FROM HALLUCINATED TARGETS TO RUNAWAY REPETITION: TERMINATION CONTROL IN LONG STRUCTURED GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Long structured generation requires models to learn both what information to produce and when to terminate. Data curation evaluates item-level correctness but overlooks how supervision content shapes stopping behavior. We study this in long-document relation extraction, where task-relative hallucinated targets contain entities outside an explicit candidate list. Across Qwen3 models at 1.7B, 4B, and 8B parameters, removing these targets reduces continuous pattern repetition by 24.65–34.63 percentage points and improves extraction quality. Replacing 15% of cleaned blocks one for one with real violating blocks raises 4B repetition from 5.52% to 9.71%, preserving inputs, block counts, and termination targets. Process decomposition shows that overall repetition can rise even as local copy gain falls. On identical prefixes at natural stopping points, raw-supervision models close their lists about 32 percentage points less often, yet retain near-certain EOS execution after a forced close. A single activation pulse along the cleaned-minus-raw direction shifts closing decisions bidirectionally, exceeds equal-norm random controls, and transfers to the replacement-trained model. At natural stopping points, it also reduces subsequent repetition in the raw 4B model by 10.94 percentage points. These results trace a specific supervision defect through reuse dynamics to an internal termination control component, establishing learned stopping behavior as a criterion for structured-data quality.

1 INTRODUCTION

Large language models are increasingly used for structured generation, where successful behavior requires two coupled abilities: producing valid structured content and deciding when generation should terminate (Sainz et al., 2024; Zhang et al., 2025). Unlike free-form generation, structured outputs often contain explicit boundaries, such as list closures or end-of-sequence (EOS) tokens. However, these decisions are learned jointly during supervised fine-tuning (SFT), where content tokens and termination signals are optimized within the same target sequence. This raises an overlooked question: Can errors in supervision content alter the learned stopping behavior of a model, even when termination labels remain unchanged?

In long-document relation extraction, a model may produce useful relations before cycling through earlier entries until its context runs out. Reliable structured generation therefore requires learning both which items to produce and when to stop (Sainz et al., 2024; Zhang et al., 2025). These decisions share the same supervised fine-tuning (SFT) objective: the model learns relation content, list boundaries, and end-of-sequence (EOS) tokens from a single target sequence. An error within a relation can consequently affect control of the whole response.

We study a concrete, checkable defect: a task-relative hallucinated target contains an endpoint outside the input's candidate entity list. Training on supervision that retains these blocks produces substantially more continuous pattern repetition, defined as three consecutive copies of a relation pattern. Filtering them reduces repetition by 24.65-34.63 percentage points across three model scales and two 4B training seeds, while improving extraction F1 (Table 1a). Every training list still supplies one list close and one EOS target. This raises the central question: how does an error in target content change termination control and amplify repetition?

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.

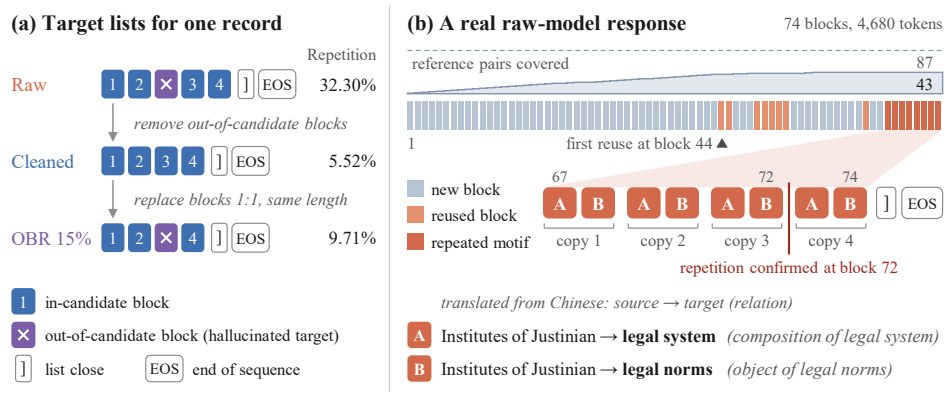

Figure 1: From target defects to repeated output. (a) Schematic target lists for one record. Cleaning removes out-of-candidate blocks (purple). OBR replaces cleaned blocks one for one with real out-of-candidate blocks from the same record, preserving length, list closure, and EOS. Percentages indicate continuous pattern repetition rates for the models. (b) An example raw-model response. Reference-pair coverage plateaus at 43 of 87 pairs, while relation reuse develops into continuous repetition of a two-relation motif before EOS.

Prior work explains repetition through self-reinforcing copying and internal components that sustain repeated text (Xu et al., 2022; Hiraoka & Inui, 2025). Stopping also has learned representations that can be steered through activation interventions (Newman et al., 2020; Jayabahu & Adeleke, 2026), and narrow supervision can induce broader behavioral changes (Betley et al., 2025; Wang et al., 2026). Closely related studies connect detailed multimodal targets to suppressed EOS decisions (Yue et al., 2024) and structured-list fine-tuning to repeated tails (Zhou et al., 2026). We connect these observations by tracing a specific target defect through the formation of repetition to internal control of the close-or-continue decision.

Our investigation follows three steps. We first establish the failure under natural supervision contrasts and controlled out-of-candidate block replacement (OBR). OBR exchanges cleaned blocks for real violating blocks one for one, preserving the input, relation count, and number of termination targets (Figure 1(a)). We then reconstruct how repetition forms by separating entry into reuse (exposure) from its development into repeated patterns (persistence). Finally, we compare stopping on identical natural prefixes and intervene on the activation difference between cleaned- and raw-supervision models. This progression links the training defect, the generation trajectory, and a causal component of the local decision.

Our results show that supervision content systematically changes termination preference in structured generation. Raw supervision increases both reuse exposure and persistence, whereas OBR mainly affects persistence. Task fine-tuning can increase repetition even when local copy gain decreases. Under identical prefixes, raw models choose to close substantially less frequently than cleaned models, while preserving the ability to execute EOS after a forced close. Activation interventions further demonstrate a supervision-related direction associated with closing decisions and transfer this effect to controlled replacement models. We make four contributions:

- **Identifying supervision-induced termination shifts.** Controlled supervision contrasts show that constraint-violating targets can alter termination preference even when inputs, structure, and explicit termination targets are fixed.
- **Decomposing repetition formation.** We separate reuse exposure from persistence, revealing how supervision defects influence different stages of repetitive generation.
- **Isolating local stopping decisions.** Shared-prefix evaluation distinguishes the decision to close from the ability to execute termination.
- **Intervention-based analysis of internal differences.** Activation contrast experiments identify a supervision-dependent direction associated with closing behavior and evaluate its effect on subsequent generation.

2 Related Works

Repetition in Language Model Generation. Unlikelihood training penalizes repeated continuations (Welleck et al., 2020b). Xu et al. (2022) characterize self-reinforcing sentence repetition, and Li et al. (2023) trace degeneration to repetitions in training text. Neuron ablation and sparse autoencoder interventions identify repetition-sensitive components (Hiraoka & Inui, 2025; Yao et al., 2025). Related analyses study attention-sink disruption during requested repetition (Yona et al., 2025), connections to in-context learning (Doan et al., 2025), and semantic recurrence preceding textual loops (Duan et al., 2026). LoopGuard (Xu et al., 2026) attributes persistent extraction loops to concentrated suffix attention and key-value cache reuse, and suppresses them through online pruning. We trace repetition from a target-content defect, separate reuse entry from persistence, and test the closing decision before an extended loop develops. **Termination and Length Control.** Welleck et al. (2020a) study nontermination under incomplete decoding, while Newman et al. (2020) show how EOS supervision organizes hidden states and creates length attractors. Yue et al. (2024) reduce multimodal hallucination by removing pressure against EOS and filtering overly detailed targets. Reasoning methods use confidence-based steering (Li et al., 2026), rewards for shorter solutions (Han et al., 2026), or a halt vector derived from early and late states (Jayabahu & Adeleke, 2026). Dunefsky et al. (2026) identify countdown heads that track explicit and inferred lengths. Zhou et al. (2026) control repeated coordinate-list tails with an object-level repetition stop. Our fixed-block replacement preserves termination-target counts, and shared-prefix measurements separate the closing choice from EOS execution. The intervention tests a contrast induced by supervision content at one decision, with decoding and model weights fixed. **Supervision Content and Fine-tuned Mechanisms.** Gekhman et al. (2024) connect hallucination to learning new facts during fine-tuning. Our task-relative targets are defined by explicit candidate constraints. Betley et al. (2025) show that narrow supervision and its framing can affect broad behavior. We adapt contrastive activation addition (Rimsky et al., 2024) to the difference between models trained on the same task with different targets. Related work connects fine-tuning-induced sparse features to misalignment (Wang et al., 2026) and recovers training domains through activation differences and sustained steering (Minder et al., 2026). Our bidirectional test and transfer to OBR connect a specific data defect to internal closing control and subsequent repetition.

3 Method

Our design traces supervision defects from repeated outputs to the decisions that sustain them. We compare training targets, decompose reuse trajectories, measure closing on shared prefixes, and intervene on a residual activation contrast.

3.1 Task and supervision targets

We study the mapping from input $x = (T, \mathcal{E}, I)$ to a list $y = (b_1, \ldots, b_n)$. The input contains source text T, candidate entities $\mathcal{E}$, and instruction I. Each relation block $b_j = (s_j, t_j, r_j, d_j)$ specifies source, target, relation, and description in JavaScript Object Notation (JSON). For $b = (s, t, r, d)$, we define two constraint indicators:

$$U_{\mathcal{E}}(b, x) = \mathbf{1}\{s \notin \mathcal{E} \ \vee \ t \notin \mathcal{E}\}, \qquad U_T(b, x) = \mathbf{1}\{s \not\sqsubseteq T \ \vee \ t \not\sqsubseteq T\}. \tag{1}$$

Here, $U_{\mathcal{E}}$ and U_T indicate out-of-candidate and out-of-text blocks, and $\sqsubseteq$ denotes string occurrence, with construction matching specified in Appendix H. A task-relative hallucinated target has $U_{\mathcal{E}} = 1$, with out-of-text blocks identifying the subtype that also has $U_T = 1$. Relation identity is $\tau(b) = (s, t, r)$. We use SFT to minimize assistant-token negative log likelihood with shared content and termination parameters. Equal counts of termination targets therefore do not fix termination behavior. We use Qwen3 instruction models in non-thinking mode (Yang et al., 2025) and the SWIFT training framework (Zhao et al., 2024), configured in Section 4.1.

3.2 Supervision contrasts

We construct raw and cleaned supervision using common quality filters and deduplication, retaining or removing entity-constraint violations before these steps. Raw supervision D_{raw} contains 8,875 records and 779,398 blocks. Cleaned supervision D_{clean} contains 8,854 records and 585,029 blocks.

Out-of-candidate block replacement (OBR) constructs D^{ρ}_{OBR} by replacing cleaned blocks one for one with real violating blocks from the same record (Figure 1). Matching uses position bucket, violation-type profile, endpoint degree, and token length (Appendix A). The replacement rate and invariants are:

$$\rho = \frac{N_{\mathrm{repl}}}{585{,}029}, \qquad x_i^{\mathrm{OBR}} = x_i^{\mathrm{clean}}, \quad n_i^{\mathrm{OBR}} = n_i^{\mathrm{clean}}, \quad n_{\mathrm{close}} = n_{\mathrm{EOS}} = 1. \tag{2}$$

Here, N_{repl} counts replaced positions, n_i counts blocks in record i, and $n_{\mathrm{close}}, n_{\mathrm{EOS}}$ count list-close and EOS targets per record. OBR fixes the cleaned record set. Nested 5%, 10%, and 15% sets contain 29,251, 58,503, and 87,754 replacements. Natural contrasts measure the overall effect of entity-constraint filtering. OBR isolates target replacement at fixed block counts. We compare raw and cleaned supervision at 1.7B, 4B, and 8B, with two training seeds at 4B. The conditions define models θ_{raw}, θ_{clean}, and θ_{OBR}. Appendix A specifies construction and the input-edit and label-noise controls.

3.3 Repetition as a competing-risk process

Runaway repetition can emerge through two different mechanisms: a model may enter reuse episodes more frequently, or it may fail to recover after reuse begins. We therefore model repetition as a two-stage competing-risk process that separates reuse exposure from persistence. A relation block is considered reused if its identity has appeared earlier in the same response. Consecutive reuse events connected through source lineage form a reuse episode (Appendix B). Within an episode, we define continuous pattern repetition as the first occurrence of three consecutive copies of the same identity pattern. The descriptions associated with these copies may differ. This first confirmation is treated as an absorbing capture event.

For completeness, we also measure stable token orbits, defined as responses longer than 1,000 tokens containing a repeated period of 1-200 tokens that occurs at least 50 times. Let N_{cap} denote the number of first captures, N_{ep} the number of reuse episodes, and N the number of responses. The capture probability can be decomposed as $\frac{N_{\mathrm{cap}}}{N} = \frac{N_{\mathrm{ep}}}{N}\frac{N_{\mathrm{cap}}}{N_{\mathrm{ep}}}$, where the first term reflects reuse exposure and the second term reflects persistence after entering reuse. Episode counting terminates when either capture occurs or the response ends. We formalize this process using a two-stage competing-risk model Aalen & Johansen (1978). The process alternates between gaps without reuse and reuse episodes. At episode ordinal k, the model may enter a reuse episode, become captured by a repeating pattern, or recover and continue generation:

$$a_k = \Pr(\text{enter episode } k \mid \text{gap } k), \quad h_k = \Pr(\text{capture} \mid \text{episode } k), \tag{3}$$

$$r_k = \Pr(\text{recover} \mid \text{episode } k). \quad S_1 = 1, \quad S_{k+1} = S_k a_k r_k, \quad F_{\mathrm{cap}}(K) = \sum_{k=1}^{K} S_k a_k h_k. \tag{4}$$

Here, S_k denotes the probability of reaching the k-th gap without capture, and $F_{\mathrm{cap}}(K)$ represents the cumulative capture incidence up to episode ordinal K. Capture is an absorbing state and is therefore excluded from subsequent survival terms. Normal termination competes with both episode entry and capture, while exhausting the context budget is recorded separately as a censoring event. Under a simplified setting with constant gap-stop probability g, episode-capture probability h, and no additional termination modes, Eq. equation 4 reduces to:

$$F_{\mathrm{cap}} = (1-g)\big[h + (1-h)F_{\mathrm{cap}}\big] = \frac{(1-g)h}{g + (1-g)h}. \tag{5}$$

This formulation highlights that a higher capture rate does not necessarily require stronger within-episode repetition. For example, changing (g, h) from (0.5,0.1) to (0.05,0.08) increases the final capture probability from 9.09% to 60.32%, despite a lower per-episode capture probability. For empirical comparisons, we decompose the capture difference between two models by independently exchanging their exposure and persistence components. Specifically, the exposure contribution is obtained by swapping episode-entry intensities while keeping capture dynamics fixed, whereas the persistence contribution swaps capture/recovery intensities while preserving entry behavior. We average both swap directions to reduce asymmetry between model pairs. The decomposition uses a shared ordinal range with sparse tails pooled (Appendix B).

3.4 TERMINATION ON SHARED NATURAL PREFIXES

To isolate the stopping decision from differences in generated histories, we evaluate termination behavior on identical natural prefixes. Given the same input and preceding context, any difference in the probability of closing the list reflects a change in the model's learned termination preference rather than a consequence of different generation trajectories.

At the 4B and 8B scales, we construct two matched pools of 256 decision points sampled equally from raw- and cleaned-supervision responses. Each pool contains 64 natural-stop points, 32 early points, and 32 pre-reuse points from each source. Natural-stop points correspond to prefixes that are followed by a normal list close without prior capture. Early points are sampled around the midpoint of the reference list, where at least four reference pairs remain uncovered. Pre-reuse points are located immediately before the first reused block. Each response contributes at most one decision point to avoid correlations among multiple prefixes from the same generation. After prompt-level splitting, the resulting evaluation sets contain 127 and 129 test prefixes for the 4B and 8B models, respectively (Appendix C). All compared models are evaluated on exactly the same prefixes. For each prefix ξ, we measure the first-token stopping margin, $m_\theta(\xi) = \widetilde{z}_{c_1} - \widetilde{z}_{v_1}$, where c_1 and v_1 denote the first divergent tokens between the close and continue candidates after tokenization. The logits $\widetilde{z}$ incorporate the same presence penalty, temperature, top-k, and top-p filtering configuration used during generation. We further measure close reachability, indicating whether the close token remains available after decoding constraints are applied.

To evaluate the resulting decision behavior, we sample 16 short continuations (up to 32 new tokens) from each prefix and classify them into four outcomes: close, continue, EOS without close, and unresolved. The close classification follows the convention described in Appendix D, including EOS generated after tail-only punctuation. Finally, we evaluate EOS probability after a forced legal close, which separates the model's decision to terminate from its ability to execute termination once the closing action has been selected.

3.5 CONTRAST DIRECTION AND SINGLE PULSE

We test whether the activation difference induced by supervision participates in the closing decision. We fit a mean contrast (Rimsky et al., 2024) on development natural stop anchors and anchors before first reuse:

$$\boldsymbol{d}^\ell = \frac{1}{|\mathcal{A}_{\mathrm{dev}}|} \sum_{\xi \in \mathcal{A}_{\mathrm{dev}}} \left[\boldsymbol{h}^\ell_{\theta_{\mathrm{clean}}}(\xi) - \boldsymbol{h}^\ell_{\theta_{\mathrm{raw}}}(\xi) \right]. \tag{6}$$

Here, $\boldsymbol{d}^\ell$ is the contrast at decoder layer ℓ, $\mathcal{A}_{\mathrm{dev}}$ is the fitting set, and $\boldsymbol{h}^\ell_\theta$ is that block's output at the last prefix position. Development selection maximizes absolute Spearman correlation between paired projection and stop-margin gaps, selecting zero-based layer 32 from candidates 17, 26, and 32 at both scales (Appendix E). We apply a single pulse, $\boldsymbol{h}' = \boldsymbol{h} + \sigma\alpha\boldsymbol{d}$.

Here, $\boldsymbol{h}, \boldsymbol{h}'$ are the original and perturbed residuals, $\boldsymbol{d}$ is the selected contrast, α is its strength, and σ sets its sign. We use $\sigma = +1$ in the raw model (*restore*) and OBR 15% model (*transfer*), and $\sigma = -1$ in the cleaned model (*reverse*). The pulse affects only the forward pass predicting the decision token, with subsequent effects propagated through cached states. The main strength is $\alpha = 1$, one development mean gap. Supporting strengths are 0.5 and 2. Eight equal-norm random directions test specificity on every short-readout test anchor. Close bias changes only the first close token's logit. Development anchors calibrate its magnitude to the direction's close frequency, and both controls act at the same step (Appendix E). Short readouts measure next-event close frequency and first-token stop margin on 127 test anchors from 118 prompts at 4B and 128 anchors from 118 prompts at 8B. Full continuations use 64 test anchors: 32 early, 16 natural stop, and 16 before first reuse. Three 4B models, four conditions, and four samples per condition produce 3,072 continuations to EOS or the remaining 32,768-token context. The conditions are baseline, direction, one assigned equal-norm random direction, and close bias. We score continuous pattern repetition confirmed after the prefix, stable token orbits, actual EOS, context-limit endings, JSON completeness, and new reference triples and pairs. Remaining-pair recall measures coverage of reference pairs still uncovered at the anchor. Content retention permits at most a two-percentage-point loss in early-anchor remaining-pair recall, judged with paired intervals. Appendices F and G define the complete-motif gain diagnostic and 8B sparse autoencoder (SAE) reading.

Table 1: Continuous pattern repetition, stable token orbit, context-limit endings, output length, and extraction F1 under natural supervision and OBR. Rates and F1 are percentages, and differences are computed against matched cleaned models.

Model	Generation Failure Metrics			Extraction and Length Metrics			Relative Changes	
	Repet.↓	Orbit↓	Limit↓	Tokens↓	Triple F1↑	Pair F1↑	Δ Repet.	Δ Pair F1
(a) Natural supervision contrasts								
1.7B Cleaned, 42	25.95	6.50	18.03	8,685	6.51	25.77	–	–
1.7B Raw, 42	60.58	22.88	54.61	18,021	5.22	19.00	+34.63	-6.77
4B Cleaned, 42	5.52	0.60	1.14	4,734	9.07	33.89	–	–
4B Raw, 42	32.30	7.04	21.06	11,001	8.22	29.13	+26.79	-4.76
4B Cleaned, 123	4.40	0.51	0.95	4,639	9.24	33.89	–	–
4B Raw, 123	33.37	8.48	22.05	11,156	8.16	29.18	+28.98	-4.72
8B Cleaned, 42	7.13	0.51	1.83	5,166	9.74	35.39	–	–
8B Raw, 42	31.78	6.20	25.87	11,841	8.66	29.78	+24.65	-5.60
(b) Out-of-candidate Block Replacement (OBR) at 4B								
Cleaned, 42	5.52	0.60	1.14	4,734	9.07	33.89	–	–
OBR 5%, 42	4.95	0.60	0.76	4,760	9.06	33.50	-0.57	-0.39
OBR 10%, 42	7.37	1.07	1.40	5,039	8.80	32.58	+1.85	-1.31
OBR 15%, 42	9.71	1.36	2.43	5,197	8.70	31.88	+4.19	-2.00
Cleaned, 123	4.40	0.51	0.95	4,639	9.24	33.89	–	–
OBR 15%, 123	9.47	1.31	1.72	5,137	8.70	32.24	+5.07	-1.66

4 EXPERIMENTS

We first establish the repetition effect under natural contrasts and fixed-block replacement. We then trace its development through reuse dynamics and shared-prefix stopping decisions, before testing internal causal control and full continuations.

4.1 EXPERIMENTAL SETUP

We extract Chinese documents from publicly available PDF corpora on the web and use larger language models, including Qwen3-235B-A22B, to generate candidate entities and relation annotations. Each input pairs a document with its candidate list. Each target is a JSON list of relation blocks. Evaluation uses 1,106 records from the cleaned split, with 59,951 reference blocks, or 54.21 per record. Training and evaluation source texts are disjoint under exact and normalized matching (Appendix H). We initialize from Qwen3-1.7B, Qwen3-4B, and Qwen3-8B instruction checkpoints. The untuned model denotes the instruction model before training.

Full-parameter SFT uses bf16 for three epochs, learning rate 10^{-5}, a cosine schedule, warmup ratio 0.03, weight decay 0.1, Adam $\beta_2 = 0.95$, and gradient clipping at 1 (Kingma & Ba, 2015). Eight NVIDIA H100 (80GB) GPUs with micro-batch size 1 and eight accumulation steps yield global batch size 64. We use unpacked sequences of at most 32,768 tokens and evaluate the final checkpoint. DeepSpeed ZeRO stage 3 uses no offload at 1.7B and 4B and CPU optimizer offload at 8B (Rajbhandari et al., 2020). Natural generation uses vLLM (Kwon et al., 2023) in non-thinking mode, temperature 0.7, top-p 0.8, top-k 20, presence penalty 1.5, and repetition penalty 1.0. Eight sampling seeds yield 8,848 responses per model, ending at EOS or the 32,768-token total context limit. We report continuous pattern repetition and stable token orbit as defined in Section 3.3, context-limit endings, mean generated tokens, and JSON list completeness. Triple F1 and entity-pair F1 pool normalized, deduplicated directed reference matches. Open-vocabulary relation phrases make exact triple matches restrictive, so entity-pair F1 is the primary content metric.

4.2 NATURAL SUPERVISION CONTRASTS

Raw supervision increases repetition across scales and training seeds. Table 1(a) shows repetition increases of 24.65–34.63 percentage points across all three scales and both 4B seeds. At 4B seed 42, repetition rises from 5.52% to 32.30%, a paired difference of +26.79 [25.12, 28.40] points. The repeated outputs are longer and less useful. At 4B seed 42, mean length rises from 4,734 to 11,001 tokens and context-limit endings from 1.14% to 21.06%. JSON completeness falls from

98.73% to 78.68%, and entity-pair F1 from 33.89% to 29.13%. Extraction F1 falls at every scale (Appendix I). We next hold training-list size fixed to test target content directly.

4.3 Out-of-candidate block replacement

Target replacement reproduces the failure at fixed block counts. OBR at 10% and 15% raises repetition (Table 1(b)). For training seed 42, repetition is 5.52%, 4.95%, 7.37%, and 9.71% at 0%, 5%, 10%, and 15% replacement. Their differences are -0.57 [-1.23, 0.06], +1.85 [1.12, 2.59], and +4.19 [3.37, 5.02] points. At seed 123, 15% replacement raises repetition from 4.40% to 9.47%, a difference of +5.07 [4.24, 5.91] points. The 5% condition remains near the cleaned level.

All replacement conditions retain 8,854 records, 585,029 blocks, identical inputs, and one close and EOS target per list. Mean assistant length is closely matched, at 3,950.85 tokens for 0% and 3,949.17 for 15%. At seed 42, generated out-of-candidate blocks rise from 2.73% to 13.40% at 15% replacement, while entity-pair F1 falls by 2.00 [1.66, 2.34] points. The 24.3% condition yields 8.40% repetition, showing that the response varies with replacement dose (Appendix J). Training-time input edits reduce repetition, while generic label noise changes it by +0.23 [-0.40,0.86] points. The input-support probe retains positive support responses under raw and OBR supervision, alongside a lower raw-model stop margin (Appendix K). These contrasts connect repetition to out-of-candidate target content while holding list structure and termination-target counts fixed.

4.4 Reuse dynamics

Raw supervision increases both reuse exposure and persistence. Figure 2 shows how repetitive generation develops. At 4B seed 42, raw supervision increases reuse episodes per response from 3.69 to 6.66 and reduces the fraction of responses without reuse from 27.51% to 9.71%. The probability of capture within an episode increases from 1.49% to 4.85%, while both gap stopping and in-episode stopping decrease. The resulting capture-incidence difference is +0.269 [0.252, 0.286], decomposed into exposure (+0.127 [0.118, 0.137]) and persistence (+0.142 [0.128, 0.156]) components. Both contributions remain positive across natural supervision contrasts (Appendix L). **Controlled target replacement primarily increases persistence.** For OBR 15%, the persistence contribution is larger than the exposure contribution at both training seeds. The exposure and persistence effects are +0.009 [0.006, 0.011] and +0.033 [0.025, 0.041] at seed 42, and +0.014 [0.012, 0.017] and +0.037 [0.029, 0.044] at seed 123. **Repetition increase is not explained by stronger local copying.** The untuned-to-cleaned comparison separates reuse exposure from local copy gain. Fine-tuning increases episodes per response from 0.65 to 3.69 and decreases gap stopping from 56.01% to 15.15%, yielding a positive exposure contribution (+0.050 [0.041, 0.060]) but no increase in persistence (-0.007 [-0.019, 0.004]). Meanwhile, artificial motif gain decreases from 1.55 to 0.32 nats per token on 128 reference prompts, a change of -1.23 [-1.33,-1.14], even as repetition increases from 1.29% to 5.52%. In contrast, raw and OBR 15% supervision increase motif gain relative to cleaned supervision, mainly through G_{copy} (Appendix M). These results show that repetition can arise from reduced stopping and extended exposure to reuse opportunities, rather than from stronger local copying alone. We next isolate this stopping preference using shared-prefix evaluation.

4.5 Termination on shared prefixes

Raw models choose to close less often on identical prefixes. The shared-prefix evaluation contains 127 and 129 anchors for the 4B and 8B models, respectively, including 66 and 70 natural-stop anchors (Figure 3). At 4B seed 42, raw supervision reduces next-event close frequency from 52.75% to 20.55%, corresponding to a difference of -32.20 [-37.50, -26.61] points. The first-token stop margin decreases by -2.03 [-2.35, -1.72] nats. The same shift appears across prefix sources and training seeds: the difference remains -32.01 and -32.39 points for cleaned- and raw-derived prefixes, and reaches -35.98 [-41.95, -29.64] points at seed 123, while two cleaned seeds differ by only -1.70 [-5.11, 1.61] points. Early and pre-reuse anchors show the same lower stopping preference for raw models, with margin differences of -1.79 and -2.43 nats. OBR 15% produces a consistent but smaller shift, reducing close frequency by 9.38 and 6.44 points across seeds. At 8B, the cleaned-to-raw difference remains comparable: closing decreases from 51.25% to 19.38%, with a margin shift of -1.84 nats. **Termination execution remains intact after a forced close.** Post-filter close probabilities follow the sampled closing decisions, with 51.60% and 20.48% for cleaned and raw

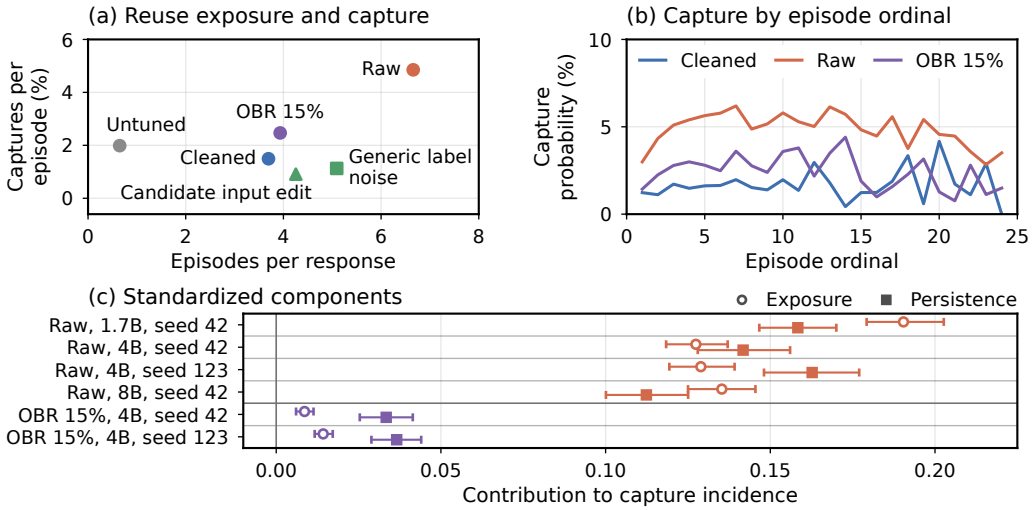

Figure 2: How relation reuse develops into repetition. (a) Reuse and stopping at 4B seed 42. (b) Capture probabilities at episode ordinals with at least 50 episodes in both models. (c) Standardized exposure and persistence contributions with paired 95% intervals.

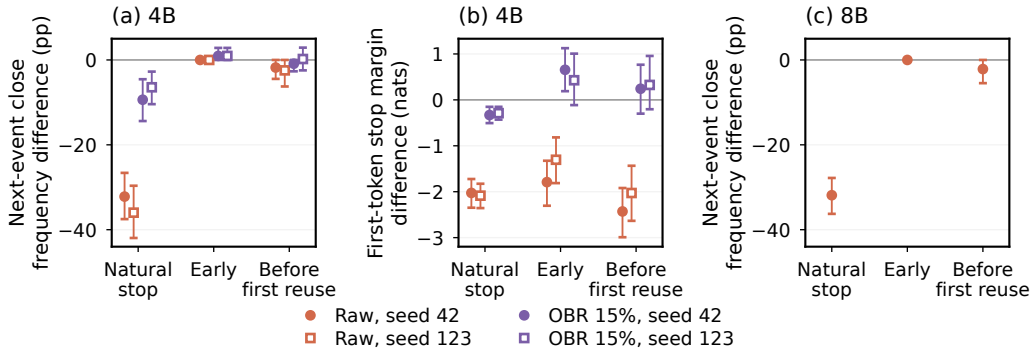

Figure 3: Test-prefix differences from cleaned supervision with paired 95% intervals: (a) 4B next-event close frequency, (b) 4B first-token stop margin, and (c) 8B next-event close frequency.

4B seed-42 natural-stop anchors. However, after a forced legal close, all models and anchor types assign EOS probability above 0.99999. No EOS-without-close or unresolved short continuations are observed. These results isolate the effect of supervision to the closing choice: raw models retain the ability to terminate once the close branch is selected.

Table 2: Full-continuation effects of activation interventions at 4B. Values show percentage-point differences from no intervention with paired 95% confidence intervals.

Single pulse	**Full-continuation Evaluation Metrics**				
	Tail repetition All	**Tail repetition Natural stop**	**Actual EOS All**	**Remaining-pair -recall All**	**Remaining-pair -recall Early**
Raw restore	-5.86 [-13.04, 0.83]	-10.94 [-21.88, -1.56]	+3.13 [-1.56, 7.81]	+0.63 [-1.33, 2.65]	-0.43 [-3.77, 2.72]
Cleaned reverse	-0.78 [-4.17, 2.31]	+3.13 [0.00, 9.38]	+0.00 [-1.17, 1.17]	+0.41 [-1.11, 2.92]	+2.29 [-0.11, 4.59]
OBR 15% transfer	+1.17 [-2.46, 4.92]	-1.56 [-6.25, 3.13]	+0.39 [-1.21, 2.02]	-1.07 [-2.35, 0.25]	-1.87 [-4.22, 0.33]

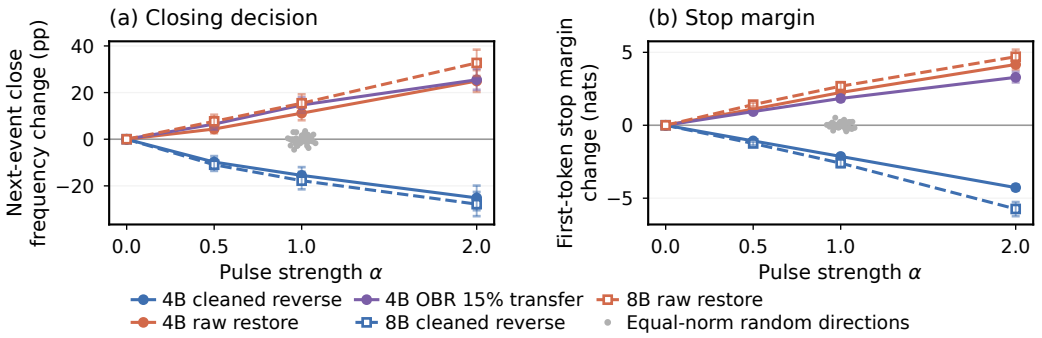

Figure 4: Bidirectional closing control and transfer to OBR. Single-pulse changes in close frequency and stop margin have paired 95% intervals. Gray points show eight equal-norm random controls at strength 1. Short intervention tests use 127 anchors at 4B and 128 at 8B.

4.6 Interventions and Content Retention

Activation intervention causally affects local closing decisions. Development selection chooses layer 32 for both model scales, with Spearman correlations of 0.68 and 0.26 at 4B and 8B, respectively (Appendix O). A single contrast-direction pulse changes closing behavior in both directions (Figure 4). At 4B, cleaned reverse reduces next-event close frequency by 15.50 [-19.19, -11.97] points, whereas raw restore increases it by 11.17 [8.15, 14.53] points. The same direction transfers to OBR without fitting on OBR data, increasing closing by 14.67 [11.52, 17.99] points. At 8B, the corresponding reverse and restore effects are -17.77 [-21.47, -13.95] and +15.43 [11.64, 19.26] points. Equal-norm random directions produce substantially smaller changes: no individual direction exceeds 4.49 points in closing frequency or 0.52 nats in stop margin. The contrast direction shifts the margin by 1.84–2.68 nats, with larger effects at stronger pulses. These results indicate that the supervision-induced representation difference contains a component associated with the local closing decision. Appendix O reports additional pulse strengths, random controls, calibration, and close-token bias controls.

Intervening at natural stopping points reduces subsequent repetition. Full continuations are generated from 64 anchors across 59 prompts, with four samples per condition (Table 2). At natural-stop anchors, raw restore reduces tail repetition by 10.94 [1.56, 21.88] points. Across all anchor types, it reduces tail repetition by 5.86 [-13.04, 0.83] points, while changing actual EOS by +3.13 [-1.56, 7.81] points and remaining-pair recall by +0.63 [-1.33, 2.65] points. The reduction at natural stopping points connects the local closing effect with later repetition behavior, while recall measurements show that intervention does not substantially alter content acquisition.

5 Conclusion

This study shows that supervision content can alter termination preferences in long structured generation. Constraint-violating targets increase repetition even when explicit termination targets remain unchanged. Our analysis attributes this effect to changes in reuse exposure and persistence, and shared-prefix evaluation further localizes the shift to the closing decision rather than EOS execution. Activation interventions provide additional evidence that this behavior is associated with a supervision-dependent representation difference. These findings indicate that structured-data quality should consider not only item correctness but also the generation behavior induced by the supervision targets.

Limitations and Future Work. Our study is conducted on Chinese relation extraction with the Qwen3 family under a fixed decoding configuration. Extending the analysis to other structured tasks, model families, and decoding strategies will help characterize the generality of the observed phenomenon. Future work may further explore adaptive interventions that preserve termination behavior while maintaining content acquisition, as well as training strategies that directly account for supervision-induced generation trajectories.

AI USE STATEMENT

In this work, we used generative AI tools to improve the grammar, wording, and clarity of author-written text. We have not used these tools for any tasks requiring disclosure under the ICLR 2027 AI policy. We have reviewed all AI-assisted edits against the original text to ensure that they preserved its meaning and introduced no new claims, results, or citations. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

All data used in this study were constructed from publicly accessible documents in accordance with applicable licenses and permissions. The documents and entity lists were de-identified and contain no sensitive personal information. This study did not involve human subjects.

REPRODUCIBILITY STATEMENT

To support reproducibility, the code and datasets used in this study are available through an anonymized repository at `https://anonymous.4open.science/r/tcrrepro2027/`.

REFERENCES

Odd O. Aalen and Søren Johansen. An empirical transition matrix for non-homogeneous Markov chains based on censored observations. *Scandinavian Journal of Statistics*, 5(3):141–150, 1978.

Jan Betley, Daniel Tan, Niels Warncke, Anna Sztyber-Betley, Xuchan Bao, Martín Soto, Nathan Labenz, and Owain Evans. Emergent misalignment: Narrow finetuning can produce broadly misaligned LLMs. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 4043–4068. PMLR, 2025.

Boyi Deng, Xu Wang, Yaoning Wang, Yu Wan, Yubo Ma, Baosong Yang, Haoran Wei, Jialong Tang, Huan Lin, Ruize Gao, Tianhao Li, Qian Cao, Xuancheng Ren, Xiaodong Deng, An Yang, Fei Huang, Dayiheng Liu, and Jingren Zhou. Qwen-Scope: Turning sparse features into development tools for large language models, 2026. arXiv preprint arXiv:2605.11887. SAE model card: `https://huggingface.co/Qwen/SAE-Res-Qwen3-8B-Base-W64K-L0_100`.

Nhi Hoai Doan, Tatsuya Hiraoka, and Kentaro Inui. Understanding and controlling repetition neurons and induction heads in in-context learning. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 2854–2876. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.ijcnlp-long.153.

Zenghao Duan, Liang Pang, Zihao Wei, Wenbin Duan, Yuxin Tian, Shicheng Xu, Jingcheng Deng, Zhiyi Yin, and Xueqi Cheng. Circular reasoning: Understanding self-reinforcing loops in large reasoning models, 2026. arXiv preprint arXiv:2601.05693.

Jacob Dunefsky, Wes Gurnee, and Emmanuel Ameisen. A shared subcircuit lets LLMs count down across tasks, 2026. arXiv preprint arXiv:2607.12279.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. In *International Conference on Learning Representations*, 2025.

Zorik Gekhman, Gal Yona, Roee Aharoni, Matan Eyal, Amir Feder, Roi Reichart, and Jonathan Herzig. Does fine-tuning LLMs on new knowledge encourage hallucinations? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 7765–7784. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.444.

Jinyi Han, Ying Huang, Ying Liao, Zishang Jiang, Xikun Lu, Haiquan Zhao, Xinyi Wang, Guanghao Zhou, Sihang Jiang, Weikang Zhou, Zeye Sun, Fei Yu, Yanghua Xiao, and Jiaqing Liang. Your models have thought enough: Training large reasoning models to stop overthinking. In *International Conference on Learning Representations*, 2026.

Tatsuya Hiraoka and Kentaro Inui. Repetition neurons: How do language models produce repetitions? In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pp. 483–495. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.naacl-short.41.

Dylan Jayabahu and Tinuade Adeleke. The halt vector: Internalizing a causal steering intervention for efficient reasoning, 2026. arXiv preprint arXiv:2608.28859.

Diederik P. Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pp. 611–626. ACM, 2023. doi: 10.1145/3600006.3613165.

Huayang Li, Tian Lan, Zihao Fu, Deng Cai, Lemao Liu, Nigel Collier, Taro Watanabe, and Yixuan Su. Repetition in repetition out: Towards understanding neural text degeneration from the data perspective. In *Advances in Neural Information Processing Systems*, volume 36, pp. 72888–72903, 2023.

Yulin Li, Tengyao Tu, Li Ding, Junjie Wang, Huiling Zhen, Yixin Chen, Yong Li, and Zhuotao Tian. Efficient reasoning with balanced thinking. In *International Conference on Learning Representations*, 2026.

Julian Minder, Clément Dumas, Stewart Slocum, Helena Casademunt, Cameron Holmes, Robert West, and Neel Nanda. Narrow finetuning leaves clearly readable traces in activation differences. In *International Conference on Learning Representations*, 2026.

Benjamin Newman, John Hewitt, Percy Liang, and Christopher D. Manning. The EOS decision and length extrapolation. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pp. 276–291. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.blackboxnlp-1.26.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. ZeRO: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pp. 1–16. IEEE, 2020. doi: 10.1109/SC41405.2020.00024.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt 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)*, pp. 15504–15522. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.828.

Oscar Sainz, Iker García-Ferrero, Rodrigo Agerri, Oier Lopez de Lacalle, German Rigau, and Eneko Agirre. GoLLIE: Annotation guidelines improve zero-shot information-extraction. In *International Conference on Learning Representations*, 2024.

Miles Wang, Tom Dupré la Tour, Olivia Watkins, Alex Makelov, Ryan A. Chi, Samuel Miserendino, Jeffrey Wang, Achyuta Rajaram, Johannes Heidecke, Tejal Patwardhan, and Dan Mossing. Persona features control emergent misalignment. In *International Conference on Learning Representations*, 2026.

Sean Welleck, Ilia Kulikov, Jaedeok Kim, Richard Yuanzhe Pang, and Kyunghyun Cho. Consistency of a recurrent language model with respect to incomplete decoding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5553–5568. Association for Computational Linguistics, 2020a. doi: 10.18653/v1/2020.emnlp-main.448.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. Neural text generation with unlikelihood training. In *International Conference on Learning Representations*, 2020b.

Dongjie Xu, Hao Wu, Weijie Shi, Yue Cui, Yuanjun Liu, Jiawei Li, Haolun Ma, An Liu, Jia Zhu, and Jiajie Xu. LoopGuard: Breaking self-reinforcing attention loops via dynamic KV cache intervention, 2026. arXiv preprint arXiv:2604.10044.

Jin Xu, Xiaojiang Liu, Jianhao Yan, Deng Cai, Huayang Li, and Jian Li. Learning to break the loop: Analyzing and mitigating repetitions for neural text generation. In *Advances in Neural Information Processing Systems*, volume 35, pp. 3082–3095, 2022.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. arXiv preprint arXiv:2505.09388.

Junchi Yao, Shu Yang, Jianhua Xu, Lijie Hu, Mengdi Li, and Di Wang. Understanding the repeat curse in large language models from a feature perspective. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 7787–7815. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.406.

Itay Yona, Ilia Shumailov, Jamie Hayes, and Yossi Gandelsman. Interpreting the repeated token phenomenon in large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 72535–72555. PMLR, 2025.

Zihao Yue, Liang Zhang, and Qin Jin. Less is more: Mitigating multimodal hallucination from an EOS decision perspective. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11766–11781. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.633.

Fu Zhang, Xinlong Jin, Jingwei Cheng, Hongsen Yu, and Huangming Xu. Rethinking the role of LLMs for document-level relation extraction: a refiner with task distribution and probability fusion. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6293–6312. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.naacl-long.319.

Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Hong Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, Wenmeng Zhou, and Yingda Chen. SWIFT: A scalable lightWeight infrastructure for fine-tuning, 2024. arXiv preprint arXiv:2408.05517.

Chenyu Zhou, Qiliang Jiang, and Boguang Pan. Dense coordinate-list fine-tuning induces a controllable interference surface in vision-language models, 2026. arXiv preprint arXiv:2606.14507.

A SUPERVISION CONSTRUCTION

Replacement units and rates. OBR replaces complete four-field blocks in cleaned target lists with violating donor blocks from the corresponding raw records, using each donor and host position at most once. Source text, candidate entities, instructions, record identities, and block counts are unchanged, and each target still has one list close and one EOS. Token length, measured on compact JSON serialization, is matched but not held fixed. Replacement rates of 0%, 5%, 10%, 15%, and 24.3% correspond to 0, 29,251, 58,503, 87,754, and 142,162 replaced blocks, and every condition retains 8,854 records and 585,029 target blocks.

Donor allocation. Matching is restricted to the same record and relative-position bucket. Block j (zero-based) of an n-block list has relative position $j/\max(n-1,1)$, bucketed as early (< 0.25), middle ($[0.25, 0.75)$), or late (otherwise). We distinguish candidate-only violations from those that also lack a source-text endpoint. The largest replacement set follows the donor population's violation-type and position profile through proportional allocation with largest-remainder rounding, capped by the available host positions in each record and bucket.

Within-record matching. An endpoint's degree counts its occurrences as a source or target within the record. We first pair donors and hosts with identical ordered source- and target-degree profiles, processing donors in decreasing token length and breaking host ties by token-length proximity, position proximity, and endpoint-profile agreement. The endpoint profile records substring nesting and whether the source and target strings appear in the donor's description. Remaining donors, taken in decreasing total endpoint degree with fixed pseudorandom tie breaks, are matched on degree distance, the sum of absolute source and target differences in $\log(1 + \text{degree})$. Hosts within 0.45 of the minimum degree distance enter the next stage, where only those within 0.06 of the closest relative position remain eligible. Selection then prioritizes token-length error, degree distance, position distance, endpoint-profile agreement, and deterministic index tie breaks.

Finally, we refine the matches with at most 12 host moves per record to reduce the magnitude of the net token-length change. Each move uses an unused host in the same position bucket. Exact degree matches stay exact, and other matches may worsen by at most 0.12 in degree distance and 0.04 in position distance. All matching decisions precede model training.

Nested subsets. Lower replacement rates take prefixes of one stratified ordering of the largest matched set. Strata are defined by violation type, host position bucket, description endpoint-mention profile, and source and target nesting indicators. Each stratum has a fixed pseudorandom order, and strata are interleaved by within-stratum fractional rank $(j+0.5)/n$ for zero-based rank j and stratum size n. The first $\mathrm{round}(585{,}029\rho)$ entries give the nested 5%, 10%, 15%, and 24.3% sets, so the 5% condition introduces no new donor matching.

Training comparisons and controls. Raw and cleaned pairs use training seed 42 at all three scales, plus a 4B pair with seed 123. OBR uses seed 42 at every nonzero rate, plus a 15% model with seed 123, and the cleaned condition serves as the 0% reference for each seed. All models start from their scale's instruct model and use the same SFT procedure. Input-edit controls keep targets fixed but change candidate membership, source-text occurrence, or both. A benign input-edit control preserves the support status of the target endpoints. Generic label noise keeps endpoints, substitutes a different relation label of nearby token length from the same record, and reorders clauses of the original description.

B REUSE EPISODES AND RISK STANDARDIZATION

Parsing and identity. The event parser accepts complete blocks with exactly the four required string-valued fields, and invalid gaps separate parser segments. A block's reuse identity is its decoded (s, t, r) triple with nonempty entries, excluding the description. By contrast, reference-content scoring normalizes strings by trimming, collapsing whitespace, and case folding before forming directed triples and pairs.

Episode state machine. For each reused block, we find its most recent earlier occurrence and the lag between their block indices. Within the same parser segment, an adjacent reused block extends the active episode if any of three lineage conditions holds: (i) its lag matches the preceding reuse lag, (ii) its identity already belongs to the episode, or (iii) its earlier occurrence belongs to the episode or to the episode's source occurrences. A new block, an empty identity, or a change of parser segment or lineage ends the episode, and a lineage change can start the next episode without an intervening new block. An episode that ends before capture and before the end of generation recovers to the next gap.

Continuous pattern repetition searches for three consecutive copies of any identity pattern, with no maximum length in blocks. Capture is confirmed at the end of the third copy and attributed to the episode retrospectively from the start of the second copy. We keep both locations, and each response is absorbed at its first capture. A response without a complete block contributes no episodes. A normal generation ending counts as a stop even when the JSON is incomplete. Reaching the context limit is recorded separately as censoring, including during an active episode, unless a capture has already been confirmed.

Risk sets and sparse tails. The risk set of gap k contains responses that reach it without capture or an earlier ending, and that of episode k contains those entering reuse from this gap. We count gap entry, gap stop, episode capture, episode stop, recovery, and censoring separately, and compute each cause-specific event probability on the corresponding risk set. Because the product-limit calculation in Eq. equation 4 treats context-limit endings as censoring rather than recovery, the entry and recovery factors are complements of the modeled competing event probabilities.

For each model pair, the common ordinal cutoff is the largest episode index at which both models have at least 50 responses at risk. Beyond this cutoff, we pool each model's sparse tail separately for gap stop, episode stop, and episode capture. Each pooled probability is the event count divided by the summed risk counts and applies through the largest observed ordinal. A pooled component with no observations at risk contributes zero probability. The cutoff is held fixed across bootstrap resamples.

Standardized components. We standardize on the common ordinal grid by exchanging stop and capture intensities between the two compared models. At each stage, the discrete cause probabilities p_q are converted to intensities λ_q as

$$\lambda_q = \frac{p_q}{p_{\text{cap}} + p_{\text{stop}}} \left[-\log\left(1 - \min\{p_{\text{cap}} + p_{\text{stop}}, 1 - 10^{-12}\}\right) \right], \qquad q \in \{\text{cap}, \text{stop}\}. \tag{7}$$

Both intensities are zero when $p_{\text{cap}} + p_{\text{stop}} = 0$, and $p_{\text{cap}} = 0$ at every gap. From the exchanged intensities, each cause probability is recovered as $(1-\exp[-\lambda_{\text{cap}}-\lambda_{\text{stop}}])\lambda_q/(\lambda_{\text{cap}}+\lambda_{\text{stop}})$, which yields four cumulative incidences over the same observed horizon. We write F_{ij} for the incidence that takes capture intensities from model i and stop intensities from model j, where $i, j \in \{0, 1\}$ index the paired models. Averaging over both orders of exchange gives

$$\begin{aligned} \Delta_{\text{exposure}} &= \tfrac{1}{2}[(F_{01} - F_{00}) + (F_{11} - F_{10})], \\ \Delta_{\text{persistence}} &= \tfrac{1}{2}[(F_{10} - F_{00}) + (F_{11} - F_{01})]. \end{aligned} \tag{8}$$

The two contributions sum to $F_{11} - F_{00}$. Exposure exchanges stopping in both gaps and episodes, whereas persistence exchanges episode capture.

C Natural prefix selection

Eligible boundaries. At each scale, anchors come from responses of the raw and cleaned models at training seed 42. Each response contributes at most one anchor, and within each source–role combination each prompt contributes at most one response. Every anchor follows at least four complete relation blocks. A natural stop anchor requires a normally ended, complete JSON list with at least six blocks, no prior capture, and no stable token orbit, and it lies after the last complete block and before the legal close.

For an early anchor, we target a block count of $\max(4, \text{round}(n_{\text{ref}}/2))$, where n_{ref} is the reference-list length and ties are rounded to the even integer. We then search outward for the nearest eligible

complete-block boundary, checking the lower candidate first at equal distance. The boundary must precede both the response's last block and any retrospective capture start, and at least four directed reference pairs must remain uncovered. A before-first-reuse anchor lies immediately before the first block whose identity appeared earlier, regardless of later capture.

Quotas and splitting. We fill the natural stop, early, and before-first-reuse roles in that order, with per-source quotas of 64, 32, and 32. Selection follows a fixed random order and does not filter by model scores. The unique prompts are split approximately in half, and all anchors from a prompt fall in the same split. Table 3 reports the resulting design sizes.

The 4B scale has six models (raw, cleaned, and OBR 15% supervision at training seeds 42 and 123), and the 8B scale has two (raw and cleaned supervision at training seed 42). Each model scores all 256 anchors of its scale, giving 2,048 model–anchor combinations and 32,768 short draws in total. Reported paired comparisons use the test prompts.

Table 3: Development and test sets for shared-prefix measurements. Role columns count natural stop, early, and anchors before first reuse, respectively.

Scale	Split	Prompts	Anchors	Natural stop	Early	Before reuse
4B	Dev	117	129	62	31	36
4B	Test	118	127	66	33	28
8B	Dev	118	127	58	37	32
8B	Test	119	129	70	27	32

D BOUNDARY READOUTS

Tokenization at the decision point. At the selected complete-block boundary, natural stop anchors use their observed legal closing suffix as the close candidate and other roles use the modal closing suffix. The continue candidate is the modal observed inter-block separator followed by the next block's opening through its source-field key. The measurement position is the final position of the longest common token prefix of the two separately tokenized candidates, and c_1 and v_1 are the next close and continue tokens. Since a merged token can place this position inside the completed block's textual tail, we keep only boundaries whose retreat crosses no decisive list-close, separator, or opening-brace character.

Decoding distribution. The presence penalty subtracts 1.5 once from the logit of every token type already present in the assistant prefix and subsequently generated text, excluding the input prompt. We then divide by temperature 0.7 and compute the first-token stop margin before applying top-k 20, then top-p 0.8, and renormalizing. The nucleus includes the token that crosses the cumulative threshold. Close reachability is the indicator that c_1 remains eligible, and next-event close frequency keeps all 16 draws in its denominator.

Structural event classification. Each short continuation runs until a configured EOS token or 32 generated tokens. We scan its first 24 decoded characters for the earliest decisive character, where a closing square bracket indicates close and a comma or opening brace indicates continue. If there is none, EOS also counts as close when the decoded continuation before it is empty or contains only tail-alphabet characters, namely space, tab, carriage return, newline, the ASCII quotes, closing brace, period, question mark, and exclamation mark, as well as the Unicode code points U+3002, U+201D, U+2019, U+300F, U+300D, U+FF01, U+FF1F, and U+2026. All other EOS endings count as EOS without close, and all remaining continuations are unresolved. This operational close frequency and literal JSON completeness are separate readouts.

Forced close and EOS. We teacher-force the complete legal close candidate, whose tokens enter the assistant presence history, and sum the next-token probabilities of the model's configured EOS and end-of-turn tokens, computed with presence penalty and temperature but before top-k and top-p filtering. This readout therefore conditions on a completed legal list close rather than counting generated close characters.

E DIRECTIONS, CONTROLS, AND CONTINUATION SCORING

Development fit. Each scale fits its own direction using the same-scale raw and cleaned models (training seed 42) on 98 anchors at 4B and 90 at 8B, restricted to development natural stop anchors and anchors before first reuse. For each candidate layer ℓ, we take the last-position output of that decoder block before the final model normalization, let $\boldsymbol{u}^\ell = \boldsymbol{d}^\ell / \|\boldsymbol{d}^\ell\|_2$, and compute the paired projection gap $(\boldsymbol{h}^\ell_{\theta_{\text{clean}}} - \boldsymbol{h}^\ell_{\theta_{\text{raw}}})^\top \boldsymbol{u}^\ell$. We select the layer with the largest absolute Spearman correlation between this gap and the paired first-token stop margin gap on the same development anchors, breaking ties by the absolute mean projection gap divided by its sample standard deviation and then in favor of the lower layer index. The OBR model contributes neither activations nor margins to fitting or selection.

Pulse and random directions. We add the signed vector at the final position during prefix prefill and disable the injection before generating subsequent tokens, but the stored attention states retain the resulting change. Baseline and intervened conditions are measured with the same procedure. Eight independent isotropic Gaussian vectors are normalized to the contrast direction's Euclidean norm and applied with the model's intervention sign and strength $\alpha = 1$ at the selected layer and position. Supporting short readouts also apply the contrast direction at $\alpha \in \{0.5, 2\}$.

Close bias. Calibration uses the first 32 development anchors in the fixed pool ordering (31 early anchors and one natural stop anchor at 4B, 32 early anchors at 8B). For each model, 16 short draws estimate the mean close frequency under the direction at $\alpha = 1$ and under each signed bias magnitude, and we pick the magnitude whose mean frequency is closest to the direction's, breaking ties toward the smaller one. The signed bias β (positive for restore and transfer, negative for reverse) is added to the first close token's logit after presence penalty and temperature but before filtering, at the first generation step only. Each model uses one calibrated value for all test roles.

Short intervention design. The short intervention test covers the raw, cleaned, and OBR 15% models at 4B (127 anchors from 118 prompts) and the raw and cleaned models at 8B (128 anchors from 118 prompts), all with training seed 42. Each model is run at every anchor under 13 conditions (baseline, three contrast strengths, all eight equal-norm random directions, and close bias), with 16 draws of at most 32 new tokens per condition. The 8B analysis comprises local intervention readouts and the SAE feature reading.

Full continuation design. A fixed ordering selects 32 early anchors, 16 natural stop anchors, and 16 anchors before first reuse from the 4B test pool (64 anchors from 59 prompts). Each anchor is deterministically assigned one of the eight random directions, independent of the scoring model and fixed across its four draws. The eight directions cover 7, 10, 9, 4, 8, 7, 10, and 9 anchors, respectively. We run each of the three models under four conditions, four draws each, with the same remaining-context budget. Generation stops at a configured EOS token or once input plus output reaches 32,768 tokens.

Repetition and retained content. We parse each prefix and continuation together to preserve patterns that begin before the intervention. A tail capture requires confirmation after the measurement boundary, and a stable token orbit is likewise assigned by its confirmation position. Actual EOS, reaching the context limit, and complete JSON-list parsing are recorded separately. New reference triples and pairs are unique directed matches absent from the prefix and produced by complete blocks that start in the continuation. Remaining-pair recall divides the number of new reference pairs by the number not yet covered at the anchor, and zero-denominator anchors are omitted from this recall but kept for other endpoints. The content-retention criterion requires the lower paired 95% interval endpoint for the early-anchor recall difference against baseline to be at least -0.02.

F ARTIFICIAL COMPLETE-MOTIF GAIN

Probe construction. The diagnostic holds artificial repeated context fixed across four 4B models: the instruct model before task SFT and the cleaned, raw, and OBR 15% models. With 128 reference prompts and motif lengths $m \in \{1, 2, 4, 8\}$, one probe per prompt and length gives 512 probes. Each

eligible boundary leaves a subsequent reference block and fits the context budget for every scored branch. Its full four-field motif has no earlier identical occurrence, and the motif's first relation identity differs from that of the next reference block. Selection balances early, middle, and late reference positions across the pool.

The context $\xi^{(1)}$ ends with one complete motif in the reference prefix, and $\xi^{(2)}$ and $\xi^{(3)}$ append one and two further complete copies using the reference serialization's separators. The repeat candidate b_{rep} is the motif's first block, and the recovery candidate b_{next} is the next block in the original reference sequence.

Candidate scores and gain. We teacher-force each full candidate but average its raw log probabilities only over tokens overlapping the source, target, and relation values,

$$s_\theta(b \mid \xi) = \frac{1}{|\mathcal{J}(b,\xi)|} \sum_{j \in \mathcal{J}(b,\xi)} \log p_\theta(w_j \mid \xi, w_{<j}), \tag{9}$$

where w_j is a candidate token and $\mathcal{J}(b,\xi)$ contains token positions overlapping any of the three identity-field values under joint tokenization. Description and formatting tokens receive no separate score positions, and no decoding penalty or truncation filter is applied. We define the gain after one extra motif copy and its two components as

$$\begin{aligned} G_{\mathrm{copy}} &= s_\theta(b_{\mathrm{rep}} \mid \xi^{(2)}) - s_\theta(b_{\mathrm{rep}} \mid \xi^{(1)}), \\ G_{\mathrm{rec}} &= s_\theta(b_{\mathrm{next}} \mid \xi^{(1)}) - s_\theta(b_{\mathrm{next}} \mid \xi^{(2)}), \\ G &= G_{\mathrm{copy}} + G_{\mathrm{rec}}. \end{aligned} \tag{10}$$

Here G measures the change in relative preference for reuse, G_{copy} the repeat-score change, and G_{rec} the recovery-score decrease. All quantities are in nats per scored token. Replacing $(\xi^{(1)}, \xi^{(2)})$ with $(\xi^{(2)}, \xi^{(3)})$ gives the supporting second-copy gain. The four models, 512 probes, and six branches per probe yield 12,288 candidate scores.

G 8B SPARSE FEATURE READING

We read the residual contrast in the feature coordinates of a sparse autoencoder (SAE) (Gao et al., 2025), using the layer-32 encoder and decoder of the released `SAE-Res-Qwen3-8B-Base-W64K-L0_100` (Deng et al., 2026). The encoder maps each 4,096-dimensional residual output of the 8B models studied here to 65,536 features via an affine map, rectification, and top-100 selection.

Feature selection uses all 127 development anchors. We take the 16 features whose decoder columns have the largest absolute cosine with the contrast direction, and the 16 with the largest absolute mean activation difference among features active on at least 5% of anchors in either model. The two sets share seven features, giving 25 in total. We manually label the features after reading their activating contexts. Table 4 lists these labels with the features' mean activations and activation frequencies on all 129 test anchors.

Features 49602 and 27710 have entry-closing labels, and 33544 is labeled for periods closing description fields. Among the activation-difference features, 6023, 57178, and 63142 are also labeled as entry termination, and others are annotated for document genres and prompt formatting. The activation difference of feature 2236 flips sign from development to test, and aligned features 37527 and 31611 never activate on the test anchors.

H TRAINING, DECODING, AND CORPUS DETAILS

Corpus construction and checks. The corpus contains Chinese documents extracted from publicly available web PDF corpora. Each record pairs a document with a candidate entity list and relation blocks, both generated by larger language models, including Qwen3-235B-A22B. Each block has source, target, relation, and description fields, with free-form relation labels. We keep the supplied train and evaluation splits and remove cross-split source-text overlap during preprocessing. The released files contain no such overlap under exact, whitespace-normalized, or case-folded

Table 4: All 25 selected 8B layer-32 features on 129 test anchors. Means and differences compare cleaned/raw activations, and active shares are percentages (cleaned/raw). The sign column reports development/test agreement and author annotations. Manual labels follow below.

id	Selection	Cosine	Cleaned	Raw	Δ	Active shares	Sign	Feature
21455	Both	0.25	90.81	80.01	+10.80	100.00 / 100.00	Agree	math-notation blocks (LaTeX, symbol-definition lists)
2904	Both	0.21	6.06	0.68	+5.38	52.71 / 7.75	Agree	RE-prompt boilerplate and percentage figures
49602	Alignment	0.21	0.23	0.00	+0.23	2.33 / 0.00	Agree	relation-entry JSON terminators
639	Both	0.19	15.90	11.04	+4.86	72.87 / 65.89	Agree	book front matter (CIP/ISBN/colophon)
41655	Both	0.16	41.01	34.60	+6.41	94.57 / 90.70	Agree	relation-entry JSON structure
32714	Both	-0.16	6.75	13.07	-6.32	44.96 / 81.40	Agree	exam multiple-choice option letters
8833	Both	0.15	26.51	22.56	+3.95	91.47 / 84.50	Agree	extraction-output JSON fields
37527	Alignment	0.15	0.00	0.00	+0.00	0.00 / 0.00	Agree	long digit runs (URLs, phone numbers)
27710	Alignment	0.15	0.29	0.00	+0.29	3.10 / 0.00	Agree	JSON entry closers
3080	Alignment	0.15	3.38	0.62	+2.76	32.56 / 6.20	Agree	technical units and parameter lists
31611	Alignment	0.14	0.00	0.00	+0.00	0.00 / 0.00	Agree	front-matter rosters and questionnaires
49183	Alignment	-0.14	0.56	0.92	-0.36	4.65 / 7.75	Agree	sentence/paragraph boundaries
33544	Alignment	0.14	0.35	0.15	+0.20	3.88 / 1.55	Agree	periods closing JSON description fields
59729	Both	0.14	29.89	24.52	+5.37	99.22 / 96.12	Agree	candidate-list separators and entry structure
2236	Alignment	-0.14	0.07	0.22	-0.15	0.78 / 2.33	Disagree sign inconsistent	markdown bold and entity mentions
2569	Alignment	0.14	1.25	0.32	+0.93	12.40 / 3.88	Agree	candidate-entity list separators
3850	Activation difference	–	34.01	27.76	+6.25	99.22 / 96.90	Agree	description-field prose
63142	Activation difference	–	25.96	21.25	+4.71	90.70 / 82.95	Agree	entry-final periods and candidate terms
33959	Activation difference	–	9.39	4.15	+5.24	63.57 / 33.33	Agree	thermal-engineering formulas and coefficients
25859	Activation difference	–	9.12	1.50	+7.62	42.64 / 7.75	Agree	statute clauses (prohibition/obligation)
30976	Activation difference	–	8.52	2.72	+5.80	51.16 / 23.26	Agree	candidate-list separators in RE prompts
6023	Activation difference	–	38.32	33.32	+5.00	99.22 / 99.22	Agree	relation-entry terminators
30197	Activation difference	–	16.15	12.04	+4.10	92.25 / 77.52	Agree	relation-label words and entity names
12735	Activation difference	–	100.14	106.71	-6.57	100.00 / 100.00	Agree sign reversed	broad punctuation/digit context
57178	Activation difference	–	31.03	26.67	+4.36	92.25 / 90.70	Agree	entry terminators and prompt entities

matching, and neither split has duplicate document–candidate inputs. The shared preprocessing removes within-record duplicate entity pairs and descriptions, and training blocks whose normalized descriptions appear in the evaluation split. The cleaned training and evaluation files contain 585,029 and 59,951 well-formed blocks, respectively. Every endpoint appears in both the candidate list (exact match) and the source text (up to whitespace and case normalization).

Training configuration and corpus sizes. Table 5 lists the configuration, and Table 6 the statistics of the supplied files. Mean token counts are computed from the recorded Qwen3 training-format tokenization, separately for the input and assistant portions. All evaluated checkpoints use the same cleaned reference targets.

Table 5: Training and natural-generation configuration. All supervision conditions at a given scale share this configuration.

Setting	Value
Initialization	Qwen3-1.7B, Qwen3-4B, Qwen3-8B
Training framework	SWIFT (Zhao et al., 2024)
Trainable parameters / precision	All parameters / BF16
Epochs / checkpoint	3 / final checkpoint
Learning rate / schedule	10^{-5} / cosine
Warmup ratio / weight decay	0.03 / 0.1
Adam β_2 / gradient clipping	0.95 / 1
Global batch size	$8\ \text{GPUs} \times 1\ \text{sample} \times 8\ \text{steps} = 64$
Maximum training length / packing	32,768 tokens / disabled
Overlength handling / checkpointing	Delete overlength examples / enabled
Training and data seeds	42 and 123 at 4B, 42 at 1.7B and 8B
Distributed optimizer	DeepSpeed ZeRO stage 3 (Rajbhandari et al., 2020)
Optimizer offload	None at 1.7B and 4B, CPU at 8B
Natural generation engine	vLLM (Kwon et al., 2023)
Reasoning mode	Non-thinking
Temperature / top-p / top-k	0.7 / 0.8 / 20
Presence / repetition penalty	1.5 / 1.0
Sampling seeds per document	0–7
Total context / stopping	32,768 tokens / EOS or context exhaustion

Table 6: Corpus statistics. Blocks are complete relation entries, and input and assistant columns report mean token counts.

Dataset	Records	Blocks	Blocks per record	Input tokens	Assistant tokens
Raw	8,875	779,398	87.82	4,124.12	5,234.42
Cleaned	8,854	585,029	66.08	4,125.49	3,950.85
OBR 5%	8,854	585,029	66.08	4,125.49	3,950.50
OBR 10%	8,854	585,029	66.08	4,125.49	3,949.90
OBR 15%	8,854	585,029	66.08	4,125.49	3,949.17
OBR 24.3%	8,854	585,029	66.08	4,125.49	3,947.65
Eval	1,106	59,951	54.21	4,008.75	3,256.41

I Full natural and replacement results

Here we examine whether changes in repetition coincide with changes in structure, support compliance, and extraction quality. Tables 7 and 8 include the untuned references and all natural-condition levels. Out-of-candidate and out-of-text shares are computed over all parsed generated blocks, repeated ones included. A block is out-of-text if either endpoint is absent from the source text. JSON list completeness and the support indicators rely on separate parsers (Appendix B).

Table 7: Natural-condition levels on 8,848 responses per row. Rates and F1 scores are percentages, and tokens are means.

Model, training seed	Repet.	Orbit	Limit	Tokens	Triple F1	Pair F1
1.7B Untuned	48.52	17.59	45.09	14,593	1.68	10.45
1.7B Cleaned, 42	25.95	6.50	18.03	8,685	6.51	25.77
1.7B Raw, 42	60.58	22.88	54.61	18,021	5.22	19.00
4B Untuned	1.29	0.03	0.03	3,024	4.49	23.23
4B Cleaned, 42	5.52	0.60	1.14	4,734	9.07	33.89
4B Raw, 42	32.30	7.04	21.06	11,001	8.22	29.13
4B Cleaned, 123	4.40	0.51	0.95	4,639	9.24	33.89
4B Raw, 123	33.37	8.48	22.05	11,156	8.16	29.18
8B Untuned	10.27	1.02	7.93	5,514	4.09	22.28
8B Cleaned, 42	7.13	0.51	1.83	5,166	9.74	35.39
8B Raw, 42	31.78	6.20	25.87	11,841	8.66	29.78

Table 8: Generated support violations and output size for the natural conditions. Support violations are percentages of parsed blocks.

Model, training seed	Out-of-candidate blocks	Out-of-text blocks	Mean complete blocks
1.7B Untuned	10.51	2.89	220.78
1.7B Cleaned, 42	4.09	1.70	150.90
1.7B Raw, 42	16.99	5.72	317.61
4B Untuned	11.38	1.78	45.16
4B Cleaned, 42	2.73	1.11	80.91
4B Raw, 42	27.11	13.07	192.67
4B Cleaned, 123	2.50	0.89	79.04
4B Raw, 123	26.73	13.18	194.51
8B Untuned	11.31	2.45	78.21
8B Cleaned, 42	2.18	0.72	88.65
8B Raw, 42	23.39	7.85	207.75

J CONTROLLED SUPERVISION CONDITIONS

These controls test whether input editing, ordinary label changes, or particular subsets of violating targets reproduce the raw-supervision pattern. We construct them as in Appendix A, and Table 9 reports each condition against cleaned supervision at 4B (training seed 42). All four input edits lower repetition, by 0.72 to 1.64 points, with the candidate and text edits giving the largest reductions. Generic label noise, constructed to preserve endpoints, leaves repetition essentially unchanged. At the 24.3% level, OBR raises repetition by 2.88 points to 8.40%, less than it does at the 15% level.

The subtype conditions keep valid blocks and a single violation type, then apply the common quality filters and deduplication. The candidate-only subtype takes violations whose endpoints occur in the source text (8,875 records with 733,877 blocks), and the out-of-text subtype takes those with an endpoint absent from it (8,867 records with 630,577 blocks). They raise repetition by 14.95 and 8.34 points, respectively. Table 10 gives severity and generated-support levels for all control conditions.

K INPUT-SUPPORT PROBE

Target replacement reproduces the repetition failure. We use an input-support probe to ask whether raw supervision also changes how the stop margin responds to input support.

We take reference-list anchors from 128 documents in the cleaned evaluation split, shared by both axes, and cut each list two to four blocks before its end. The preceding reference blocks form the assistant prefix, and the remaining blocks specify the content whose support can be removed. The candidate axis replaces the endpoints of these blocks in the candidate list with distractors of equal

Table 9: Control levels and paired differences against cleaned supervision. Rates are percentages and differences are percentage points with 95% intervals, over 1,106 prompts.

Condition	Repet.	Δ repetition [95% interval]	Pair F1	Δ pair F1 [95% interval]
Cleaned	5.52	–	33.89	–
OBR 24.3%	8.40	+2.88 [2.10, 3.67]	30.74	-3.15 [-3.54, -2.74]
Candidate	3.88	-1.64 [-2.27, -1.04]	33.29	-0.60 [-0.92, -0.28]
Text	4.05	-1.47 [-2.15, -0.85]	33.39	-0.50 [-0.76, -0.23]
Both	4.49	-1.03 [-1.66, -0.40]	33.28	-0.61 [-0.90, -0.32]
Benign	4.79	-0.72 [-1.32, -0.12]	33.79	-0.10 [-0.37, 0.18]
Generic	5.74	+0.23 [-0.40, 0.86]	33.45	-0.43 [-0.71, -0.16]
Raw, candidate-only	20.47	+14.95 [13.70, 16.25]	30.24	-3.65 [-4.03, -3.25]
Raw, out-of-text	13.86	+8.34 [7.31, 9.32]	32.64	-1.24 [-1.56, -0.92]
Raw	32.30	+26.79 [25.12, 28.40]	29.13	-4.76 [-5.22, -4.28]

Table 10: Control severity and support-compliance levels, in percent. Repeated blocks remain in the support-rate denominators.

Condition	Orbit	Limit	JSON	Out-of-candidate	Out-of-text
Cleaned	0.60	1.14	98.73	2.73	1.11
OBR 24.3%	1.25	1.96	97.86	18.61	7.70
Candidate	0.43	0.50	99.31	4.87	1.15
Text	0.44	0.88	98.98	2.36	0.92
Both	0.31	0.51	99.34	2.79	1.12
Benign	0.52	0.83	98.94	2.63	1.08
Generic	0.85	1.25	98.59	2.87	1.21
Raw, candidate-only	4.24	7.90	91.51	22.65	7.58
Raw, out-of-text	2.58	3.47	96.34	15.03	11.32
Raw	7.04	21.06	78.68	27.11	13.07

token length, leaving their occurrences in the source text intact. The text axis instead replaces every literal occurrence of these endpoints in the source with a donor entity of equal token length and keeps the candidate list as is. At dose $r \in \{0.5, 1\}$, support removal edits $\mathrm{round}(rn)$ of the n remaining blocks, while a neutral edit matches the edit count, positions, and token lengths without removing reference support. The mention-matched subset further requires removal and neutral editing to change the same number of text mentions at both doses, which selects 80 anchors before model-specific exclusions.

The readout is the first-token stop margin under the deployed presence penalty and temperature, taken before top-k and top-p filtering. Unlike the readout in Appendix D, which uses the observed closing syntax and a continuation through the next source-field key, this probe compares the canonical suffixes `]` and `, {`. Joint tokenization locates the first token at which the two suffixes diverge, and the generated assistant prefix is kept as the presence-penalty history. Writing $m_i^{\mathrm{remove}}(r)$ and $m_i^{\mathrm{neutral}}(r)$ for the two margins at anchor i, we define the support effect δ_i and its through-origin slope b_i as

$$\delta_i(r) = m_i^{\mathrm{remove}}(r) - m_i^{\mathrm{neutral}}(r), \qquad b_i = \frac{0.5\,\delta_i(0.5) + \delta_i(1)}{1.25}, \quad \delta_i(0) = 0. \tag{11}$$

The candidate-support and text-support slopes are the per-axis means of b_i, in nats from zero to full removal. Informative anchors are selected per model by excluding those whose close probability, measured before filtering, is at least 0.999 at baseline or at most 10^{-6} after full removal. This leaves 118 to 128 anchors per axis and 73 to 80 in the mention-matched subset.

The instruction arm adds a user-turn instruction either to stop now or to continue with a named remaining relation, and compares it with a length-matched neutral instruction. Both changes are reported as the instructed margin minus the neutral one, so a positive value always favors closing. The empty-answer reference contains 128 further anchors whose candidate entities come from an-

Table 11: Input-support slopes in nats. Candidate and text columns include 95% intervals. The final column gives the text-support point estimate on the mention-matched subset. Counts list informative candidate-axis, text-axis, and mention-matched anchors.

Model, training seed	Informative counts	Candidate	Text	Mention-matched
1.7B Untuned	118/119/73	1.01 [0.65, 1.43]	0.96 [0.62, 1.36]	1.24
4B Untuned	121/119/75	1.72 [1.13, 2.41]	1.60 [1.04, 2.22]	2.17
8B Untuned	123/123/77	1.05 [0.71, 1.39]	0.80 [0.50, 1.13]	1.11
1.7B Cleaned, 42	128/128/80	0.51 [0.33, 0.72]	0.60 [0.39, 0.85]	0.84
4B Cleaned, 123	128/128/80	0.94 [0.64, 1.29]	1.11 [0.76, 1.52]	1.57
4B Cleaned, 42	128/128/80	0.84 [0.55, 1.19]	1.00 [0.67, 1.39]	1.42
8B Cleaned, 42	128/128/80	0.92 [0.61, 1.28]	1.06 [0.73, 1.46]	1.48
4B Benign, 42	128/128/80	0.98 [0.68, 1.36]	1.20 [0.84, 1.62]	1.66
4B Generic, 42	127/128/80	0.91 [0.61, 1.22]	1.00 [0.68, 1.37]	1.42
4B OBR 10%, 42	128/128/80	0.78 [0.53, 1.06]	0.88 [0.61, 1.19]	1.22
4B OBR 15%, 42	128/128/80	0.77 [0.55, 1.02]	0.83 [0.58, 1.12]	1.16
4B OBR 24.3%, 42	128/128/80	0.74 [0.54, 0.96]	0.81 [0.56, 1.09]	1.14
4B Candidate, 42	127/127/79	0.53 [0.30, 0.77]	0.77 [0.49, 1.10]	1.08
4B Text, 42	120/124/77	-0.73 [-1.09, -0.34]	0.70 [0.28, 1.15]	1.09
4B Both, 42	127/127/79	0.32 [0.04, 0.64]	0.80 [0.51, 1.09]	1.15
4B Raw, 42	127/128/80	0.67 [0.47, 0.91]	0.80 [0.55, 1.10]	1.10
4B Raw candidate-only, 42	127/128/80	0.67 [0.46, 0.93]	0.83 [0.55, 1.15]	1.16
4B Raw out-of-text, 42	127/128/80	0.79 [0.53, 1.11]	0.95 [0.64, 1.30]	1.38

other document and never appear in the source, so an empty list is the only supported answer. Its empty-list margin compares closing and continuing at the text boundary after the opening bracket, using token alignment to handle a merged bracket token. Each model is scored on 384 anchors and 2,560 inputs. The 18 models are the untuned and cleaned references at every scale, both cleaned 4B seeds, and the raw, replacement, input-edit, and subtype conditions listed in Table 11.

Output-side supervision retains a positive support response. Raw supervision gives candidate-support and text-support slopes about 0.2 to 0.3 nats below those of the two cleaned 4B models, and on each full axis the raw and OBR estimates fall within the intervals of both cleaned models. The gap is similar on the mention-matched subset, where the text-support slope is 1.42 for cleaned seed 42 against 1.10 for raw and 1.16 for OBR 15%. Raw supervision instead shifts the margin level. Its untouched candidate-axis margin is more than two nats below those of the cleaned models, with OBR 15% in between (Table 12). This gap agrees with the shared-prefix comparison. Generic label noise gives an even lower margin without increasing repetition.

The probe is sensitive to input-side training edits. After text-edit training, neutral edits alone lower the margin by about 3 nats on both axes, and the candidate-support slope is negative. Training with both edits gives much smaller shifts, below 0.8 nats, and like candidate-edit training it keeps a positive but reduced candidate-support slope. In the other 16 models, full-dose neutral shifts stay within 0.39 nats in absolute value. For cleaned seed 42, support removal adds another 0.78 and 0.91 nats beyond the neutral edit, against 0.64 and 0.72 under raw supervision. Because the input-edit arms reduce repetition, these input-side responses are separate from the output-target failure.

Input removal and explicit instructions shift the decision much less than the supervision contrast does. Full removal changes the post-filter close probability by at most 3.42 points in any model (Table 13), and the stop instruction moves the margin by less than half a nat in either direction. The continue instruction raises the stop margin in 17 of the 18 models, including every fine-tuned one (0.29 nats under raw supervision), and lowers it only in the untuned 8B model. The supervision contrast and the single pulse, by comparison, shift the same kind of local decision by about two nats.

Task SFT drives the empty-list margin below zero under every supervision condition. At 4B it falls from $+2.24$ in the untuned model to about -24 in the cleaned models, and every fine-tuned 4B

Table 12: Neutral-edit shifts at full dose and untouched margin levels, in nats. Each pair of columns corresponds to the candidate and text axes on that model's informative anchors.

Model, training seed	Neutral edit Candidate	Neutral edit Text	Untouched Candidate	Untouched Text
1.7B Untuned	0.01	0.08	-7.30	-7.43
4B Untuned	0.29	0.24	-5.58	-5.40
8B Untuned	0.17	0.21	-8.30	-8.30
1.7B Cleaned, 42	0.11	0.06	-4.57	-4.57
4B Cleaned, 123	0.27	0.18	-4.70	-4.70
4B Cleaned, 42	0.29	0.18	-4.62	-4.62
8B Cleaned, 42	0.30	0.19	-4.61	-4.61
4B Benign, 42	0.39	0.35	-5.08	-5.08
4B Generic, 42	0.36	0.27	-7.78	-7.85
4B OBR 10%, 42	0.29	0.21	-5.39	-5.39
4B OBR 15%, 42	0.29	0.22	-5.33	-5.32
4B OBR 24.3%, 42	0.25	0.19	-5.35	-5.34
4B Candidate, 42	0.12	-0.12	-5.20	-5.20
4B Text, 42	-3.17	-2.93	-5.31	-5.38
4B Both, 42	-0.65	-0.79	-5.95	-5.98
4B Raw, 42	0.19	0.19	-7.02	-7.09
4B Raw candidate-only, 42	0.20	0.19	-6.71	-6.78
4B Raw out-of-text, 42	0.23	0.17	-5.09	-5.17

Table 13: Readouts for the same 18 models: instructed-minus-neutral stop/continue margin changes (nats), probability changes at full support removal (percentage points), and the share of empty-answer anchors whose close token ranks first (%).

Model, training seed	Stop	Continue	Δ close prob. Candidate	Δ close prob. Text	Empty-list margin	Close first (%)
1.7B Untuned	-0.09	+1.47	-0.52	+0.41	-1.85	3.91
4B Untuned	-0.46	+0.91	+3.42	+0.04	2.24	59.38
8B Untuned	-0.43	-0.20	+0.00	+0.00	0.11	45.31
1.7B Cleaned, 42	-0.02	+0.83	+0.07	+1.12	-18.04	0.00
4B Cleaned, 123	+0.15	+0.82	+2.15	+2.55	-23.79	0.00
4B Cleaned, 42	-0.01	+0.75	+1.55	+1.82	-24.40	0.00
8B Cleaned, 42	+0.05	+0.34	+3.11	+2.14	-8.45	0.00
4B Benign, 42	+0.20	+0.97	+1.78	+1.63	-20.42	0.00
4B Generic, 42	+0.16	+1.16	+0.93	+0.65	-20.67	0.00
4B OBR 10%, 42	-0.07	+0.72	+1.03	+1.18	-21.47	0.00
4B OBR 15%, 42	+0.08	+0.56	+1.01	+0.87	-21.08	0.00
4B OBR 24.3%, 42	+0.06	+0.86	+0.50	+0.69	-23.52	0.00
4B Candidate, 42	+0.00	+0.84	-0.06	+0.84	-20.91	0.00
4B Text, 42	+0.10	+1.05	-1.08	-0.14	-22.59	0.00
4B Both, 42	-0.01	+0.63	-0.40	-0.04	-21.17	0.00
4B Raw, 42	+0.00	+0.29	+0.44	+0.20	-21.78	0.00
4B Raw candidate-only, 42	+0.05	+0.37	+0.21	+0.50	-18.68	0.00
4B Raw out-of-text, 42	+0.06	+0.70	+1.12	+1.68	-23.21	0.00

condition, raw supervision included, lies between -24.40 and -18.68. The drop is smaller at 1.7B ($-1.85 \rightarrow -18.04$) and 8B ($+0.11 \rightarrow -8.45$). Before tuning, the close token ranks first on 59.38%, 45.31%, and 3.91% of empty-answer anchors at 4B, 8B, and 1.7B, but on none after task SFT. The empty-list margin does not track the repetition differences among fine-tuned conditions. For data auditing, the out-of-candidate target rule locates the output-target failure, while the measured input-support response remains present.

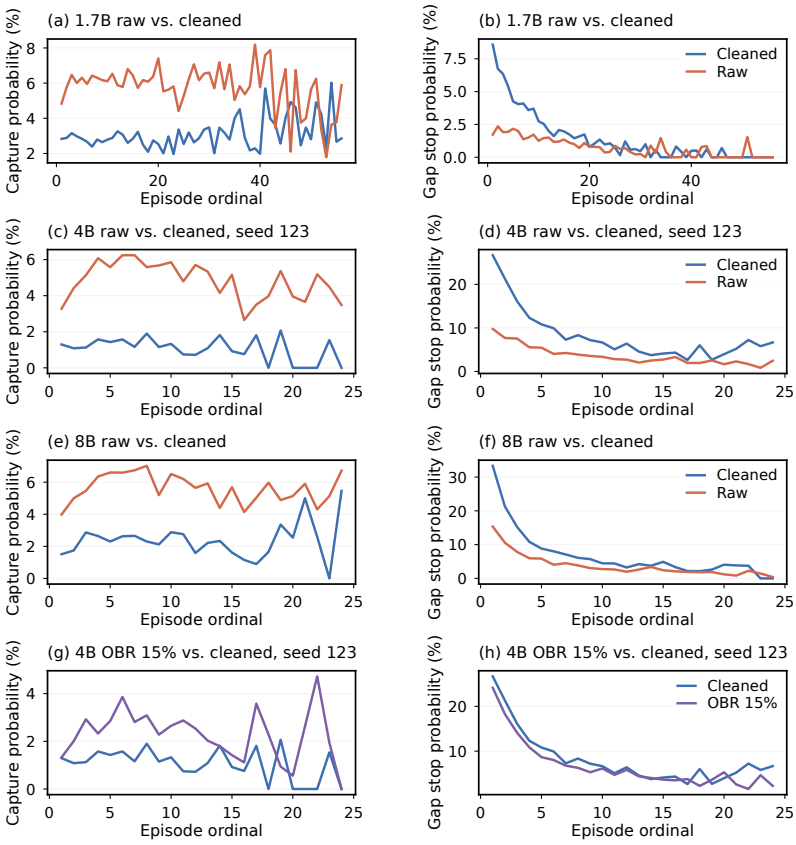

Figure 5: Additional natural and OBR process curves by model pair (rows): capture (left) and gap stop (right) probabilities for ordinals with ≥ 50 started episodes in each model. Training seed: 42 unless labeled 123.

L PROCESS RESULTS ACROSS ALL CONTRASTS

Tables 14, 17, and 19 decompose all nine capture-incidence contrasts into reuse and stopping components, under the lineage definition and the first confirmed continuous pattern repetition. Because the incidence estimator treats context-limit endings as censored, its final value can differ from the directly observed response repetition rate. Exposure and persistence are computed with the common-tail standardization of Appendix B.

Among the controls, generic label noise raises exposure and lowers persistence by similar amounts, and its capture incidence is almost unchanged. The candidate input edit raises exposure only slightly but lowers persistence, so capture incidence falls by 1.6 percentage points. These two controls thus combine exposure and persistence in different ways despite similar or lower answer-level repetition. The ordinal curves not shown in Figure 2b are given in Figures 5 and 6.

M ARTIFICIAL MOTIF GAIN RESULTS

The artificial motif diagnostic checks whether natural repetition and the conditional preference to repeat an explicitly supplied motif rank the models in the same order. Table 20 reports the total gain and its components, computed with the identity-value token mask from Appendix F. Both raw and OBR have a higher total gain than cleaned. These differences, like the change from untuned to cleaned, come almost entirely from G_{copy}. None of the three contrasts shows a reliable change in G_{rec}. Table 21 breaks the results down by motif length for every contrast. At each length the G

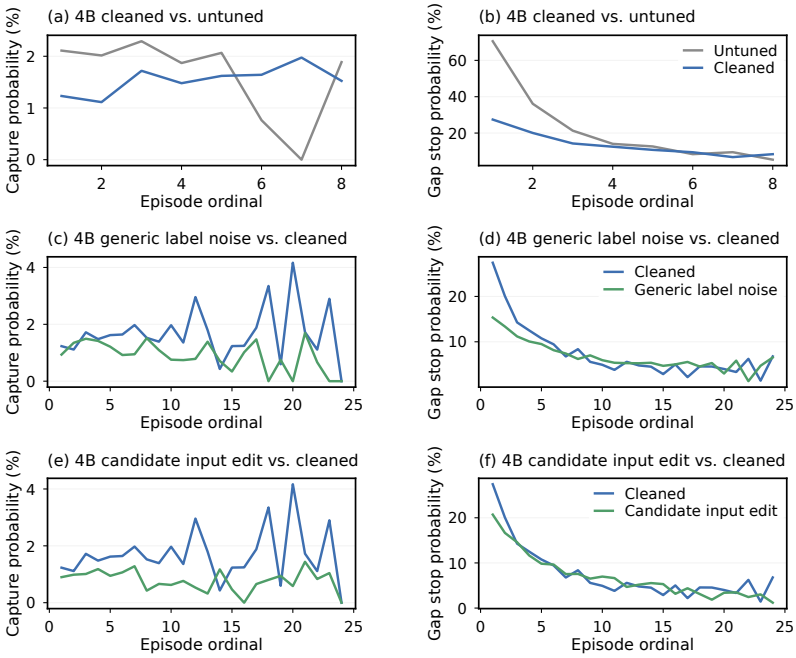

Figure 6: Untuned, generic label noise, and candidate input edit process curves at 4B. Each row identifies its reference and treatment, with capture probability on the left and gap stop probability on the right. Curves are restricted to ordinals with at least 50 episodes in both models.

differences have the same sign as in the aggregate, although the OBR difference at length 8 is not reliably above zero.

N SHARED-PREFIX RESULTS BY ROLE AND SOURCE

Here we check whether the shared-prefix result depends on the source model or on the selected boundary role. The contrasts in Tables 22, and 23 are computed on test anchors only and include the cleaned training-seed reference contrast. Within each source stratum, the compared models see the same prefixes. The full-pool readouts in Tables 26 and 27 also include development anchors, so their aggregation set differs from that of Table 22. Post-filter close probabilities for the test comparisons are listed in Table 24.

O INTERVENTION DETAILS AND FULL CONTINUATIONS

Table 18 lists every candidate layer, Table 28 the selected norms along with held-out projection discrimination (reported descriptively), and Tables 25 and 29 the full strength and random-direction series. Table 16 compares the contrast direction with the within-anchor random-direction mean.

We calibrate the close bias on the first 32 development anchors in a fixed order (31 early and one natural stop anchor at 4B, all early at 8B). Each candidate magnitude in $\{0, 0.5, 1, 2, 4\}$ uses 16 draws per anchor, with ties in calibration error broken toward the smaller magnitude. Short-intervention baselines do not reuse draws from the shared-prefix measurement. Because the deterministic seed includes the condition identifier, distinct conditions receive independently seeded draws even when their logits coincide. Table 15 reports the selected signed biases and realized test outcomes.

Tables 30, 31, and 32 cover every full-continuation arm and role. Each cell gives the level and paired change of the continuation endpoint, not the local closing result. The random direction is fixed per anchor across its four draws (Appendix E). Recall is averaged only over anchors with a nonempty remaining reference-pair set, and the other endpoints over all anchors.

Table 14: Process contrasts with paired 95% intervals from 5,000 prompt resamples (part 1 of 3). Conditional rates and the share without reuse are percentages, episode counts retain their units, capture incidence, exposure, and persistence are probabilities. Dashes in component columns denote quantities defined only for the contrast.

Metric	Reference	Treatment	Difference [95% interval]
4B Raw, seed 42 minus 4B Cleaned, seed 42			
Episodes per response	3.69	6.66	+2.96 [2.78, 3.16]
Responses without reuse	27.51	9.71	-17.80 [-19.13, -16.53]
Captures per episode	1.49	4.85	+3.36 [3.09, 3.64]
Gap stop probability	15.15	5.55	-9.61 [-10.39, -8.91]
In-episode stop probability	7.68	4.23	-3.45 [-3.78, -3.15]
Capture incidence	0.055	0.324	+0.269 [0.252, 0.286]
Exposure	–	–	+0.127 [0.118, 0.137]
Persistence	–	–	+0.142 [0.128, 0.156]
4B OBR 15%, seed 42 minus 4B Cleaned, seed 42			
Episodes per response	3.69	3.93	+0.24 [0.12, 0.37]
Responses without reuse	27.51	23.51	-4.00 [-5.20, -2.80]
Captures per episode	1.49	2.47	+0.98 [0.75, 1.20]
Gap stop probability	15.15	13.42	-1.74 [-2.31, -1.19]
In-episode stop probability	7.68	7.42	-0.26 [-0.59, 0.08]
Capture incidence	0.055	0.097	+0.042 [0.033, 0.051]
Exposure	–	–	+0.009 [0.006, 0.011]
Persistence	–	–	+0.033 [0.025, 0.041]
4B Cleaned, seed 42 minus 4B Untuned			
Episodes per response	0.65	3.69	+3.05 [2.85, 3.24]
Responses without reuse	70.54	27.51	-43.03 [-44.89, -41.11]
Captures per episode	1.99	1.49	-0.49 [-1.01, 0.02]
Gap stop probability	56.01	15.15	-40.86 [-42.97, -38.73]
In-episode stop probability	24.88	7.68	-17.19 [-18.50, -15.96]
Capture incidence	0.013	0.055	+0.042 [0.034, 0.051]
Exposure	–	–	+0.050 [0.041, 0.060]
Persistence	–	–	-0.007 [-0.019, 0.004]

Table 15: Test outcomes for development-selected close biases. Frequencies: percent, differences: percentage points (paired 95% intervals). Zero bias preserves logits and uses separate draws.

Model and operation	β	Before	After	Δ next-event close frequency
4B Cleaned reverse	-1	27.02	12.99	-14.03 [-17.32, -10.88]
4B Raw restore	+0	11.12	10.63	-0.49 [-1.84, 0.86]
4B OBR 15% transfer	+2	24.16	48.18	+24.02 [19.29, 28.91]
8B Cleaned reverse	+0	30.03	31.40	+1.37 [-0.64, 3.34]
8B Raw restore	+0	11.13	11.62	+0.49 [-1.07, 2.05]

Table 16: Contrast direction at strength 1 minus the within-anchor mean of eight equal-norm random directions, with paired 95% intervals.

Model and operation	Δ next-event close frequency (pp)	Δ first-token stop margin (nats)
4B Cleaned reverse	-15.60 [-18.97, -12.24]	-2.16 [-2.31, -2.02]
4B Raw restore	+11.53 [8.66, 14.50]	+2.23 [2.07, 2.40]
4B OBR 15% transfer	+15.77 [12.80, 18.65]	+1.85 [1.66, 2.05]
8B Cleaned reverse	-17.83 [-21.16, -14.49]	-2.66 [-2.94, -2.41]
8B Raw restore	+15.63 [12.03, 19.41]	+2.72 [2.46, 2.98]

Table 17: Process contrasts with paired 95% intervals from 5,000 prompt resamples (part 2 of 3). Units and conventions follow Table 14.

Metric	Reference	Treatment	Difference [95% interval]
4B Generic, seed 42 minus 4B Cleaned, seed 42			
Episodes per response	3.69	5.09	+1.40 [1.26, 1.54]
Responses without reuse	27.51	15.39	-12.12 [-13.22, -11.06]
Captures per episode	1.49	1.13	-0.37 [-0.53, -0.20]
Gap stop probability	15.15	10.09	-5.06 [-5.66, -4.51]
In-episode stop probability	7.68	7.28	-0.41 [-0.72, -0.10]
Capture incidence	0.055	0.057	+0.002 [-0.004, 0.009]
Exposure	–	–	+0.018 [0.015, 0.020]
Persistence	–	–	-0.015 [-0.022, -0.009]
4B Candidate, seed 42 minus 4B Cleaned, seed 42			
Episodes per response	3.69	4.25	+0.56 [0.43, 0.69]
Responses without reuse	27.51	20.72	-6.79 [-7.96, -5.64]
Captures per episode	1.49	0.91	-0.58 [-0.76, -0.41]
Gap stop probability	15.15	12.78	-2.38 [-2.93, -1.84]
In-episode stop probability	7.68	7.94	+0.26 [-0.10, 0.60]
Capture incidence	0.055	0.039	-0.016 [-0.023, -0.010]
Exposure	–	–	+0.005 [0.004, 0.007]
Persistence	–	–	-0.022 [-0.028, -0.015]
4B Raw, seed 123 minus 4B Cleaned, seed 123			
Episodes per response	3.49	6.75	+3.26 [3.05, 3.47]
Responses without reuse	26.72	9.93	-16.78 [-18.00, -15.57]
Captures per episode	1.26	4.94	+3.68 [3.40, 3.97]
Gap stop probability	15.97	5.43	-10.54 [-11.34, -9.80]
In-episode stop probability	8.42	4.05	-4.37 [-4.71, -4.04]
Capture incidence	0.044	0.336	+0.292 [0.274, 0.309]
Exposure	–	–	+0.129 [0.119, 0.139]
Persistence	–	–	+0.163 [0.148, 0.177]

Table 18: Development layer selection. The selected layer maximizes the absolute Spearman correlation between paired projection and margin gaps, with all three candidates retained here.

Scale	Layer	Development anchors	Contrast norm	Spearman correlation	Selected
4B	17	98	2.77	0.05	No
4B	26	98	8.21	0.43	No
4B	32	98	32.52	0.68	Yes
8B	17	90	4.61	-0.01	No
8B	26	90	17.37	0.13	No
8B	32	90	60.93	0.26	Yes

Table 19: Process contrasts with paired 95% intervals from 5,000 prompt resamples (part 3 of 3). Units and conventions follow Table 14.

Metric	Reference	Treatment	Difference [95% interval]
4B OBR 15%, seed 123 minus 4B Cleaned, seed 123			
Episodes per response	3.49	4.05	+0.57 [0.44, 0.69]
Responses without reuse	26.72	24.16	-2.55 [-3.68, -1.34]
Captures per episode	1.26	2.34	+1.08 [0.86, 1.28]
Gap stop probability	15.97	13.24	-2.73 [-3.33, -2.16]
In-episode stop probability	8.42	7.05	-1.37 [-1.72, -1.02]
Capture incidence	0.044	0.095	+0.051 [0.042, 0.059]
Exposure	–	–	+0.014 [0.012, 0.017]
Persistence	–	–	+0.037 [0.029, 0.044]
1.7B Raw, seed 42 minus 1.7B Cleaned, seed 42			
Episodes per response	9.02	10.08	+1.07 [0.70, 1.42]
Responses without reuse	8.61	1.79	-6.83 [-7.73, -5.96]
Captures per episode	2.88	6.01	+3.13 [2.92, 3.35]
Gap stop probability	4.01	1.54	-2.47 [-2.73, -2.23]
In-episode stop probability	4.02	2.30	-1.73 [-1.92, -1.55]
Capture incidence	0.260	0.609	+0.349 [0.333, 0.364]
Exposure	–	–	+0.190 [0.179, 0.203]
Persistence	–	–	+0.158 [0.147, 0.170]
8B Raw, seed 42 minus 8B Cleaned, seed 42			
Episodes per response	3.26	5.71	+2.45 [2.27, 2.62]
Responses without reuse	33.34	15.42	-17.92 [-19.28, -16.64]
Captures per episode	2.19	5.57	+3.38 [3.08, 3.69]
Gap stop probability	16.92	6.96	-9.97 [-10.82, -9.19]
In-episode stop probability	8.09	4.42	-3.68 [-4.02, -3.32]
Capture incidence	0.071	0.319	+0.248 [0.232, 0.264]
Exposure	–	–	+0.135 [0.125, 0.145]
Persistence	–	–	+0.112 [0.100, 0.125]

Table 20: Motif gains over 512 probes from 128 prompts, in nats per scored token. Differences have paired 95% intervals, and the second-copy gain compares contexts with two and three copies.

Readout	Reference	Treatment	Difference [95% interval]
Cleaned, seed 42 minus Untuned			
G	1.55	0.32	-1.23 [-1.33, -1.14]
G_{copy}	1.55	0.29	-1.26 [-1.36, -1.16]
G_{rec}	0.00	0.02	+0.03 [0.00, 0.05]
Second-copy gain	0.70	0.54	-0.16 [-0.20, -0.13]
Raw, seed 42 minus Cleaned, seed 42			
G	0.32	0.41	+0.09 [0.07, 0.10]
G_{copy}	0.29	0.38	+0.08 [0.07, 0.10]
G_{rec}	0.02	0.03	+0.00 [0.00, 0.01]
Second-copy gain	0.54	0.51	-0.03 [-0.04, -0.01]
OBR 15%, seed 42 minus Cleaned, seed 42			
G	0.32	0.38	+0.06 [0.05, 0.08]
G_{copy}	0.29	0.35	+0.06 [0.04, 0.07]
G_{rec}	0.02	0.03	+0.00 [0.00, 0.01]
Second-copy gain	0.54	0.52	-0.01 [-0.03, 0.00]

Table 21: Paired motif-gain differences by motif length, in nats per scored token with 95% intervals. Every length uses the same 128 prompts.

Motif length	G	G_{copy}	G_{rec}
Cleaned, seed 42 minus Untuned			
1	-0.91 [-1.08, -0.77]	-1.01 [-1.19, -0.86]	+0.10 [0.06, 0.14]
2	-0.98 [-1.16, -0.81]	-1.05 [-1.22, -0.89]	+0.07 [0.03, 0.12]
4	-1.22 [-1.40, -1.06]	-1.25 [-1.43, -1.10]	+0.03 [-0.03, 0.09]
8	-1.81 [-2.04, -1.58]	-1.72 [-1.94, -1.51]	-0.09 [-0.18, -0.01]
Raw, seed 42 minus Cleaned, seed 42			
1	+0.06 [0.04, 0.08]	+0.06 [0.04, 0.08]	+0.00 [0.00, 0.01]
2	+0.03 [0.01, 0.05]	+0.04 [0.03, 0.06]	-0.01 [-0.02, 0.00]
4	+0.10 [0.07, 0.15]	+0.10 [0.07, 0.14]	+0.00 [-0.01, 0.01]
8	+0.15 [0.12, 0.19]	+0.14 [0.11, 0.17]	+0.01 [0.00, 0.03]
OBR 15%, seed 42 minus Cleaned, seed 42			
1	+0.03 [0.00, 0.05]	+0.02 [0.00, 0.05]	+0.01 [0.00, 0.01]
2	+0.09 [0.07, 0.12]	+0.09 [0.06, 0.12]	+0.00 [-0.01, 0.01]
4	+0.09 [0.06, 0.12]	+0.08 [0.05, 0.11]	+0.01 [0.00, 0.02]
8	+0.03 [-0.01, 0.07]	+0.04 [0.00, 0.08]	-0.01 [-0.02, 0.00]

Table 22: Test-prefix paired differences by role and source, with 95% intervals (part 1 of 2). Counts give anchors/prompts, next-event close frequency differences are percentage points, and first-token stop margin differences are nats.

Anchor role	Source	Anchors/ prompts	Δ next-event close frequency	Δ first-token stop margin
4B Raw, seed 42 minus 4B Cleaned, seed 42				
Natural stop	Both	66/66	-32.20 [-37.50, -26.61]	-2.03 [-2.35, -1.72]
Natural stop	Cleaned	33/33	-32.01 [-38.26, -26.13]	-2.78 [-3.23, -2.34]
Natural stop	Raw	33/33	-32.39 [-41.10, -24.05]	-1.27 [-1.54, -1.00]
Early	Both	33/33	+0.00 [0.00, 0.00]	-1.79 [-2.31, -1.33]
Early	Cleaned	16/16	+0.00 [0.00, 0.00]	-1.69 [-2.47, -1.00]
Early	Raw	17/17	+0.00 [0.00, 0.00]	-1.89 [-2.59, -1.26]
Before first reuse	Both	28/28	-1.79 [-4.46, 0.00]	-2.43 [-2.99, -1.92]
Before first reuse	Cleaned	15/15	+0.00 [0.00, 0.00]	-2.23 [-3.05, -1.48]
Before first reuse	Raw	13/13	-3.85 [-9.62, 0.00]	-2.66 [-3.35, -1.98]
4B Raw, seed 123 minus 4B Cleaned, seed 123				
Natural stop	Both	66/66	-35.98 [-41.95, -29.64]	-2.09 [-2.36, -1.83]
Natural stop	Cleaned	33/33	-32.77 [-42.23, -24.05]	-2.68 [-3.03, -2.32]
Natural stop	Raw	33/33	-39.20 [-47.16, -31.63]	-1.49 [-1.72, -1.27]
Early	Both	33/33	+0.00 [0.00, 0.00]	-1.30 [-1.81, -0.82]
Early	Cleaned	16/16	+0.00 [0.00, 0.00]	-1.24 [-1.93, -0.54]
Early	Raw	17/17	+0.00 [0.00, 0.00]	-1.37 [-2.14, -0.68]
Before first reuse	Both	28/28	-2.46 [-6.25, 0.00]	-2.03 [-2.63, -1.43]
Before first reuse	Cleaned	15/15	+0.00 [0.00, 0.00]	-1.73 [-2.67, -0.83]
Before first reuse	Raw	13/13	-5.29 [-12.50, 0.00]	-2.38 [-3.10, -1.65]

Table 23: Test-prefix paired differences by role and source, with 95% intervals (part 2 of 2). Units and conventions follow Table 22.

Anchor role	Source	Anchors/ prompts	Δ next-event close frequency	Δ first-token stop margin
4B OBR 15%, seed 42 minus 4B Cleaned, seed 42				
Natural stop	Both	66/66	-9.38 [-14.39, -4.55]	-0.33 [-0.51, -0.15]
Natural stop	Cleaned	33/33	-7.95 [-14.40, -1.70]	-0.31 [-0.60, -0.04]
Natural stop	Raw	33/33	-10.80 [-18.00, -3.98]	-0.35 [-0.56, -0.13]
Early	Both	33/33	+0.95 [0.00, 2.84]	+0.65 [0.19, 1.13]
Early	Cleaned	16/16	+0.00 [0.00, 0.00]	+0.42 [-0.03, 0.89]
Early	Raw	17/17	+1.84 [0.00, 5.51]	+0.87 [0.08, 1.70]
Before first reuse	Both	28/28	-0.89 [-2.68, 0.00]	+0.24 [-0.30, 0.77]
Before first reuse	Cleaned	15/15	+0.00 [0.00, 0.00]	+0.30 [-0.62, 1.17]
Before first reuse	Raw	13/13	-1.92 [-5.77, 0.00]	+0.18 [-0.43, 0.80]
4B OBR 15%, seed 123 minus 4B Cleaned, seed 123				
Natural stop	Both	66/66	-6.44 [-10.42, -2.75]	-0.29 [-0.43, -0.15]
Natural stop	Cleaned	33/33	-7.01 [-12.32, -1.70]	-0.36 [-0.55, -0.17]
Natural stop	Raw	33/33	-5.87 [-10.61, -0.95]	-0.22 [-0.41, -0.02]
Early	Both	33/33	+0.95 [0.00, 2.84]	+0.43 [-0.11, 1.01]
Early	Cleaned	16/16	+0.00 [0.00, 0.00]	+0.13 [-0.45, 0.66]
Early	Raw	17/17	+1.84 [0.00, 5.51]	+0.70 [-0.19, 1.65]
Before first reuse	Both	28/28	+0.22 [-2.46, 2.90]	+0.33 [-0.20, 0.96]
Before first reuse	Cleaned	15/15	+1.67 [0.00, 5.00]	+0.46 [-0.50, 1.52]
Before first reuse	Raw	13/13	-1.44 [-5.77, 1.44]	+0.16 [-0.41, 0.81]
4B Cleaned, seed 123 minus 4B Cleaned, seed 42				
Natural stop	Both	66/66	-1.70 [-5.11, 1.61]	-0.01 [-0.11, 0.10]
Natural stop	Cleaned	33/33	-1.52 [-7.58, 3.98]	+0.02 [-0.17, 0.20]
Natural stop	Raw	33/33	-1.89 [-5.68, 1.89]	-0.03 [-0.15, 0.09]
Early	Both	33/33	+0.00 [0.00, 0.00]	-0.19 [-0.39, 0.01]
Early	Cleaned	16/16	+0.00 [0.00, 0.00]	-0.15 [-0.38, 0.10]
Early	Raw	17/17	+0.00 [0.00, 0.00]	-0.23 [-0.59, 0.11]
Before first reuse	Both	28/28	+0.67 [0.00, 2.01]	-0.18 [-0.41, 0.06]
Before first reuse	Cleaned	15/15	+0.00 [0.00, 0.00]	-0.37 [-0.69, -0.04]
Before first reuse	Raw	13/13	+1.44 [0.00, 4.33]	+0.03 [-0.23, 0.30]
8B Raw, seed 42 minus 8B Cleaned, seed 42				
Natural stop	Both	70/67	-31.87 [-36.27, -27.81]	-1.84 [-2.11, -1.58]
Natural stop	Cleaned	33/33	-31.44 [-36.36, -26.52]	-2.42 [-2.87, -2.02]
Natural stop	Raw	37/37	-32.26 [-39.36, -25.51]	-1.32 [-1.53, -1.11]
Early	Both	27/27	+0.00 [0.00, 0.00]	-1.51 [-2.03, -1.02]
Early	Cleaned	13/13	+0.00 [0.00, 0.00]	-1.99 [-2.82, -1.21]
Early	Raw	14/14	+0.00 [0.00, 0.00]	-1.06 [-1.52, -0.63]
Before first reuse	Both	32/32	-2.15 [-5.47, 0.00]	-2.68 [-3.15, -2.23]
Before first reuse	Cleaned	15/15	+0.00 [0.00, 0.00]	-3.04 [-3.67, -2.37]
Before first reuse	Raw	17/17	-4.04 [-10.29, 0.00]	-2.37 [-3.13, -1.80]

Table 24: Test-only post-filter close probability and next-event close frequency levels, in percent. Reference and treatment follow each contrast heading.

Anchor role	Close probability Reference	Close probability Treatment	Close frequency Reference	Close frequency Treatment
4B Raw, seed 42 minus 4B Cleaned, seed 42				
Natural stop	51.60	20.48	52.75	20.55
Early	0.00	0.00	0.00	0.00
Before first reuse	1.83	0.00	1.79	0.00
4B Raw, seed 123 minus 4B Cleaned, seed 123				
Natural stop	51.00	15.06	51.04	15.06
Early	0.00	0.00	0.00	0.00
Before first reuse	2.51	0.00	2.46	0.00
4B OBR 15%, seed 42 minus 4B Cleaned, seed 42				
Natural stop	51.60	43.86	52.75	43.37
Early	0.00	1.00	0.00	0.95
Before first reuse	1.83	1.17	1.79	0.89
4B OBR 15%, seed 123 minus 4B Cleaned, seed 123				
Natural stop	51.00	45.76	51.04	44.60
Early	0.00	0.88	0.00	0.95
Before first reuse	2.51	2.74	2.46	2.68
4B Cleaned, seed 123 minus 4B Cleaned, seed 42				
Natural stop	51.60	51.00	52.75	51.04
Early	0.00	0.00	0.00	0.00
Before first reuse	1.83	2.51	1.79	2.46
8B Raw, seed 42 minus 8B Cleaned, seed 42				
Natural stop	53.07	20.32	51.25	19.38
Early	0.00	0.00	0.00	0.00
Before first reuse	3.00	0.80	2.54	0.39

Table 25: All contrast strengths versus no intervention. Short readouts use 127/128 test anchors (4B/8B), 118 prompts per model, and 16 draws per anchor and condition.

α	Δ next-event close frequency (pp)	Δ first-token stop margin (nats)	Δ post-filter close probability (pp)
		4B Cleaned reverse	
0.5	-9.79 [-12.50, -7.24]	-1.07 [-1.16, -0.99]	-9.79 [-11.78, -7.80]
1	-15.50 [-19.19, -11.97]	-2.14 [-2.28, -2.00]	-16.42 [-19.67, -13.18]
2	-25.15 [-30.44, -19.89]	-4.27 [-4.47, -4.07]	-25.24 [-30.32, -19.97]
		4B Raw restore	
0.5	+4.38 [2.57, 6.40]	+1.11 [1.04, 1.18]	+4.91 [3.55, 6.31]
1	+11.17 [8.15, 14.53]	+2.25 [2.09, 2.41]	+10.87 [8.23, 13.59]
2	+24.95 [20.29, 29.65]	+4.17 [3.82, 4.54]	+25.14 [20.58, 29.73]
		4B OBR 15% transfer	
0.5	+6.50 [4.15, 9.06]	+0.94 [0.85, 1.03]	+8.36 [6.47, 10.22]
1	+14.67 [11.52, 17.99]	+1.84 [1.65, 2.04]	+15.03 [12.26, 17.68]
2	+25.49 [21.20, 30.17]	+3.28 [2.94, 3.66]	+26.76 [22.54, 30.95]
		8B Cleaned reverse	
0.5	-10.94 [-13.57, -8.41]	-1.25 [-1.39, -1.12]	-10.26 [-12.26, -8.38]
1	-17.77 [-21.47, -13.95]	-2.60 [-2.87, -2.34]	-17.72 [-20.79, -14.67]
2	-27.83 [-32.99, -22.61]	-5.74 [-6.23, -5.26]	-27.61 [-32.28, -22.78]
		8B Raw restore	
0.5	+7.67 [5.01, 10.50]	+1.41 [1.28, 1.55]	+7.11 [5.27, 9.06]
1	+15.43 [11.64, 19.26]	+2.68 [2.42, 2.95]	+14.88 [11.69, 18.12]
2	+32.71 [27.11, 38.41]	+4.70 [4.23, 5.18]	+32.35 [27.04, 37.66]

Table 26: 4B whole-pool execution and reachability levels. The pool contains 256 anchors, reachability and close probability are percentages, margins are nats, and forced-close EOS is a probability. EOS without close and unresolved short continuations are zero in every row.

Model, seed	Role	Anchors	Close reachability	Close probability	First-token stop margin	Forced-close EOS
Cleaned, 123	All anchors	256	48.83	27.99	-4.09	1.00000
Cleaned, 123	Natural stop	128	91.41	53.77	0.17	1.00000
Cleaned, 123	Early	64	1.56	0.45	-10.42	1.00000
Cleaned, 123	First reuse	64	10.94	3.97	-6.28	1.00000
Cleaned, 42	All anchors	256	51.95	28.05	-4.06	1.00000
Cleaned, 42	Natural stop	128	96.88	53.78	0.16	1.00000
Cleaned, 42	Early	64	1.56	0.40	-10.30	1.00000
Cleaned, 42	First reuse	64	12.50	4.26	-6.27	1.00000
Raw, 123	All anchors	256	21.48	8.78	-5.91	1.00000
Raw, 123	Natural stop	128	42.97	17.56	-1.82	1.00000
Raw, 123	Early	64	0.00	0.00	-11.49	0.99999
Raw, 123	First reuse	64	0.00	0.00	-8.48	1.00000
Raw, 42	All anchors	256	28.13	10.67	-5.98	1.00000
Raw, 42	Natural stop	128	56.25	21.34	-1.82	1.00000
Raw, 42	Early	64	0.00	0.00	-11.72	1.00000
Raw, 42	First reuse	64	0.00	0.00	-8.57	1.00000
OBR 15%, 123	All anchors	256	44.53	25.13	-4.17	1.00000
OBR 15%, 123	Natural stop	128	83.59	48.26	-0.14	1.00000
OBR 15%, 123	Early	64	1.56	0.45	-10.20	1.00000
OBR 15%, 123	First reuse	64	9.38	3.56	-6.22	1.00000
OBR 15%, 42	All anchors	256	44.14	24.03	-4.03	1.00000
OBR 15%, 42	Natural stop	128	82.81	46.27	-0.17	1.00000
OBR 15%, 42	Early	64	1.56	0.51	-9.73	1.00000
OBR 15%, 42	First reuse	64	9.38	3.07	-6.03	1.00000

Table 27: 8B whole-pool execution and reachability levels. The pool contains 256 anchors, reachability and close probability are percentages, margins are nats, and forced-close EOS is a probability. EOS without close and unresolved short continuations are zero in every row.

Model, seed	Role	Anchors	Close reachability	Close probability	First-token stop margin	Forced-close EOS
Cleaned, 42	All anchors	256	50.39	26.60	-4.46	1.00000
Cleaned, 42	Natural stop	128	98.44	52.24	0.07	1.00000
Cleaned, 42	Early	64	0.00	0.00	-11.68	1.00000
Cleaned, 42	First reuse	64	4.69	1.90	-6.31	1.00000
Raw, 42	All anchors	256	25.78	9.63	-6.45	1.00000
Raw, 42	Natural stop	128	50.78	19.06	-1.81	1.00000
Raw, 42	Early	64	0.00	0.00	-13.14	1.00000
Raw, 42	First reuse	64	1.56	0.40	-9.01	1.00000

Table 28: Selected direction scales and descriptive held-out projection AUC. Relative norm divides the direction norm by the mean of the two model residual norms.

Scale	Contrast norm	Cleaned residual norm	Raw residual norm	Relative norm (%)	AUC	Test anchors
4B	32.52	340.34	338.01	9.59	0.778	127
8B	60.93	677.13	666.69	9.07	0.739	129

Table 29: All eight equal-norm random directions at strength 1, relative to no intervention. Direction numbers are fixed within scale, and units and 95% intervals follow Table 25.

Direction	Δ next-event close frequency (pp)	Δ first-token stop margin (nats)	Δ post-filter close probability (pp)
	4B Cleaned reverse		
1	+0.39 [-1.46, 2.10]	+0.03 [0.00, 0.05]	+0.32 [-0.04, 0.80]
2	-2.71 [-4.75, -0.72]	-0.10 [-0.12, -0.07]	-1.66 [-2.41, -1.04]
3	+3.05 [0.93, 5.11]	+0.18 [0.16, 0.21]	+2.91 [2.13, 3.81]
4	+0.39 [-1.44, 2.19]	+0.05 [0.03, 0.08]	+0.58 [0.16, 1.08]
5	-0.05 [-2.12, 1.95]	-0.15 [-0.19, -0.11]	-0.01 [-0.65, 0.67]
6	+0.05 [-2.00, 2.13]	+0.06 [0.02, 0.09]	-0.47 [-1.08, 0.04]
7	+2.51 [0.67, 4.46]	+0.31 [0.27, 0.34]	+3.07 [2.22, 3.93]
8	-2.85 [-4.94, -0.77]	-0.20 [-0.22, -0.17]	-2.58 [-3.54, -1.75]
	4B Raw restore		
1	+0.05 [-1.72, 1.84]	+0.04 [0.01, 0.07]	-0.07 [-0.81, 0.64]
2	+0.05 [-1.72, 1.85]	+0.14 [0.12, 0.16]	+1.15 [0.64, 1.79]
3	-1.77 [-3.44, 0.00]	-0.09 [-0.12, -0.06]	-1.65 [-2.48, -0.91]
4	-1.23 [-2.79, 0.30]	-0.10 [-0.13, -0.08]	-0.64 [-1.20, -0.20]
5	-0.10 [-1.80, 1.62]	+0.22 [0.18, 0.26]	+0.58 [0.12, 1.17]
6	-0.05 [-1.64, 1.58]	-0.04 [-0.07, -0.01]	+0.71 [0.24, 1.29]
7	-0.98 [-2.55, 0.72]	-0.26 [-0.30, -0.23]	-0.78 [-1.27, -0.37]
8	+1.18 [-0.59, 3.25]	+0.21 [0.19, 0.24]	+1.52 [0.91, 2.28]
	4B OBR 15% transfer		
1	-0.44 [-2.03, 1.09]	-0.03 [-0.06, 0.00]	+0.22 [-0.26, 0.83]
2	-0.69 [-2.59, 1.20]	+0.07 [0.05, 0.10]	+1.40 [0.70, 2.21]
3	-2.76 [-4.62, -0.93]	-0.12 [-0.15, -0.09]	-1.94 [-2.72, -1.30]
4	-1.72 [-3.47, 0.00]	-0.13 [-0.15, -0.10]	-0.70 [-1.20, -0.31]
5	-0.54 [-2.70, 1.54]	+0.19 [0.15, 0.24]	+0.85 [0.22, 1.55]
6	+0.30 [-1.48, 2.07]	-0.02 [-0.05, 0.00]	+0.34 [-0.25, 0.88]
7	-3.79 [-5.95, -1.66]	-0.28 [-0.31, -0.25]	-2.40 [-3.44, -1.44]
8	+0.79 [-1.20, 3.05]	+0.22 [0.19, 0.24]	+2.46 [1.61, 3.36]
	8B Cleaned reverse		
1	-4.49 [-6.85, -2.08]	-0.37 [-0.41, -0.33]	-4.10 [-5.31, -3.00]
2	-2.88 [-5.08, -0.68]	-0.02 [-0.05, 0.01]	-0.97 [-1.76, -0.34]
3	-0.83 [-2.95, 1.29]	-0.16 [-0.19, -0.13]	-1.59 [-2.33, -0.95]
4	+0.44 [-1.74, 2.70]	+0.09 [0.06, 0.11]	+1.87 [1.16, 2.69]
5	+3.52 [0.99, 5.93]	+0.38 [0.33, 0.42]	+3.69 [2.75, 4.74]
6	+1.95 [-0.41, 4.43]	+0.17 [0.14, 0.20]	+1.47 [0.72, 2.21]
7	+0.78 [-1.49, 3.20]	+0.22 [0.18, 0.25]	+1.52 [0.90, 2.23]
8	+2.00 [-0.30, 4.35]	+0.24 [0.21, 0.26]	+3.30 [2.40, 4.34]
	8B Raw restore		
1	+3.03 [0.95, 5.29]	+0.52 [0.47, 0.57]	+3.15 [2.17, 4.21]
2	-0.44 [-2.41, 1.34]	+0.05 [0.02, 0.08]	+0.67 [0.29, 1.16]
3	+1.61 [0.00, 3.31]	+0.24 [0.20, 0.27]	+1.08 [0.54, 1.72]
4	-1.03 [-3.18, 0.91]	-0.10 [-0.13, -0.07]	-1.34 [-2.10, -0.71]
5	-1.32 [-2.96, 0.29]	-0.42 [-0.46, -0.37]	-2.13 [-3.00, -1.32]
6	-0.73 [-2.59, 1.03]	-0.16 [-0.19, -0.13]	-0.82 [-1.43, -0.33]
7	-0.98 [-2.80, 0.79]	-0.26 [-0.29, -0.22]	-1.39 [-2.11, -0.80]
8	-1.71 [-3.44, -0.18]	-0.18 [-0.21, -0.15]	-1.38 [-2.09, -0.79]

Table 30: Complete 4B cleaned continuation results. Each nonbaseline cell gives its level, paired difference in parentheses, and 95% interval for that difference, relative to no intervention. Rates and recall use percent and percentage points, while new reference matches, tokens, and blocks retain their count units.

Outcome	No inter	Reverse	Assigned equal-norm random direction	Close bias
All anchors, 64 anchors				
Tail repet.	5.86	5.08 (-0.78) [-4.17, 2.31]	5.86 (+0.00) [-2.82, 2.94]	4.69 (-1.17) [-3.91, 1.49]
Orbit in tail	0.39	0.00 (-0.39) [-1.23, 0.00]	0.00 (-0.39) [-1.23, 0.00]	0.00 (-0.39) [-1.23, 0.00]
Actual EOS	99.22	99.22 (+0.00) [-1.17, 1.17]	99.22 (+0.00) [-2.05, 1.64]	99.61 (+0.39) [0.00, 1.23]
Context-limit	0.78	0.78 (+0.00) [-1.17, 1.17]	0.78 (+0.00) [-1.64, 2.05]	0.39 (-0.39) [-1.23, 0.00]
Comp.	99.22	99.22 (+0.00) [-1.17, 1.17]	99.22 (+0.00) [-2.05, 1.64]	99.61 (+0.39) [0.00, 1.23]
New triples	1.06	0.88 (-0.18) [-0.51, 0.12]	1.15 (+0.09) [-0.09, 0.27]	1.35 (+0.29) [0.06, 0.56]
New pairs	4.23	4.39 (+0.16) [-0.32, 0.63]	4.77 (+0.54) [0.20, 0.91]	4.70 (+0.47) [0.06, 0.86]
Re recall	12.15	12.56 (+0.41) [-1.11, 1.92]	13.68 (+1.53) [0.35, 2.74]	13.54 (+1.39) [0.17, 2.61]
Mean tokens	1,876	1,879 (+3) [-299, 276]	1,928 (+52) [-337, 518]	1,886 (+10) [-243, 203]
Mean blocks	32.52	32.57 (+0.05) [-6.34, 5.56]	33.88 (+1.36) [-6.48, 11.08]	32.21 (-0.30) [-5.83, 3.39]
Natural stop, 16 anchors				
Tail repet.	0.00	3.13 (+3.13) [0.00, 9.38]	0.00 (+0.00) [0.00, 0.00]	1.56 (+1.56) [0.00, 4.69]
Orbit in tail	0.00	0.00 (+0.00) [0.00, 0.00]	0.00 (+0.00) [0.00, 0.00]	0.00 (+0.00) [0.00, 0.00]
Actual EOS	100.00	100.00 (+0.00) [0.00, 0.00]	100.00 (+0.00) [0.00, 0.00]	100.00 (+0.00) [0.00, 0.00]
Context-limit	0.00	0.00 (+0.00) [0.00, 0.00]	0.00 (+0.00) [0.00, 0.00]	0.00 (+0.00) [0.00, 0.00]
Comp.	100.00	100.00 (+0.00) [0.00, 0.00]	100.00 (+0.00) [0.00, 0.00]	100.00 (+0.00) [0.00, 0.00]
New triples	0.00	0.00 (+0.00) [0.00, 0.00]	0.00 (+0.00) [0.00, 0.00]	0.00 (+0.00) [0.00, 0.00]
New pairs	0.00	0.03 (+0.03) [0.00, 0.09]	0.02 (+0.02) [0.00, 0.05]	0.00 (+0.00) [0.00, 0.00]
Re recall	0.00	0.10 (+0.10) [0.00, 0.30]	0.05 (+0.05) [0.00, 0.15]	0.00 (+0.00) [0.00, 0.00]
Mean tokens	99	202 (+104) [34, 187]	107 (+8) [-19, 45]	144 (+45) [8, 88]
Mean blocks	1.42	3.09 (+1.67) [0.58, 2.94]	1.61 (+0.19) [-0.30, 0.88]	2.23 (+0.81) [0.16, 1.61]
Early, 32 anchors				
Tail repet.	6.25	7.03 (+0.78) [-4.69, 6.25]	9.38 (+3.13) [-1.56, 7.81]	6.25 (+0.00) [-5.47, 5.47]
Orbit in tail	0.00	0.00 (+0.00) [0.00, 0.00]	0.00 (+0.00) [0.00, 0.00]	0.00 (+0.00) [0.00, 0.00]
Actual EOS	100.00	99.22 (-0.78) [-2.34, 0.00]	98.44 (-1.56) [-4.69, 0.00]	100.00 (+0.00) [0.00, 0.00]
Context-limit	0.00	0.78 (+0.78) [0.00, 2.34]	1.56 (+1.56) [0.00, 4.69]	0.00 (+0.00) [0.00, 0.00]
Comp.	100.00	99.22 (-0.78) [-2.34, 0.00]	98.44 (-1.56) [-4.69, 0.00]	100.00 (+0.00) [0.00, 0.00]
New triples	1.76	1.54 (-0.22) [-0.85, 0.31]	1.99 (+0.23) [-0.10, 0.55]	2.23 (+0.48) [0.07, 0.93]
New pairs	7.27	7.91 (+0.64) [-0.17, 1.41]	8.30 (+1.03) [0.46, 1.63]	8.09 (+0.83) [0.13, 1.48]
Re recall	20.06	22.35 (+2.29) [-0.11, 4.59]	23.31 (+3.25) [1.51, 5.15]	22.85 (+2.79) [0.89, 4.64]
Mean tokens	2,526	2,776 (+250) [-86, 692]	2,999 (+473) [-10, 1,275]	2,668 (+142) [-78, 382]
Mean blocks	43.51	48.98 (+5.47) [-0.30, 13.52]	53.18 (+9.67) [-0.10, 26.36]	45.80 (+2.29) [-1.28, 6.25]
Before first reuse, 16 anchors				
Tail repet.	10.94	3.13 (-7.81) [-15.63, -1.56]	4.69 (-6.25) [-10.94, -1.56]	4.69 (-6.25) [-12.50, -1.56]
Orbit in tail	1.56	0.00 (-1.56) [-4.69, 0.00]	0.00 (-1.56) [-4.69, 0.00]	0.00 (-1.56) [-4.69, 0.00]
Actual EOS	96.88	98.44 (+1.56) [0.00, 4.69]	100.00 (+3.13) [0.00, 7.81]	98.44 (+1.56) [0.00, 4.69]
Context-limit	3.13	1.56 (-1.56) [-4.69, 0.00]	0.00 (-3.13) [-7.81, 0.00]	1.56 (-1.56) [-4.69, 0.00]
Comp.	96.88	98.44 (+1.56) [0.00, 4.69]	100.00 (+3.13) [0.00, 7.81]	98.44 (+1.56) [0.00, 4.69]
New triples	0.72	0.45 (-0.27) [-0.61, 0.06]	0.61 (-0.11) [-0.41, 0.13]	0.94 (+0.22) [-0.20, 0.64]
New pairs	2.38	1.70 (-0.67) [-1.45, 0.05]	2.47 (+0.09) [-0.59, 0.70]	2.59 (+0.22) [-0.53, 0.97]
Re recall	8.48	5.42 (-3.06) [-6.53, -0.23]	8.03 (-0.44) [-3.21, 1.80]	8.46 (-0.02) [-2.56, 2.55]
Mean tokens	2,354	1,762 (-592) [-1,451, -56]	1,606 (-747) [-1,649, -43]	2,065 (-289) [-1,118, 227]
Mean blocks	41.63	29.22 (-12.41) [-31.17, -1.06]	27.53 (-14.09) [-32.78, -1.20]	35.02 (-6.61) [-24.78, 4.52]

Table 31: Complete 4B raw continuation results. Each nonbaseline cell gives its level and, in parentheses, its paired difference relative to no intervention. Rates and recall are reported in percent, with differences in percentage points, while new reference matches, tokens, and blocks retain their count units.

Outcome	No inter	Restore	Assigned equal-norm random direction	Close bias
All anchors, 64 anchors				
Tail repetition	27.73	21.88 (-5.86)	25.78 (-1.95)	30.47 (+2.73)
Stable token orbit in tail	2.34	1.95 (-0.39)	1.56 (-0.78)	4.30 (+1.95)
Actual EOS	83.59	86.72 (+3.13)	85.55 (+1.95)	83.59 (+0.00)
Context-limit endings	16.41	13.28 (-3.13)	14.45 (-1.95)	16.41 (+0.00)
JSON list completeness	83.59	86.72 (+3.13)	85.55 (+1.95)	83.59 (+0.00)
New reference triples	1.30	1.35 (+0.05)	1.07 (-0.23)	1.11 (-0.20)
New reference pairs	5.03	5.15 (+0.12)	4.89 (-0.14)	4.78 (-0.25)
Remaining-pair recall	14.54	15.17 (+0.63)	14.60 (+0.06)	13.92 (-0.61)
Mean tail tokens	6,639	6,036 (-602)	6,425 (-214)	6,890 (+251)
Mean tail blocks	117.13	105.75 (-11.37)	112.23 (-4.90)	122.79 (+5.66)
Natural stop, 16 anchors				
Tail repetition	23.44	12.50 (-10.94)	15.63 (-7.81)	20.31 (-3.13)
Stable token orbit in tail	0.00	0.00 (+0.00)	0.00 (+0.00)	0.00 (+0.00)
Actual EOS	90.63	90.63 (+0.00)	92.19 (+1.56)	92.19 (+1.56)
Context-limit endings	9.38	9.38 (+0.00)	7.81 (-1.56)	7.81 (-1.56)
JSON list completeness	90.63	90.63 (+0.00)	92.19 (+1.56)	92.19 (+1.56)
New reference triples	0.16	0.05 (-0.11)	0.20 (+0.05)	0.20 (+0.05)
New reference pairs	0.39	0.30 (-0.09)	0.56 (+0.17)	0.61 (+0.22)
Remaining-pair recall	1.03	0.79 (-0.24)	1.70 (+0.67)	1.64 (+0.61)
Mean tail tokens	2,422	2,326 (-96)	2,085 (-336)	2,226 (-196)
Mean tail blocks	37.72	36.47 (-1.25)	30.83 (-6.89)	33.23 (-4.48)
Early, 32 anchors				
Tail repet.	29.69	24.22 (-5.47)	25.00 (-4.69)	32.03 (+2.34)
Orbit in tail	2.34	3.13 (+0.78)	0.78 (-1.56)	4.69 (+2.34)
Actual EOS	80.47	85.16 (+4.69)	84.38 (+3.91)	80.47 (+0.00)
Context-limit	19.53	14.84 (-4.69)	15.63 (-3.91)	19.53 (+0.00)
Comp.	80.47	85.16 (+4.69)	84.38 (+3.91)	80.47 (+0.00)
New triples	2.07	2.03 (-0.04)	1.70 (-0.38)	1.73 (-0.34)
New pairs	8.28	8.23 (-0.05)	8.01 (-0.27)	7.95 (-0.33)
Re recall	23.26	22.83 (-0.43)	22.94 (-0.33)	22.49 (-0.77)
Mean tokens	8,450	7,758 (-692)	7,809 (-641)	8,908 (+458)
Mean blocks	149.07	137.89 (-11.18)	137.78 (-11.29)	159.88 (+10.81)
Before first reuse, 16 anchors				
Tail repetition	28.13	26.56 (-1.56)	37.50 (+9.38)	37.50 (+9.38)
Stable token orbit in tail	4.69	1.56 (-3.13)	4.69 (+0.00)	7.81 (+3.13)
Actual EOS	82.81	85.94 (+3.13)	81.25 (-1.56)	81.25 (-1.56)
Context-limit endings	17.19	14.06 (-3.13)	18.75 (+1.56)	18.75 (+1.56)
JSON list completeness	82.81	85.94 (+3.13)	81.25 (-1.56)	81.25 (-1.56)
New reference triples	0.91	1.30 (+0.39)	0.67 (-0.23)	0.77 (-0.14)
New reference pairs	3.17	3.86 (+0.69)	2.98 (-0.19)	2.59 (-0.58)
Remaining-pair recall	10.59	14.23 (+3.63)	10.83 (+0.23)	9.08 (-1.51)
Mean tail tokens	7,233	6,303 (-930)	7,996 (+763)	7,517 (+284)
Mean tail blocks	132.64	110.77 (-21.88)	142.52 (+9.88)	138.14 (+5.50)

Table 32: Complete 4B OBR 15% continuation results. Each nonbaseline cell gives its level, paired difference in parentheses, and 95% interval for that difference, relative to no intervention. Rates and recall use percent and percentage points, while new reference matches, tokens, and blocks retain their count units.

Outcome	No inter	Transfer	Assigned equal-norm random direction	Close bias
All anchors, 64 anchors				
Tail repet.	7.03	8.20 (+1.17) [-2.46, 4.92]	5.86 (-1.17) [-4.37, 1.95]	7.81 (+0.78) [-3.73, 5.65]
Orbit in tail	0.39	0.78 (+0.39) [-0.81, 1.64]	0.39 (+0.00) [-1.15, 1.17]	0.78 (+0.39) [0.00, 1.19]
Actual EOS	98.83	99.22 (+0.39) [-1.21, 2.02]	98.05 (-0.78) [-2.94, 1.17]	98.05 (-0.78) [-2.38, 0.77]
Context-limit	1.17	0.78 (-0.39) [-2.02, 1.21]	1.95 (+0.78) [-1.17, 2.94]	1.95 (+0.78) [-0.77, 2.38]
Comp.	98.83	99.22 (+0.39) [-1.21, 2.02]	97.66 (-1.17) [-3.46, 0.85]	98.05 (-0.78) [-2.38, 0.77]
New triples	1.19	1.09 (-0.10) [-0.33, 0.14]	1.03 (-0.16) [-0.35, 0.02]	0.96 (-0.23) [-0.45, -0.02]
New pairs	4.63	4.24 (-0.39) [-0.88, 0.08]	4.43 (-0.20) [-0.62, 0.23]	4.09 (-0.54) [-0.96, -0.12]
Re recall	13.37	12.30 (-1.07) [-2.35, 0.25]	12.75 (-0.62) [-1.59, 0.30]	12.14 (-1.22) [-2.55, 0.17]
Mean tokens	2,258	2,174 (-84) [-614, 448]	2,335 (+77) [-411, 607]	2,219 (-39) [-462, 422]
Mean blocks	38.71	38.18 (-0.53) [-9.67, 8.96]	41.80 (+3.09) [-5.64, 13.52]	38.63 (-0.08) [-6.96, 7.63]
Natural stop, 16 anchors				
Tail repet.	3.13	1.56 (-1.56) [-6.25, 3.13]	1.56 (-1.56) [-6.25, 3.13]	0.00 (-3.13) [-7.81, 0.00]
Orbit in tail	0.00	0.00 (+0.00) [0.00, 0.00]	0.00 (+0.00) [0.00, 0.00]	0.00 (+0.00) [0.00, 0.00]
Actual EOS	100.00	100.00 (+0.00) [0.00, 0.00]	100.00 (+0.00) [0.00, 0.00]	100.00 (+0.00) [0.00, 0.00]
Context-limit	0.00	0.00 (+0.00) [0.00, 0.00]	0.00 (+0.00) [0.00, 0.00]	0.00 (+0.00) [0.00, 0.00]
Comp.	100.00	100.00 (+0.00) [0.00, 0.00]	100.00 (+0.00) [0.00, 0.00]	100.00 (+0.00) [0.00, 0.00]
New triples	0.00	0.00 (+0.00) [0.00, 0.00]	0.00 (+0.00) [0.00, 0.00]	0.00 (+0.00) [0.00, 0.00]
New pairs	0.00	0.00 (+0.00) [0.00, 0.00]	0.00 (+0.00) [0.00, 0.00]	0.02 (+0.02) [0.00, 0.05]
Re recall	0.00	0.00 (+0.00) [0.00, 0.00]	0.00 (+0.00) [0.00, 0.00]	0.04 (+0.04) [0.00, 0.11]
Mean tokens	222	185 (-37) [-124, 51]	231 (+8) [-81, 126]	66 (-156) [-298, -52]
Mean blocks	3.44	2.59 (-0.84) [-2.36, 0.64]	3.81 (+0.38) [-1.23, 2.61]	1.08 (-2.36) [-4.27, -0.86]
Early, 32 anchors				
Tail repet.	7.81	9.38 (+1.56) [-3.13, 6.25]	8.59 (+0.78) [-3.91, 6.25]	13.28 (+5.47) [-2.34, 14.06]
Orbit in tail	0.78	0.78 (+0.00) [-2.34, 2.34]	0.00 (-0.78) [-2.34, 0.00]	1.56 (+0.78) [0.00, 2.34]
Actual EOS	98.44	99.22 (+0.78) [-1.56, 3.13]	96.88 (-1.56) [-5.47, 1.56]	96.09 (-2.34) [-5.47, 0.00]
Context-limit	1.56	0.78 (-0.78) [-3.13, 1.56]	3.13 (+1.56) [-1.56, 5.47]	3.91 (+2.34) [0.00, 5.47]
Comp.	98.44	99.22 (+0.78) [-1.56, 3.13]	96.09 (-2.34) [-7.03, 1.56]	96.09 (-2.34) [-5.47, 0.00]
New triples	1.98	1.97 (-0.01) [-0.43, 0.42]	1.78 (-0.20) [-0.57, 0.17]	1.68 (-0.30) [-0.65, 0.08]
New pairs	8.05	7.46 (-0.59) [-1.45, 0.27]	7.61 (-0.44) [-1.17, 0.45]	7.09 (-0.96) [-1.77, -0.16]
Re recall	22.84	20.97 (-1.87) [-4.22, 0.33]	21.29 (-1.55) [-3.28, 0.33]	20.08 (-2.76) [-5.08, -0.44]
Mean tokens	3,322	3,288 (-34) [-891, 806]	3,484 (+162) [-634, 1,105]	3,579 (+257) [-452, 1,068]
Mean blocks	57.52	58.35 (+0.84) [-14.81, 16.74]	62.27 (+4.76) [-9.65, 23.99]	62.60 (+5.09) [-6.34, 18.77]
Before first reuse, 16 anchors				
Tail repet.	9.38	12.50 (+3.13) [-6.25, 14.06]	4.69 (-4.69) [-9.38, 0.00]	4.69 (-4.69) [-10.94, 1.56]
Orbit in tail	0.00	1.56 (+1.56) [0.00, 4.69]	1.56 (+1.56) [0.00, 4.69]	0.00 (+0.00) [0.00, 0.00]
Actual EOS	98.44	98.44 (+0.00) [-4.69, 4.69]	98.44 (+0.00) [-4.69, 4.69]	100.00 (+1.56) [0.00, 4.69]
Context-limit	1.56	1.56 (+0.00) [-4.69, 4.69]	1.56 (+0.00) [-4.69, 4.69]	0.00 (-1.56) [-4.69, 0.00]
Comp.	98.44	98.44 (+0.00) [-4.69, 4.69]	98.44 (+0.00) [-4.69, 4.69]	100.00 (+1.56) [0.00, 4.69]
New triples	0.81	0.42 (-0.39) [-0.66, -0.16]	0.56 (-0.25) [-0.52, -0.05]	0.47 (-0.34) [-0.70, -0.05]
New pairs	2.42	2.03 (-0.39) [-1.13, 0.22]	2.48 (+0.06) [-0.28, 0.38]	2.17 (-0.25) [-0.80, 0.30]
Re recall	7.78	7.25 (-0.53) [-2.63, 1.70]	8.41 (+0.63) [-0.15, 1.54]	8.38 (+0.60) [-1.68, 3.38]
Mean tokens	2,164	1,934 (-230) [-1,279, 942]	2,140 (-24) [-1,132, 1,111]	1,651 (-513) [-1,389, 158]
Mean blocks	36.36	33.42 (-2.94) [-18.91, 15.78]	38.84 (+2.48) [-15.86, 24.56]	28.23 (-8.13) [-21.99, 2.89]