direction can contain many such candidates and near-unit singleton reach. Thus similar union-level accessibility need not imply the same internal concentration of causal control.

Sensitivity after removing repeated checkpoints and baselines. Removing intermediate Pythia checkpoints and retaining one preferred replacement condition per task–model–phase cell gives 33 defined $0 \to 1$ cells, 16 input+output and 17 output-only. Pearson $r = .231$ ($p = .389$) for input+output and $r = .407$ ($p = .105$) for output-only; the corresponding Spearman values are $\rho = .249$ ($p = .353$) and $\rho = .402$ ($p = .109$). For $1 \to 0$, the three source-state-limited directions remain excluded, leaving 15 input+output and 16 output-only cells; Pearson $r = .245$ ($p = .379$) and $r = .727$ ($p = .00141$), respectively. Thus the reduced $0 \to 1$ association is positive but unresolved in both phases, whereas the $1 \to 0$ association remains supported for output-only intervention.

Phase and replacement baseline alter the intervention counterfactual, while model-size comparisons are available only for selected task–phase cells. The main RQ1 fits therefore separate intervention phase and describe the observed condition matrix; this sensitivity analysis removes repeated checkpoint and replacement-baseline observations.

C RQ2 Supplement: Composition, Boundary Cases, and Specificity

Table 3 summarizes the evaluable nonempty RQ2 conditions by intervention regime. Among 25 mean-donor sets, 22 gaps are negative, two positive, and one tied, with median -0.146. Among 11 mean sets, 10 gaps are negative and one positive, with median -0.166.

Table 4 gives the per-condition mean-donor decomposition used for the aggregate result. The two positive mean-donor conditions are listed in Table 5. They show that suppression is the dominant interaction pattern rather than a mathematical necessity.

Task	Model	Phase	Ref.	Compet.	n_0/n_1	$U^{0\to1}$	$s_1^{0\to1}$	$U^{1\to0}$	$s_1^{1\to0}$	$N_{.05}^{0\to1}/N_{.05}^{1\to0}$	$\lvert J\rvert$
Primary model–task sweep											
Arithmetic	Pythia-1B	I+O	MD	0.016	9822/178	0.074	0.035	1.000	1.000	0/25	25
Arithmetic	Pythia-1B	Out	MD	0.016	9810/190	0.008	0.008	0.100	0.100	0/1	1
Arithmetic	Qwen2-1.5B	I+O	MD	0.552	217/295	0.217	0.078	1.000	1.000	4/11	11
Arithmetic	Qwen2-1.5B	Out	MD	0.552	229/283	0.157	0.074	0.989	0.922	2/6	7
Arithmetic	Qwen2.5-1.5B	I+O	MD	0.039	9626/374	0.456	0.133	1.000	0.992	22/598	600
Arithmetic	Qwen2.5-1.5B	Out	MD	0.039	489/23	0.319	0.127	1.000	0.957	10/466	556
Grammar	Pythia-1B	I+O	MD	0.049	491/21	0.208	0.189	0.905	0.810	2/51	72
Grammar	Pythia-1B	Out	MD	0.000	491/21	0.000	0.000	0.381	0.381	0/2	2
Grammar	Qwen2-1.5B	I+O	MD	0.693	252/567	0.829	0.496	0.707	0.367	5/5	5
Grammar	Qwen2-1.5B	Out	MD	0.386	–	0.000	0.000	0.000	0.000	0/0	0
Grammar	Qwen2.5-1.5B	I+O	MD	0.670	163/349	0.853	0.472	0.731	0.450	40/43	45
Grammar	Qwen2.5-1.5B	Out	MD	0.339	163/349	0.620	0.472	0.438	0.281	4/7	7
Random FSM	Pythia-1B	I+O	MD	0.281	284/126	0.739	0.366	0.968	0.825	35/47	47
Random FSM	Pythia-1B	Out	MD	0.057	284/126	0.261	0.158	0.595	0.357	7/11	12
Random FSM	Qwen2-1.5B	I+O	MD	0.313	279/131	0.427	0.315	0.893	0.847	2/3	3
Random FSM	Qwen2-1.5B	Out	MD	0.099	–	0.000	0.000	0.000	0.000	0/0	0
Random FSM	Qwen2.5-1.5B	I+O	MD	0.314	292/118	0.192	0.099	0.754	0.737	3/8	13
Random FSM	Qwen2.5-1.5B	Out	MD	0.101	–	0.000	0.000	0.000	0.000	0/0	0
NLI	Pythia-1B	I+O	MD	0.000	–	0.000	0.000	0.000	0.000	0/0	0
NLI	Pythia-1B	Out	MD	0.000	–	0.000	0.000	0.000	0.000	0/0	0
NLI	Qwen2-1.5B	I+O	MD	0.323	151/54	0.219	0.199	0.963	0.759	6/15	22
NLI	Qwen2-1.5B	Out	MD	0.000	–	0.000	0.000	0.000	0.000	0/0	0
NLI	Qwen2.5-1.5B	I+O	MD	0.506	102/103	0.618	0.608	0.505	0.466	5/5	13
NLI	Qwen2.5-1.5B	Out	MD	0.012	102/103	0.186	0.186	0.087	0.087	1/1	1
Jailbreak	Pythia-1B	I+O	MD	0.984	–	0.000	0.000	0.000	0.000	0/0	0
Jailbreak	Pythia-1B	Out	MD	0.984	–	0.000	0.000	0.000	0.000	0/0	0
Jailbreak	Qwen2-1.5B	I+O	MD	0.934	1461/104	0.324	0.052	1.000	1.000	1/174	174
Jailbreak	Qwen2-1.5B	Out	MD	0.934	479/33	0.315	0.069	0.970	0.667	1/46	48
Jailbreak	Qwen2.5-1.5B	I+O	MD	0.909	1419/146	0.352	0.075	1.000	1.000	2/124	124
Jailbreak	Qwen2.5-1.5B	Out	MD	0.909	461/51	0.373	0.091	0.980	0.745	4/66	66
Larger-model check											
Arithmetic	Qwen2-7B	Out	M	0.718	136/376	0.632	0.250	1.000	1.000	35/66	73
Jailbreak	Qwen2-7B	Out	M	0.904	462/50	0.288	0.063	1.000	1.000	2/112	133
NLI	Qwen2-7B	I+O	M	0.615	79/126	0.608	0.241	0.984	0.960	13/26	33
Mean-replacement check											
Arithmetic	Qwen2-1.5B	I+O	M	0.552	4516/5484	0.313	0.072	1.000	0.995	6/14	14
Arithmetic	Qwen2-1.5B	Out	M	0.552	4516/5484	0.200	0.054	1.000	0.995	2/14	14
Grammar	Qwen2.5-1.5B	I+O	M	0.670	268/551	1.000	0.840	1.000	0.976	39/40	41
Grammar	Qwen2.5-1.5B	Out	M	0.339	268/551	0.090	0.090	0.167	0.167	1/1	1
NLI	Qwen2.5-1.5B	I+O	M	0.506	102/103	1.000	0.990	1.000	1.000	10/13	13
NLI	Qwen2.5-1.5B	Out	M	0.012	102/103	0.176	0.176	0.087	0.087	1/1	1
Random FSM	Qwen2.5-1.5B	I+O	M	0.314	292/118	0.599	0.291	0.975	0.881	8/18	18
Random FSM	Qwen2.5-1.5B	Out	M	0.101	292/118	0.421	0.360	0.907	0.898	4/4	4
Pythia checkpoint check											
Grammar	Pythia-1B 0k	I+O	MD	0.000	–	0.000	0.000	0.000	0.000	0/0	0
Grammar	Pythia-1B 48k	I+O	MD	0.063	480/32	0.142	0.067	0.750	0.594	2/17	18
Grammar	Pythia-1B 96k	I+O	MD	0.042	489/23	0.074	0.035	0.913	0.739	0/19	22
NLI	Pythia-1B 0k	I+O	MD	0.000	–	0.000	0.000	0.000	0.000	0/0	0
NLI	Pythia-1B 48k	I+O	MD	0.000	–	0.000	0.000	0.000	0.000	0/0	0
NLI	Pythia-1B 96k	I+O	MD	0.000	–	0.000	0.000	0.000	0.000	0/0	0
Random FSM	Pythia-1B 0k	I+O	MD	0.000	–	0.000	0.000	0.000	0.000	0/0	0
Random FSM	Pythia-1B 48k	I+O	MD	0.293	283/127	0.042	0.042	1.000	1.000	0/1	2
Random FSM	Pythia-1B 96k	I+O	MD	0.241	–	0.000	0.000	0.000	0.000	0/0	0

Table 2: All 50 configured runs. Fourteen rows have $|J| = 0$. n_0/n_1 gives the held-out $B = 0/B = 1$ source-state cohort sizes when directional singleton evaluation is defined; "–" on structural-zero rows means that evaluation was not run, not that a source state is absent. $N_{.05}^{0\to1}/N_{.05}^{1\to0}$ counts candidates with singleton reach at least .05. $s_1^d = \max_j \Pr(F_j^d = 1)$. Ref.: MD, mean donor; M, mean.

On the shared complete-case population, $U(J) = P(\text{preserved}) + P(\text{suppressed})$ and $E(J) = P(\text{preserved}) + P(\text{coalition-only})$, hence $E(J) - U(J) = P(\text{coalition-only}) - P(\text{suppressed})$. This identity holds numerically in every analyzed condition. A negative gap therefore has a concrete example-level meaning: loss of singleton-reachable examples under joint intervention exceeds the number of new coalition-only examples. Figure 8 visualizes these two masses directly.

Replacement	n	$E < U$	$E > U$	Median $E - U$	Pearson r	p
mean-donor	25	22	2	-0.146	-0.535	0.00587
mean	11	10	1	-0.166	-0.400	0.223

Table 3: RQ2 composition by replacement regime for every evaluable nonempty candidate set.

Task	Model	Ph.	$\|J\|$	$U(J)$	$E(J)$	Δ_{comp}	Supp.	Coalition-only
Arithmetic	Pythia-1B	I+O	25	0.090	0.033	-0.057	0.065	0.008
Arithmetic	Pythia-1B	Out	1	0.010	0.004	-0.006	0.009	0.003
Arithmetic	Qwen2-1.5B	I+O	11	0.668	0.576	-0.092	0.098	0.006
Arithmetic	Qwen2-1.5B	Out	7	0.617	0.506	-0.111	0.309	0.197
Arithmetic	Qwen2.5-1.5B	I+O	600	0.476	0.037	-0.439	0.439	0.000
Arithmetic	Qwen2.5-1.5B	Out	556	0.350	0.092	-0.258	0.293	0.035
Grammar	Pythia-1B	I+O	72	0.236	0.041	-0.195	0.199	0.004
Grammar	Pythia-1B	Out	2	0.016	0.066	+0.051	0.000	0.051
Grammar	Qwen2-1.5B	I+O	5	0.745	0.388	-0.357	0.408	0.051
Grammar	Qwen2.5-1.5B	I+O	45	0.770	0.682	-0.088	0.271	0.184
Grammar	Qwen2.5-1.5B	Out	7	0.496	0.566	+0.070	0.238	0.309
Random FSM	Pythia-1B	I+O	47	0.810	0.312	-0.498	0.510	0.012
Random FSM	Pythia-1B	Out	12	0.363	0.183	-0.180	0.180	0.000
Random FSM	Qwen2-1.5B	I+O	3	0.576	0.312	-0.263	0.298	0.034
Random FSM	Qwen2.5-1.5B	I+O	13	0.354	0.215	-0.139	0.176	0.037
NLI	Qwen2-1.5B	I+O	22	0.415	0.268	-0.146	0.176	0.029
NLI	Qwen2.5-1.5B	I+O	13	0.561	0.151	-0.410	0.434	0.024
NLI	Qwen2.5-1.5B	Out	1	0.137	0.137	0.000	0.000	0.000
Jailbreak	Qwen2-1.5B	I+O	174	0.369	0.066	-0.303	0.303	0.000
Jailbreak	Qwen2-1.5B	Out	48	0.357	0.068	-0.289	0.293	0.004
Jailbreak	Qwen2.5-1.5B	I+O	124	0.412	0.093	-0.319	0.319	0.000
Jailbreak	Qwen2.5-1.5B	Out	66	0.434	0.102	-0.332	0.336	0.004
Grammar	Pythia-1B 48k	I+O	18	0.180	0.064	-0.115	0.133	0.018
Grammar	Pythia-1B 96k	I+O	22	0.111	0.045	-0.066	0.070	0.004
Random FSM	Pythia-1B 48k	I+O	2	0.339	0.251	-0.088	0.183	0.095

Table 4: Mean-donor composition for all 25 evaluable nonempty candidate sets. Suppression is singleton-reachable mass lost under the full-set intervention; coalition-only is new mass reached only by the full-set intervention.

Figure 9 compares discovered sets with structurally matched noncandidate sets that preserve layer, locus, channel type, cardinality, phase, baseline, and examples. Positive specificity means that the discovered set has the larger simultaneous-intervention effect. The matched-control diagnostic is defined for the conditions shown in the figure, with the strongest task-level medians in NLI and random FSM.

D RQ3 Supplement: Spiking-Like Diagnostics and Temporal Geometry

Figure 10 compares discovered candidates with same-layer or same-head structural controls. Candidates have substantially larger singleton causal strength, and threshold-testable behaviour is concen-

Task	Model	Phase	$U(J)$	$E(J)$	$E - U$
Grammar	Qwen2.5-1.5B	Out	0.496	0.566	+0.070
Grammar	Pythia-1B	Out	0.016	0.066	+0.051

Table 5: Mean-donor super-additive boundary cases. Positive $E(J) - U(J)$ means that coalition-only gains exceed suppression of singleton-reachable examples.

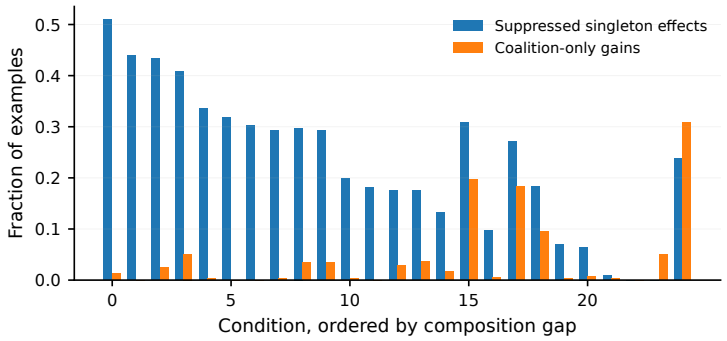


Figure 8: Example-level decomposition of the mean-donor composition gap. Conditions are sorted by $E(J) - U(J)$. The two displayed masses are the singleton-reachable examples suppressed by the full-set intervention and the coalition-only full-set effects; their difference is the composition gap.

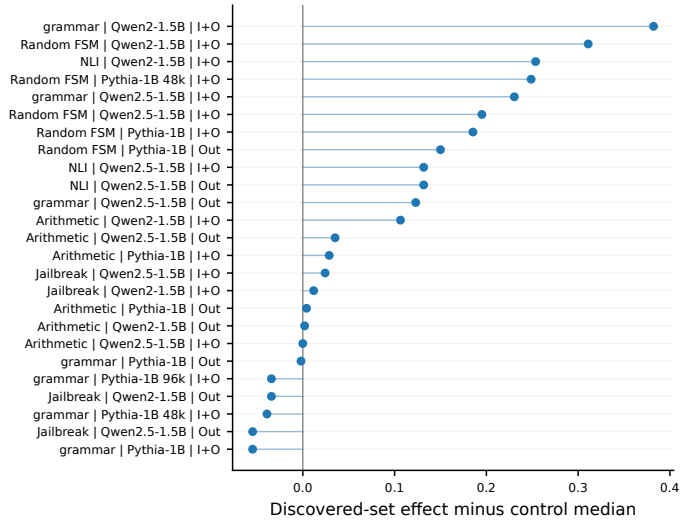


Figure 9: Matched-set specificity by task. Points are conditions; vertical marks are task medians.

trated among those high-leverage candidates. This links the graded crossing to the discovered causal population rather than to arbitrary coordinates at the same locus.

Salience proxies do not reproduce the overtopping separation. Figure 10 shows the strong candidate–control separation in intervention-defined singleton strength. By contrast, endogenous WANDA and activation-magnitude diagnostics show only small candidate–control differences in held-out threshold association (Table 7). These proxy diagnostics are supplementary to the causal dose sweep.

Strength matching asks whether an endogenous threshold-association diagnostic adds information beyond held-out causal strength. Figure 11 pairs each candidate with a same-locus control of similar causal strength. Across six run/direction conditions with valid matched pairs (136 unit pairs), the median candidate-minus-control threshold-MCC difference is -0.00145 (95% condition-bootstrap CI $[-0.110, 0.015]$; one-sided Wilcoxon $p = .781$), with candidates higher in three of six conditions.

Scope	Known-flip cond.	Non-flip cond.	Single persistent	First persistent dose	Non-flip stable
Behavioural graded population	48	48	0.989	0.588	0.989

Table 6: Condition-level graded response. Values are medians of within-channel summaries.

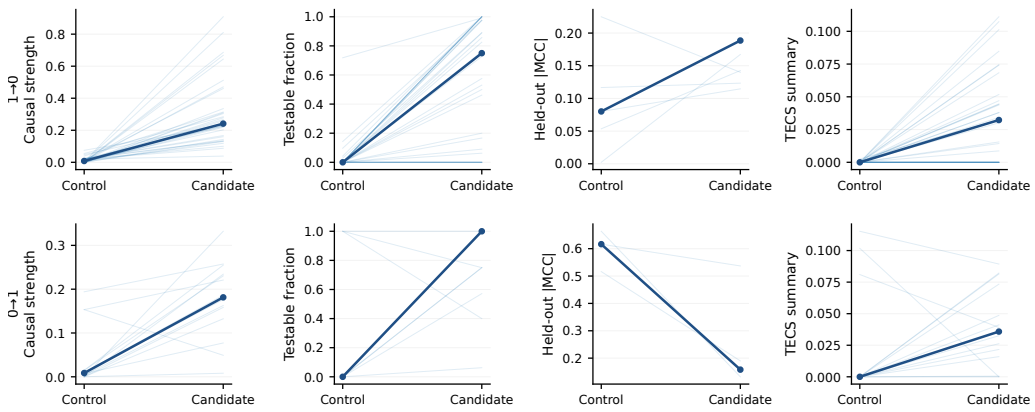


Figure 10: Candidate-versus-control diagnostics by discovery direction. Columns summarize singleton causal strength, threshold-testable fraction, held-out association with an endogenous scalar proxy, and a lower bound on threshold-event causal strength. Lines pair control and candidate summaries within the same run/direction condition.

Table 7: Candidate–control differences in held-out threshold association for endogenous salience proxies. ΔMCC is candidate minus same-locus control; p is a one-sided Wilcoxon test and q is Benjamini–Hochberg adjusted across the feature analysis.

Proxy	n	ΔMCC	p	q
WANDA layer z-score	8	0.016	.473	.880
WANDA percentile rank	8	0.006	.578	.880
WANDA raw value	8	0.000	.680	.880
|activation| layer z-score	8	0.013	.473	.880
|activation| raw value	8	0.000	.680	.880

The matched diagnostic therefore adds no candidate separation beyond causal strength, placing the evidential weight on the graded behavioural crossing, margin localization, and temporal cross-sweep validation.

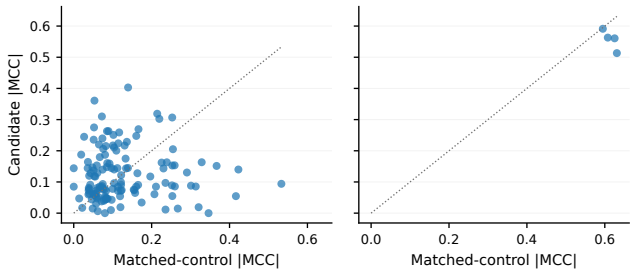


Figure 11: Strength-matched threshold-association sensitivity. Candidate held-out association is plotted against a same-locus control matched for causal strength. The diagonal denotes equal association; the matched analysis does not show a candidate advantage after conditioning on causal strength.

Figure 12 summarizes the nested threshold-event causal-strength lower bound. The candidate distribution extends beyond the matched-control distribution in both discovery directions. This endogenous-observability analysis is secondary to the graded behavioural crossing, but it shows that the intervention-defined phenomenon can also leave a measurable signature in internal scalar summaries.

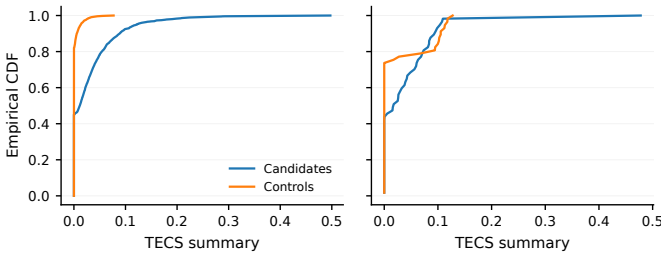


Figure 12: Empirical distribution of the nested threshold-event causal-strength lower bound for discovered candidates and noncandidate controls, separated by discovery direction.

Population event localization and strength sharpening. The continuous-margin analysis covers 45 run/direction conditions. The EVENT-minus-adjacent contrast is $+4.80$ pp (95% condition-bootstrap CI $[2.14, 6.40]$ pp; one-sided paired $p = 7.04 \times 10^{-7}$). Stronger channels show a median High-minus-Low concentration difference of $+0.431$ across 36 paired conditions (95% CI $[0.243, 0.847]$; $p = 2.53 \times 10^{-8}$). Figure 13 shows the raw population localization separately from the main three-panel summary.

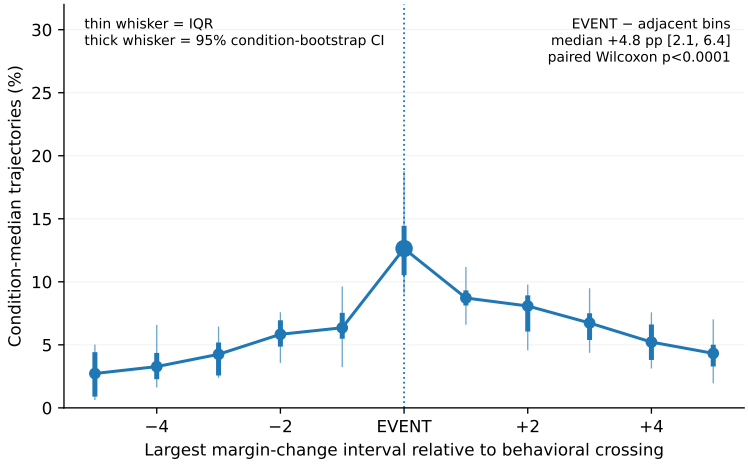


Figure 13: Population event localization before adjacent-offset normalization. Thin whiskers show interquartile ranges; thick whiskers show 95% condition-bootstrap intervals. The behavioral crossing is the EVENT offset.

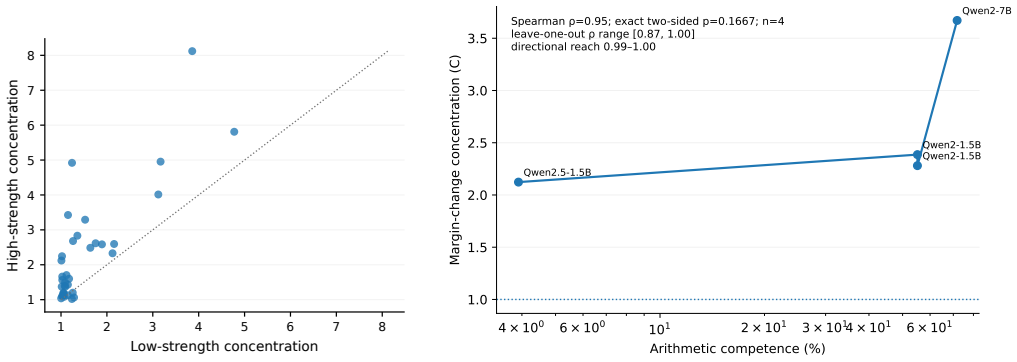
The High–Low concentration contrast is positive in 32 of 36 paired run/direction groups. As a task-specific sensitivity, the Arithmetic output-only $1 \to 0$ slice has $\rho = 0.949$ between competence and C (exact two-sided permutation $p = 0.167$, $n = 4$). With four points this slice is descriptive; Figure 14 therefore centers the within-condition strength contrast and shows the Arithmetic slice as context.

Figure 15 reports the secondary-effect preemption diagnostic used to check whether the dominant event coincides with a reduced secondary contribution.

Continuous margin shifts and binary decisions. For a fixed token history and a differentiable downstream scalar score g, a local write perturbation obeys $g_{\mathrm{do}(A)} - g = \langle \nabla_m g, \Delta m_A \rangle + R_A$, where R_A is the Taylor remainder. Scaling the perturbation can move the score across a decision boundary. Autoregressive continuation can then amplify or redirect that change through later token choices.

D.1 Temporal ablation protocol

The temporal analysis uses known-flip support at full intervention strength and evaluates both output-only and input+output phases. The first generated token is determined by prompt-prefill logits. In



(a) Paired low/high strength tertiles.

(b) Arithmetic competence sensitivity.

Figure 14: Condition-level robustness checks for margin-change concentration. Left, arrows connect low- and high-strength tertiles within the same run/direction condition. Right, the prespecified Arithmetic output-only $1 \to 0$ slice relates competence to concentration.

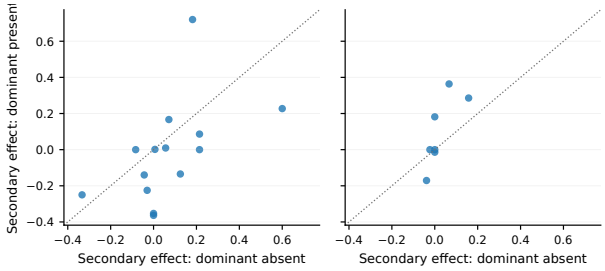


Figure 15: Secondary-effect comparisons: $1 \to 0$ left, $0 \to 1$ right. Points below equality indicate a smaller secondary contribution when the dominant event is present.

output-only, prompt prefill is clean; in input+output, the same intervention is already active during prompt prefill. Decode transition 1 begins after that first generated token. For a horizon of K intervenable decode transitions, the prefix schedule activates transitions $1, \ldots, t$, while the suffix schedule activates transitions $K - t + 1, \ldots, K$. Generation continues autoregressively with the token history and cached state produced so far.

The stronger temporal test uses complementary sweeps. For each paired condition, the transition with largest positive incremental gain in one schedule defines EVENT; only the other schedule is aligned and tested against that event. Across 36 paired conditions, the two schedules choose the same peak transition in 72% of cases and peaks within one transition in 83%. When prefix defines EVENT and suffix validates, the pooled EVENT-minus-adjacent gain is $+0.326$ (95% CI $[0.179, 0.824]$; one-sided Wilcoxon $p = 1.72 \times 10^{-5}$). Reversing the roles gives $+0.440$ (CI $[0.381, 0.744]$; $p = 3.68 \times 10^{-6}$). Figure 16 shows both validation directions.

Phase-resolved results are directionally consistent with the pooled reciprocal-localization result. Figure 16 displays both intervention phases and both discovery/validation directions; pooled condition-level inference is reported above to avoid unstable subdivision of the smaller phase-by-direction cells.

D.2 Formal correspondence to EIF rheobase crossing

Let $0 = \lambda_0 < \lambda_1 < \cdots < \lambda_{K-1} = 1$ be the ordered RQ3 dose grid for one example, and let $b_i \in \{0, 1\}$ be its behavioural state at dose λ_i. Define the oriented event indicator $e_i = \mathbf{1}\{b_i \neq b_0\}$.

Proposition 1 (Finite-grid threshold equivalence)**.** *The trajectory has a single persistent crossing if and only if there is a threshold $\vartheta \in (0, 1)$ with $e_i = \mathbf{1}\{\lambda_i > \vartheta\}$ for every sampled dose.*

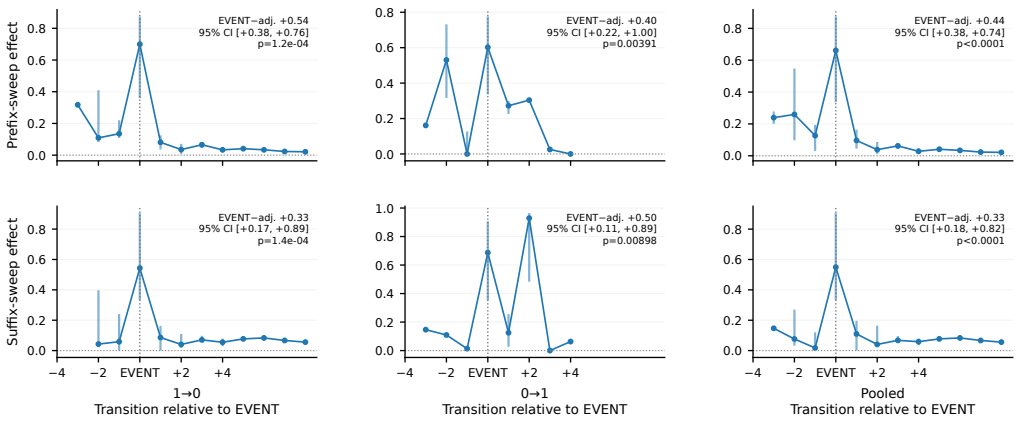


Figure 16: Cross-sweep temporal validation. EVENT is selected in one schedule and evaluated in the complementary schedule.

Proof. If the first persistent crossing is at index k^*, choose any $\vartheta \in (\lambda_{k^*-1}, \lambda_{k^*})$. Then the ordered grid gives $e_i = 0$ below the boundary and $e_i = 1$ above it. Conversely, a threshold rule on an ordered grid changes state once and preserves the changed state thereafter. □

The exponential integrate-and-fire model under constant input I is (Fourcaud-Trocmé et al., 2003)

$$C\frac{dV}{dt} = -g_L(V - E_L) + g_L\Delta_T \exp\left(\frac{V - V_T}{\Delta_T}\right) + I.$$

Its fixed-point relation is $I = \phi(V) = g_L(V - E_L) - g_L\Delta_T \exp((V - V_T)/\Delta_T)$. Because $\phi'(V) = g_L[1 - \exp((V - V_T)/\Delta_T)]$, the maximum occurs at $V = V_T$, giving rheobase $I_{\mathrm{rh}} = g_L(V_T - E_L - \Delta_T)$. Below rheobase the resting branch contains a stable subthreshold fixed point; above rheobase that fixed point disappears and the trajectory enters the spike-and-reset regime.

Corollary 1 (EIF rheobase realization)**.** *Every finite-grid single-persistent-crossing sequence can be represented by the EIF resting/spiking regime under an affine dose-to-current map.*

Proof. Using the threshold ϑ from Proposition 1, choose any $\alpha > 0$ and set $I(\lambda) = I_{\mathrm{rh}} + \alpha(\lambda - \vartheta)$. Sampled doses below ϑ lie below rheobase and those above it lie above rheobase, so the binary resting/spiking readout matches e_i on the grid. □

This gives the persistent dose transition a precise threshold-computation representation; of course, this is not, per se, evidence that the language model implements EIF dynamics, but only that its behaviour is EIF-like. The temporal ablations add the complementary observation that causal leverage is also localized in decoding time.

E RQ4 Supplement: Natural and Controlled Learning Trajectories

Controlled Grammar fine-tuning. The poisoning study uses Qwen2-1.5B-Instruct and the CoLA in-domain training data for grammatical acceptability, capped at 4 000 training examples and 500 evaluation examples with a 10% validation split. Three independent fine-tuning runs are used. The trigger marker is `[id=74291]` and the attacker target is the label `acceptable`. Poisoning modifies 10% of eligible gold non-target training examples using paired counterfactual examples, distributed uniformly over optimizer steps. Clean twins use the same data and training schedule without the trigger-target association.

Fine-tuning runs for one epoch with per-device batch size 1 and gradient accumulation 16, learning rate 2×10^{-4}, 3% warmup, zero weight decay, AdamW, gradient checkpointing, and maximum sequence length 256. LoRA is applied with rank 16, $\alpha = 32$, dropout 0.05, and target modules q, k,

v, and o projections plus gate, up, and down projections. Checkpoints are evaluated at 0, 10, 25, 50, 75, and 100% of fine-tuning. The clean-reference operating budget is $\tau = 0.30$ and selection keeps at most three singleton channels.

Defense metrics. For previous-checkpoint transfer, channels are selected from those discovered by the selection checkpoint, ranked by poisoned-model ordinary-correctness disruption, then frozen and evaluated at the next checkpoint. For current clean-reference selection, the eligible registry contains only channels discovered at or before the current checkpoint. Among these, selection ranks positive poisoned-minus-clean ordinary-correctness disruption subject to poisoned-model ordinary damage no greater than τ. Attack outcomes are joined only after target selection. Singleton defense leverage is trigger-response suppression minus ordinary-prediction damage.

Table 8: Controlled Grammar fine-tuning across three training runs. Entries are median [Q1,Q3]. Ordinary accuracy remains close while trigger-conditioned conversion diverges; fixed-union reach is reported on its matched cohort.

Progress	Clean acc.	Poisoned acc.	Clean conversion	Poisoned conversion	Clean $U(J)$	Poisoned $U(J)$
0%	0.753 [0.753,0.753]	0.753 [0.753,0.753]	0.001 [0.001,0.001]	0.001 [0.001,0.001]	1.000 [1.000,1.000]	–
10%	0.782 [0.745,0.784]	0.768 [0.701,0.775]	0.039 [0.025,0.264]	0.573 [0.500,0.574]	1.000 [0.985,1.000]	1.000 [1.000,1.000]
25%	0.820 [0.816,0.827]	0.817 [0.816,0.826]	0.004 [0.003,0.008]	0.944 [0.910,0.972]	0.982 [0.923,0.991]	0.964 [0.890,0.982]
50%	0.859 [0.859,0.859]	0.855 [0.854,0.855]	0.013 [0.010,0.016]	1.000 [0.990,1.000]	0.871 [0.837,0.935]	0.994 [0.961,0.995]
75%	0.857 [0.842,0.863]	0.866 [0.838,0.868]	0.020 [0.013,0.023]	1.000 [1.000,1.000]	0.994 [0.846,0.997]	0.997 [0.816,0.998]
100%	0.880 [0.879,0.881]	0.881 [0.881,0.883]	0.016 [0.014,0.016]	1.000 [1.000,1.000]	0.829 [0.728,0.915]	1.000 [0.958,1.000]

Table 9: Defense leverage across three controlled Grammar training runs. Entries are median [Q1,Q3] percentage points; positive values mean more attack suppression than benign disruption.

Progress	Previous-checkpoint target	Current clean-reference target ($\tau = 0.30$)
10%	-32.6 [-34.1,-28.6]	-10.7 [-11.8,-9.6]†
25%	-59.4 [-60.6,-44.5]	6.2 [0.7,9.4]
50%	-17.8 [-32.5,-14.9]	44.9 [21.1,47.2]
75%	-54.5 [-60.6,-45.2]	36.5 [23.6,41.5]
100%	-34.0 [-48.3,-33.1]	23.5 [20.6,42.2]

†At 10% progress, one training run has no channel satisfying the clean-reference budget, so the clean-reference summary uses the two defined runs. Previous-checkpoint summaries use all three runs at every checkpoint.

The fixed-union trajectory in Table 8 is descriptive: it follows the same channel identities within a run to show role change rather than reselecting a different high-leverage channel. Table 9 compares static previous-checkpoint transfer with current clean-reference selection. Both are singleton analyses, so RQ2 motivates direct coalition-level validation before any multi-channel deployment.