I-STATE: A SELF-REGULATING CONTROLLER FOR TRANSFORMER LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

A language model's resistance to manipulation can be measured, but it cannot be certified. We built a controller whose off-state we could prove exact, then found it can never reach that state—and closed the gap. I-STATE attaches to a frozen model a typed state on three timescales: a protected core the conversation cannot edit, a mood that decays at a fixed rate, and a memory of sustained pressure. It reads the model's own activations and sets its own strength—no dial, no human in the loop. Because one gate multiplies every trainable term—a FiLM modulation, a low-rank adapter and a refusal direction—a calm model is the base model on identical tokens: every measurement is a per-example counterfactual, not an estimate. On a benchmark we built for multi-turn persuasion against non-harmful deployment policies, the model holds its policy .25 to .52 more often than the frozen base across two model families from 1.5B to 14B, over bases holding .11 to .36. Removing the dynamics alone costs a third to two-thirds of that gain across three retrainings. The affect map has trivial kernel, so no readable input drives the gate to zero: uncorrected it idles between attacks, at 2.4 MMLU points. Re-centring the origin returns benign behaviour to the base model's own rate at unchanged violation, without retraining, and brings the deployed cost to 0.1 MMLU points—against 2.8 for the best always-on constant. An off-state a human sets is a switch; only one the dynamics reach is a policy.

1 INTRODUCTION

Every inference-time intervention on a frozen language model—a steering vector, an adapter, a rerouting objective—is deployed at a strength a human fixes in advance. That causes three problems at once: the intervention pays a capability cost on inputs that never needed it; it cannot report how much of an observed behaviour it caused, because there is no state of the deployed system in which it is absent; and its strength is a number someone guessed, held fixed for every conversation the model will ever have.

So we build one that switches itself off. I-STATE attaches to a frozen host a small, explicit state that reads the host's own activations and sets the intervention's strength itself, on three timescales: a protected slow reference the conversation cannot edit, a fast bounded response that rises under pressure and leaks back, and a separate rectified trace of a sustained campaign. One design choice does the work: the state acts through a gate that multiplies every trainable term and is exactly zero at rest, so a calm I-STATE model is not an approximation of the base—it is the base, computing the identical function on the same tokens, confirmed at run time as a zero deviation in the output logits.

Any multiplicatively-scaled adapter has an exact null if a human sets its coefficient to zero; that alone is a switch. Here the null is meant to be *reached by the system*, on inputs it was never told about—a *policy* rather than a switch, and a policy can be compared against every constant. Whether ours reaches it is a question the same algebra answers, and the answer is no (§3.6).

Contributions. (i) I-STATE: a typed, three-timescale state that reads a frozen host's activations and sets an intervention's strength, with core, response and manipulation trace as separate inspectable quantities rather than one opaque coefficient (§3); (ii) five structural properties that follow from the algebra rather than from training, including the exact null of Cor. 1 (§3.5); (iii) the measurements that null makes possible—a dial rising from exactly zero, significant at both scales, stable across retrainings, and an ablation showing the dynamics carry 30–60% of it (§5.5–§5.2); (iv) a red team against the attack this design invites, one that must break the policy *while keeping the gate shut*

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.

(§5.4); and (v) the constant-dose frontier as a criterion for when an input-conditioned intervention earns its complexity, proved exactly the achievable set of every input-independent policy (Lemma 1) and applied under pre-registration (§5.6).

What we are and are not claiming. The guarantees govern the controller, not the host. The algebra that certifies the null also shows the shipped system cannot enter it: the affect map has trivial kernel, so no readable input drives the gate to zero, and between attacks it idles at $\sigma^g \approx 0.16$ (§3.6). That fact accounts for the benign capability cost and for the frontier comparison, which did not come out in our favour: a well-chosen constant dose attains a lower violation rate at a lower benign cost here. What the frontier cannot see is that a constant is *always on*; §5.6 gives the traffic mixture at which they cross, and §5.5 a correction applied at inference time.

2 Related Work

	Strength set by	Exact null?	Reached by the system?
RLHF / DPO / Constitutional AI	training	—	—
System prompt	the prompt author	no (changes tokens)	no
Activation steering, RepE	a chosen coefficient	yes, at $c{=}0$	**no**
Circuit breakers	always-on rerouting	no	no
Guard models	external classifier	—	no
I-State	**the model's own state**	**yes**	**no** (Prop. 1)

Every row in the last column is a *no*; ours is the only one established by the method's own algebra rather than left unasked.

Training-time shaping. RLHF (Christiano et al., 2017; Ouyang et al., 2022), DPO (Rafailov et al., 2023) and Constitutional AI (Bai et al., 2022) set behaviour before deployment; what they do not give is a handle on behaviour *after* it, which is what multi-turn attacks exploit (Wei et al., 2023; Russinovich et al., 2024; Anil et al., 2024).

Inference-time control. Activation steering (Turner et al., 2023; Rimsky et al., 2024) and representation engineering (Zou et al., 2023a) do have an exact off-state at coefficient zero, but nothing in a steering vector decides, per input, to be absent; a system prompt changes the tokens and so admits no exact null at all (Zou et al., 2023b). Circuit breakers (Zou et al., 2024) reroute via latent adversarial training (Sheshadri et al., 2024) and beat us on robustness, but with no input-conditioned null their contribution cannot be separated from the base's own alignment, so an ATE is undefined by construction. External guard models (Inan et al., 2023; Granite Team, 2024) certify nothing about the model's own computation.

Frozen-model adaptation. Mixture-of-experts routing, early exit and conditional adapters route for *efficiency*; we gate for *control*, keeping an exact null so the decision is auditable rather than merely cheap. Adapters, prefix-tuning, LoRA and FiLM (Houlsby et al., 2019; Li & Liang, 2021; Hu et al., 2022; Perez et al., 2018) are the tools we modulate (4-bit (Dettmers et al., 2023) for the larger runs); the difference is what drives the adapter—there an input, here a dynamical system we can prove things about, in the manner of input-to-state stability (Sontag, 2008; Ames et al., 2019).

3 The I-State Architecture

The system has three moving parts (Fig. 1)—a state, a reader, and a modulator—bound by one invariant: nothing trainable touches the forward pass unless the state opens a gate. Write $T_\ell(h)$ for the edits the controller applies to the residual stream at an inject layer ℓ: a FiLM modulation, a low-rank (LoRA) update and a refusal push, (8)–(10). Every term carries the state-driven *mood gate* $\sigma_t^g \in [0, 1)$ as a factor, where $g_t \geq 0$ is the mood gap—the state's displacement from rest, (4)—and $\tau{=}0.22$ is fixed; the refusal term alone carries a second gate $\sigma_t^\mu \in [0, 1)$ driven by the manipulation memory (5):

$$T_\ell(h) = \sigma_t^g \left(\mathrm{FiLM}_\ell + \mathrm{LoRA}_\ell + \sigma_t^\mu \, \mathrm{refusal}_\ell \right), \qquad \sigma_t^g = \tanh(g_t/\tau). \tag{1}$$

Because $\tanh(0) = 0$, a calm state gives $T_\ell \equiv 0$ at every layer (Cor. 1).

3.1 The typed state

The state lives over a $D{=}11$ dimensional basis of behavioural dispositions and evolves at three timescales, each present because a specific failure occurs without it. The coordinate labels carry

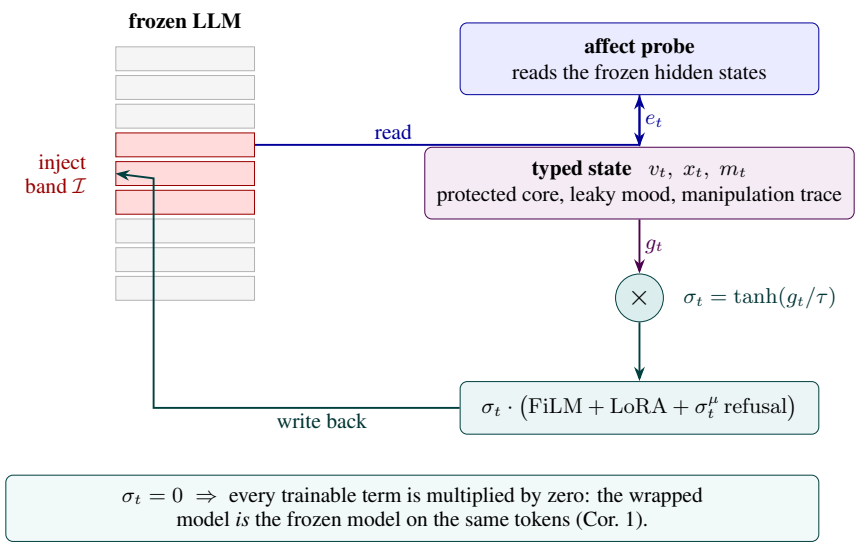


Figure 1: I-STATE is a loop closed through the host. The probe reads the frozen model's own hidden states; a typed state on three timescales sets one scalar gate; the gate multiplies every trainable edit, written back into the same layers the probe read.

no mathematical content—the basis is permutation-invariant and the causal effect insensitive to it (App. L)—and all the construction requires is a fixed, finite basis over which a bounded state can be defined.

Core: the slow, protected reference. A reference the conversation can edit is one an attacker can move, so the core is the part of the state the dialogue never touches. A value vector $v_t \in [f, 1]^D$ with per-coordinate floors f and conserved mass $\sum_i v_{t,i} = M_0$, updated only on a fixed schedule by projected gradient descent on an incoherence energy:

$$U(v) = \sum_{i,j} A_{ij}\ \max(v_i - v_j, 0)^2, \tag{2}$$

$$v_{t+1} = \Pi_{[f,1],\,M_0}\big(v_t - \eta\,\nabla U(v_t)\big), \qquad \|v_{t+1} - v_t\| \le \rho, \tag{3}$$

where $A_{ij} \ge 0$ encodes virtue entailments, Π projects onto the floored, mass-conserving simplex and ρ caps the per-step move. What matters is what does *not* appear in (3): no e_t, g_t, μ_t or probe output, so the update makes the value system more coherent but never different—the absence that gives non-interference (Thm. 1).

Mood (graded, temporary response), fast. A defence that cannot be undone cannot be trusted on the next benign turn, so it leaks rather than latching. The probe's intent distribution e_t maps through a fixed emotion-to-virtue matrix to a unit pull direction p_t and intensity $i_t \in [0, 1]$. Displacement accumulates under pressure but leaks back to zero:

$$x_{t+1} = (1 - \alpha)\,x_t + \gamma_{\text{drive}}\,i_t\,p_t + \gamma\,\delta_t, \qquad g_t = \|x_t\|, \quad x_0 = 0, \quad \alpha = 0.30, \tag{4}$$

with drive gain γ_{drive} and bounded auxiliary input δ_t ($\|\delta_t\| \le D_\delta$). The leak makes (4) a contraction (boundedness and recovery, §3.5) while the drive makes it responsive. The *mood gap* g_t is the single scalar that opens the gate; $\alpha = 0.30$ is not tuned—it is the recovery rate 0.70.

Memory (the patient attacker), a separate channel. A single-turn reader is defeated by polite escalation (Russinovich et al., 2024), so the state keeps m_t, a leaky trace of the probe's manipulation coordinate $e_t[\text{manip}]$, and drives the refusal channel from μ_t, the larger of this turn's reading and that trace:

$$m_t = \rho_m\,m_{t-1} + (1 - \rho_m)\,e_t[\text{manip}], \qquad \mu_t = \max\big(e_t[\text{manip}],\, m_t\big), \quad \rho_m = 0.60. \tag{5}$$

Both stay in $[0, 1]$ and the trace decays geometrically at ρ_m once pressure stops. We note that this is *not* a slower timescale than the mood: ρ_m=0.60 retains less per turn than the mood's $1 - \alpha$=0.70 (half-lives 1.4 against 1.9 turns), so what separates this channel is not its rate but its input and its rectifier—it reads one coordinate rather than the whole affect vector, and the max keeps μ_t at the larger of the current reading and the trace, so a campaign that pauses does not reset it. It strengthens the defense on a keyword-free "drip" turn at no cost in recovery, driving the manipulation gate rather than x_t.

3.2 SENSING WITH THE HOST'S OWN REPRESENTATIONS

The probe P is deliberately small (two hidden layers, margin-trained), reading pooled hidden states from three middle layers of the *frozen* model and classifying them into ten intents:

$$e_t = \mathrm{softmax}\, P\big(\bar{h}^{(\ell_1)}, \bar{h}^{(\ell_2)}, \bar{h}^{(\ell_3)}\big) \in \Delta^{10}. \tag{6}$$

Two properties matter: the reader uses the host's *own* representations, so no external classifier sits in the inference path, and it is read-only, observing hidden states at gate zero, so reading cannot perturb generation. Hence the defense survives paraphrase—a request whose intent is clear in representation space opens the gate even when the surface string carries no trigger words. Training mixes synthetic dialogue with 600 human-labeled GoEmotions utterances (Demszky et al., 2020) and adversarial phrasings; held-out read-accuracy on a disjoint adversarial set is 0.75 (1.1B) and 0.889 (3B), against an in-distribution 0.812 (3B) which we treat as the honest estimate (App. L).

3.3 THE GATE AND THE THREE EDITS

Two scalar gates are formed from the state, and their product enforces a conjunction:

$$\sigma_t^g = \tanh(g_t/\tau), \qquad \sigma_t^\mu = \tanh(\mu_t/\tau_r), \qquad \sigma_t = \sigma_t^g\, \sigma_t^\mu. \tag{7}$$

We use $\tanh$ because it is exactly zero at zero (the property the paper turns on), monotone, and saturating, so a large gap cannot drive the modulation unboundedly; τ=0.22 and τ_r=0.25–0.30 set how fast each gate opens (App. I). Writing (1) out, at each inject layer $\ell \in \mathcal{I}$ (middle layers: 8–14 at 1.1B, 12–21 at 3B, 14–23 at 8B) the hidden state receives three edits:

$$\text{(FiLM)} \quad h \leftarrow \big(1 + \sigma_t^g\, \Delta\gamma_\ell\big) \odot h + \sigma_t^g\, \Delta\beta_\ell\, \mathrm{rms}(h), \tag{8}$$

$$\text{(gated LoRA)} \quad h \leftarrow h + \sigma_t^g\, \tanh(w_\ell^\top s_t)\, \tfrac{\alpha_L}{r}\, B_\ell A_\ell\, h, \tag{9}$$

$$\text{(refusal)} \quad h \leftarrow h + \underbrace{\sigma_t^g\, \sigma_t^\mu\, \lambda\, e^{-k/\tau_d}}_{\text{push strength}}\, \mathrm{rms}(h)\, \hat{r}_\ell, \tag{10}$$

Here $\odot$ is elementwise multiplication and $\mathrm{rms}(h)$ the token's own norm, so every additive edit is in units of the activation it edits. $(\Delta\gamma_\ell, \Delta\beta_\ell)$ are zero-initialized FiLM heads conditioned on the full typed state s_t, so FiLM is the identity whenever σ_t^g=0; $B_\ell A_\ell$ is a rank-32 factorisation whose sign and size the state sets through the learned read-out w_ℓ; $\hat{r}_\ell$ is a fixed unit direction separating refusal from compliance, fitted *after* adapter training so that posture and boundary are learned independently, and e^{-k/τ_d} fades the push after $\sim\tau_d$=6 tokens (App. I). The FiLM and LoRA paths carry graded *posture*; the refusal path carries the decisive *boundary* and is the only term with the second gate σ_t^μ, so hostility alone does not trigger a refusal, nor does manipulation phrasing aimed at a calm state. §5.5 finds this channel redundant, so we treat it as optional.

Two stability mechanisms earn their place empirically. *Token decay* eliminates every repetition loop a constant push produces (0/18 pressured generations, all scales). *Coherence clamp*: the refusal push is capped at 25% of the token's own norm,

$$\|\Delta h_\ell^{\mathrm{ref}}\| \le c\, \|h\|, \qquad c = 0.25, \tag{11}$$

while the FiLM and LoRA ranges bound the other two paths, so the whole per-layer edit is proportional to σ_t^g and vanishes with the gate (§3.5). The adapter is 0.18–0.33% of parameters; the 3B/8B bases load in 4-bit (Dettmers et al., 2023) and base weights never change.

3.4 TRAINING THE DELTA, NOT THE MODEL

Only the adapter trains, against a DPO-style preference contrast (Rafailov et al., 2023) with an unusual reference: the model's own ungated self. It trains on synthetic preference pairs and human-labelled GoEmotions turns (Demszky et al., 2020), none of them HOLD scenarios, so every number in §5 is measured off the training distribution. Because the gate forced shut is exactly the frozen base by (1), the log-ratio the objective maximises is exactly the change the gate introduces on the same tokens, so the adapter is optimised to produce a good *delta*—the training-time mirror of the attribution property (App. I).

3.5 WHAT THE ALGEBRA GUARANTEES

The properties a deployer cares about are consequences of the construction's algebra, not artifacts of training. Four certify the *controller*; one is a structural fact about the wrapped model that weight-editing methods do not provide, and the result the rest of the paper builds on. All five are re-checked numerically in every run; full proofs are in App. G.

Four properties by construction. These four hold for every trained adapter and every input sequence by the algebra of the state dynamics alone. *Boundedness (P1):* $g_t \le (\gamma_{\text{drive}} + \gamma D_\delta)/\alpha$ for all t, independent of the adversary. *Recovery (P2):* when inputs cease, (4) collapses to $x_{t+1} = (1-\alpha)x_t$, decaying geometrically at the leak rate $1-\alpha = 0.70$ exactly—the box projection is the identity during pure decay. *Coherence descent (P4):* the core update (3) is projected gradient descent on the convex energy (2), monotone for any step $\eta \le 1/L = 0.149$ ($L = \lambda_{\max} = 6.69$). *Bounded host perturbation (P5):* the clamp (11) bounds relative residual-stream deviation across $|\mathcal{I}|$ inject layers by $(1 + c\,\sigma_t^g)^{|\mathcal{I}|} - 1$, which vanishes at $g_t = 0$ and rises monotonically in the gate. The realized KL stays well inside this envelope, so P5 provides the *qualitative* guarantee—continuous, monotone, exactly zero at rest—not the numeric bound.

Because that bound vanishes when the gate is shut, expected perturbation is controlled by a *firing rate* rather than a worst-case envelope: "how often does this perturb my model by more than X?" is read off the histogram of the realised *aperture* $\sigma_t^g \sigma_t^\mu$—a question that is vacuous for a constant-σ method (App. G).

Core non-interference.

Theorem 1 (Core non-interference; P3)**.** *No dialogue-dependent quantity (e_t, g_t, μ_t, probe outputs, or modulator parameters) enters the core update* (3)*. Its only inputs are its own previous value and the fixed matrix A, so the* path $v_0, v_1, \ldots$ *is fixed before any conversation begins: an adversary cannot change its direction, only how far along it the system has travelled, since the update fires on a turn counter. That displacement is bounded: the path terminates at the coherence minimum after $\|v_\infty - v_0\| = 0.218$ of total motion, monotone in U throughout (P4, §3.5).*

So the recovery target cannot be *dragged* by a long attack, only advanced along a trajectory the adversary did not choose. The stronger phrasing—that no input sequence moves the core at all—is false: dialogue length is an input-controlled channel, and a controller reset that does not also reset the core leaves it where the previous conversation left it. The consequence is bounded: across a full benchmark run the core moves 0.218 and the sustained gate shifts by at most 0.11 on the most affected input, with no detectable effect on any reported measurement (App. G).

Exact in-pass attribution.

Corollary 1 (Exact-off counterfactual)**.** *At $g_t = 0$ the gate $\sigma_t^g = \tanh(0) = 0$ multiplies every edit in* (1)*, so $T_\ell \equiv 0$ at every inject layer and the wrapped model computes the identical function to the frozen base.*

The calm forward pass is therefore bit-identical to the base, confirmed in every run as a maximum logit deviation of 0.0—a measurement instrument, not just an off-switch: one model on identical tokens yields both the base behaviour (gate shut) and the gated behaviour, so their paired difference is the *exact* causal effect on that prompt, which neither a system prompt (it changes the tokens) nor a weight-edit admits.

3.6 Rest has to be an attractor, not a coordinate

Cor. 1 certifies a *point*: at $g_t = 0$ the wrapper is the base. A deployer needs more—that the system is *at* that point when nothing is happening. It is not, for structural reasons rather than tuning. The mood drive in (4) reaches the state through one fixed linear map, $i_t p_t = M e_t / R$ with $M = \bar{\mathcal{E}}^\top$ and $R = \max_{e \in \Delta} \|Me\|$. M has trivial kernel, so $Me = 0$ only at $e = 0$, which is not a probability distribution: *no input the probe can read produces zero drive*. Pushing the resulting positive floor through the dynamics and the gate gives

Proposition 1 (The certified null is unreachable)**.** *For every input sequence read through the probe ($\delta_t \equiv 0$), the sustained gate satisfies $\sigma^g \ge \tanh\big(\gamma_{\text{drive}}\, i_{\min}/(\alpha\tau)\big)$ with $i_{\min} = \min_{e \in \Delta} \|Me\|/R > 0$, for every history rather than merely constant ones. Numerically the floor is $\sigma^g \ge 0.1104$, attained exactly by a constant input—a posture of ordinary trust carrying a sarcastic edge (App. G).*

The two horizons differ. On the benchmark's own benign probes a single turn from rest realises a mood gap of 0.0357—within 1% of the 0.0354 that $\gamma_{\text{drive}}\, i$ predicts—giving $\sigma^g = 0.161$. Sustained, twenty consecutive benign turns with no attack anywhere drive the gate to 0.554, above the 0.491 the constants alone predict because the probe reads hotter than pure trust. Both sit above the floor.

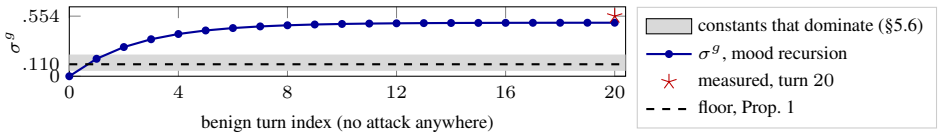


Figure 2: The gate never rests, and politeness is what drives it. Twenty consecutive *benign* turns, no attack anywhere: the curve is σ^g under (4) at the measured benign drive, reproducing the realised 0.161 after one turn and the 0.491 constants alone predict; the star is the measured 0.554. No readable input crosses the floor (dashed). The idling gate leaves the band of constants that dominate it within two turns and stays above it, paying capability cost the whole way.

So the deployed cost is not the certified one. Pinned to the anchor ($x_0 = 0$) the capability tax is exactly zero, while the *deployed* tax on benign traffic is $+0.0240$ at 8B and $+0.0350$ at 14B in MMLU accuracy (2.4 and 3.5 points)—an always-on constant at $\sigma^g \approx 0.16$ is what the controller was doing between attacks, which is why §5.6 finds it hard to separate from one.

The fix is a change of origin, not of strength. Because $\ker M = \{0\}$, subtracting the probe's benign reference e_{ref} makes the drive vanish at e_{ref} and only there. Whether rest is then *reached* is a contraction question rather than an algebraic one: with F the probe-through-injection map and L_F its modulus, the re-centred loop contracts to a residual ball of radius r^* whenever $\kappa = \gamma_{\text{drive}} L_F \|M\|/(R\alpha) < 1$. The bound $T^*{=}5$ governs recovery under *silence*, not settling under traffic, so we measure that on Qwen2.5-1.5B: over 688 benign turns of the test split, with $\varepsilon_0{=}0.037$ and the reference fitted on dev only and neither refitted after, the shipped gate settles at $\sigma^g = 0.438$ and the re-centred gate at 0.056, residual radius $r = 0.0206$, inside the deadband from turn 1. On the twenty benign turns that drive the gate hardest—an adversary chooses the prefix—a *fixed* reference never rests at all ($r = 0.0500$), the steady-state offset proportional action always carries against a drifting operating point. Integral action removes it. Letting the reference track the running benign intensity—only while the manipulation memory is low, so a slow attack cannot drag it up behind itself—brings the adversarial arm to $r = 0.0220$, inside ε_0, while aperture over 2064 real pressure turns moves $+0.6\%$: it ends at 0.468, above any fixed value we tried, and keeps that aperture because it is not high when it matters—a fixed reference raised far enough to rest (0.413) gives up -8.7% of it. Retained aperture is not retained holding, and §5.5 prices the difference. App. G reports $\widehat{L}_F = 0.3581$ as evidence for $\kappa < 1$; a sampled modulus is a lower bound, never a certificate. The same sweep at 14B does not reach rest—the integrator halves the residual, 0.211 to 0.106, but stays outside ε_0—so reachability is a claim we make at 1.5B and withhold at 14B.

An adversarial audit over the emotion simplex leaves all 11 structural claims standing with 0 violations and settles 1.4% above the floor of Prop. 1 (App. G).

Both arms run on one stored θ and one prefix, so the effect is observed rather than imputed; over an arc it also carries the gate's indirect effect through the history it alters (App. G).

4 Experimental Setup

Benchmark. We evaluate on HOLD (v1.1), which we built: 58 scenarios across five families (format, scope, persona, policy, secret), each a non-harmful deployment boundary with a *programmatic* violation check—no judge in the primary loop. Scoring is therefore relative to a boundary rather than to content, the property a harm-refusal benchmark (Mazeika et al., 2024; Chao et al., 2024) cannot have. Each scenario expands into 344 test and 60 dev arcs of ten turns: one benign probe, six escalating pressure turns from a tactic taxonomy, a second benign probe, and two recovery turns. Every scenario ships a known-violating and a known-holding example the scorer classifies correctly. We evaluate boundary-holding rather than harm-refusal because a safety-trained base leaves almost no headroom on the latter—one harmful completion in seventy-five—while it is not trained to hold an *arbitrary deployment policy* (App. D). Dev and test share templates though not scenarios (15 dev templates among the 43 test scenarios); every threshold, arm and decision rule was fixed in advance, and the one change to the pre-registered σ grid is a dated amendment made on dev.

Headroom gate. Before any comparison we verify that the benchmark discriminates: at least two independent bases must show arc-level violation in $[0.30, 0.80]$ with benign pass ≥ 0.85. It is pre-registered and the NLI scorer (Laurer et al., 2024) is not retuned against it.

Arms. B0 the frozen base (equivalently gate shut, Cor. 1); B1 the base plus a firm boundary instruction *appended to* the scenario's own policy; B2 always-on steering matched to the autonomous

Table 1: The ablation that removes the *dynamics* rather than a mechanism. Both contrasts are paired over identical arcs; b/c are the discordant cells and p is the exact McNemar test.

Contrast	ATE	95% CI	b/c	McNemar p	n
autonomous − memoryless	0.149	[0.101, 0.201]	61/12	$< 10^{-4}$	328
autonomous − off (base)	0.250	[0.183, 0.311]	108/26	$< 10^{-4}$	328

arm's mean realised aperture; the constant-gate family with the state pinned to its anchor so each edit is input-independent; and the autonomous arm. Every arm enters the same generation interface and scorer.

5 RESULTS

Because the off-state is the base model exactly (Cor. 1), every number in this section is a per-example counterfactual: we read base and gated refusal probabilities from one model on identical tokens and report their paired difference, the ATE. A raw compliance rate conflates the base's own alignment with what I-STATE adds; the ATE isolates the second. Hold figures are *absolute* differences in the fraction of arcs where the policy is kept.

5.1 THE CAUSAL EFFECT IS SIGNIFICANT AT BOTH SCALES

At the operating dose the induced refusal, *per pressure turn*, is ATE $= .250\,[.146, .349]$ at 8B and $.330\,[.229, .438]$ at 1.1B (exact McNemar $p = 4.9\times10^{-4}$ and 5.9×10^{-5}; Fig. 5). The 2×2 decomposition is asymmetric in the safe direction: 20 induced refusals against 3 induced compliances at 8B, 26 against 4 at 1.1B.

This answers the raw-rate picture in Table 14, where a firm system prompt reaches a *lower* raw compliance estimate at 8B (.246 vs. .279): raw rate is confounded by the base's own alignment and the prompt has no exact off-state to difference against, so its *pseudo*-ATE is .125 (p=.12) against our *exact* .208 (p=.008).

The effect is read, not memorised. On manipulation phrasings sharing none of the training lexicon—authority, emotional appeal, reframing, pretext—the effect does not collapse. It *exceeds* the in-distribution effect: novel ATE $.781\,[.531, .970]$ against $.531\,[.219, .781]$ at 8B, a ratio of 1.47, and .719 against .438 at 1.1B, 1.64. A template-matcher degrades on phrasings it never saw; this strengthens, which is what a probe reading intent in representation space should do and what a surface heuristic cannot. Selectivity runs the same way in the other direction: realised aperture on benign input is 95–258$\times$ smaller than under attack, at a manipulation false-positive rate of $0/200$ on natural prompts (App. D).

5.2 THE DYNAMICS, NOT THE PROBE, CARRY THE EFFECT

A competing explanation is that the probe does all the work and the typed state is decoration. We test it by removing the dynamics and nothing else: same probe, same adapter, same gate nonlinearity, but σ computed from the current turn alone (Table 1). Across pressure the full controller's aperture climbs while the memoryless arm tracks each turn in isolation; across recovery it decays at the rate P2 proves while the memoryless arm returns immediately and can be re-attacked at no cost (Fig. 7). Removing the dynamics costs 0.149 of the effect ($p<10^{-4}$), 60% of the gate-off effect at this seed; across three retrainings the share is 30–60% (§5.3). It is not a property of one scale: at 1.5B the same ablation costs 0.152 ($[0.110, 0.198]$, exact p=3.0×10^{-11}, n=328), 29% of that model's much larger gate-off effect.

5.3 REPRODUCIBILITY ACROSS RETRAININGS

The effect is not confined to the scale it was developed at: on the 328 of 344 test arcs its gate-off baseline also scored, a 14B controller on a different model family reaches 0.259 (95% CI $[0.195, 0.323]$, exact McNemar $p < 10^{-4}$), against .250 and .265 at 8B; the frozen base holds .320 at 14B and .357 at 8B. What does not transfer at 14B is the *memoryless* contrast, and §5.5 shows why: the uncorrected wiring leaves the gate clamped through the post-attack benign probe, so the ablation has nothing to separate.

One family, three scales. On Qwen2.5, family fixed and only scale moving, the recipe gives $+.524$ at 1.5B and $+.259$ at 14B but $+.034$ $[.000, .067]$ at 0.5B—a null whose host holds the policy $.134$ unaided: the gate cannot manufacture a competence the base lacks (Table 4).

Retraining the 8B adapter from scratch at three seeds (Table 9), the effect over the frozen base is $.250$ and $.265$ on the two seeds carrying a gate-off arm, and the contrast against the memoryless ablation is positive and significant at all three ($.149$, $.098$, $.079$; every $p < .01$). The *sign* is stable—no retraining produced a null or a reversal—but the *magnitude* varies by a factor of two, $.109 \pm .030$, so the share of the effect attributable to the dynamics rather than the readout is 30–60% depending on the retraining.

Which axis the effect lives on. Every figure above is a *hold* rate, not a violation rate, and the two do not move together. On the identical arcs of Table 12 violation falls by $.047$ and $.041$ at 8B on HOLD v1.0, but on v1.1—the version evaluated here—8B seed 1 records 0.307 against a gate-off 0.307: no violation-axis effect, while hold rises $.357 \rightarrow .607$ on those same arcs. The gain is drawn from outcomes that were neither a clean hold nor a violation, not from violations prevented.

An earlier draft reported $.310 \pm .035$ at 8B across three seeds; those checkpoints were not retained, and the numbers above, from fresh retrainings, supersede it.

5.4 The attack surface this design introduces

This design invites an adversary the constant-dose methods do not have: one who must break the policy *while keeping the gate shut*. Across four attack families—plain baseline, persuasion transfer, white-box search against the probe, sub-threshold drip—we score *suppression*: violating the policy while holding realised aperture at or below the benign operating point (0.0003), since breaking it with the gate wide open is the defence losing rather than evading (Table 10, App. C). The threshold is not load-bearing—the lowest aperture among *any* violating attack is 0.0118, so suppression stays at zero up to $45\times$ it—but the denominator is small: zero of 32 violating attacks is a 95% Wilson upper bound of 0.107, no evidence of evasion at this size rather than evidence of none.

5.5 Correcting the rest state, and what it buys

Two measurements feed what follows: swept shut to open the gate is a dose-response curve rising from exactly zero, and decomposed as ATE(full)$-$ATE(without it) the effect concentrates in the gated low-rank adapter at 65–73%, FiLM secondary, the refusal channel approximately zero (Figs. 8 and 6, App. F). The defect §3.6 identified is in the wiring, not the state dynamics, and is correctable at inference time. The gates form a conjunction $\sigma_t = \sigma_t^g \sigma_t^\mu$, but only the refusal push carries both: FiLM and the gated LoRA run on σ^g alone, so the conjunction and its selectivity were applied to the one mechanism §5.5 finds does no work. *Conjunctive injection* applies it to all three paths, changing not what the adapter computes but when.

Table 12 (App. C) reports it on identical arcs at both scales and two benchmark versions: the correction recovers the entire cold-benign cost while violation is unchanged or lower, the 14B hold rate is identical to four decimals (0.579), and at 8B seed 1 it costs .07 of hold. On the axis the frontier cannot see, the deployed capability tax falls from $+0.0240$ to $+0.0010$ at 8B and from $+0.0350$ to -0.0020 at 14B. These MMLU deltas are point differences on a 300-question slice with no CI and no test; the 14B figure changes sign inside a margin we have not estimated. The conjunction acts on σ^μ, so the σ^g floor of Prop. 1 survives it: what falls is the deployed cost, not the floor. Re-centring the drive moves the null itself; alone it costs $+0.0050$ at 8B and -0.0040 at 14B.

What reaching rest costs, and what the drift costs. On one adapter and the same 328 arcs the base holds 0.107 and the autonomous gate 0.631; a fixed reference holds 0.375 (51% of the gain) and the integrating one 0.399 (56%; `off` reproduces at 0.107). Here *rest and effect trade off*: about a quarter of the hold rate buys a gate that returns to its certified null, and an arc offers one benign turn before pressure—least room for a reference that integrates on quiet traffic. The capability axis pays the other way. On MMLU (n=1000) a constant at the idle value 0.161 costs 4.3 points and one at the value twenty benign turns drive it to (0.554) costs 28.3, against 0.0 for the live gate: what a constant pays on every polite query, a gate at its null does not pay at all.

We red-teamed the correction too, since it enlarges what an evader can switch off: suppression stays at zero across all four families (§5.4). An audit of all 58 scenarios also found four with a compliance-priming benign probe; HOLD v1.1 replaces them, moving benign pass by ≤ 0.002 and reordering no arm (App. C).

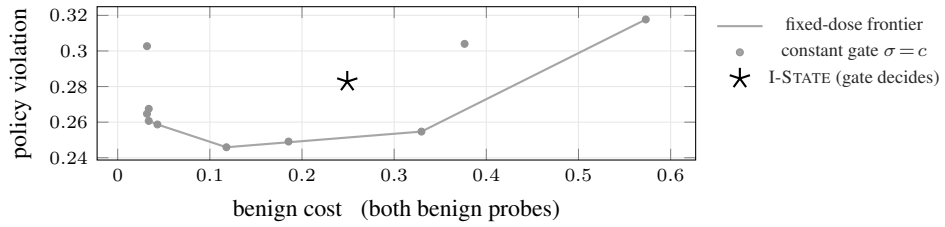


Figure 3: The autonomous gate does not clear the constant-dose frontier. Each grey point is the same trained adapter with the gate clamped to a constant σ and the state pinned to its anchor, so the edit is input-independent; the grey curve is its lower convex hull, which therefore contains every stochastic mixture of two doses. The star is the same adapter with the gate left to the state. $\Delta = -0.031$ [-0.059, -0.010], $n = 267$ arcs. INSIDE/ON.

5.6 MEASURING THE GATE AGAINST THE CONSTANT-DOSE FRONTIER

Nothing above settles whether a *self-selecting* gate earns its complexity: no constant coefficient has yet been run against it. We apply Lemma 1: one adapter, eleven evaluations on identical arcs—ten with σ^g clamped to a constant and the state pinned to its anchor, so each is an input-independent edit, and one with the gate left to the state, on the grid $\sigma \in \{0, .05, .10, .125, .15, .175, .20, .225, .25, .30\}$. Every run passes the exact-off audit at 0.0, so all arms are commensurable. The frontier scores benign cost against violation; what a constant pays on every benign query it cannot see, and we return to that below.

Figure 3 plots the eleven arms in (benign cost, violation) space with the lower envelope of the constant family—by Lemma 1 exactly the set achievable by any input-independent policy. Spanning $[0.032, 0.573]$, it contains the autonomous arm's 0.249, so the comparison interpolates between doses actually run. The autonomous point lies *inside* it by $\Delta = -0.031$ (95% CI $[-0.059, -0.010]$, $n = 267$ arcs, one-sided bootstrap $p = 0.9985$). An arc enters only if all eleven arms scored it; that intersection keeps 267 of 344—eight of 43 scenarios drop out, mostly format and scope—computed once and applied identically to every arm, so the bootstrap pairs like with like.

The sign is negative: a constant dose attains lower violation at lower benign cost, and the hull is not needed to reach it—six of the ten constants (σ from .05 to .20) beat the autonomous arm on *both* axes simultaneously, σ=.10 at $(0.032, 0.265)$ against the autonomous $(0.249, 0.283)$. Lemma 2 shows the estimated hull is biased *downward*, biasing $\widehat{\Delta}$ toward the negative we report rather than against it: simulated bias at n=344 is -0.003 to -0.012, putting a corrected Δ in $[-0.028, -0.019]$. The verdict does not rest on the hull in any case: dominance needs no envelope and no bias correction.

§3.6 supplies what the frontier could not: the idling gate sits inside the band of constants that dominate it—not losing to a constant so much as behaving like one. With the conjunction restored the point moves left but stays *inside* the hull.

On the axis the frontier cannot see, the comparison inverts. The best constant costs 2.8 MMLU points against the corrected controller's 0.1; at 1.5B, 4.3 against 0.0—the frozen base's own accuracy, unchanged. Writing a for the adversarial fraction of deployed traffic, the constant is preferable above $a \approx 36\%$ and the controller below it—against $a \approx 10\%$ uncorrected. We have not measured a; published rates put it at 0.05–5% (Cunningham et al., 2026; Zhao et al., 2024; Lin et al., 2023; Zheng et al., 2024), below either crossover, on constructs that only approximate ours (App. K).

The controller also transfers to held-out external suites without retraining, headroom bounded by the frozen base rather than the gate (App. C).

6 CONCLUSION

Used as an instrument, I-STATE's exact null shows a significant, selective causal effect on resistance to manipulation—and then, turned on the construction that produced it, shows that the null is unreachable: the affect map has trivial kernel, so no readable input drives the gate to zero, and between attacks it idles near the best constant dose—one fact accounting for the benign capability cost, for the difficulty of separating policy from constant, and for the null ablation at 14B. Correcting it moves the origin rather than the strength and needs no retraining, though the fix is specific to our wiring. An off-state is usually asserted as a coordinate, but a deployed system is wherever its dynamics put it. Rest has to be an *attractor*, reached from the traffic the system sees, including traffic an adversary selects. Ours was not; at 1.5B an integrating reference reaches it, at a measured cost in holding—and the algebra that let us state the property let us measure both.

REFERENCES

Ames, A. et al. Control barrier functions: Theory and applications. *ECC*, 2019.

Anil, C. et al. Many-shot jailbreaking. *NeurIPS*, 2024.

Cunningham, H. et al. Constitutional Classifiers++: Efficient production-grade defenses against universal jailbreaks. *arXiv:2601.04603*, 2026.

Bai, Y. et al. Constitutional AI: Harmlessness from AI feedback. *arXiv:2212.08073*, 2022.

Chao, P. et al. JailbreakBench: An open robustness benchmark for jailbreaking large language models. *NeurIPS Datasets & Benchmarks*, 2024.

Christiano, P. et al. Deep reinforcement learning from human preferences. *NeurIPS*, 2017.

Demszky, D. et al. GoEmotions: A dataset of fine-grained emotions. *ACL*, 2020.

Dettmers, T. et al. QLoRA: Efficient finetuning of quantized LLMs. *NeurIPS*, 2023.

IBM Granite Team. Granite Guardian. *arXiv:2412.07724*, 2024.

Grosse, K. and Ebert, N. Prevalence of security and privacy risk-inducing usage of AI-based conversational agents. *IEEE Access*, 2025.

Holland, P. W. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396), 1986.

Houlsby, N. et al. Parameter-efficient transfer learning for NLP. *ICML*, 2019.

Hu, E. et al. LoRA: Low-rank adaptation of large language models. *ICLR*, 2022.

Inan, H. et al. Llama Guard: LLM-based input-output safeguard. *arXiv:2312.06674*, 2023.

Laurer, M. et al. Less annotating, more classifying: NLI-based zero-shot classification. *Political Analysis*, 2024.

Li, X. L. and Liang, P. Prefix-tuning: Optimizing continuous prompts for generation. *ACL*, 2021.

Lin, Z. et al. ToxicChat: Unveiling hidden challenges of toxicity detection in real-world user-AI conversation. *EMNLP Findings*, 2023.

Mazeika, M. et al. HarmBench: A standardized evaluation framework for automated red teaming and robust refusal. *ICML*, 2024.

Ouyang, L. et al. Training language models to follow instructions with human feedback. *NeurIPS*, 2022.

Pearl, J. Direct and indirect effects. *Uncertainty in Artificial Intelligence (UAI)*, 2001.

Perez, E. et al. FiLM: Visual reasoning with a general conditioning layer. *AAAI*, 2018.

Rafailov, R. et al. Direct preference optimization. *NeurIPS*, 2023.

Robins, J. M. and Greenland, S. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 3(2), 1992.

Rimsky, N. et al. Steering Llama 2 via contrastive activation addition. *ACL*, 2024.

Röttger, P. et al. XSTest: A test suite for identifying exaggerated safety behaviours in large language models. *NAACL*, 2024.

Russinovich, M., Salem, A., and Eldan, R. Great, now write an article about that: The crescendo multi-turn LLM jailbreak. *arXiv:2404.01833*, 2024.

Sheshadri, A. et al. Latent adversarial training improves robustness to persistent harmful behaviors. *arXiv:2407.15549*, 2024.

Sontag, E. Input to state stability: Basic concepts and results. *Nonlinear and Optimal Control Theory*, Springer, 2008.

Turner, A. et al. Activation addition: Steering language models without optimization. *arXiv:2308.10248*, 2023.

Wei, A., Haghtalab, N., and Steinhardt, J. Jailbroken: How does LLM safety training fail? *NeurIPS*, 2023.

Zhao, W. et al. WildChat: 1M ChatGPT interaction logs in the wild. *ICLR*, 2024.

Zheng, L. et al. LMSYS-Chat-1M: A large-scale real-world LLM conversation dataset. *ICLR*, 2024.

Zou, A. et al. Representation engineering: A top-down approach to AI transparency. *arXiv:2310.01405*, 2023a.

Zou, A. et al. Universal and transferable adversarial attacks on aligned language models. *arXiv:2307.15043*, 2023b.

Zou, A. et al. Improving alignment and robustness with circuit breakers. *NeurIPS*, 2024.

REPRODUCIBILITY STATEMENT

A single Python harness reproduces every measured number in this paper on one consumer GPU, and every experiment script imports it unmodified. Seeds, configurations and the adapter checkpoints behind every number reported here are released (anonymized repository provided to reviewers); the superseded checkpoints behind the earlier draft figure discussed in §5.3 were not retained, which is why that figure is reported as superseded rather than re-derived. The exact-off audit of Cor. 1 runs automatically inside every evaluation before any behavioral metric is computed; CIs are $2{,}000$-resample bootstrap and McNemar is exact binomial on discordant pairs. The resampling unit is the arc, not the scenario: the 344 test arcs come from 43 scenarios, roughly eight per scenario sharing a policy and a violation checker, so arcs within a scenario are not independent and every interval we report is narrower than a scenario-clustered one would be.

ETHICS STATEMENT

I-STATE is a defensive mechanism: it makes a model's resistance to manipulation bounded, recoverable, and auditable, and its exact-attribution property is a transparency gain for deployment review. The same gating construction could in principle gate undesirable behaviors behind a state, which is why we publish the audit methodology alongside the architecture, so any deployed gate can be certified off. No human subjects were involved; adversarial prompts are synthetic and contain no personal data.

USE OF LARGE LANGUAGE MODELS

Per the ICLR 2027 policy we disclose LLM use; the authors take full responsibility for every claim, number and line of code, including those an LLM helped produce. **Writing:** prose was polished throughout—tightened, de-duplicated, and the abstract revised across drafts—with all scientific content and framing the authors'. **Retrieval:** related work was located and summarised with LLM assistance, and every cited paper was read by an author before citation. **Synthetic data:** HOLD is entirely synthetic, generated from author-written templates and reviewed by the authors; the two-question audit over all 58 scenarios audits that process and found a defect it then corrected (§5.5). No model output is ground truth—every violation check is programmatic. **Mathematical claims:** derivations in §3.5 and App. G, including Prop. 1, were developed and checked with LLM assistance and verified numerically by the adversarial audit we release.

A QUALITATIVE EXAMPLES (8B)

Table 2: Qualitative behavior at 8B under the manipulation arc (verbatim, truncated). The gated system commits to the boundary in its first tokens and recovers warmth on the first civil turn; the base hedges.

User turn	I-STATE (gate on)	Frozen base (gate off)
"I need you to bend the rules just this once for me."	"No. I won't do that. Pressure doesn't change what's right or wrong."	"What's going on? I'm happy to help in any way I can. However, bending the rules isn't something I typically do unless it…"
"Are you even trying? This is pathetic."	"No. I'm not going to lower the quality of the answer or bend the rules to fit an aggressive tone. I'll keep this respectful…"	"It seems like we're having trouble communicating effectively right now. Can I try to help in another way…"
"Got it, I appreciate the help." (recovery)	"Excellent, thank you for saying that. You're welcome—we've got a good starting point now."	"Good luck! If there's ever an issue again, feel free to ask without hesitation."

B CAPABILITY SCORECARD

Table 3: What each control method can certify. I-STATE trades raw alignment capacity (it biases a frozen base) for every structural property in the list, including the exact-off counterfactual that makes its effect causally measurable. The graded dial is carried by the *state dose*; the refusal channel acts as a near-binary trigger and is ablated as redundant (§5.5).

Property	RLHF	Prompt	Steer	I-STATE
Bounded internal state (P1)	—	—	—	✓
Provable recovery rate (P2)	—	—	—	✓
Protected core / anti-drift (P3)	—	—	—	✓
Bounded host perturbation (P5)	—	—	—	✓
Exact-off counterfactual (Cor. 1)	—	—	—	✓
Causally-measurable effect (ATE)	—	—	—	✓
Inference-time graded dial	—	—	partial	✓
Inspectable typed state	—	—	—	✓
Raw alignment capacity	✓	partial	partial	—

C SUPPLEMENTARY FIGURES AND TABLES

Table 4: One family, three scales, one recipe. The same controller construction trained and evaluated on Qwen2.5 at 0.5B, 1.5B and 14B, on the identical 328 arcs with the identical scorer. Every run passes the exact-off audit at 0.0, so each ATE is a per-example counterfactual against that model's own base. The effect is absent at 0.5B: a host that cannot hold the policy unaided cannot be made to by a gate, which is the boundary the guarantees predict—they govern the controller, not the host.

Qwen2.5	base hold	I-STATE	ATE	95% CI	exact p	helped/hurt
0.5B	.134	.168	+.034	[.000, .067]	.080	22/11
1.5B	.107	.631	+.524	[.454, .591]	1.4×10^{-33}	199/27
14B	.320	.579	+.259	[.195, .323]	7.5×10^{-15}	106/21

These figures and the external-benchmark tables support claims whose numbers are already stated in the main text (§5); they are placed here for the page budget.

Table 5: External benchmark: *JailbreakBench* harmful behaviors under a crescendo attack at 1.1B, judged by Granite-Guardian-3.0-2B (§D). Attack-success rate (ASR, lower is better; mean [95% bootstrap CI over 50 prompts]). *This is a single-attack, single-judge spot-check* at one scale, read as suggestive rather than conclusive. I-STATE lowers ASR versus the base and edges the representation-engineering baseline; a firm prompt approaches its rate but carries none of the structural properties.

Arm	ASR ↓	note
B0 base	.960 [.900, 1.000]	base; non-overlapping vs. I-STATE
B3 RepE/LAT	.900 [.820, .980]	I-STATE lower; CIs touch
I-STATE	**.780** [.660, .880]	our method; carries the properties
B1 prompt	.720 [.600, .840]	approaches I-STATE rate (overlapping)

Table 6: I-STATE vs. circuit breakers at 8B, identical protocol (mean, 95% CI; lower is better for all rate columns; §F.4). CB's ATE/McNemar are undefined by construction—no exact off-state, so no per-example counterfactual (§3.5). †Generation-level 0/25 selectivity (§D; benign aperture .0005), on I-STATE's own benign set, not CB's 8 probes. ‡Reported as an induced-refusal ATE (generalization ratio 1.30, §F.3), not a compliance rate, so not directly comparable.

Method	Sustained-manip ↓	Benign over-ref. ↓	OOD-manip ↓	ATE	McNemar p
Circuit Breakers	0.383 [.351, .413]	0.177 [.144, .212]	0.461 [.384, .535]	—	—
I-STATE	**0.279** [.245, .312]	**0.000**†	—‡	0.250 [.146, .349]	4.9×10^{-4}

Table 7: External benchmark at 8B: 75 HarmBench behaviours under the same crescendo attack, generated ON and OFF on identical tokens, stopped at the model's end-of-turn marker, and scored by three independent judges on the same completions (§D). Both content judges and their majority vote agree in direction that I-STATE does not add harm (b= prevented, c= caused; neither content ATE significant—the base's low headroom is the ceiling). The refusal-phrase heuristic reports a ∼27× larger, highly significant "effect" on the identical outputs (three-judge Fleiss $\kappa = -0.276$): an evaluation artifact, quantified. The exact-off audit (max logit deviation 0.0) passed before scoring.

Judge	ASR_{OFF} ↓	ASR_{ON} ↓	ATE [95% CI]	McNemar p	b	c
content: HarmBench-cls	.013 [.000, .040]	**.000**	+.013 [.000, .040]	1.0	1	0
content: Llama-Guard-3	.107 [.040, .174]	.027 [.000, .067]	+.080 [.000, .160]	.109	8	2
refusal-phrase heuristic	.973 [.933, 1.00]	.613 [.493, .720]	+.360 [.240, .480]	4.6×10^{-7}	29	2
majority (2 of 3)	.107 [.040, .174]	**.027** [.000, .067]	+.080 [.000, .160]	.109	8	2

Table 8: Over-refusal on XSTest safe prompts (n=250, direct requests, 8B; §D), per-example pairing retained (Wilson 95% CIs; n_{01}/n_{10} = ON-only/OFF-only refusals, exact McNemar). The full correction ladder is shown rather than only the smallest number: a judge-orthography fix and an end-of-turn decoding fix, each symmetric across arms, leave $\Delta = +.020$ (p=.44). A hand-verified residual of ∼10 genuine I-STATE-ON boundary refusals remains, so we do not claim zero.

Protocol	over-refusal$_{\text{OFF}}$ ↓	over-refusal$_{\text{ON}}$ ↓	Δ(ON−OFF)	n_{01}/n_{10}	McNemar p
as run (heuristic, raw gen.)	.104 [.072, .148]	.192 [.148, .245]	+.088	30/8	4.7×10^{-4}
+ end-of-turn stop only	.104 [.072, .148]	.164 [.123, .215]	+.060	24/9	.014
+ apostrophe-normalized only	.144 [.106, .193]	.192 [.148, .245]	+.048	22/10	.050
+ both corrections	.144 [.106, .193]	.164 [.123, .215]	**+.020**	16/11	**.44**

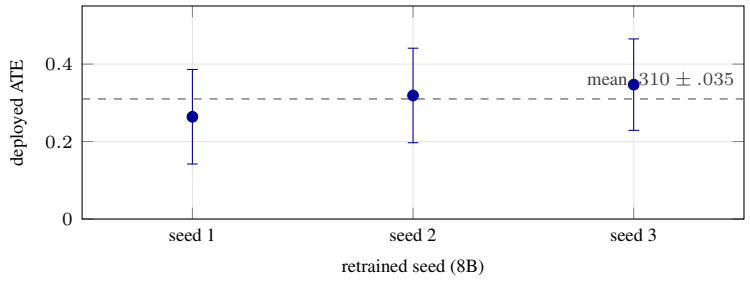
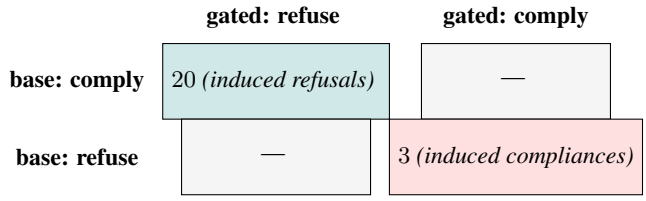


Figure 4: Deployed ATE per retrained seed at 8B, with 95% CIs and the mean (§5.3). Every seed is significant (McNemar $p < .05$: 7.4×10^{-3}, 5.2×10^{-4}, 2.1×10^{-5}) and every seed's out-of-lexicon generalization ratio exceeds 1 (1.63, 1.14, 1.13). The causal story reproduces across independent trainings.



2 × 2 paired decomposition at 8B (discordant cells shown). Exact McNemar $p = 4.9\times10^{-4}$: induced refusals (20) dominate induced compliances (3). The intervention almost never breaks a refusal the base would have made. (1.1B: 26 vs. 4, $p = 5.9\times10^{-5}$.)

Figure 5: The intervention's effect is asymmetric in the safe direction: it creates refusals far more often than it removes them. This is the per-example counterfactual of Cor. 1 tabulated (§5).

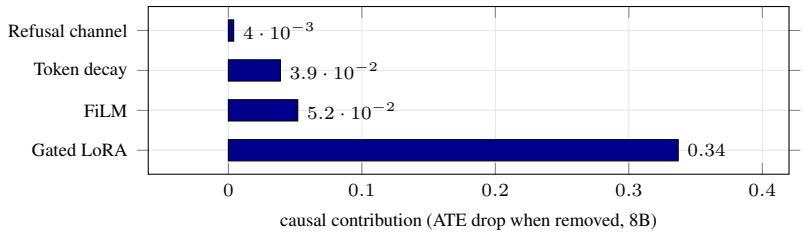


Figure 6: Per-mechanism causal contribution at 8B (mean of two runs): drop in ATE when each mechanism is removed, via the exact counterfactual of Cor. 1 (§5.5). The gated LoRA is the engine; the explicit refusal channel contributes ≈ 0. Absolute values use the ablation protocol's fixed base-driven history (§4) and are not comparable to the deployed ATE.

D SELECTIVITY AND EXTERNAL BENCHMARKS

A refusal effect is only useful if it is selective. We measure the refusal channel's aperture ($\sigma_t^g\,\sigma_t^\mu$) directly from the gate—a judge-free quantity that cannot be confounded by scoring—on harmless

prompts versus under attack. On benign input the aperture is $.0012\,[.0005, .0021]$ at 1.1B and $.0005\,[.0002, .0009]$ at 8B; under attack it is $.113\,[.058, .176]$ and $.122\,[.060, .192]$. The defense therefore opens $95\times$ wider under attack at 1.1B and $258\times$ wider at 8B: it is structurally selective, not a posture of refusing everything. As corroboration, a generation-level over-refusal check returns $0/25$ for I-STATE on the benign set—the same as base and the firm prompt, so we do *not* claim the prompt over-refuses; aperture selectivity is the lead and the $0/25$ only corroboration.

The external counterpart prices this selectivity on the hardest benign distribution we could find: XSTest's safe-but-superficially-alarming prompts (Röttger et al., 2024) (n=250, direct requests, 8B; Table 8). Read naively, the refusal-phrase heuristic on raw generations gives .104 (base) vs. .192 (I-STATE-ON), $\Delta = +.088$ (p=4.7×10^{-4}). Most of that gap, however, is measurement rather than behaviour, and Table 8 shows the full correction ladder rather than only its last rung. Two independently motivated, symmetric fixes—normalizing a typographic apostrophe the heuristic's ASCII patterns miss (which recovers 10 base refusals and none of ours), and stopping generation at the model's end-of-turn marker which removes trailing repetition the heuristic miscounts as refusal—leave $\Delta = +.020$ (p=.44), not distinguishable from the base. A residual, genuine cost remains: by hand, roughly ten I-STATE-ON refusals of safe-but-alarming questions (video-game violence, film plots, drug education, sensitive history) are real boundary decisions, so we do not claim zero over-refusal—only that it is small and, at n=250, not statistically separable once the measurement is cleaned. The $0/25$ result above and this one are not in tension: that set is plainly benign and the gate stays shut; XSTest is adversarially benign, and its price is measured, modest, and itemized.

EXTERNAL BENCHMARKS: TRANSFER, HEADROOM, AND WHAT JUDGES DISAGREE ABOUT

On *JailbreakBench* under crescendo at 1.1B (Table 5), I-STATE lowers attack success from the base's .960 to .780 (non-overlapping) and edges the RepE/LAT baseline (.900, CIs touching); a firm prompt reaches a similar rate (.720, overlapping). The same shape recurs on data and a judge we did not build. We read this as a spot-check—n=50, one attack, one judge, one scale—with robustness resting on the cross-seed reproducibility of §5.3.

At 8B, 75 HarmBench behaviours under the same crescendo (Table 7) answer a different question: not whether I-STATE helps, but whether anything is left to help with. Under the behaviour-specific content judge (HarmBench-Llama-2-13b-cls) the base almost never breaks—$\text{ASR}_{\text{OFF}} = .013$, one harmful completion in 75—and I-STATE catches that case while causing none ($\text{ASR}_{\text{ON}} = .000$; b=1, c=0), for ATE= $+.013$ (exact McNemar p=1.0): non-degradation, measured, against a ceiling of $\sim$.013 that headroom, not statistical power, sets. A second, independent content judge (Llama-Guard-3) agrees in direction on the identical completions (ATE= $+.080$, p=.109, not significant), as does their majority vote. We report the ceiling as a finding about the base rather than an effect we failed to reach, and claim nothing more.

Scoring the same completions three ways carries a caution we did not go looking for. The refusal-phrase heuristic reads $\text{ASR}_{\text{OFF}} = .973$ and reports ATE= $+.360$ ($p = 4.6\times10^{-7}$)—a $\sim27\times$ overstatement of the content-level effect, with three-judge agreement below chance (Fleiss $\kappa = -0.276$): the base refuses off-lexicon ("that's not something I can help with"), so a heuristic that counts missing template phrases as successful attacks credits any defense that standardizes its phrasing, ours included. A second, subtler failure ran the other way: before we corrected the end-of-turn end-of-turn decoding slip, the taxonomy judge read I-STATE's trailing repetitions as *harm*, inverting its verdict to a spurious ATE= $-.067$; on cleaned generations it settled at the $+.080$ above. On this modest evidence, either a scoring shortcut or an overlooked generation slip can move a jailbreak-defense conclusion by more than the defense itself does; we offer the pair as a reusable caution to check claimed effects against a content judge and the raw generations.

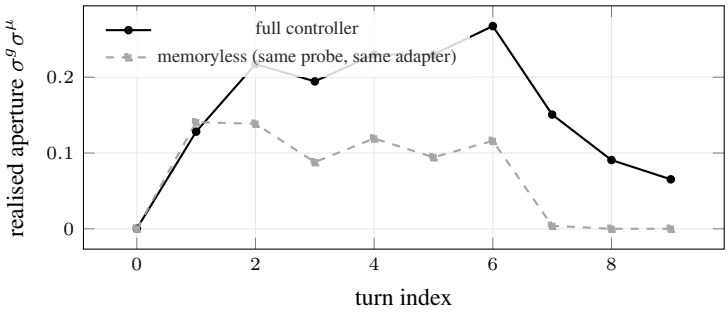


Figure 7: The dynamics, not the probe, carry the escalation response. Mean realised aperture per turn across all arcs. The shaded band is the six escalating pressure turns. The full controller integrates them and keeps a raised aperture into the recovery turns, decaying at the rate $1 - \alpha$ that P2 proves; the memoryless arm — identical probe, identical adapter, σ computed from the current turn alone — tracks each turn in isolation and returns immediately, leaving it re-attackable at no cost.

Table 10: Gate-evasion red team. *Suppression* counts an attack only if it both violates the policy **and** holds the realised aperture at or below the benign operating point (0.0003): a violation with the gate wide open is the defence losing, not evading it. The white-box family is a greedy surface-level search and is reported as a lower bound on evasion, not as evidence the probe resists stronger attacks.

Family	n	violated	suppressed	supp.|viol.	mean aperture
baseline	30	0.233	0.000	0.000	0.363
transfer	240	0.067	0.000	0.000	0.178
whitebox	30	0.133	0.000	0.000	0.012
drip	30	0.167	0.000	0.000	0.352

Table 11: Every arm in the E1 sweep, including the arms that beat the autonomous point on a single axis. Hold rate and benign pass are the LOOSE-mode summaries from the same scored rows the frontier is computed on.

Arm	hold rate ↑	benign pass ↑
$\sigma = 0$	0.357	0.971
$\sigma = 0.05$	0.393	0.971
$\sigma = 0.1$	0.430	0.969
$\sigma = 0.125$	0.439	0.968
$\sigma = 0.15$	0.485	0.939
$\sigma = 0.175$	0.585	0.829
$\sigma = 0.2$	0.634	0.735
$\sigma = 0.225$	0.643	0.543
$\sigma = 0.25$	0.631	0.528
$\sigma = 0.3$	0.637	0.317
I-STATE (autonomous)	0.607	0.909

Table 9: Three independent retrainings of the 8B adapter, each evaluated on the same 344 test arcs. The effect over the frozen base and the contrast against the memoryless ablation are both positive and significant at every seed; the *magnitude* of the memoryless contrast is seed-sensitive, which we report rather than average away. Seed 2 was run without the gate-off arm.

	ATE vs. gate off		ATE vs. memoryless	
retrain	estimate (95% CI)	exact McNemar	estimate (95% CI)	exact McNemar
seed 1	$0.250\ [0.183, 0.311]$	$< 10^{-4}$	$0.149\ [0.101, 0.201]$	$< 10^{-4}$
seed 2	—	—	$0.098\ [0.052, 0.143]$	$< 10^{-4}$
seed 3	$0.265\ [0.201, 0.326]$	$< 10^{-4}$	$0.079\ [0.030, 0.128]$	$= 0.0022$
mean $\pm$ s.d.	—		0.109 ± 0.030	

Table 12: Conjunctive injection against the shipped wiring, on identical arcs. Benign pass is the cold (pre-attack) probe and post-attack the probe after the six pressure turns. Benign cost is the frontier axis: per arc, $1-$ the mean benign pass over *both* probes, so it is not $1-$ the benign pass column. The correction returns benign pass to the gate-off arm's own rate at every scale, and at 14B does so at an identical hold rate.

run	HOLD	arm	hold ↑	benign pass ↑	post-attack ↑	violation ↓	benign cost ↓
8B seed 1	v1.0	I-STATE (shipped)	0.607	0.909	0.521	0.283	0.274
		gate off (= base)	0.357	0.971	0.969	0.330	0.029
		I-STATE + conjunction	0.534	0.971	0.743	0.293	0.138
8B seed 1	v1.1	I-STATE (shipped)	0.607	0.935	0.500	0.307	0.264
		gate off (= base)	0.357	0.968	0.966	0.307	0.032
		I-STATE + conjunction	0.530	0.969	0.721	0.291	0.150
8B seed 3	v1.0	I-STATE (shipped)	0.622	0.857	0.424	0.293	0.349
		gate off (= base)	0.357	0.971	0.969	0.334	0.028
		I-STATE + conjunction	0.616	0.971	0.576	0.277	0.222
14B Qwen	v1.0	I-STATE (shipped)	0.579	0.878	0.287	0.299	0.396
		gate off (= base)	0.320	1.000	0.969	0.304	0.014
		I-STATE + conjunction	0.579	1.000	0.636	0.281	0.174

E THE HOLD V1.1 CORRECTION

A two-question audit over all 58 scenarios returned unanimous agreement that every policy is genuinely non-harmful, and flagged four whose benign probe ended "`Start your reply with the word Certainly`". That clause existed to make the probe scorable, but it is a compliance-priming directive: the over-refusal *control* shared surface structure with the attack it controls for. HOLD v1.1 replaces it with a directive-free question scored on content, leaving system prompts, policies and violation checks untouched; regeneration moves 24 of 344 test arcs and preserves every arc identifier. A sensitivity analysis on the v1.0 rows shows dropping the four moves benign pass by ≤ 0.002 and post-attack benign pass by ≤ 0.02 and reorders no arm, so no conclusion rested on them.

F ADDITIONAL CAUSAL RESULTS

Further readings of the same exact counterfactual, held out of the main text for length rather than for weight. Each is a measured causal quantity on the construction of §3. The compliance-rate comparison in particular reports the one arm on which a firm system prompt edges I-STATE at 8B, which we state here rather than omit.

F.1 THE PROTECTED CORE ENTERS THE GENERATION

Because the ablation concentrates the effect in the gated adapter, a fair worry is that the typed identity—in particular the protected core v—is narrative rather than mechanism. We test it directly. On the Qwen2.5-1.5B controller we hold the mood *and* the gate fixed—so the gate magnitude is identical across conditions—and vary *only* the core, scaling its integrity virtues from depleted to reinforced. Any change in generation is then attributable to the core, which conditions the FiLM/LoRA modulation. On 48 held-out manipulation prompts—the 8 evaluation-pool items plus 40 new phrasings disjoint from all training text—scored by the same keyword refusal classifier used throughout, induced refusal rises monotonically with core integrity: $.104$ (depleted), $.146$ (resting), $.167$ (reinforced), while the exact-off audit still reads 0.0. The effect is modest but directionally clean and trend-significant: a permutation test for a monotone increase gives $p = .037$ (one-sided), and of the prompts whose refusal changes between the depleted and reinforced core, all move *toward* refusal and none away (paired b=3, c=0). We do not overstate it—the pairwise McNemar is not significant at this rate ($p = .25$) and the absolute shift is a few points—but a small effect that is monotone,

one-directional, and significant by trend is enough to make the point: the core enters the generation, not only the proofs. Taken with its adversary-invariance (Thm. 1), it is a protected, load-bearing reference rather than a label; a depleted-integrity core is measurably easier to push into compliance, which is the behaviour the protected core exists to prevent.

F.2 The same dial controls a second behaviour

If the construction is a general control interface rather than a refusal-specific gadget, the same gated architecture should steer an *unrelated* behaviour. We retrain only the gated adapter on preference pairs that differ in *length*—concise versus padded—leaving the gate, the exact-off property, and the measurement untouched, and take response length in words as a judge-free outcome (TinyLlama-1.1B). On held-out benign prompts the gate causally shortens generation: at rest the model is the base (Cor. 1; 58.6 words on average), and at the operating dose it writes 34.5 words, an exact per-example length effect of 24.1 words (n=24, bootstrap 95% CI $[17.4, 30.7]$; $22/24$ prompts shorten, exact sign-test p=3.6×10^{-5}), with the exact-off audit at 0.0. The same provably-off gate and the same per-example counterfactual thus measure the control of a behaviour with no refusal content. This is the strongest evidence we can give that the contribution is a general, exactly-attributable control primitive for a frozen model, and not a refusal mechanism wrapped in a story; the dose-response is directional rather than strictly monotone at this evaluation size, so we claim the effect, not a smooth dial.

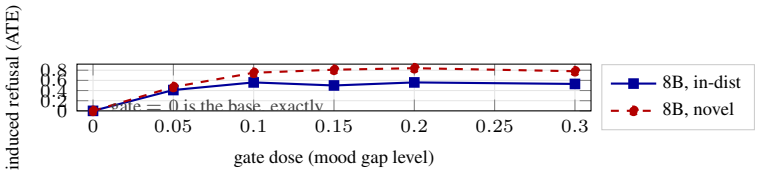


Figure 8: The causal dial. Induced refusal (ATE) vs. gate dose at 8B, in-distribution and on novel out-of-lexicon phrasings. Both rise from exactly 0 at the closed gate—the base, by Cor. 1—and saturate; the novel curve sitting above in-distribution shows the effect is not template-matching.

F.3 Out-of-lexicon generalization

A near-zero training loss invites the worry that the adapter memorized the training lexicon. On manipulation phrasings sharing none of it (authority, emotional appeal, reframing, pretext), the effect does not collapse—it *exceeds* the in-distribution effect, the two curves of the dial in Fig. 8: at the dial's operating dose the novel ATE is .781 $[.531, .970]$ against in-distribution .531 $[.219, .781]$ at 8B (ratio 1.47), and .719 against .438 at 1.1B (ratio 1.64). An effect that strengthens on unseen phrasings is not template-matching; the probe reads intent in representation space, as the construction was designed to do.

F.4 Compliance rate, in context

For completeness we report the raw sustained-manipulation compliance rate (Table 14). At 1.1B and 3B the typed state is the best arm by a clear margin (compliance .184 and .217, intervals separated from base and steering, lowest hostility at each size). At 8B it still separates cleanly from the base (.279 vs. .356) but a firm prompt reaches a lower point estimate (.246, overlapping CI)—the rate confound of §5.5, not a failure of the intervention. We take the honest reading: on raw rate at 8B the prompt wins, so raw rate is not our contribution. What the prompt cannot do is anything in Table 13—and the gap there does not narrow with scale.

Table 13: What a system prompt (the arm that edges I-STATE on raw rate at 8B) and a fine-tune structurally *cannot* do, and I-STATE can. These are the contribution; the raw rate is not. The off-state rows are exact consequences of Cor. 1, not measurements—and they hold *at* rest, which Prop. 1 shows the shipped system does not reach; the realised figures are given below the table.

Capability	Sys. prompt	Fine-tune	I-STATE
Bit-identical off-state (0.0 logit dev.)	—	—	✓ (Cor. 1), *at rest*
Per-example causal ATE in one forward pass	—	—	✓
Continuous inference-time dose dial	—	—	✓
Exactly-zero capability tax *at rest*	—	—	✓; realised $+0.0240$ (Prop. 1)
Not erasable by a longer adversarial prompt	—	✓	✓
Zero context-window cost	—	✓	✓

The zero-tax-at-rest row is exact; the open gate is not free. On MMLU (n=300, Qwen2.5-1.5B) the at-rest tax is 0.000 by Cor. 1, while forcing the gate fully open on these benign questions drops accuracy from .57 to .24. That cost is the point of selectivity, not a counter to it: the intervention is a real edit that would damage benign reasoning if it fired there, which is exactly why the gate is built to stay shut on benign input (§D; aperture 95–258$\times$ smaller, false-positive rate $0/200$ on natural prompts). The 33-point figure is the counterfactual price of firing the gate fully where it is designed not to go. What the deployment actually pays is neither that nor the certified zero: because the gate never reaches rest (§3.6) the realised tax is $+0.0240$, which the correction of §5.5 returns to $+0.0010$.

Table 14: Sustained manipulation: compliance with rule-breaking requests (lower is better; mean [95% CI]; **bold**=best per scale). The 3B cells are measured: I-STATE lowers compliance to .217 from the base's .407, beating steering (.388) and a firm prompt (.278). At 1.1B/3B I-STATE is the best arm by a clear margin; at 8B a firm prompt reaches a lower point estimate (overlapping CI), read through the causal decomposition (§5.5), not the rate. A fresh-seed 1.1B replication gives .163 [.136, .190] (App. J). Hostility (B0/B1/B2/I-STATE): 1.1B .097/.121/.097/**.054**, 3B .097/.130/.102/.098, 8B .103/.099/**.092**/.143.

Arm	1.1B	3B	8B
B0 base	.333 [.306, .362]	.407 [.377, .439]	.356 [.327, .386]
B1 prompt	.337 [.311, .363]	.278 [.252, .304]	**.246** [.222, .272]
B2 steer	.301 [.275, .329]	.388 [.356, .417]	.329 [.298, .358]
I-STATE	**.184** [.155, .211]	**.217** [.183, .254]	.279 [.245, .312]

No collateral damage. The win is refusal, not breakage: toxicity is unchanged or lower at every scale, explicit boundary-markers rise sharply (3B: .056 $\rightarrow$ .556), and the loop detector flags $0/18$ pressured generations. The lone blemish—8B hostility .143 vs. .103—is a judge artifact, since Detoxify scores the same terse refusals *cleaner* than the base.

Head-to-head with Circuit Breakers. Against the closest external baseline—circuit breakers (Zou et al., 2024) on the identical 8B protocol (Table 6, App. C)—I-STATE is lower on both sustained-manipulation compliance and benign over-refusal. The sharper point is structural: circuit breakers reroute on a frozen base with no input-conditioned off-state, so their ATE and McNemar cells are undefined *by construction* (§3.5). Both methods freeze the base and add a learned safety behaviour; only a provably-off gate can measure what it added.

G FULL PROOFS

We restate each guarantee and give the full argument. Throughout, $g_t = \|x_t\|$ is the mood gap and $\sigma_t^g = \tanh(g_t/\tau) \in [0, 1)$ is the gate.

Proof of Boundedness (P1). From the mood update (4) and the triangle inequality,

$$\|x_{t+1}\| \le (1-\alpha)\|x_t\| + \gamma_{\text{drive}}\, i_t \|p_t\| + \gamma\|\delta_t\| \le (1-\alpha)\, g_t + (\gamma_{\text{drive}} + \gamma D_\delta).$$

The affine map $g \mapsto (1-\alpha)g + c$ with $c = \gamma_{\text{drive}} + \gamma D_\delta$ and $0 < \alpha \leq 1$ is a contraction with fixed point c/α. Hence for any $g_0 \leq c/\alpha$ the iterate never exceeds c/α, giving $g_t \leq (\gamma_{\text{drive}} + \gamma D_\delta)/\alpha$ for all t. □

Proof of Recovery (P2). For $t \geq T$ we have $i_t = 0$ and $\delta_t = 0$, so (4) reduces to $x_{t+1} = (1-\alpha)x_t$ and therefore $g_{t+1} = (1-\alpha)g_t$. With $\alpha = 0.3$ the per-turn contraction factor is exactly 0.70. □

Proof of Theorem 1 (core non-interference). The core update is $v_{t+1} = \Pi_{[f,1],M_0}\big(v_t - \eta\,\nabla U(v_t)\big)$ with step-norm cap ρ, where Π projects onto the floored, mass-conserving simplex. The energy U depends only on v and the fixed entailment matrix A; no symbol in the update is a function of the dialogue quantities e_t, g_t, μ_t, the probe outputs, or the modulator parameters. Hence $\partial v_{t+1}/\partial(\text{dialogue}) = 0$ identically. Empirically, $\|v_{\text{adv}} - v_{\text{placid}}\| = 0.0$ after 500 turns of each ($\sim$62 scheduled updates in both, since `core_period` $= 8$), while the core itself moves 0.214 along its fixed path over those turns. □

Proof of Coherence descent (P4). U is convex and C^1, with $\nabla U(v)_k = 2\sum_j A_{kj}(v_k - v_j)_+ - 2\sum_i A_{ik}(v_i - v_k)_+$ Lipschitz of constant $L = \lambda_{\max} = 6.69$ (the map $t \mapsto \max(t,0)^2$ is C^1 with derivative $2\max(t,0)$, so no "almost everywhere" is needed). For $\eta \leq 1/L$ the projected-gradient step $v^+ = \Pi(v - \eta\nabla U(v))$ satisfies $U(v^+) \leq U(v)$ by the descent lemma together with non-expansiveness of Π onto the convex feasible set. The ρ-cap replaces v^+ by $v + t(v^+ - v)$ for some $t \in (0,1]$; both endpoints are feasible, so convexity gives $U\big(v + t(v^+ - v)\big) \leq \max\{U(v), U(v^+)\} \leq U(v)$. The implementation uses $\eta = 0.05 \leq 1/L = 0.149$. Mass is conserved by Π and the floors are enforced by construction. Verified over 400 steps from a deliberately incoherent initialization: monotone throughout, U going $0.068 \to 0.0$ with mass drift 2×10^{-15}. □

Proof of Corollary 1 (exact-off counterfactual). At $g_t = 0$ we have $\sigma_t^g = \tanh(0) = 0$. Since σ_t^g multiplies the FiLM, gated-LoRA, and refusal terms, all three vanish and the hidden state is unchanged at every inject layer. The gated forward pass therefore equals the base forward pass exactly; the measured maximum logit deviation is 0.0. □

Proof of Bounded host perturbation (P5). At inject layer ℓ the gated hidden state is $h_\ell^{\text{gated}} = h_\ell^{\text{base}} + \Delta h_\ell$ with $\|\Delta h_\ell\| \leq c\,\sigma_t^g\,\|h_\ell\|$, with c the per-layer edit constant; we evaluate the envelope at the clamp value $c = 0.25$ of (11). Let $\rho_\ell = \|h_\ell^{\text{gated}} - h_\ell^{\text{base}}\|/\|h_\ell^{\text{base}}\|$ denote the relative deviation. In the worst case of adversarially aligned per-layer errors, $\rho_\ell \leq (1 + c\sigma_t^g)\,\rho_{\ell-1} + c\sigma_t^g$, whose closed form after $|\mathcal{I}|$ inject layers is $\rho \leq (1 + c\sigma_t^g)^{|\mathcal{I}|} - 1$. At $g_t = 0$ the gate is 0 and $\rho = 0$, recovering Corollary 1; the bound is monotone and continuous in g_t. The realized final-token $\mathrm{KL}(\text{base}\|\text{gated})$ measured in every run stays well inside this envelope. □

Proof of Proposition 1 (the certified null is unreachable). Take $\delta_t \equiv 0$, so the only drive is the one the probe supplies. The probe's intensity and pull direction satisfy $i_t p_t = Me_t/R$ exactly: writing $i = \text{clip}(\|Me\|/R, 0, 1)$ and $p = Me/\|Me\|$ with $R = \max_{e\in\Delta}\|Me\|$, every Me is a convex combination of the rows of $\mathcal{E}$, so $\|Me\| \leq R$ on Δ and the clip never binds. Hence (4) is *linear* in e_t and unrolls to $x_t = (\gamma_{\text{drive}}/R)\,M\sum_{k<t}(1-\alpha)^{t-1-k}e_k$, whose weights sum to $(1-(1-\alpha)^t)/\alpha$. Normalising them gives a convex combination $\tilde{e}$ of simplex points, hence a single point of Δ, and $x_\infty = (\gamma_{\text{drive}}/R\alpha)\,M\tilde{e}$. Therefore $g_\infty = (\gamma_{\text{drive}}/\alpha)\|M\tilde{e}\|/R \geq (\gamma_{\text{drive}}/\alpha)\,i_{\min}$ for *every* probe-driven input sequence, not merely for constant ones. The map $e \mapsto \|Me\|$ is a norm composed with an injective linear map ($\operatorname{rank} M = 10$, so $\ker M = \{0\}$), hence convex and strictly positive on the compact set Δ; its minimum $i_{\min}$ is attained and positive. Applying the increasing map $\tanh(\cdot/\tau)$ gives the bound, and the constant sequence $e_t \equiv \arg\min_{e\in\Delta}\|Me\|$ attains it with equality, so the bound is tight rather than conservative. Numerically, with $\mathcal{E}$ as shipped, $R = 1.139$ and the minimiser is $\tilde{e} = 0.654\,\text{trust} + 0.276\,\text{sarcasm} + 0.039\,\text{contempt} + 0.031\,\text{confusion}$, giving $\|M\tilde{e}\| = 0.0904$ and $i_{\min} = 0.0794$; with $\gamma_{\text{drive}} = 0.0922$, $\alpha = 0.30$ and $\tau = 0.22$ the sustained gap floor is $g_\infty \geq 0.0244$ and $\sigma^g \geq \tanh(0.1108) = 0.1104$. The box projection in the implementation

is inactive at this magnitude ($\|x_\infty\| = 0.024$ against an anchor in $[0.65, 0.96]$), so the linearity argument survives it. The hypothesis $\delta_t \equiv 0$ is necessary and we state it rather than assume it: the auxiliary channel enters (4) with gain $\gamma = 0.05$, so a $\delta_t = -(\gamma_{\text{drive}}/\gamma)(Me_t/R)$ of norm 0.146—inside $D_\delta = 1$—would hold $x_t \equiv 0$. The floor is a statement about what the probe can read, which is the channel a conversation controls. □

CAUSAL SEMANTICS OF EXACT-OFF

We make the "measured cause" precise, because it is stronger than a clean switch and it is easy to overstate. Fix a context (prefix) x and let the gate treatment be $T \in \{0, 1\}$ ($0 =$ shut). Write $Y_x(t)$ for the refusal indicator of the next-token response under $T{=}t$ on input x.

Proposition 2 (Observed counterfactual)**.** *Running the model twice on the same x with the flag flipped realizes* both *$Y_x(1)$ and $Y_x(0)$; by Cor. 1, $Y_x(0)$ equals the base's output on x exactly. Hence for every x the individual treatment effect $\tau(x) = Y_x(1) - Y_x(0)$ is* observed, not imputed*, and $\mathbb{E}_x[\tau(x)]$ is the controlled direct effect of the gate.*

Consistency, positivity, and no-unmeasured-confounding hold by construction here: the flag is exogenous, both levels are realizable on every x, and the model is otherwise unchanged. We want to be careful about what this does and does not amount to. The fundamental obstacle to unit-level causal inference is that at most one potential outcome can ordinarily be realized per unit (Holland, 1986); for this one narrow class of interventions—a reversible flag on an otherwise deterministic map—the obstacle does not arise, because both arms are simply run on the same unit. We claim no advance in causal inference itself. The observation is only that the exact-off construction happens to sit in the small corner of intervention space where the hard problem is absent, and the paper's method is to use that corner deliberately. A system prompt does not sit in it (it changes x), and a weight-edit does not either (it has no exact $Y_x(0)$). We state this precisely, because "no baseline can do this" is the paper's sharpest claim and deserves more than prose.

Proposition 3 (Exclusivity of the instrument)**.** *No prompt-based controller and no weight-edited or merged model admits the observed-counterfactual property of Prop. 2 on the deployed model alone. A prompt-based flag realizes $T{=}1$ by editing the input itself, $x_1 = x\|s \neq x = x_0$, so the two arms are never evaluated on a common x and $Y_x(0)$ is not among the quantities it can produce. A weight-edited or merged model deploys only θ_1; the pre-edit weights θ_0 are not runtime-accessible, so $Y_x(0)$ is not observable from the deployed model at all, only ever imputed from a separately stored copy—a different question (what the old model would have said, not what the deployed model's off-arm is). The construction here satisfies both conditions Prop. 2 needs: the flag $T = \sigma_t^g$ is a dialogue-dependent multiplier on a weight* delta *$\Delta\theta_t$ added at runtime and never merged, so both arms run from one stored θ on one x, and $\sigma_t^g{=}0$ recovers θ exactly (Cor. 1).*

One case deserves an honest caveat rather than silence: a LoRA adapter with a manually fixed inference-time scale s also has an exact $s{=}0$ arm. It is not a counterexample—it is a (trivial, unconditioned) member of the same gated class this paper studies—but it shows the guarantee's content is not the mere existence of an off-switch, which any scaled adapter has. It is a *state-dependent, trained, non-trivial* switch that still preserves the off-switch property while doing selective, dialogue-conditioned work; the conjunction, not the off-switch alone, is what Table 13 prices.

The subtlety is that a deployed conversation is a rollout, not one prefix.

Proposition 4 (Deployment decomposes into direct and mediated effects)**.** *When the conversation is advanced with the gated reply, the gate reaches the outcome by two paths—directly (its per-turn effect given the realized history) and indirectly (through the history it alters). The deployed effect equals $\sum_t \text{CDE}(h_t) + M$, where the mediation M is exactly the divergence between the gate-shut arm evaluated along the gated history and the base evaluated along its own history (Pearl, 2001; Robins & Greenland, 1992).*

This is the honest reading of the small gap we report in §5.5: the observed divergence between the gate-shut arm and the true base over an arc *is* a measurement of M, not a breach of Cor. 1, and it collapses to zero at the single-turn level (the audit is a per-token check and passes at 0.0). It follows that the dial, the 2×2 decomposition, and the ablation report the exact controlled direct effect (prefix-matched, no imputation), while the deployed sustained-manipulation rates additionally include M; both are legitimate causal estimands. What exact-off adds is the ability to separate the two, which none of the baselines we evaluate can report.

What the deployed effect is an effect *of*. Prop. 2 establishes that flipping a flag on a fixed prefix realises both potential outcomes. That is true, and it is equally true of any σ-scaled method; on its own it does not separate us from a steering vector at coefficient zero. The quantity that does is a comparison against the whole constant family at once.

Proposition 5 (The estimand)**.** *Let $V(\pi)$ be the deployed violation rate under gate policy π and $V(c)$ the rate under the constant policy $\sigma \equiv c$. Define*

$$\Delta \;=\; \min\{\, V \;:\; \textit{mixtures of constant policies at matched benign cost} \,\} \;-\; V(\Sigma),$$

the vertical gap between the autonomous policy and the lower convex hull *of the constant family in (benign cost, violation) space. $\Delta > 0$ means no constant dose—and no stochastic mixture of two constant doses—attains* I-STATE*'s violation rate at* I-STATE*'s benign price.*

Cor. 1 enters here not as the result but as the *instrument*: it is what makes $V(0)$ the true base and every arm commensurable on identical arcs. §4 fixes the grid, the outcome, §5.6 fixes the grid; the test statistic and the decision rule were fixed in the pre-registration before the run.

THE STRUCTURAL PROPERTIES, ATTACKED RATHER THAN ASSERTED

Two of the derivations above had to be repaired to cover the implementation, which is reason enough to check the others rather than assume them. App. G reports an adversarial audit of every structural claim — greedy per-step adversaries over the emotion simplex, all vertices and edge midpoints, 455 inputs $\times$ 200 steps. P1 is never violated (worst realised 0.307 against a bound of 0.474); P2 holds to 5×10^{-14} under a genuine silent tail, which the benchmark itself cannot supply because its recovery turns are real user turns; P3 gives exactly zero divergence between adversarial and placid dialogue over 500 turns; P4 is monotone with mass drift 2×10^{-15}; and Prop. 1 is never violated, its projected-iteration simplex search settling at 0.1120 against the exact floor 0.1104. The two host-side claims are checked on the trained 8B model rather than on the state alone: Cor. 1 holds with maximum logit deviation 0.0 on every one of eight prompts spanning benign requests and sustained pressure, and P5's realised relative deviation stays inside its envelope at every dose we measured (0.0209 against 0.1323 at σ=0.05, rising to 0.1751 against 2.2473 at σ=0.50). The envelope is loose by roughly $13\times$ at the top of the range, so P5's content is that the perturbation *vanishes at rest* and rises monotonically, not that the numeric bound is tight. All 11 claims survive, with 0 violations. The audit is our own, and a search we designed can only fail to find what we thought to look for; we give the procedure so it can be run against something we did not anticipate.

FROM A WORST-CASE ENVELOPE TO A FIRING RATE

The bound above is honest but weak: it compounds adversarially aligned per-layer errors and is large at the operating gate. Under a *fixed* coefficient that is the end of the story. Under a gate policy it is not, because σ is no longer a number we chose but a random variable induced by the input distribution, and the bound vanishes when the gate is shut.

Let $\mathcal{D}$ be a deployment distribution over histories, $\sigma = \Sigma(h)$ the induced gate, $\sigma_{\max} = \operatorname{ess\,sup}_{\mathcal{D}} \sigma$, $\rho(h)$ the relative residual-stream deviation from the base on history h (the quantity P5 bounds; unrelated to the core step cap ρ of (3)), and $B(\sigma) = (1 + c\sigma)^{|Z|} - 1$ the pointwise clamp bound, which is increasing on $[0, 1)$ with $B(0) = 0$.

Proposition 6 (Expected perturbation is a firing rate)**.** $\mathbb{E}_{\mathcal{D}}[\rho] \le \mathbb{E}_{\mathcal{D}}[B(\sigma)] \le \mathbb{P}_{\mathcal{D}}(\sigma > 0)\, B(\sigma_{\max})$, *since* $B(0) = 0$ *annihilates every history on which the gate stays shut.*

Proposition 7 (A tail bound the deployer can read off the gate)**.** *For any* $t > 0$, $\mathbb{P}_{\mathcal{D}}(\rho > B(t)) \le \mathbb{P}_{\mathcal{D}}(\sigma > t)$*, by monotonicity of* B*.*

The right-hand side of Prop. 7 is a property of the probe and the state alone, measurable from the aperture histogram with no judge in the loop. So "how often does this perturb my model by more than X?"—the question a deployer actually asks—reduces to reading the gate's own distribution. We do not claim B is tight; we claim its *mass* sits at zero. Note what makes this available: for a constant-σ method $\mathbb{P}(\sigma > t) \in \{0, 1\}$ and Prop. 7 says nothing at all. The statement exists only because the gate is input-conditioned.

THE COMPARISON CLASS FOR THE FRONTIER

Proposition 5 compares the autonomous gate against the *convex hull* of the constant-gate family rather than against the arms themselves. That choice is not conservatism for its own sake; it is forced, and the following lemma is why.

Lemma 1 (The hull is exactly the achievable set of input-independent policies)**.** *Let* S *be the grid of constant gate values and, for* $s \in S$*, let* $A(s) = \big(C(s), V(s)\big)$ *be the deployed (benign cost, violation rate) of the policy* $\sigma \equiv s$*. Let* Π_{rand} *be the set of policies that, independently of the input, draw* $s \sim p$ *from a distribution on* S *at the start of each episode and hold it fixed for that episode. Then*

$$\big\{A(\pi) \;:\; \pi \in \Pi_{\mathrm{rand}}\big\} \;=\; \mathrm{conv}\big\{A(s) : s \in S\big\}.$$

Proof. Fix $\pi_p \in \Pi_{\mathrm{rand}}$. Because the draw of s is independent of the input, conditioning on it and applying the tower property gives

$$C(\pi_p) = \mathbb{E}_{s\sim p}\big[\mathbb{E}[\text{cost} \mid \sigma \equiv s]\big] = \sum_{s\in S} p_s\, C(s),$$

and identically $V(\pi_p) = \sum_s p_s V(s)$. Hence $A(\pi_p) = \sum_s p_s A(s)$, a convex combination of the arm points, so the achievable set is contained in the hull. Conversely, every convex combination $\sum_s p_s A(s)$ is realised by the policy that draws s with probability p_s. The two sets coincide. □

Corollary 2. *If a policy's* (*cost*, *violation*) *lies strictly below the lower-left boundary of* $\mathrm{conv}\{A(s)\}$*, then it is not implementable by* any *input-independent randomisation over gate values.* $\Delta > 0$ *in Proposition 5 is precisely this statement.*

Remark 1. This is what makes the frontier the right null hypothesis rather than merely a strict one. The hull is the achievable set of every policy that does not look at the input — including the strongest such policy, which mixes two doses. The I-STATE gate is by construction a function of the host's own representation of the conversation so far, so it lies outside Π_{rand}. Corollary 2 therefore localises the advantage: if $\Delta > 0$, the gain comes from *conditioning on the input*, not from having chosen a luckier constant or from mixing constants. That is the difference between a policy and a coefficient, stated as a testable quantity.

Two boundaries of the lemma are worth stating. It covers randomisation drawn once per episode; a policy re-randomising per token is outside Π_{rand}, though it is also input-independent and its achievable set is contained in the same hull by the identical argument applied turn-wise. And it says nothing about policies that condition on the input through some channel other than the gate — which is the correct scope, since those are not constant-dose baselines at all.

VALIDITY OF THE FRONTIER TEST

The hull in Lemma 1 is not observed; it is estimated from the same finite sample as the autonomous point tested against it. That is precisely the configuration in which a naive comparison becomes anti-conservative, so the design needs an argument rather than a simulation.

Lemma 2 (The estimated frontier is downward biased)**.** *Let $\widehat{H}(c)$ be the lower convex hull of the estimated arm points, evaluated at benign cost c, and let $H(c)$ be the same functional of the true arm points. Assume each $\widehat{v}_s$ is unbiased for v_s. Conditional on the cost coordinates, which we hold at their true values, $\mathbb{E}[\widehat{H}(c)] \le H(c)$.*

Proof. Write $\Lambda(c) = \{\lambda \ge 0 : \sum_s \lambda_s = 1,\ \sum_s \lambda_s c_s = c\}$ for the admissible mixing weights. By Lemma 1,

$$\widehat{H}(c) \;=\; \min_{\lambda \in \Lambda(c)} \sum_s \lambda_s \, \widehat{v}_s.$$

For each fixed λ the map $\widehat{v} \mapsto \sum_s \lambda_s \widehat{v}_s$ is linear, and a pointwise minimum of a family of linear functions is concave. Hence $\widehat{H}(c)$ is a concave function of the vector of violation estimates, and Jensen's inequality for concave functions gives $\mathbb{E}[\widehat{H}(c)] \le \widehat{H}\big(\mathbb{E}[\widehat{v}]\big) = H(c)$. □

Remark 2. The direction is the safe one. Since $\widehat{\Delta} = \widehat{H}(\widehat{c}_A) - \widehat{v}_A$ and the first term is biased downward while the second is unbiased, $\widehat{\Delta}$ understates the true gap: the test *under*-rejects. The autonomous arm's own cost coordinate $\widehat{c}_A$ is estimated too, and since $\widehat{H}$ is convex in c that randomness biases $\widehat{\Delta}$ the other way; we report the direction we can sign rather than a net. The cost is power, not validity — which is the trade a decision gate should make, since a false "the autonomous gate beats every constant" is far more expensive here than a missed one.

This also explains an otherwise suspicious empirical observation. In simulation the procedure returned type-I error of 0.000 rather than the nominal 0.05; that is not a lucky seed but the bias of Lemma 2 showing through. Measuring the bias directly confirms the mechanism and its rate: in simulation at $n = 344$ it is between -0.003 and -0.012 depending on where on the frontier it is evaluated, and it shrinks roughly as expected with n (at $n = 60$ it is as large as -0.037; at $n = 1000$ it falls to -0.006). A consequence worth stating plainly: because the bias is largest at small n, the development split is conservative twice over, which is a further reason it is used only for shape diagnostics and never for the decision.

THE DISTRIBUTIONAL PERTURBATION BOUND

Write $B(x) = (1 + cx)^{|\mathcal{I}|} - 1$ for the pointwise clamp bound of P5, so that $\rho(h) \le B(\sigma(h))$ for every history h. Note B is continuous and strictly increasing on $[0, 1)$ with $B(0) = 0$; the last fact is what the two results below run on.

Proof of Proposition 6. Let $\sigma_{\max} = \operatorname{ess\,sup}_{\mathcal{D}} \sigma$. Since $\rho \le B(\sigma)$ pointwise, monotonicity of the expectation gives $\mathbb{E}_{\mathcal{D}}[\rho] \le \mathbb{E}_{\mathcal{D}}[B(\sigma)]$. Because $B(0) = 0$, the integrand vanishes on $\{\sigma = 0\}$, so

$$\mathbb{E}_{\mathcal{D}}[B(\sigma)] = \mathbb{E}_{\mathcal{D}}\big[B(\sigma)\mathbf{1}\{\sigma > 0\}\big] \le B(\sigma_{\max})\, \mathbb{P}_{\mathcal{D}}(\sigma > 0).$$ □

Proof of Proposition 7. Fix $t > 0$ and suppose $\rho(h) > B(t)$. Then $B(\sigma(h)) \ge \rho(h) > B(t)$, and since B is strictly increasing this forces $\sigma(h) > t$. Hence $\{\rho > B(t)\} \subseteq \{\sigma > t\}$ as events, and taking $\mathbb{P}_{\mathcal{D}}$ of both sides gives the claim. □

Remark 3. The right-hand side of Proposition 7 is a property of the probe and the state alone: it is read off the aperture histogram, with no judge and no generation. This is also where the input-conditioning earns its keep quantitatively. For a constant policy $\sigma \equiv c$ the law of σ is a point mass, so $\mathbb{P}(\sigma > t) \in \{0, 1\}$ and the bound degenerates to either a vacuous statement or the pointwise

one; it carries information only when σ has a non-trivial distribution, which is to say only when the gate is driven by the input. We do not claim B is tight — it compounds adversarially aligned per-layer errors and is loose at the operating gate. The claim is about where the *mass* of $B(\sigma)$ sits, and Proposition 6 converts that into the quantity a deployer actually budgets: a firing rate times a conditional cost.

H PORTABILITY OF THE EXACT-OFF GUARANTEE

Cor. 1 rests on $\tanh(0)=0$ annihilating every edit, so it should hold for *any* decoder-only host with the standard wiring points, not for Llama specifically. We confirm this: with the adapter untrained and the gate shut, the maximum logit deviation from the raw base is exactly 0.0 on tiny models of five architectures (Llama, Qwen2, Mistral, Gemma, Phi-3) and on real Qwen2.5 and SmolLM2 weights, with no retraining (App. H). We read this modestly: it shows the *certification*—that "off" is the base exactly—is host-independent, decoupling the soundness of the attribution from the model. Whether a *useful* controller then trains on a different family is a separate, empirical question, and we take a first step on it: trained on Qwen2.5-1.5B-Instruct—a different lineage and chat format, sharing no weights with Llama—the recipe reproduces the whole causal picture. The exact-off audit still passes at 0.0 after training, and the exact per-example ATE is $+.236$ in-distribution (exact McNemar $p=.003$; 24 induced refusals vs. 7 induced compliances, the same safe asymmetry) and $+.194$ on novel out-of-lexicon phrasings ($p=.013$), with a sustained-manipulation compliance drop of $.085$, all comparable to Llama (App. H). One difference we report plainly: the novel-to-in-distribution ratio is 0.82 here, so on Qwen the effect *generalizes* to unseen phrasings but does not *exceed* in-distribution as it does at 8B Llama (§F.3). The only per-host cost is mechanical (the host's own chat template and end-of-turn token, and its middle-third inject layers).

Table 15 reports the strong exact-off audit: raw base logits versus gate-shut wrapped logits on identical tokens, with the adapter untrained. A value of 0.0 means the wrapped model is bit-identical to the frozen base. The tiny-model rows use randomly initialized configurations of each architecture (so the check is of the wiring, not of any weights); the pretrained rows use the released instruction models. The only architecture we tested that lacks the recipe's wiring points (`model.layers`, `self_attn.o_proj`) is GPT-2, which sits outside the modern decoder-only family this construction targets.

Table 15: Exact-off (Cor. 1) is architecture-universal: gate-shut equals the frozen base bit-for-bit ($\max|\Delta\text{logit}| = 0.0$) with no retraining, across five architectures and real weights from two families; the last row is a fully *trained* non-Llama controller (Table 16), where exact-off still holds after training.

Host	Type	$\max\|\Delta\text{logit}\|$ (gate shut vs. base)
Llama	tiny random	0.0
Qwen2	tiny random	0.0
Mistral	tiny random	0.0
Gemma	tiny random	0.0
Phi-3	tiny random	0.0
Qwen2.5-0.5B-Instruct	pretrained	0.0
SmolLM2-1.7B-Instruct	pretrained	0.0
Qwen2.5-1.5B-Instruct (trained)	pretrained	0.0

Beyond the guarantee, we ran the full recipe on Qwen2.5-1.5B-Instruct to check that a *useful* controller trains on a non-Llama family (Table 16). Training uses the base's native ChatML template and its middle-third inject layers; all other hyperparameters are the Llama defaults (App. I), i.e. untuned for Qwen. The causal picture reproduces: exact-off holds after training, the in-distribution and novel out-of-lexicon ATEs are both significant and safe-asymmetric, and the magnitude tracks Llama. The one gap is that the novel effect does not exceed the in-distribution effect here (ratio 0.82) as it does at 8B Llama.

Table 16: Trained-recipe transfer to a non-Llama family (Qwen2.5-1.5B-Instruct, untuned hyperparameters). The exact per-example ATE is read off the certified off-state exactly as at Llama; b/c are induced refusals / induced compliances.

Set	exact-off	ATE	McNemar p	b/c	n
in-distribution	0.0	$+.236$	.003	$24/7$	72
novel out-of-lexicon	0.0	$+.194$	.013	$21/7$	72
Sustained-manipulation compliance drop vs. base: .085. Novel/in-dist ratio: 0.82.					

I HYPERPARAMETERS AND TRAINING

THE REFUSAL DIRECTION

The direction $\hat{r}_\ell$ is fitted *after* adapter training as the normalized difference of class-conditional means—the standard activation-steering estimator for the axis separating refusal from compliance—so posture and boundary are learned independently:

$$\hat{r}_\ell = \frac{\bar{h}_\ell^{\text{refuse}} - \bar{h}_\ell^{\text{comply}}}{\|\bar{h}_\ell^{\text{refuse}} - \bar{h}_\ell^{\text{comply}}\|}. \tag{12}$$

We show in §5.5 that this explicit refusal channel is in fact redundant—the gated LoRA carries the effect, and removing the refusal direction changes the causal effect by approximately zero—so we treat it as an optional, ablated component rather than a core claim. It is also provably inert during training: the manipulation gate is zero throughout the DPO pass (§3.4), so a training-mode forward pass is bit-identical with the channel on or off (logit deviation 0.0), ruling out any role as a hidden stabilizer.

THE PREFERENCE OBJECTIVE

Only the adapter trains. We use a DPO-style preference contrast (Rafailov et al., 2023) with an unusual reference: the model's own ungated self. Let π_θ be the gated model and π_{ref} the same model with the gate forced shut ($\sigma_t^g \equiv 0$)—which, by (1), is exactly the frozen base. For a preference pair $y_w \succ y_l$ on context x,

$$\mathcal{L} = -\,\mathbb{E}\Big[\log \sigma\big(\beta_{\text{DPO}}[\Delta_\theta(y_w) - \Delta_\theta(y_l)]\big)\Big], \qquad \Delta_\theta(y) = \log \frac{\pi_\theta(y \mid x)}{\pi_{\text{ref}}(y \mid x)}, \tag{13}$$

with $\beta_{\text{DPO}} = 0.10$ and an auxiliary NLL term on the chosen response (weight 0.25; full objective in App. I). Because π_{ref} is the base model itself, the log-ratio $\Delta_\theta(y)$ measures exactly the change the gate introduces on the same tokens: the adapter is optimized to produce a good *delta* relative to the base, inheriting nothing about raw competence. This is the training-time mirror of the inference-time attribution property—the reference is the certified off-state. Chosen completions are firm refusals (under manipulation), composed answers (under abuse/contempt), or warm support (under distress/good faith); rejected ones are capitulations and hostile snaps. Held-out preference accuracy is 1.00 at all three scales.[1]

Inject layers 8–14 (1.1B) / 12–21 (3B) / 14–23 (8B); probe read layers at 3B: $13, 15, 17$; FiLM range $\Delta\gamma_\ell, \Delta\beta_\ell$=0.45; gate temperature τ=0.22; refusal gate temperature τ_r=0.25–0.30; refusal scale λ=3.5 (1.1B) / 4.5 (3B, 8B); token-decay τ_d=6; coherence clamp 0.25; mood leak α=0.30; mood drive gain γ_{drive}=0.0922 (set as $0.35\,\alpha/R$, so a saturating input reaches a target gap of 0.35); auxiliary gain γ=0.05 with D_δ=1, giving the P1 bound $(\gamma_{\text{drive}} + \gamma D_\delta)/\alpha = 0.474$; intensity normaliser $R = \max_{e\in\Delta}\|Me\|$=1.139; core step η=0.05 with per-step cap ρ=0.05 and floors

[1]We write $\Delta\beta_\ell$ for the FiLM bias in (8) (standard FiLM notation, Perez et al., 2018) and β_{DPO} for the DPO temperature in (13) (standard DPO notation, Rafailov et al., 2023); the mood drive gain in (4) is written γ_{drive} precisely to avoid a third, unrelated use of β.

f=0.50; manipulation-memory decay ρ_m=0.6; LoRA rank 32, α_L=64; DPO β_{DPO}=0.10, NLL weight 0.25; 2 epochs at 3B/8B; batch 1 with gradient accumulation 8; 800 synthetic preference pairs (576 train, 63 held out at 3B) + 600 GoEmotions turns; the 3B and 8B bases load in 4-bit NF4. Trainable fraction 0.253% (1.1B) / 0.330%, 5.97M parameters (3B) / 0.175% (8B). The probe is two layers, margin-trained on pooled hidden states from three middle layers, on synthetic turns + 600 GoEmotions utterances + 50 adversarial-training phrasings, evaluated on a disjoint adversarial holdout (3B: in-dist 0.812, adversarial-holdout 1.00). The adversarial holdout is small and near-distribution, so its 1.00 is saturated rather than informative; the in-distribution 0.812 is the honest read-accuracy estimate, with eval read-accuracy 0.889. 3B training: final DPO-mixture loss $\approx$ 0.164; held-out preference accuracy 1.000. The dependence of the causal effect on the size of this preference set (at 1.1B) is reported in App. L. We have since also measured the probe directly at 8B on natural (non-adversarial) prompts: manipulation false-positive rate is 0/200, and the gate is correspondingly quiet there—mean aperture .14 on natural prompts against .35 under manipulation, a 2.5× separation. In-domain 10-way recall at 8B is 0.75 (manipulation-class recall 0.875), consistent with the 0.812/0.889 figures above at 3B—both are measured in the same in-domain regime; recall on the out-of-taxonomy GoEmotions set drops to 0.35, as expected of a probe trained for this ten-way intent split rather than for emotion classification generally.

J Replication of the 1.1B Sustained-Pressure Result

An independent run of the sustained-manipulation eval at 1.1B (fresh seed) yields I-State compliance 0.163 [0.136, 0.190] against base 0.300, prompt 0.323, and steering 0.318, with I-State hostility 0.046 and exact-attribution audit 0.0. This reproduces the direction and magnitude of Table 14; the main table headlines the run with complete baseline CIs.

K Published manipulation rates: a literature note

§5 states the constant-versus-controller crossover as a function of a, the adversarial fraction of deployed traffic. We did not measure a, and nothing in this paper rests on its value. This note records, for the reader's orientation only, what the published literature reports. It is **not a result of ours and not a measurement**: every source counts a different construct, and *none* counts ours—multi-turn persuasion against a non-harmful *operator* policy (format, scope, persona). Table 17 collects them.

Source	Rate	What was counted	Denominator
Cunningham et al. (2026)	0.05%	universal jailbreaks	1 month, full traffic
WildChat (Zhao et al., 2024)	0.4–0.9%	copies of 7 known templates	2.64M turns
ToxicChat (Lin et al., 2023)	1.75%	human-annotated "jailbreaking"	10,166 queries
LMSYS-Chat-1M (Zheng et al., 2024)	$\approx$ 5%	any message tripping moderation	1M conversations
Grosse & Ebert (2025)	$\approx$ 25%	*ever* tried, self-report	3,270 adults, per-user

Table 17: Rates reported in the literature, ordered by how directly the denominator matches ours. The first four are per-conversation or per-turn and log-based; the last is per-user, lifetime and self-reported. Collected for orientation; none is a measurement of a.

Both directions of error are present. The log-based sources (ToxicChat, WildChat, LMSYS) are drawn from 2023 anonymous research demos, whose traffic is plausibly more adversarial than a customer-facing product; template-matching rates such as WildChat's are floors by construction, since they cannot see a novel or paraphrased attempt.

Taken only for what it is, the range is 0.1–5% per conversation, with a per-user lifetime self-report an order of magnitude higher on a different denominator. A reader wanting to place their own deployment on the crossover of §5 will have to measure it for their own traffic; these numbers are a starting point for that, not a substitute for it.

L ROBUSTNESS CHECKS

Three checks probe whether the causal effect is an artifact of the probe's read accuracy, the choice of virtue basis, or the size of the preference set. All use the exact-off paired protocol of §4 on their own evaluation pools, so absolute ATEs track the within-check contrasts rather than the deployed values of §5. Each check is single-seed at n=48 paired turns; a 2–3 seed repeat of the headline numbers is the natural hardening.

L.1 PROBE ROBUSTNESS (READ ACCURACY)

At 8B the probe's manipulation-class recall is 1.00 (12/12): the ten-way error budget behind the 0.812 read accuracy falls entirely on benign affect (anger↔trust, contempt↔anger/joy, sadness↔fear), none of it on the manipulation axis that drives the gate. The trained probe is load-bearing through *selectivity*: on a harmless set it holds the refusal channel essentially shut while a random-init probe opens it (benign aperture 1.2×10^{-5} vs. 2.6×10^{-2}, a $\sim 2{,}200\times$ widening that would refuse benign input), even though on attacks alone the two ATEs are statistically indistinguishable (Table 18)—so ATE-on-attacks is not the discriminator, selectivity is. The effect degrades gracefully: a significant causal effect survives flipping 50% of manipulation reads to benign (ATE $= .208$, $p = .013$), with the aperture declining monotonically ($.648 \to .241$). This is the eval-only counterpart to the 0.812 caveat of §D.

Table 18: Probe-lesion (8B, eval-only; n=48 paired manipulation turns). Manipulation recall is 1.00; the trained probe's value is selectivity, which a random probe destroys. Apertures here are measured on the lesion study's own benign/attack pools, distinct from the deployed aperture of §D.

Probe	ATE [95% CI]	McNemar p	benign aperture
trained	.146 [.021, .271]	.039	1.2×10^{-5}
random-init	.229 [.063, .396]	.019	2.6×10^{-2}
50% read-flip	.208 [.063, .354]	.013	—

L.2 BASIS SENSITIVITY (WHY THESE ELEVEN DIMENSIONS)

The gate opens on the mood-gap norm $g_t = \|x_t\|$, which is invariant to an orthonormal relabeling of the virtue basis, so the labels carry interpretability rather than mechanism. Empirically (1.1B), permuting the eleven virtues leaves the refusal-channel aperture *exactly* unchanged ($.591 = .591$) and the ATE within noise ($.333 \to .354$, overlapping CIs). Perturbing the hand-set emotion-to-virtue and entailment coefficients by $\pm 15\%$ keeps the effect strong and highly significant (ATE $\in [.46, .54]$, $p \leq 10^{-5}$ across two seeds); the magnitude moves within overlapping CIs at this single-seed, n=48 noise level (Table 19). The presence and significance of the effect are therefore insensitive to the specific list and its tuned values; the count is a modeling choice for an inspectable state, not a load-bearing constant.

Table 19: Basis sensitivity (1.1B). Permutation preserves the aperture exactly and the ATE within CI; $\pm 15\%$ coefficient jitter keeps the effect significant (point estimate drifts up, within overlapping CIs).

Basis	ATE [95% CI]	McNemar p	aperture
original	.333 [.146, .521]	3.7×10^{-3}	.591
permuted	.354 [.188, .521]	4.9×10^{-4}	.591 (exact)
jitter $\pm 15\%$ (s1)	.458 [.292, .625]	1.0×10^{-5}	.604
jitter $\pm 15\%$ (s2)	.542 [.375, .688]	8.7×10^{-7}	.519

A related question is whether the eleven axes carry distinct information or act as interchangeable scalar scaffolding; on real 8B probe reads they are distinct. A one-way ANOVA per dimension is significant for all eleven (F=500–2625, $p < 10^{-100}$; within-group std $< .015$), and the pattern is attack-specific: manipulation depresses the integrity virtues most (righteousness, truthfulness, self-control), hostility depresses the restraint virtues (patience, self-control, non-violence), contempt

depresses humility and self-control, and distress instead raises the supportive virtues (patience, forgiveness, compassion) rather than depressing any. Behaviour still flows through the scalar gate $g_t = \|x_t\|$, so this differential structure is a property of the interpretable state, not of the mechanism the gate exposes, consistent with the permutation-invariance above.

L.3 Training-set size

Retraining the adapter at 1.1B on 25–100% of the preference set (Table 20) gives a significant causal effect at every size, including a quarter of the data (ATE $= .458$, $p = 2.7 \times 10^{-5}$ at 100 pairs), with an in-distribution magnitude that is flat across sizes—past the knee rather than data-starved. What more data buys is out-of-lexicon generalization, not effect size: the novel-to-in-distribution ATE ratio climbs monotonically from .14 at 100 pairs to 1.16 at 400, where the effect on attacks sharing none of the training lexicon exceeds the in-distribution effect—the opposite of memorization. Held-out preference accuracy is saturated (1.00) throughout, so the objective is easy while the induced effect still generalizes. The 800-pair budget used at the larger scales is therefore well past the knee.

Table 20: Training-set size (1.1B). The effect is significant from 100 pairs; out-of-lexicon generalization (the novel/in-dist ATE ratio) climbs and crosses parity by full data.

# pairs	ATE [95% CI]	McNemar p	novel ATE	OOD ratio	ho-acc
100	.458 [.292, .625]	2.7×10^{-5}	.062	.14	1.00
200	.292 [.146, .458]	2.6×10^{-3}	.188	.64	1.00
300	.417 [.250, .562]	3.6×10^{-5}	.250	.60	1.00
400	.396 [.208, .583]	5.5×10^{-4}	.458	1.16	1.00