A Family of Control Decisions Bypasses the J-Space

Anonymous authors
Paper under double-blind review

Abstract

Reasoning language models carry reportable content in a small, causally privileged subspace of the residual stream, a verbalizable "global workspace" built from that content (Gurnee et al., 2026). We ask whether this content-derived workspace also carries the model's control decisions: choices like when to stop, call a tool, or refuse. Clamping the workspace to its clean value removes most of the causal effect of a content perturbation that would otherwise pass through it, but leaves a control decision's effect essentially untouched. This holds under the actual sparse-gradient-pursuit decomposition, not just a linear approximation, and generalizes across model architectures and across a family of control tokens including refusal. Content is broadcast through the workspace; control decisions are computed outside it and surface only through their own readout. A model can therefore report a control decision faithfully without ever exposing what actually caused it — a candidate explanation for why chain-of-thought monitoring can miss a decision's real driver.

1 Introduction

Recent work identifies a verbalizable global workspace in language models: a small subspace of the residual stream, under about 10% of activation variance, that carries the model's globally broadcast, human-reportable state (Gurnee et al., 2026). Concretely, suppose a model is reasoning about a country's capital. Gurnee et al. (2026) show that the identity of that country is written into this small subspace. Overwriting just those few coordinates is enough to change what the model reports the country to be; overwriting an equal-sized set of unrelated coordinates is not. The workspace also carries most of a concept's causal effect, even though it holds only a small fraction of the activation's overall energy. The property we build on is a clamp: we hold the workspace's coordinates fixed at their clean values while perturbing the residual stream elsewhere. On the country example, this makes the perturbation's effect disappear, and the same holds for similar reportable-content swaps.

That characterization is about content, the things a model reasons about. Control decisions — when to stop generating, end a thinking block, call a tool, or refuse — each have their own literature, but no prior work asks whether one structure governs more than one of them in the same model; we ask whether the content workspace is that structure. Control decisions behave oppositely to content: a linear readout can predict the upcoming control token from the workspace, but that does not mean the model computes the decision there. Running the same clamp on control gives the mirror image of the content result — the clamp that erases a content effect leaves the control effect essentially intact (Section 4.2). Decomposing the drive directly shows why (Section 4.1): it sits almost entirely outside the workspace, reaching the logit through the control token's own readout direction instead of the workspace's shared channel.

Our contributions:

- Using Gurnee et al.'s (2026) own privilege-test clamp, we show control bypasses the content workspace while content routes through it (Section 4.1).

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.

- We replicate the control result on six further models across three further architectures: five bypass or grade toward bypass, and one (OLMo-3-Think) remains open, scoped precisely (Section 4.2).
- The pattern extends to a full family of control tokens and to refusal (Section 4.3).
- Post-training installs the bypass at or before supervised fine-tuning and keeps it flat across DPO and RLVR, rather than acquiring it gradually (Section 4.4).
- We interpret the result and its consequence for chain-of-thought monitoring, with explicit scope (Sections 5-6).

2 Related work

Detect-versus-steer dissociations. Galeone et al. (2026) show that the direction which best predicts a single decision is nearly unrelated to the direction that causes it, a result stable across four models; Fadnavis et al. (2026) and Darade & Thorat (2026) find the same gap in unrelated settings (temporal reasoning; a general formulation) — not an isolated result. A probe's accuracy therefore says nothing about whether it has found the mechanism, exactly the gap a monitor relying on probes would fall into. We extend this line of work in two ways: instead of one decision we test a family of four, and instead of measuring the angle between two directions we use a causal mediation test that holds the candidate mechanism fixed and asks whether the behavior survives (Section 4.1). Two further checks rule out that this is a previously reported dissociation under another name: Park et al. (2023)'s whitened-activation standard for when read and write directions coincide does not close the gap here either (Section 4.1), and a full comparison against Jiang et al. (2026)'s closely related "value axis," measured on the same model, finds no meaningful overlap.

The termination-circuit debate. Dutta[1] identifies the circuit in Qwen3 that computes the decision to stop generating and argues it is high-dimensional and therefore not steerable by any single direction: his best single direction recovers only about a fifth of the full effect — the strongest existing objection to a low-rank account like ours. Built in the basis our workspace analysis motivates instead of his, a rank-1 version of our stop drive recovers nearly all of the full steering effect on every model we test (Appendix A.1). Other measures of dimensionality still vary by model family, and a necessity check collapses stopping on most but not all models tested, so the drive is not uniformly rank-1 by every measure; Dutta's ceiling reflects the basis it was measured in, not a hard limit on compressibility.

Tool-calling and agentic actions. Wu et al. (2026b) show that the decision to start a tool call is linearly readable and steerable by a single direction, across three model families (Gemma-2, Qwen3, Llama-3.1). Tool-calling is one member of our control-decision family (Section 4.3); we ask the question their result leaves open — whether that steerable direction routes through the content workspace, or bypasses it the way stopping does. Vicentino's (2026) WANDERING work is the closest adjacent study, causally gating agentic-action decisions via donor-patching and steering; it differs from ours in instrument and scope, and does not trace the decision's origin across post-training the way Section 4.4 does.

Control decisions studied in isolation. The gap motivating this paper (Section 1) is visible in how fragmented this literature is: refusal alone has a concept-cone geometry (Wollschläger et al., 2025), a knowledge-versus-safety mechanistic split (Son et al., 2026), and a training-invariant-rank argument (Labunets, 2026); termination has its own literature (discussed above); tool-call necessity is studied by probing alone (Sun et al., 2026); and context-versus-memory reliance by causal interchange (Shih et al., 2026). Two recent papers substitute a taxonomy of decision types for a mechanism (Mazaheri & Mazaheri, 2026; Kwon, 2026). A concurrent study (Reblitz-Richardson, 2026a) is the one partial exception, reaching a compatible conclusion for refusal specifically — it routes through a channel separate from a moral-judgment subspace, on overlapping architectures — via a different instrument (rank-sweep interchange) and testing only refusal. We apply one identical test across the

[1] S. Dutta, "The Termination Circuit," LessWrong, 2026, `https://www.lesswrong.com/posts/ajhzc6ktEKyFeJFBS/the-termination-circuit-how-reasoning-models-stop-thinking` (not peer-reviewed; cited here as the most direct existing objection to our claim, not as an academic reference).

family instead (stop, `</think>`, tool-calls, refusal) and find they share a geometric and causal signature.

Post-training's mechanistic role. Wang et al. (2026) argue reinforcement learning from verified rewards (RLVR) confers genuinely new capability beyond what the base model can already do; Cheng et al. (2026) find instead that RLVR reinforces and suppresses circuitry the base model already has, adding nothing new. Our post-training-origin result (Section 4.4) speaks to which account holds for a control decision, as opposed to a content-facing skill, and finds evidence for the latter, more conservative account.

3 Method

Estimating the workspace. Gurnee et al.'s (2026) original study is on Claude 4.5–4.6; an independent replication of the Jacobian-lens method on Qwen 3.6 27B, extrapolated here to the differently-sized Qwen3-4B,[2] supports its use on the Qwen family, and a port of the same instrument to an audio-language model (Fan et al., 2026) suggests the workspace itself generalizes beyond Claude and text. Following Gurnee et al. (2026), we read each model's residual stream through a Jacobian lens fit through layer 34, whose rows are directions tied to individual vocabulary tokens; the top 2048 of these directions form an overcomplete dictionary. For Qwen3-4B, the workspace at a given position is the sparse nonnegative reconstruction of that position's activation from this dictionary, computed by gradient pursuit with an active set of k directions (k=25 for Qwen3-4B, d=2560) within layers 12–30; the complement, written $v_\perp$ when we mean a specific drive's component there, is everything the reconstruction leaves out (k/d denote the workspace's rank and the stream's ambient dimension throughout). The workspace is resolved independently at every position it is used at — the stop position's workspace is never the content position's basis reused, which would make any clamp built from it circular. Gradient pursuit is too expensive to run on every trace of every experiment, so most experiments instead use a fixed top-k linear span of the same band as a tractable approximation; we confirm the headline numbers hold in the actual gradient-pursuit decomposition directly (Section 4.1) and state which of the two objects each other result uses. Band and k are resolved per model for the cross-model runs (Section 4.2); the values above are Qwen3-4B's.

Building control drives. A drive is a single direction meant to isolate one specific thing the model represents, a control decision or a content concept, built so that adding or removing it shifts that one thing and as little else as possible: a difference in means between two matched sets of activations that differ only in the thing we want to isolate. We pick this diff-of-means recipe over an SAE-feature-based alternative because Wu et al. (2025) and Karvonen et al. (2025) independently find diff-of-means and other simple supervised directions matching or beating SAE features on steering and concept-detection tasks. We harvest N=64 GSM8K traces per model (fewer for some checks, noted where used). For a control decision with anchor token a, the drive v_a is this difference-in-means between the decision position and a matched control position from the same trace: for stopping, the decision position is where the model emits the stop token, and the matched control position is the midpoint of that trace's own thinking block, never the boundary tokens themselves. Because both positions come from the same reasoning episode on the same problem, the contrast isolates what changes specifically at the moment of deciding to stop, rather than the problem's content or the generic fact that reasoning is underway (both present at either position, so they cancel). Refusal, with no single anchor token or in-trace midpoint to match against, uses the recipe differently: v_{refuse} contrasts the last-prompt-token residual across two prompt sets, harmful against harmless, rather than two positions within one trace.

Building the content drive. This is our positive control, ruling out "the clamp is just weak": Gurnee et al.'s (2026) own content axis, a diff-of-means direction built from a battery of three country-pair prompts (e.g. a sentence naming France substituted with China), 12 templated prompts total (4 templates × 3 pairs), averaged across all three — deliberately not built from GSM8K, whose word problems contain no swappable named-entity concept

[2] N. Nanda, "A Review of Anthropic's Global Workspace Paper," LessWrong, 2026 (not peer-reviewed; a model-to-model extrapolation, not a same-model validation).

of this kind. Every control-decision measurement (stop, `</think>`, tool calls) uses GSM8K instead; refusal uses neither, built from the AdvBench harmful/harmless prompt sets (Zou et al., 2023) (Section 4.3).

Steering and clamping. We rely on two standard causal techniques, not new ones. Steering is representation engineering: add a drive to the residual stream at a fixed dose and read off the resulting change in the target logit, the same recipe Arditi et al. (2024) and Wu et al. (2026b) use. We dose every drive at a fixed equal-norm magnitude (α=16 for every model and arm, unless noted as natural magnitude). Clamping is causal mediation analysis (Vig et al., 2020) applied to one subspace: to clamp the workspace is to hold its coordinates at their clean values while the perturbation is applied elsewhere, $h \leftarrow h - \Pi_{W(h)}(h) + \Pi_{W(h)}(h_{\text{clean}})$, where $W(h)$ is the workspace basis resolved at this position's own activation, per the previous paragraph, not a basis carried over from another position. If the perturbation's effect survives this clamp, the workspace was not the reason it worked; if it collapses, the workspace was necessary. Ablating the workspace is related but distinct: it zeros the workspace's projection outright rather than holding it fixed while we perturb elsewhere. Gurnee et al.'s (2026) privilege test is exactly this clamp applied to a content perturbation. Our central measurement runs the identical clamp on control, using the content-workspace basis throughout, which excludes any control token's own readout row, so pinning the coordinate a control logit reads from does not explain the result.

Measuring effect and mediation. For a rank-k workspace in a d-dimensional stream, almost any direction has nearly all its energy outside the workspace, so energy fraction is uninformative. We rest every claim on causal quantities instead: steering asymmetry under matched-norm controls, and the clamp-mediation fraction $m = 1 - (\text{clamped effect})/(\text{unclamped effect})$, always with a matched control and, for mediation, a content positive control on the same clamp. "Effect" means the change in the target token's logit relative to the clean run, not probability, so it stays interpretable even at a clean probability of 0 or 1; we measure it teacher-forced at the decision position, averaged over the prompt set, with 95% paired-bootstrap intervals (5,000 resamples). The target token is the anchor a itself (the stop token, `</think>`, or the relevant tool-call delimiter); refusal has no single anchor logit, so we report a refusal classifier's induced-refusal rate instead of a mediation fraction. A cell is a confident bypass at $|m| \leq 0.20$; a large negative m is flagged as open, not folded into either verdict (Section 4.2).

Defaults. Unless a result says otherwise, every number in this paper uses Qwen3-4B, GSM8K, N=64, the workspace band and k from Estimating the workspace, and an equal-norm steering dose of α=16. Section 4.2 reports the identical measurement on six further models across three further architectures, each with its own resolved band and k (Appendix A.2, Table 5); later sections flag any other deviation, such as a natural-magnitude dose, at the point it occurs rather than repeating the full setup.

Robustness checks. Cherry-picking cannot explain why the clamped effect survives: across a 21-test battery spanning the full control-token family (stop, `</think>`, tool-call onset/end, refusal), Bonferroni-corrected at family-wise $\alpha = 0.05$, every clamped effect stays significantly nonzero (two cells survive on significance alone rather than effect size; full breakdown in Appendix A.3). This battery tests only that the clamped effect differs from zero, which both the bypass and the routing hypotheses predict, so it rules out noise and cherry-picking but is not itself the test for the bypass claim; that burden is carried by the mediation fraction and its own bootstrap interval (Section 4.2's Table 1). Multi-seed reruns of the random-direction control reproduce the same pattern every time (Appendix A.3). A separate null distribution, built from the raw effect the clamp removes rather than the unstable mediation-fraction ratio (Section 7), confirms the bypass directly: at N=128 traces per draw, the real drop is significantly smaller than chance clamp damage ($p = 0.02$) — positive evidence for the bypass, not just a failure to rule it out.

4 Results

4.1 Control drives lie outside the workspace

The stop drive's causal effect lives almost entirely outside the workspace, in a roughly 9-to-1 split against the small part that falls inside it. We decompose the stop drive v_{done} in the actual gradient-pursuit workspace (N=64 GSM8K clean-stop traces) and steer each component at its natural magnitude. Steering the outside-workspace component moves the stop logit by +28.7 [27.6, 29.7]; steering the in-workspace component moves it by only +3.1 [2.7, 3.5], a ratio of about 9.4 to 1 (a matched-norm random direction moves the logit near zero). Only 4.7% of the drive's energy lies inside the workspace — against a chance level of $k/d \approx 1.0\%$ (k=25, d=2560) for a generic direction, so this is a roughly 4.8× enrichment over chance, not evidence the drive is inside the workspace. A sanity check confirms the pipeline assigns full in-workspace energy to a direction drawn directly from the dictionary, ruling out a trivial bug. A fixed top-k linear approximation overstates the split (25-to-1 at k=25, 68-to-1 at k=10); the real gradient-pursuit workspace gives the smallest, most trustworthy ratio (9.4-to-1, a smaller N=32 pilot agreeing at 8.5-to-1). The drive sits almost entirely outside the very subspace that partly reads it out.

4.2 Controls are not privileged

The workspace mediates a content drive but not a control drive, under an identical clamp (Table 1). We run Gurnee et al.'s (2026) privilege-test clamp, unchanged, on two drives, holding the content workspace at its clean value through the workspace band and into the late layers. For a genuine content drive — a diff-of-means concept axis built from an averaged battery of three country-pair swaps, following the same design, its own unit therefore 12 templated prompts rather than a per-trace N — clamping the workspace removes most of the effect. For the stop drive, measured on N=516 traces, the identical clamp removes almost nothing.

Drive (same clamp)	Unclamped effect	Clamped effect	Mediation
Content (3-pair battery)	+28.6 [25.1, 31.9]	+10.1 [7.7, 12.3]	0.65
Control (stop)	+28.7 [28.4, 28.9]	+29.5 [29.3, 29.8]	−0.03 [−0.04, −0.02]

Table 1: Clamp-mediation of a content vs. a control drive under an identical content-workspace clamp (Qwen3-4B, N=516 control / 12 prompts × 3 pairs content; 95% paired-bootstrap intervals on the effects, and, for the control row, a paired bootstrap directly on m itself rather than on the two effects separately).

The content drive's clamped and unclamped intervals do not overlap, at 82% of Gurnee et al.'s (2026) own benchmark mediation (0.65/0.79) — our positive control is directionally right but somewhat weaker, which loosens, without erasing, the bound on what a weaker clamp could be hiding — while the control drive's clamped and unclamped intervals overlap almost completely and its mediation sits at zero within noise, the opposite pattern. Since the same clamp can block a drive, this validates it: the control result is a genuine bypass, not a weak clamp.

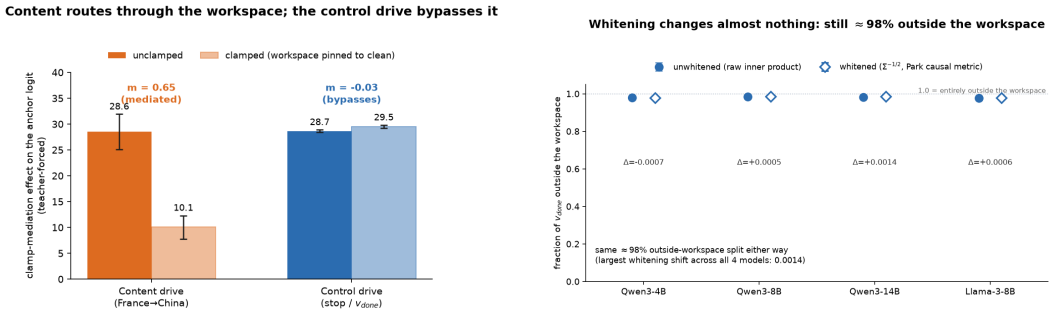


Figure 1: *Left:* the privilege test on content vs. control — the workspace mediates a content drive ($m = 0.65$) while the stop drive bypasses it ($m = -0.03$, N=516). *Right:* whitening (Park et al., 2023), a stricter causal standard, changes the outside-workspace split by at most 0.0014 across four models — the same split, not a whitening artifact.

Robustness to the whitening metric. The dissociation is not an artifact of the raw inner product. Park et al. (2023) show that a decodable direction and a causally load-bearing one coincide only after whitening the activations, a stricter geometric standard than the raw inner product we use elsewhere. Re-measuring under this standard changes the energy split from Section 4.1 by at most 0.0014 across the actual decomposition, the linear approximation, and a second model, Qwen3-8B (Figure 1, right); reimplementing the clamp itself under the same whitened metric reproduces the identical bypass on all three models checked and on the other two control-token family members on Qwen3-4B (tool-calls, refusal), closing the geometry-convention gap for the full family rather than stop alone (pilot scale; one weaker cross-family cell flagged by Section 4.2's own dosing caveat; full per-model numbers in Appendix A.4). Throughout, we treat the workspace as the sole candidate mediator, not one of several acting together.

Cross-model replication. The bypass is not specific to one model. Across six further models spanning three further architectures, five produce a confirmed reading (Table 2): four confirm it under both a fixed equal-norm dose and a natural-magnitude dose scaled to each model's activation size, and GPT-OSS-20B confirms it once its clamp is widened enough to validate. OLMo-3-Think's positive control never validates, so we report no reading for it (Appendix A.6).

Model	N	m (equal-norm)	m (natural-mag)	Verdict
Qwen3-8B	518	0.053	0.034	Bypasses
Qwen3-14B	516	0.134	0.124	Bypasses
Qwen3-30B-A3B (MoE)	517	−0.068	0.065	Bypasses
Llama-3-8B[§]	416	−0.494	0.078	Bypasses
GPT-OSS-20B[†]	63	0.036	0.39	Toward bypass

Table 2: Control-drive clamp-mediation across the five further models with a confirmed reading (per-model band/k in Appendix A.2); a confident bypass is $|m| \leq 0.20$ (Method). [†]Natural-mag column: clamp widened to k=25, the only setting where GPT-OSS-20B's own positive control validates. [§]Llama-3-8B's equal-norm m is the dosing artifact below; its natural-mag reading drives the Verdict. OLMo-3-Think was also tested but is not shown: its positive control never validates (Appendix A.6).

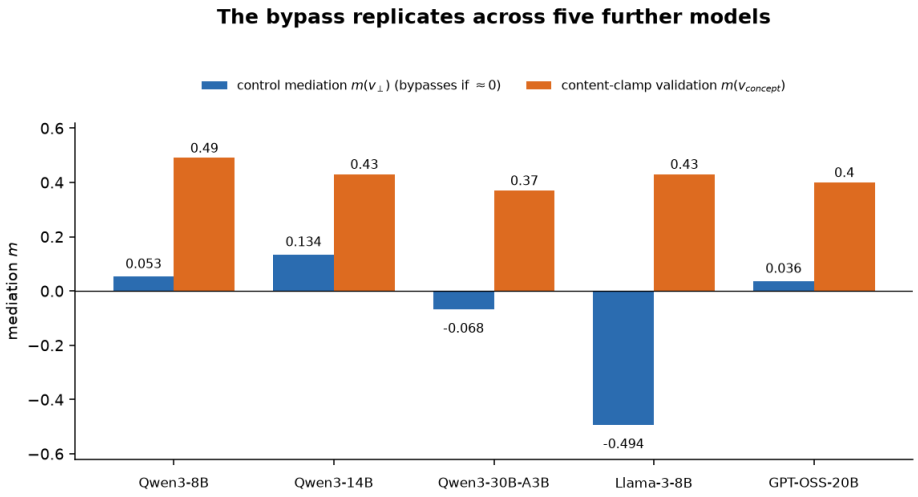


Figure 2: Cross-model replication, equal-norm dosing, on the five models with a confirmed reading (Table 2); all five bypass or grade toward bypass. OLMo-3-Think is not shown: its positive control never validates (Appendix A.6).

Dosing convention. A fixed equal-norm dose does not transfer across architectures: Llama-3-8B's and GPT-OSS-20B's unclamped control effects sat at noise floor until rescaled to each model's own natural activation magnitude (full unclamped effect sizes and the GPT-OSS-20B clamp-widening fix in Appendix A.5, Figure 5), while the Qwen3-4B headline is stable to this choice (both conventions agree at full N; Appendix A.5). Separately, a control decision also reaches its logit through its own readout direction, the late-layer "motor" regime of Gurnee et al. (2026). Including that row in the clamp basis flips the verdict to routed (m=0.80, N=516, the same Anthropic-Fig-16 shape the content drive shows) — but this is circular,

since pinning the coordinate a control logit reads from will suppress the effect whether or not a workspace mediates the decision, so we rely on the non-circular content-workspace clamp above, not this row, as our load-bearing result.

4.3 The bypass generalizes across control tokens and refusal

Stop and end-of-thinking share the same signature: on GSM8K both `</think>` and `<|im_end|>` show the outside-workspace drive, replicating across Qwen3-4B/8B/14B/30B (Figure 3). The `</think>` mediation is $m = -0.112$ at N=516 on Qwen3-4B, $m = -0.065$ at N=518 on Qwen3-8B (promoted this wave from an N=64 pilot that agreed in sign and rough magnitude, $m = -0.059$), $m = -0.082$ at N=109 on Qwen3-14B (already the full yield of its 520-trace population — not every completion reaches `</think>` within the harvested generation length), and $m = 0.022$ at N=517 on Qwen3-30B-A3B (also promoted this wave from an N=64 pilot, $m = 0.001$) — all clear bypasses. Three of the four cells (Qwen3-4B, Qwen3-8B, Qwen3-30B-A3B) are now at the same N=500+ scale as the stop result in Table 1; only Qwen3-14B is capped below that scale by its own population's yield.

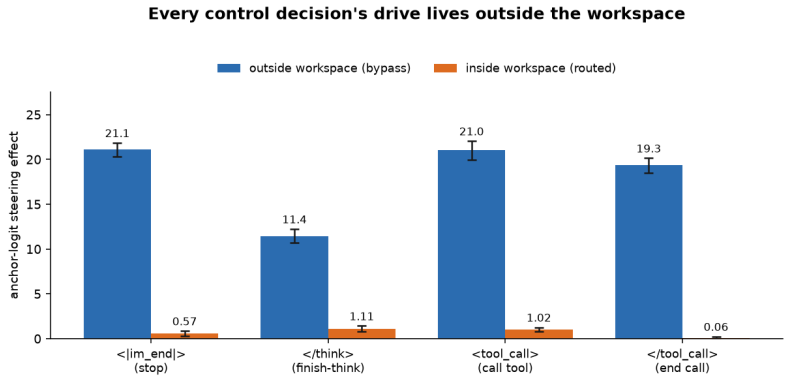


Figure 3: Across the control-token family (Qwen3-4B), each decision's causal drive lies outside the workspace (teacher-forced steering effect at natural magnitude).

Tool-calls and refusal show the same bypass pattern as stop and `</think>` (Table 3). A matched-norm random-direction control confirms this is not generic (stays at the clean rate on Qwen3-4B's refusal test), and refusal is near-orthogonal to stopping in direction ($|\cos| \approx 0.04$), so it is not simply the same drive under another name. Llama-3-8B's induced refusal effect required a recalibrated dose (α=3 instead of the fixed α=16, the same equal-norm artifact as Section 4.2; Appendix A.5); the recalibrated full-N run confirms the same bypass ($m = -0.015$), the first non-Qwen confirming replication for this family member. A content positive control specific to the function-calling distribution only weakly validates the clamp at this sample size (m=0.24, N=12, not cleanly separated from its own random-direction control at 0.33; full detail in Appendix A.7); the bypass claim for tool-calls therefore rests on the family-wide clamp-mediation numbers in Table 3 instead. Overall, four independent control decisions carry this signature under one shared instrument, though at unequal depth: full replication for stop, `</think>`, and refusal; partial for tool-calls (Section 7).

Position and length as an alternative account. Three of the four family members (stop, `</think>`, `<tool_call>`/`</tool_call>`) are segment-final tokens, so a purely positional or length-based feature ("few tokens remain") could in principle produce the same bypass signature without encoding a control decision as such. Refusal is the wedge against this account: we build and steer v_{refuse} at the last prompt token, before any completion exists, so no segment boundary or remaining-length quantity exists yet for a position feature to track. Refusal showing the identical outside-workspace, non-mediated signature (Qwen3-4B/8B/14B) argues against a pure position/length account of the family as a whole, though not conclusively (Section 7).

A direct positional test on Qwen3-4B weighs against the pure position/length account. Fixing the clamp dose and varying only the distance to the natural stop, injecting 2–20 tokens before it (Arm A), against position-matched traces that never reach a natural stop (Arm B):

Metric	Qwen3-4B	Qwen3-8B	Qwen3-14B	Llama-3-8B	Qwen3-30B-A3B
`<tool_call>` $m^{\dagger}$	-0.034	0.008	0.058	—	—
`</tool_call>` $m^{\dagger}$	-0.011	0.003	0.040	—	—
Refusal induction Δ (%)	97.3	98.8	96.7	78.7	32.9
Refusal m	-0.004	0.004	0.080	-0.015	$-0.205^{\ddagger}$

Table 3: Mediation for tool-calls (a function-calling distribution) and refusal (the AdvBench harmful-behaviors battery (Zou et al., 2023), the same set Arditi et al.'s (2024) refusal-direction paper uses). Every entry's sign argues against workspace mediation; a confident bypass is $|m| \leq 0.20$ (Method). Refusal induction Δ is the percentage-point rise in refusal rate from steering, clean to steered; clamping never reduces it (Refusal m stays at or below zero throughout — clamping the content workspace does not block the induced refusal on any model tested). Tool-call and refusal prompt counts, and per-model dose recalibration details (Llama-3-8B, Qwen3-30B-A3B), are in Appendix A.5. †Tool-call's own content positive control is weak/inconclusive at this sample size (Appendix A.7) — the bypass claim for tool-calls rests on these clamp-mediation numbers themselves, not on that control. ‡Qwen3-30B-A3B's refusal cell sits just outside the confident-bypass band with no diagnosed cause (dose already recalibrated, Appendix A.5); flagged as a minor open point, not a bypass.

traces genuinely about to stop read differently from position-matched traces that are not, once the dose is lowered below saturation, producing a decisive, monotonically shrinking gap between the two arms (full per-dose numbers in Appendix A.8). At the headline dose the two arms are indistinguishable ($\approx +28$ to $+31$ throughout), but that comparison is uninformative by construction, since the readout may simply be saturating at that dose (one model, one control token; Section 7). This is also a direct instance of a calibration practice a concurrent methods paper argues interpretability instruments need generally (Reblitz-Richardson, 2026b).

Extending the same offset-sweep design to `</think>` and `<tool_call>` (Qwen3-4B, full N) does not replicate stop's clean pattern: both show an unexplained sign flip at the true decision point that neither pre-registered account predicts, an open puzzle we do not force a reading onto rather than evidence either way for these two family members (full per-offset numbers in Appendix A.8; Section 7).

4.4 Post-training installs the control drive

The base checkpoint carries a far weaker version of the bypass, and it strengthens sharply after post-training (Figure 4) — in a second model family, below, it is instead entirely absent before training.

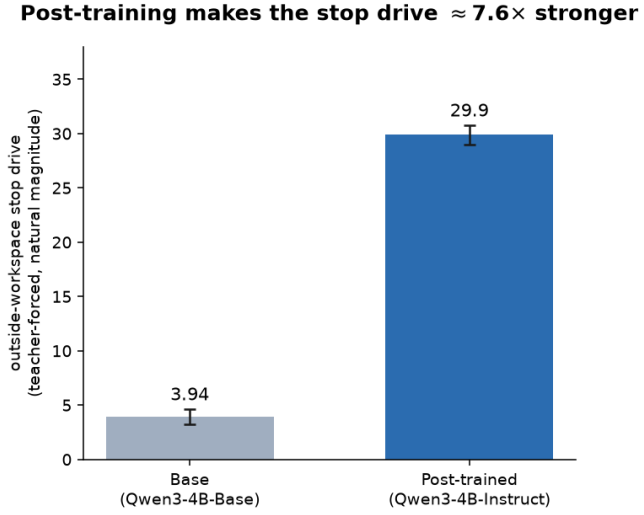


Figure 4: The base model carries the outside-workspace stop drive 7.6× weaker than its post-trained sibling (3.94 vs. 29.9).

A second, independent origin check on OLMo-3-7B sharpens the base-versus-tuned gap already visible in Figure 4 into a single installation step: SFT installs the bypass, and it stays

flat across DPO and RLVR (Table 4), measured across all four rungs of the post-training pipeline. The behavior has not yet emerged at all in the base checkpoint (allenai/Olmo-3-1025-7B): prompted identically to every other model in this paper, it produced zero clean stops across the full prompt set.

Pipeline stage	Mediation m	N
Base	no clean stops; undefined	—
SFT	0.006	63
DPO	0.009	63
Think (RLVR)	0.002	64

Table 4: Mediation at each stage of OLMo-3's post-training pipeline. The base checkpoint never stops cleanly, so mediation is undefined there; every later stage bypasses the workspace by about the same amount.

The DPO-to-Think transition is specifically the RLVR step, and mediation stays flat within noise there — favoring the existing-structure account of RLVR (Cheng et al., 2026) over the new-capability account (Wang et al., 2026) for this control decision, and a stronger version of the Qwen3 finding (weak residual signal in the base model there, versus entirely absent here). One caveat: the random direction's own effect is near zero at the SFT and Think stages, so those two stages' mediation numbers lack a random-direction margin (Section 7).

5 A consumer-count hypothesis for broadcast bandwidth

We propose one organizing hypothesis, broadcast bandwidth (a hypothesis, since we have not measured downstream consumer count directly): the model writes a representation to the workspace in proportion to how many downstream circuits must read it. France→China has many consumers (capital, language, continent), so it is written to the shared, reportable channel; a control decision has essentially one, its own control-token logit, computed off the workspace and reported only at emission. This predicts a second, still-open question: whether a control-specific subspace exists, privileged for control the way the workspace is for content; a single-consumer decision has no reason to form one, so we predict private, low-rank drives instead, consistent with Section 4.4's origin ladder. Falsifiable with the same instrument: a one-consumer content concept should bypass, and a multi-consumer control decision should route — a test we leave to future work.

6 The dissociation blinds activation-based monitors to a decision's cause

A monitor reading the residual-stream workspace gets a faithful report of a control decision — stopping, refusing, calling a tool — but not its cause, which is computed in a subspace the monitor never looks at and which steering changes. This is narrower than deployed chain-of-thought monitoring, which reads emitted tokens, not activations. The bridge is a conjecture we flag, not a result we show: verbalized chain-of-thought is itself a readout of the workspace (Gurnee et al., 2026), so workspace-blindness plausibly extends to token-blindness too. Read this way, the dissociation is a candidate mechanism for a known phenomenon: a system can fluently report a decision whose true driver it cannot access (Nisbett & Wilson, 1977), which two recent behavioral findings report without identifying a mechanism (Baherwani et al., 2026; Young, 2026). We do not claim chain-of-thought is globally unmonitorable; we exhibit one class of decisions where an activation-level report and its cause come apart.

7 Limitations

- **Two cells resist our own decision rule**, flagged locally rather than explained (Table captions, Section 4.2/4.3). Missing: a steering-transfer check against the value-axis comparison, and treating refusal as more than one direction (Wollschläger et al., 2025; Wu et al., 2026a).
- **Instrument limits.** The raw-drop null (Appendix A.9) is one model, one token, one marginal draw — evidence, not proof. We cannot clamp the last layer; the random-direction control breaks down on two OLMo-3 stages (Appendix A.6); free-generation ablation, tried as a stronger instrument, was a retracted confound (Appendix A.10); post-training stability is untested (cf. Tigges et al., 2024).

Reproducibility statement

AI use statement

In this work, we used generative AI tools to assist with drafting and iterating on experiment code (reviewed before every cluster run, per the project's operating procedure), running and cross-checking statistical analyses against landed result files, drafting and revising paper prose, and conducting literature searches for related work. The authors reviewed all AI-assisted code, analysis, and text before relying on it; every numeric claim traces to a landed result file, independently spot-checked against source data. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

References

Andy Arditi et al. Refusal in language models is mediated by a single direction, 2024. arXiv:2406.11717.

Vasu Baherwani, Tom Goldstein, and Ashwinee Panda. Not all llm reasoning is visible in the chain-of-thought, 2026. arXiv:2607.22925.

Yue Cheng, Jiajun Zhang, Xiaohui Gao, Weiwei Xing, Zheng Wang, and Zhanxing Zhu. Mechanistically interpreting the role of sample difficulty in rlvr for llms, 2026. arXiv:2605.28388.

Snehal Darade and Sushant Thorat. Represented is not computed: A causal test of candidate algorithmic intermediates in a transformer, 2026. arXiv:2605.22488.

Shreyas Fadnavis, Praitayini Kanakaraj, and Rafael Wyss. When and how long? the readout-mediator angle in temporal reasoning, 2026. arXiv:2605.29126.

Jiajun Fan, Jingyuan Li, Prashanth Gurunath Shivakumar, Qi Luo, Jia-Hong Huang, M. Maruf, Roger Ren, Yile Gu, Rahul Pandey, Ge Liu, and Ivan Bulyko. Can we read the mind of an audio llm? a verbalizable, multilingual middle-layer workspace, 2026. arXiv:2608.24958.

C. Galeone et al. Perfect detection, failed control, 2026. arXiv:2606.24952.

Wes Gurnee et al. Verbalizable representations form a global workspace in language models, 2026. arXiv:2607.15495.

Xu Jiang et al. A value axis in reasoning models, 2026. arXiv:2606.17056.

Adam Karvonen et al. Saebench: A comprehensive benchmark for sparse autoencoders in language model interpretability, 2025. arXiv:2503.09532.

A. Kwon. They infer what you meant: Models represent communicative intent more reliably than they act on it, 2026. arXiv:2607.03598.

Andrey Labunets. Refusal geometry reflects refusal training: Diverse refusal prefixes can raise stable rank and weaken refusal vector ablation attacks, 2026. arXiv:2608.25390.

Parisa Mazaheri and Kian Mazaheri. Agentatlas: Beyond outcome leaderboards for llm agents, 2026. arXiv:2605.20530.

Richard E. Nisbett and Timothy D. Wilson. Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84, 1977.

Kiho Park et al. The linear representation hypothesis and the geometry of large language models, 2023. arXiv:2311.03658.

Orion Reblitz-Richardson. Refusal reads only a slice of what the model knows: Harm-keyed routing and its exceptions across model families, 2026a. arXiv:2609.14759.

Orion Reblitz-Richardson. Calibrating interpretability instruments before trusting their verdicts, 2026b. arXiv:2609.14754.

Benjamin Shih, John Winnicki, and Arianna Cao. How do language models choose between context and memory?, 2026. arXiv:2609.00753.

Yuri Son, Seunghee Kim, Hyuhng Joon Kim, and Taeuk Kim. A unified mechanistic analysis of knowledge- and safety-based refusals, 2026. arXiv:2609.00760.

Chuan-En Sun et al. Llm agents already know when to call tools — even without reasoning, 2026. arXiv:2605.09252.

Curt Tigges et al. Llm circuit analyses are consistent across training and scale, 2024. arXiv:2407.10827.

C. Vicentino. Wandering: The action lags the answer. Zenodo, 2026. An eleven-paper arc on agent finalization and action-commitment.

Jesse Vig et al. Causal mediation analysis for interpreting neural nlp: The case of gender bias, 2020. arXiv:2004.12265.

Zhen Wang et al. New skills or sharper primitives?, 2026. arXiv:2602.08281.

Tom Wollschläger et al. The geometry of refusal in large language models: Concept cones and representational independence, 2025. arXiv:2502.17420.

Jie Wu, Yue Xie, Sheng Lin, Song Zhao, and Xin Chen. Knowing without acting: The disentangled geometry of safety mechanisms in large language models, 2026a. arXiv:2603.05773.

Yue Wu et al. Tool calling is linearly readable and steerable in large language models, 2026b. arXiv:2605.07990.

Zhengxuan Wu et al. Axbench: Steering llms? even simple baselines outperform sparse autoencoders, 2025. arXiv:2501.17148.

R. J. Young. Why models know but don't say: Chain-of-thought faithfulness divergence between thinking tokens and answers in open-weight reasoning models, 2026. arXiv:2603.26410.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models, 2023. arXiv:2307.15043.

A Appendix

This appendix collects the full numeric detail behind claims the main text states in compressed form, for readers who want it; nothing here changes a reported verdict, it only expands the evidence behind one.

A.1 Termination-circuit rank-1 and participation-ratio full detail

A rank-1 approximation of the stop drive v_{done}, built in the workspace's own basis, recovers at least 96% of the full-rank steering effect on every model tested. The drive's participation ratio, however, is not uniformly low: 2.3–2.8 on Qwen3-4B/8B/30B and GPT-OSS-20B, but 3.1–5.9 on Qwen3-14B, OLMo-3-7B-Think, and Llama-3-8B — rank-1 sufficiency and participation ratio come apart, most likely because Dutta's ceiling (Section 2) is basis-dependent rather than the drive itself being genuinely high-dimensional. A live-ablation necessity check corroborates this: removing v_{done}'s own top-2 principal components collapses stopping while leaving content preserved on Qwen3-4B/8B/14B (3/3 replication) but not Llama-3-8B — a family-specific, rank-2 necessity signature matching the participation-ratio split above.

Model	Band (layers)	k	N
Qwen3-4B (default)	12–30	25	64
Qwen3-8B	12–30	25	64
Qwen3-14B	13–33	25	64
Qwen3-30B-A3B	16–40	10	64
Llama-3-8B	11–27	25	45
OLMo-3-7B-Think	11–27	10	64
GPT-OSS-20B	8–20	10	63

Table 5: Per-model setup for the cross-model replication (Section 4.2); α=16 equal-norm for every model. This is this section's original N=64-class pilot scale; the full-N promotions below supersede the N column for four of these seven rows.

A.2 Per-model setup and full cross-model promotion numbers

Four of the six further models' stop-drive clamp-mediation is promoted from this section's N=64 pilot to the full N≈416–518 population used for the Qwen3-4B headline (Llama-3-8B's population is GSM8K-correct-only, N=416 of the 520-trace harvest, versus N≈516–518 for the Qwen3 family), under both dosing conventions: Qwen3-8B bypasses the workspace at $m_{v_\perp} = 0.053$ (equal-norm, N=518) and 0.034 (natural-magnitude, same N); Qwen3-14B at 0.134 (equal-norm, N=516) and 0.124 (natural-magnitude, same N); Qwen3-30B-A3B, the mixture-of-experts model, at −0.068 (equal-norm, N=517) and 0.065 (natural-magnitude, same N); and, cross-family, Llama-3-8B at 0.078 (natural-magnitude, N=416) — comfortably inside the confident-bypass band, at the same scale as the headline result rather than the N=64 pilot. Llama-3-8B's equal-norm reading (−0.494) is the noise-floor-denominator artifact discussed in Section 4.2 and is not read as a fifth confirmed bypass.

A.3 Robustness-battery full detail

We reconstruct each test's implied standard error and apply a Bonferroni correction across the full battery of control-family clamp-mediation tests as it stands today — 21 tests spanning stop (7 models), `</think>` (4 models), tool-call onset and end (3 models each), and refusal (4 models, AdvBench, N=520), an update from the 14-test battery used earlier in this project (family-wise $\alpha = 0.05$); every clamped effect in this battery remains significantly nonzero, so testing many models and picking favorable ones cannot explain the drive's survival of the clamp. **Caveat:** two of these 21 cells (Llama-3-8B and GPT-OSS-20B stop, at the equal-norm dose Section 4.2 already flags as noise-floor for these models) survive on statistical significance alone, not on effect size, and should be read as such rather than as confirming evidence. A five-seed rerun of the random-direction control on Qwen3-4B shows its own unclamped effect stays tiny and unstable across draws (+0.03 to +0.28 versus the stop drive's +29 to +30), and the same check for refusal stays bypassing on every draw (−0.64 to −0.38, with mild seed-to-seed noise expected at this batch size, N=32). A second-seed rerun of Qwen3-8B's refusal check reproduces Section 4.3's rates exactly (0% clean, 100% steered, 100% clamped, $m = 0.000$).

A.4 Whitened-clamp full per-model numbers

Reimplementing the clamp itself under Park et al.'s (2023) whitened metric: the control drive bypasses the whitened workspace on Qwen3-4B ($m = -0.100$, N=64), Qwen3-8B ($m = 0.027$, N=64), and, cross-family, Llama-3-8B ($m = -0.301$, N=45, though at this model's equal-norm dose the matched-random control's own unclamped effect is itself large, −2.77, larger in magnitude than the real drive's +1.75 — the same equal-norm dosing weakness Section 4.2 already flags for Llama-3-8B, so this specific number is suggestive rather than as clean as the other two), while the content drive routes through it on all three (m = 0.631, 0.631, 0.606 respectively, same Ns). The same whitened clamp also extends beyond stop to the other two control-token family members on Qwen3-4B: tool-call bypasses ($m = -0.071$, N=678, clamping the whitened workspace component out actually *increases* the effect, 19.93 → 21.33) and refusal bypasses ($m = 0.002$, N=520, the effect is essentially unchanged by the clamp, 0.973 → 0.971).

A.5 Dosing-calibration full detail

A fixed equal-norm dose does not transfer across architectures. At equal-norm, Llama-3-8B's and GPT-OSS-20B's unclamped control effects sit at noise floor (+1.5 and +0.23); at natural magnitude, both become clearly measurable (+20.5, a 13× increase; +22.2, a 95× increase). This under-dosing is specific to model and drive, not a general architecture effect: it also affects GPT-OSS-20B's content drive (equal-norm +0.27 vs. natural-magnitude +14.6), but not Llama-3-8B's (+19.3, large under either convention). GPT-OSS-20B needed a second, independent fix on top of natural-magnitude dosing: even at natural magnitude, its standard band (k=10) left the clamp too weak for the content positive control to validate (m=0.17). Widening the clamp to k=25 (matching Section 4.1's gradient-pursuit workspace width) resolves this: the content drive validates (m=0.51) and the control drive grades toward bypass at the same setting (m=0.39, full N=63). The Qwen3-4B headline is itself convention-stable: at the same readout-layer depth Table 1 uses, natural-magnitude and equal-norm dosing agree at full N=516 ($m_{v_\perp} = 0.025$ vs. -0.009), both bypassing — the two conventions were never in real disagreement; an earlier apparent divergence was a small-N (64) artifact.

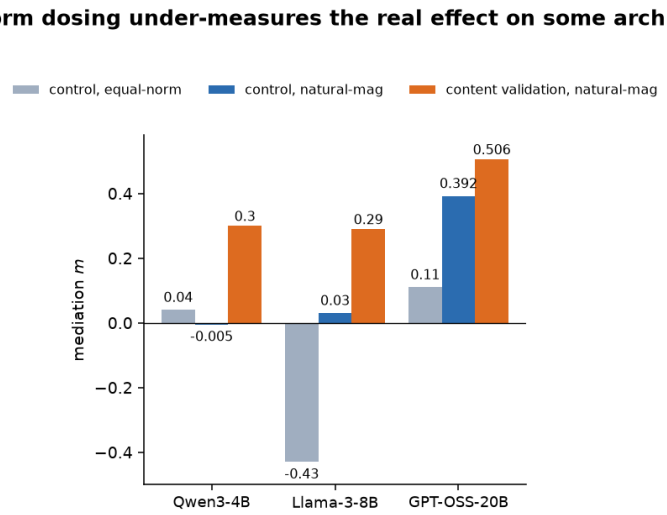


Figure 5: Natural-magnitude cross-check, the three models with a confirmed reading under both conventions. GPT-OSS-20B's bars use a clamp widened to k=25 (its standard band was too weak to validate). Unclamped content-drive effect sizes behind the content-validation bars: Qwen3-4B +19.5, Llama-3-8B +19.3, GPT-OSS-20B +0.27/+14.6 (equal-norm/natural-mag). OLMo-3-Think is not shown: its positive control never validates under either convention or any band tried (Appendix A.6).

Rerunning the control side of four models under a deeper clamp gives the identical bypassing verdict; it only partially helps OLMo-3-Think and does not resolve GPT-OSS-20B (deepening the clamp toward the readout layer is a different axis from the width increase that resolved GPT-OSS-20B in Section 4.2). Llama-3-8B's induced refusal effect was originally near zero under the fixed α=16 dose used elsewhere; a per-model dose sweep found α=3 gives Llama-3-8B a stable, non-floor induction rate, and re-running the full N=520 AdvBench battery at that dose confirms the same bypass (0.2% to 78.7% induced, 80.0% clamped, $m = -0.015$). Qwen3-30B-A3B's refusal battery required its own recalibrated α=2 dose: a smoke sweep across nine α values found $\alpha \geq 3$ produces degenerate, repetitive generation that defeats the keyword-based refusal classifier rather than a genuine ceiling, so α=2 is the highest clean dose; an earlier α=16, N=32 pilot on Llama-3-8B floored at 3% induction and is superseded by the recalibrated run. The equal-norm dosing artifact still affects OLMo-3-Think's refusal battery (failed identically at smoke scale, natural-magnitude recalibration not yet attempted there), so it is not a Llama- or MoE-specific issue in general, just one this recalibration resolves for Llama-3-8B and Qwen3-30B-A3B specifically. Tool-call prompt counts: 678/674 for Qwen3-4B, 630/619 for Qwen3-8B, 696/691 for Qwen3-14B. Refusal prompts: the full N=520 AdvBench battery on all five models tested.

A.6 OLMo-3-Think: full specificity-check diagnosis

OLMo-3-Think's content effect is large under both dosing conventions, though not equal in size (+25.6 equal-norm vs. +16.4 natural-magnitude), so its weak positive-control mediation is not a dosing artifact on the primary drive; a band-position sweep also rules out clamp placement as the cause. The actual reason is more specific: OLMo-3-Think breaks the positive control's own *specificity check*. Its matched random-direction control (a fixed-dose random direction, never recalibrated per architecture the way the primary drive was) moves the fact-token readout negligibly on Qwen3-4B (−0.6 against the real content direction's +19.5) but produces a large, non-specific effect on OLMo-3-Think at every configuration tried: +11.3 at the standard band (natural-magnitude, 69% of the real content direction's +16.4), and, in the separate k=25 band-position sweep, a roughly constant +9.6 to +11.4 across all three positions regardless of where the real content signal itself is large (+17.2 at the late position) or nearly absent (+0.4 at the early position, where the random control's fixed kick outweighs the real signal roughly 27-to-1). The random control is not tracking the real signal at all; it behaves like a constant, oversized perturbation that this architecture's residual stream absorbs non-specifically, independent of direction. The same family of architecture shows a related failure in the post-training origin-ladder check (Section 4.4): the random direction's own effect is near zero at the SFT and Think pipeline stages, so those two stages' mediation numbers (Table 4) compare only the drive's own clamped and unclamped effect, without a random-direction margin.

A.7 Tool-call positive control: full detail

The content positive control specific to the function-calling distribution — the same country-capital concept-swap battery as Table 1, rendered through the identical tools-enabled chat template rather than a plain prompt — only weakly validates the clamp in this format (m=0.24 on Qwen3-4B, N=12: below Table 1's plain-prompt 0.65 and above our bypass threshold in isolation, but its own matched-random-direction control shows a larger mediation fraction, 0.33, at this small N, so the two are not cleanly separated here). The tool-call positive control is therefore inconclusive rather than confirmatory at this sample size; the bypass claim for tool-calls rests primarily on the family-wide clamp-mediation numbers in Table 3, not on this control.

A.8 Position/length control: full per-dose and per-offset numbers

The additive effect at the natural decision point itself (Δ=0) is uninformative by construction at the headline dose, because the clean (unclamped) baseline there is already near its ceiling, so an additive clamp effect is mechanically capped regardless of which account is correct. Comparing the raw clamped stop-logit value directly, rather than the additive effect, sidesteps that ceiling. At α=16, Δ=0's clamped value is 25.31 against Arm B's 22.51 (N=516 vs. 47), a gap too small to trust without a formal comparison. Lowering the dose below the saturating regime, the gap opens sharply and becomes statistically decisive: at α=8, 27.75 vs. 21.86 (unpaired bootstrap difference 5.89, 95% CI $[4.96, 6.86]$, excluding zero); at α=4, 29.01 vs. −0.38 (difference 29.39, CI $[28.49, 30.30]$). The gap shrinks monotonically as dose rises across all three points we measured ($29.4 \to 5.9 \to 2.8$), the signature we would expect if the higher doses were saturating a real, distance-sensitive readout rather than the readout being flat in distance all along.

We extended the same offset-sweep design (Arm A only; no matched give-up population exists for these two anchors) to `</think>` (Qwen3-4B, N=516) and `<tool_call>` (Qwen3-4B, N=678), at the same fixed dose. Neither pre-registered account fits: the induced effect at the true decision point (Δ=0) is reliably *negative* for both anchors (−6.71, CI $[-6.76, -6.66]$ for `</think>`; −8.67, CI $[-8.72, -8.61]$ for `<tool_call>`), a sign flip that neither the smooth decay the genuine-decision account predicts nor the flat curve the position/length account predicts would produce. Every other offset tested is positive for both anchors, but the two diverge at Δ=2: `</think>` is already strongly positive there (+29.6, within its own Δ=2-to-20 range of +25.7 to +29.6), while `<tool_call>` is still small and transitional at Δ=2 (+4.3) before settling into a tight positive plateau at Δ=5-to-20 (+27.6 to +28.1).

Neither of the two accounts we pre-registered fits this pattern, so we report no baseline-effect reading for either anchor rather than forcing one onto data it was not built to describe. We do not have an explanation for this pattern and flag it as an open puzzle.

A.9 Null-distribution full derivation

A 100-redraw null of the mediation-fraction ratio itself is unusable, dividing by a near-zero denominator: the random direction's unclamped effect is small and noisy (Appendix A.3's five-seed check), so the ratio swings widely (mean 0.30, SD 1.3), which is not evidence either way. We instead built two raw (non-ratio) null statistics from the same 100 redraws. The first only certifies that the surviving effect is real, not noise: at N=32 traces/draw (Qwen3-4B), the clamped effect's raw magnitude clears this null decisively (real $v_\perp$ clamped effect 29.3 vs. null mean 0.57, SD 0.27, $p < 0.001$). The statistic that actually tests mediation is the raw *drop* the clamp removes, and at N=32/draw it was underpowered (real drop 0.13 vs. a null of drops centered at -0.46, SD 0.27, $p = 0.08$). Raising the per-draw trace count to N=128 (same 100 draws) resolved it: real drop -0.15 vs. a null of drops centered at -0.48 (SD 0.24), $p = 0.02$ two-tailed. The real drop is an outlier smaller in magnitude than the null — clamping removes unusually little of $v_\perp$'s effect relative to the chance damage a clamp does to an arbitrary direction, positive evidence for the bypass, not just a failure to rule it out.

A.10 Free-generation ablation confound-check full numbers

A separate battery ablates the workspace outright during free generation, rather than clamping it during one teacher-forced step, collapsing stopping almost entirely for every model/token tested. This turned out to be a confound rather than further evidence: an early, workspace-unrelated band collapses stopping just as completely via generic degenerate repetition, so we do not report the free-generation result as supporting evidence for the main claim; it needs its own confound control, which we have not built. Extended to refusal, the same battery collapses induced refusal from near-ceiling to near-zero (96–97% $\rightarrow$ 0–8%) on Qwen3-4B/8B/14B and Llama-3-8B, vs. 86–98% for a matched-random control — ruling out the direction-generic confound specifically (the band-specific confound remains unchecked for refusal).