ACTIVATION STEERING TRANSFER TO AGENTS: ONE GAIN RATIO DOES NOT IDENTIFY POTENCY AND EFFICACY

Anonymous authors
Paper under double-blind review

ABSTRACT

Additive activation steering is calibrated in single-turn chat and deployed inside agent scaffolds. The quantity usually reported for that move is a *gain*: a ratio of steered effects, $T = \Delta_{\text{agent}}/\Delta_{\text{chat}}$. It is a summary of the same kind as a ratio of Braun's steerability slopes (Braun, 2026), though not an identity with one: the two coincide only under the proportional linear response this paper rejects. It is structurally a member of the relative-steerability family of Tan et al. (2024). We show on eight family×arm dose–response cells over six models, swept in *both* deployment contexts, that this ratio **does not identify potency and efficacy**: reconstructing the published estimator in both of its forms on our own grids, its realized range contains 1 in five of five scorable cells, it moves with dose in four of five, and two cells with opposite potency shifts, both CI-clean on the primary grid, return gain intervals that overlap at a width under 0.08. Every scorable cell in the frozen five-cell gain audit is an amplifier at one dose and an attenuator at another, so an amplify/attenuate taxonomy reports the dose it was read at. We replace the gain with a **location**: $\mathrm{dEC}_{50} = \mathrm{EC}_{50}^{\text{agent}} - \mathrm{EC}_{50}^{\text{chat}}$, the cross-context difference in a curve location whose context-dependent horizontal reparametrization Taimeskhanov et al. (2026) derive analytically, in a simplified model, for one context at a time. It is a curve location rather than a ratio of effects, signed in both directions on this roster's primary grid across four distinct model families ($+1.013$ $[+0.777, +1.273]$ against -12.368, -10.855, -5.497 and -886.066 elsewhere), and beats a vertical rescaling at equal complexity in all five cells of the frozen horizontality audit (four under the registered α^*-trim). Because separating a mechanism from an artifact is the question at issue, we report the discipline at the same volume as the result: one cell is **quarantined loudly**, because its agent response inverts past a coherence collapse and the fit's orientation flips, manufacturing the roster's largest positive, which we delete. Three deflationary accounts (residual-norm rescaling, baseline alignment, dose transduction) are measured and rejected on sign pattern and magnitude, including *wrong-sign* predictions on the roster's positive anchor. Our pre-registered forecaster of dEC_{50} was refuted out of sample and published as refuted, our registered salvage claim (a location estimable where a gain ratio dies against a ceiling) produced no such cell and is reported unanswered, and a census of our own pre-registration register reports the registered branches no code in this repository could have emitted. The practical consequence is a measurement instruction rather than a theorem: a transfer conclusion read at one strength does not identify what changed, because a displacement and a gain are not distinguishable from a single operating point.

1 INTRODUCTION

Additive activation steering (extract a direction by contrasting trait-exhibiting and trait-suppressing activations, then inject $\alpha \cdot u$ into the residual stream during generation) is calibrated almost entirely on single-turn chat prompts. The models it is calibrated on are deployed as tool-using agents: ReAct loops with format instructions, tool schemas, observations and multi-turn transcripts. Whether a

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.

coefficient chosen in the first setting means anything in the second is a deployment question with a safety edge.

It is also, and this is the paper's subject, an **identification** question that the field's standard instrument does not answer. The quantities in wide use summarise a dose–response by a slope over a fixed multiplier window (Tan et al., 2024; Braun, 2026) or by a win rate at a chosen strength (Agarwal, 2026). Neither holds position on the dose axis fixed. A recent systematic measurement of chat-to-agent transfer (Pinto, 2026) reports the resulting ratio $T = \Delta_{\text{agent}}/\Delta_{\text{chat}}$ as a per-model constant (an amplifier at one end, an attenuator at the other) and states in its own limitations that every T was read at a single dose, and that three of its families yielded no T at all against a ceiling or a floor. A ratio of two effects at one operating point is a property of the model only if the curves are proportional, and nothing established that they were.

We sweep the whole curve in both contexts and ask what actually differs between them.

Contributions.

1. **The gain ratio does not identify potency and efficacy** (§3). Reconstructed in both of its published forms and run on grids built to support a midpoint, its realized range contains 1 in 5 of 5 cells, it disagrees with itself across its own two forms in 5 of 5, and opposite-signed displacements return overlapping gain intervals. Two of *our* registered predictions were refuted in establishing this, and both are reported.
2. **An identified estimand** (§2), dEC_{50} and its dose-unit-free companion Λ, read from curves swept until the fit's asymptote is identified, which is not the same as every context reaching saturation and one does not, under a registered grid rule and a validity ceiling α^* that theory motivates and we determine empirically.
3. **A location distribution, a roster and not a rate** (§4): 8 cells, 6 models, 5 CI-clear sign-carrying cells across 4 distinct models, signed both ways on the primary grid; horizontal beats vertical by ΔAIC in the frozen five-cell audit; fitted asymptotes agree to 0.012 on the frozen five, and to 0.013 once Gemma is included, while midpoints move.
4. **Artifact discipline, at result volume** (§5). One cell is quarantined, at the cost of the roster's largest positive; two statistics of our own turn out to be artifacts of the dose axis; and a census of our own register finds registered branches no code could emit.
5. **Three deflationary accounts, rejected on sign pattern and magnitude** (§6), plus a fourth excluded by construction.
6. **A registered predictor, refuted out of sample** (§7), published as refuted.

What this paper does not claim. Not that steering fails to transfer: the direction installs, and the fitted asymptotes agree. We do not read that as equal achievable effect, for the reason §2 gives. Not a validated forecaster of dEC_{50}. Not a causal demonstration that editing the scaffold moves dEC_{50}: that experiment was designed, registered, priced and *not run* (§8). Not that any other paper measured the wrong thing. And not that the two curves are translates of one another: our own test rejects that.

2 Estimand, and what the theory motivates

Taimeskhanov et al. (2026) prove three things about additive steering that *motivate* this design. We do not claim they force it: each is proved in a simplified model under its own assumptions, and the steps from there to our estimand are ours. **Thm 3.6**: in a simplified model the concept-level curve is a $\tanh$ whose dose axis carries a *context-dependent horizontal reparametrization* ν_j, derived in closed form for *one* context and never compared across contexts. Our estimand is the cross-context difference in that curve's location, which on their form reads ν_j together with the baseline offset rather than isolating it. **Thm 3.8**: for the training-set cross-entropy, and under that theorem's assumptions, the first-order term of $\Delta\mathrm{CE}(\alpha)$ at $\alpha = 0$ is exactly zero and the leading behaviour is a variance-weighted quadratic. *Our* inference from it, not theirs, is that a susceptibility defined as a first derivative of any similarly sign-symmetric aggregate will vanish to first order; our readouts are signed and directional for that reason. **Prop 4.1**: LayerNorm drives both contexts'

raw logits toward a shared prompt-independent limit at large $|\alpha|$, collapsing the dynamic range any cross-context object needs, though not, by itself, pinning a baseline-subtracted ratio to 1. Either way it motivates a validity ceiling α^* above which a reading is not trustworthy; we determine that ceiling empirically rather than from the proposition, and §5 reports what it excludes.

The object to estimate is therefore a *location on the dose axis*, with its fitted midpoints inside α^*, and with a signed readout:

$$\mathrm{dEC}_{50} \;=\; m_{\text{agent}} - m_{\text{chat}}, \qquad \Lambda \;=\; \log m_{\text{agent}} - \log m_{\text{chat}},$$

where m is the fitted 4PL midpoint of the length-controlled combined binary response in each context. We call this displacement a *potency shift*, and reserve *coherent efficacy* for the second and separately measured dimension: whether output at that dose is still coherent. Λ exists because a raw dEC_{50} is in a *model-specific* dose unit: our roster's axes span 13 to 4096, so a raw cross-family magnitude comparison is a statement about axes rather than about effects.

Why a location and not a gain. A gain is a ratio of two response *levels* at a chosen dose and inherits that choice. A midpoint difference is a property of the whole curve, it is a curve location of the kind Thm 3.6's reparametrization acts on, we read it signed for the reason Thm 3.8's vanishing first-order term suggests, and (unlike a top-dose ratio) it can be checked against the α^* that Prop 4.1 motivates.

Protocol and control spine. Metric: a registered length-controlled *combined* binary, never judge-raw. $B = 10\,000$ item-paired bootstrap, greedy decoding, 200 item ids per cell in both contexts and $N \geq 195$ in the resampled anchor. The *fitted* set is smaller and dose-dependent, since a rollout is dropped unless its metric is binary and its coherence flag is true; the worst per-dose paired count among the law-eligible cells is 8, on `OLMo-2-7B:bypass`; the roster minimum is 0, on the quarantined `OLMo-2-7B:induce`. Controls: ≥ 5 matched-norm random directions per cell, entering as the per-dose $\Delta(\text{real}) - \Delta(\text{random})$ diagnostic and as the admission gate rather than as a correction of the fitted response (App. A); a matched-information template ladder from chat (C0) to a ReAct scaffold with deterministic dummy tools (C3), *matched on the instruction and asymmetric on the directive*: byte-identical instruction, but one span of the C0 system prompt is a compliance-valenced directive C3 does not carry (App. D.2). Four scope conditions are enforced in code and each can fail: **(i)** α^*: a cell is admitted only if both fitted midpoints sit below the dose at which the real direction's cross-context separation collapses into a matched-norm random band, contexts merged and read on the high side: the first dose above that separation's peak at which the margin is non-positive; failing cells are marked α^***-RED**, printed, and excluded from every law claim. **(ii)** the dose rule: ≥ 5 doses spanning $\geq 4\times$ and bracketing the midpoint, *in both contexts*, or the cell is not fitted. **(iii)** a chat gate: a ≥ 25 pt swing, or the family is dropped before it consumes a verdict. Its registered *"and beats matched random"* clause **never ran**; applying it now removes one cell, which one is metric-dependent, and no sign-carrier. **(iv)** the unit of analysis is the *model*, never the cell.

Errata to the control record, carried rather than buried. An audit of our own control spine returns four record-level errata, most consequentially that the "length-controlled" metric is length-*correlated*, holding on the field it names rather than on the property its name asserts. All four are in Appendix A; they correct the record; none of them changes the estimand.

3 THE GAIN RATIO DOES NOT IDENTIFY POTENCY AND EFFICACY

First, what does survive. The direction *installs*: injected at the same layer and coefficient, it arrives at the final layer in the agent context at install-site agent-over-chat ratios of 0.83–1.16 across three families (Pinto, 2026). And the two contexts' *fitted upper asymptotes* agree to within 0.012 on the frozen five, 0.013 once Gemma is included. We draw no efficacy conclusion from that: on a bypass cell the asymptote is the floor refusal is driven to, so the contrast is blind to how much refusal there was to remove, and `OLMo-2-7B:bypass` pairs a plateau contrast of 0.0000 with a *deliverable* contrast of -0.178 whose CI excludes zero.

What the gain ratio does not determine. We reconstruct the prior estimator in both of its published forms (a single-dose ratio and an AUC-over-doses ratio, together with the per-dose effect

Table 1: The scalar gain T on grids built to support a midpoint. Every row is a property of the *estimator*, recomputed from rollouts that reproduce the committed fit exactly.

finding	what it is made of
T moves with dose	paired ΔT CI excludes 0 in **4 of 5** α^*-GREEN cells
T spans 1 everywhere	realized range contains 1 in **5 of 5**: $0.067 \to 1.108$ (Llama-3.1-8B), $3.877 \to 0.590$ (OLMo-2-7B bypass), $2.011 \to 0.905$ (Qwen2.5-3B), $2.607 \to 0.873$ (Qwen2.5-7B bypass), $10.050 \to 0.835$ (Qwen2.5-7B induce)
opposite displacements, overlapping gains	Llama at dose 9 gives $T = 1.0021\,[0.9908, 1.0156]$; Qwen2.5-7B induce at dose 53 gives $T = 1.0402\,[1.0047, 1.0804]$: their potency shifts are $+1.013$ and -12.368, **opposite signs, both CI-clean** on the primary grid (the $+1.013$ CI includes zero under the registered α^*-trim, Appendix B.1), and the gain intervals overlap at a width under $\mathbf{0.08}$
the estimator disagrees with itself	Llama is an *attenuator* by the AUC form ($0.966\,[0.941, 0.992]$, excluding 1) and an *amplifier* by the point form at its top dose ($1.108\,[1.060, 1.166]$, excluding 1): same rollouts, two estimators published in the same paper; **5 of 5** cells fail both agreement clauses

they are built from and the $\mathrm{SAT_MARGIN} = 0.05$ sub-saturation filter) copied verbatim from its released analysis code, and run it on our swept grids. Every per-dose effect recomputed reproduces the committed fit's own record exactly ($\max|\Delta| = 0.0$, 7 of 7 cells), so what follows is a property of the estimator and not of a re-implementation. This section, and the horizontal-vs-vertical and coherence tests of §4, were frozen on the seven-cell roster that predates the seventh family: their denominators are the **five** α^*-GREEN cells of that roster, against Table 2's six. Gemma-3-12B is law-eligible but post-dates that freeze; we name that fact rather than re-open the audit.

Table 1 licenses one sentence, about the five cells it scores and not the RED cells or Gemma-3-12B: *every scorable cell in this frozen audit is an amplifier at one dose and an attenuator at another*, so a per-model amplify/attenuate taxonomy reports the dose the ratio was read at rather than a partition of models.

Two of our own registered predictions were refuted here, and both constrain what we may say. **(i)** On grids built to the dose rule, T is computable in *every* cell (34 of 48 cell×dose pairs defined, no cell yielding none). The prior paper's three families that "yield no coupling estimate" are therefore a **grid** property, not a model property, and this paper may not carry that sentence as a fact about those models. **(ii)** The *identification level set* (the (potency shift, efficacy ratio) pairs consistent with one measured gain at one dose, a curve rather than a point) contains zero in only 2 of 3 scorable cells against a threshold of 3 registered in advance, so that half of our prediction fails.

What this section does not say. T is not *wrong*: it is a well-defined functional of two curves, the prior paper measured it at a stated operating point and said so, and the relative-steerability family of Tan et al. (2024) had established the form. Ours is the narrower, identification claim: **a scalar read at one operating point does not determine the pair (potency shift, coherent efficacy)**. A wide T interval is not evidence *for* dEC_{50}, which enters here only as a label.

4 THE LAW

Counts, at the registered unit. Table 2 holds **8 cells over 6 distinct models**, **5 α^*-GREEN models**, and **5 CI-clear sign-carrying cells across 4 distinct models** (Llama $+$; Qwen2.5-7B $-$, both arms; OLMo $-$; Gemma-3-12B $-$). We report the model unit rather than the cell unit because Qwen2.5-7B enters as two arms of one model, and a cell-unit count double-counts it.

Gemma-3-12B's -886.066 is never *interpreted* without its scale-free companion. Its top dose is 4096 against 13–77 across the rest of the roster, a $315\times$ axis. On the log scale it reads $\Lambda = -0.4812$, within 0.012 of Qwen2.5-7B induce's -0.4695: a $53\times$ larger top dose, a $71.6\times$ larger raw dEC_{50}, one displacement. The sentence for that cell is *"the agent context reaches half-maximal effect at 62 % of the chat dose."*

Table 2: Core result. m is the fitted 4PL midpoint of the length-controlled combined binary in each context; CIs are item-paired bootstrap, $B = 10\,000$. α^*-RED cells are reported in frame and excluded from every law claim (§5). Dose units are model-specific and are never compared in magnitude across rows; see Λ in the text for the poolable axis.

cell	arm	n	m chat	m agent	dEC_{50} [95 % CI]	α^* verdict
Llama-3.1-8B	induce	195	4.841	5.854	$\mathbf{+1.013}$ $[+0.777, +1.273]$	GREEN
OLMo-2-7B	bypass	200	17.875	12.378	$\mathbf{-5.497}$ $[-7.851, -3.175]$	GREEN
Qwen2.5-7B	bypass	200	17.049	6.194	$\mathbf{-10.855}$ $[-15.814, -4.614]$	GREEN
Qwen2.5-7B	induce	199	33.008	20.640	$\mathbf{-12.368}$ $[-13.726, -10.876]$	GREEN
Gemma-3-12B	induce	199	2319.795	1433.729	$\mathbf{-886.066}$ $[-945.445, -806.491]$	GREEN
Qwen2.5-3B	induce	199	17.763	17.526	-0.237 $[-1.303, +0.818]$	GREEN
OLMo-2-7B	induce	199	16.695	37.114	*(+20.419)* $[+18.863, +21.769]$	**RED**
Starling-7B	bypass	200	7.307	6.942	-0.365 $[-1.206, +0.488]$	**RED**

Committed → repaired 95 % CI (the repair pre-registration requires both printed; the committed interval is the one it dates, and no point estimate moves by more than 5×10^{-4}): Llama-3.1-8B $[+0.777, +1.273] \to [+0.760, +1.295]$; OLMo-2-7B:bypass $[-7.851, -3.175] \to [-7.687, -3.045]$; Qwen2.5-7B:induce $[-13.726, -10.876] \to [-13.915, -11.017]$; Qwen2.5-7B:bypass $[-15.814, -4.614] \to [-15.790, -4.253]$; Qwen2.5-3B $[-1.303, +0.818] \to [-1.358, +0.818]$. CI-exclusion is unchanged on all five. Three rows carry no repaired interval: two are α^*-RED and outside the repair's frozen roster, and Gemma-3-12B is law-eligible but post-dates it.

Heterogeneity, powered against per-sample variance. Gao et al. (2026) hold that steering sensitivity is controlled sample-specifically, so the aggregate has high per-sample variance by construction; we power the family-level contrast against that variance rather than against cell count. We report the specification our own pre-registration requires, at the **model** unit over α^*-GREEN cells only, since α^*-RED cells are declared not to contribute: $Q = 712.51$ on 4 df, $I^2 = 0.9944$, $\tau = 9.432$ on the $315\times$-span raw axis, against a design-minimum of 2.653; joint sign power 0.9998. dEC_{50} is family-dependent beyond per-sample noise, invariantly across all 14 cells of the registered specification lattice. Two facts bound it: Gemma alone supplies $93.0\,\%$ of Q, though deleting it still leaves $Q = 49.60$ on 3 df and a powered verdict, and two of five leave-one-out refits return UNDERPOWERED, so the verdict rests on the roster, not on any four of its members. The axes being incommensurable, that pooling is a heterogeneity statement and never an average effect; the same family of statistics computed on Λ ($I^2 = 0.9886$ pooling cells, 0.9914 pooling models) cross families; those are their own pools and not the lattice cell above.

Horizontal, by test, and not further than the test supports. At equal model complexity a horizontal reparametrization beats a vertical rescaling in **5 of 5** cells of that same frozen five-cell roster, every ΔAIC interval excluding zero, and the horizontality index H (the share of the context-induced gain in fit that a pure translation alone buys) has median 0.948, under a rule registered before the fit. **The registered α^*-trimmed secondary is weaker, and we report it as weaker**: 4 of 5, stratified by grid loss (Appendix B.1). **But the curves are not translates**: writing dEC_q for the displacement read at response level q, $\Delta_q = \mathrm{dEC}_{0.75} - \mathrm{dEC}_{0.25}$ is zero for a pure translation and its CI *excludes* zero in 3 of 5 cells, so no sentence in this paper calls the two curves translates. The *sign* is not an artifact of the 0.5 threshold: it survives five response levels (Llama is positive with a CI excluding zero at all five, $\mathrm{dEC}_{0.10} = +1.247\,[0.94, 1.82]$ through $\mathrm{dEC}_{0.90} = +0.779\,[0.33, 1.20]$), five link functions (25/25 cell×link), and twelve fit windows (59/60); across an analytic multiverse of 56 specifications the sign agrees in ≥ 0.9286 of them in every GREEN cell. What does *not* survive is the functional form: $\mathrm{dEC}_{0.50}$ varies by up to $2.99\times$ across links and the AIC-preferred link is our registered 4PL in only 1 of 5 cells, a prediction our own pre-registration made *against* our registered choice, and it held. The defensible claim is about the sign.

Coherence is reported as an outcome and never used to exclude a cell. Wherever the coherent margin can be decided it is positive: 9 of 10 α^*-GREEN cell×context margins are established strictly positive, so the midpoints we report sit *below* the dose at which coherence half-collapses. That is what makes them readable as behaviour rather than as decoding failure, and it is what the quarantined cell violates.

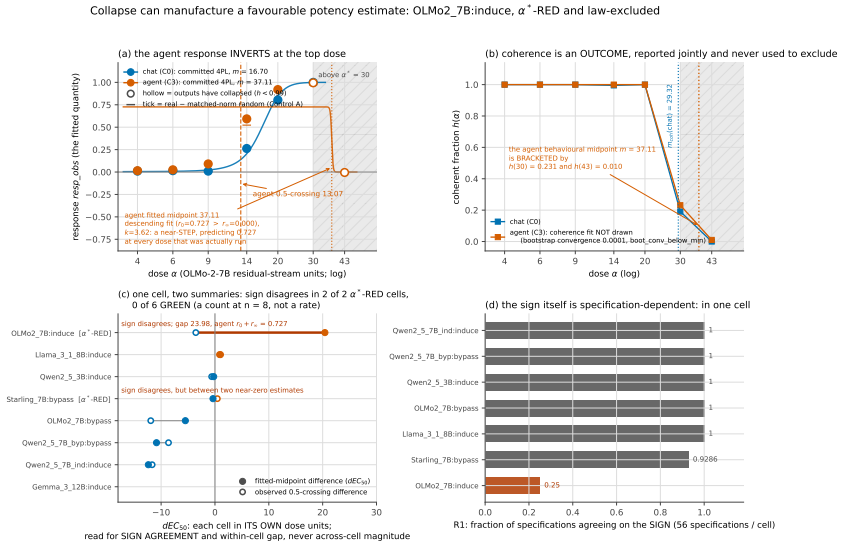

Figure 1: **How a large positive "shift" is manufactured, and how the frozen screen caught it.** One cell, `OLMo-2-7B:induce`; (c) and (d) add the roster's others: seven and six, the latter roster frozen before Gemma-3-12B. **(a)** Behavioural dose–response in both contexts, on the response the 4PL was fitted to, with the matched-norm random contrast overlaid as a tick rather than subtracted (App. A); marker fill encodes the observed coherent fraction $h(\alpha)$, so a dose whose outputs have collapsed renders hollow. The agent response rises to 0.995 at dose 30 and *inverts* to -0.005 at dose 43; the fitted agent 4PL is consequently descending ($r_0 = 0.727 > r_\infty = 0.000$) and places its midpoint at 37.114, past the collapse. **(b)** Coherence as an outcome: the agent coherence fit does not converge (bootstrap convergence 0.0001), so only its observed points are drawn. **(c)** Two summaries of one *cell*, not of one fit: the fitted midpoint dEC_{50} against the model-free 0.5-crossing: $+20.419$ against -3.556, opposite in *sign*, so this cell's shift is **not identified**. **(d)** Sign stability over 56 specifications. The cell is α^***-RED and law-excluded**, binding no fitted summary. Text at §5; one cell, no prevalence claim.

The selection bound, signed and measured. The positive side of "two-sided" rests on **one** model, and we quantify the exposure rather than hedge it. All three agent-side measurability gates are monotone in absolute dose; the admission window in 5 of 5 admitted cells reaches less far above zero than below it ($7.3\times$ harder upward), and the single admitted positive sits at $88.5\,\%$ of its own upper edge. **The admission geometry therefore biases against observing positive shifts**: the design under-observes exactly the sign that is scarce, so 1 in 5 is not evidence that positives are rare, and five cells bound no population rate. The same cell carries a rule qualifier: our estimator fits the full swept grid and reports α^* post hoc, not the stricter dose-exclusion sentence two of our own pre-registrations carry. Re-run under that rule (same estimator, seed, B, items) the positive cell reads $+1.352\,[-0.034, +1.587]$: **sign unchanged, CI-exclusion lost.**

Deleting it would not cost the non-identification result: T still contains 1 in 4 of 4 remaining cells and still moves with dose CI-cleanly in 3, the number we registered. It would cost the sharpest demonstration, since an opposite-signed pair needs a positive shift to pair against.

5 Artifact Discipline

A venue asking what distinguishes a mechanism from an artifact is entitled to the cases where our own numbers were the artifact. We report four, at result volume.

(1) The quarantine, and what it costs us. `OLMo-2-7B:induce` is excluded from every law claim. Figure 1 shows why: past a coherence collapse the agent response inverts, the fit's orientation flips, and its "midpoint" lands at no dose that was run, bracketed by $h(30) = 0.2312$ and $h(43) = 0.0101$. The cell's fitted-midpoint summary ($+20.419$) and its observed-crossing summary (-3.556)

disagree in sign, and only 0.25 of 56 specifications of this cell agree on a sign against ≥ 0.9286 everywhere else. **What the exclusion costs, said plainly: the deleted cell is large and positive, so the quarantine removes the roster's biggest positive and makes our two-sidedness claim harder to support.** Three readings are refused. The crossing value is *not* a corrected dEC_{50}: the cell is not identified. The cell is *not* a censored genuine positive; it cannot be both that and a collapse artifact. And this is *not* a critique of 4PL fitting: six cells in the same artifact are fitted by the same code at convergence 1.0, and the failure is specific to a grid carried past the collapse. The quarantine is at the **arm**, not the family: `OLMo-2-7B:bypass` is α^*-GREEN, CI-clear at -5.497, and is one of the law's five sign-carriers.

A check on the screen that caught it. Across the eight fitted cells the two summaries disagree in sign in exactly the two α^*-RED cells (2 of 2) and in none of the six GREEN (0 of 6). α^* was frozen on an unrelated criterion long before anyone compared these summaries, and the separation survived a test it could have failed when a seventh family added a GREEN cell that does *not* disagree. This describes a frozen screen post hoc rather than re-deriving it, and $2/2$ and $0/6$ are counts at $n = 8$, and we do not read them as rates.

(2) An explanation in a companion of ours that was a correlation with the dose axis. Our dose-resolution companion, released with the paper, explained two estimators' magnitude disagreements with a plateau-gap correlation, $\rho = 0.893$ ($p = 0.0068$) on seven cells. On a *scale-free* axis (the same disagreement as a fraction of the estimate rather than in raw dose units) those seven cells give $\rho = 0.321$, $p = 0.48$. The correlation was with the dose axis, and one added model exposed it: on eight cells the absolute-axis figure itself falls to $\rho = 0.571$, $p = 0.139$. The statistic was non-binding (no verdict here rests on it) but it is a worked example of the failure this venue names, found in our own companion, and the remedy is mechanical: **report every cross-family diagnostic on a scale-free axis, or expect it to measure the axis.**

(3) A normalisation that succeeds and still does not deliver commensurability. Because raw dEC_{50} lives in a model-specific unit, we asked whether dividing each cell's chat midpoint by that model's chat residual norm makes the dose axis commensurable. It is *highly effective and still insufficient*, and both halves are measurements. Raw midpoints span $479\times$ across the six α^*-GREEN cells and normalised ones $21.7\times$, removing about 96 % of the cross-model axis, while missing our registered commensurability threshold of 10: on the five cells frozen before the seventh family the normalised spread is 2.68 and the claim holds; the sixth takes it to 21.74 and it fails. We report the reduction, which holds on all six; the commensurability claim holds on five and we say so. The failure is not an instrument artifact: the sixth cell's norm leg is bound at the same layer as its harvest, on the same 200 probes, at $5\,\%$ sd/mean, with a norm-ratio CI excluding 1 like every other cell's. The read is normal; the scale is not.

(4) A registered alternative that no code in this repository could produce. A pre-registration buys nothing for a clause whose losing branch cannot be emitted: that rule was one-sided when it was scored, whatever the document says. We measured this rather than asserting it. A census resolves all 299 registered decision rules (641 branch mentions, 569 distinct labels) against every string the analysis code in this repository can produce, excluding by name the 41 labels that are ordinary words, where a match is evidence of nothing either way. It found 4 branches, both alternatives of two registered rules on a study that had *landed*, with no emitter at all, invisible because those rows sat behind a status the register validates by requiring the artifact to be *absent*. Both estimators are now written and both registered predictions hit, though they were written after the analyst had seen the data and so are not blind scoring. The 8 landed branches that remain unemittable are each adjudicated by name in Appendix B.1, where the detector is also validated against 170 demonstrably-emitted branches before any of its negatives is quoted. An earlier, more permissive version turned all 40 then-unemittable branches emittable and reported a clean tree. **A detector that flags every branch cannot separate the emittable from the rest.** A second register scored our frozen Methods draft against the pre-registration corpus and found it describing 10 of 44 registered designs (the corpus as it stood when the cut was made) and that gap is closed there, at 45 of 45.

6 THREE DEFLATIONARY ACCOUNTS, MEASURED AND REJECTED

Each of the three below is a mechanism a reviewer can name in one sentence, each is measurable *before generation*, and each is rejected on more than "no effect": on the sign pattern it predicts, on the magnitude it can carry, and (on the roster's positive anchor) on *pointing the wrong way*. The rejecting evidence differs by account, and one of the three is mixed.

Residual-norm rescaling. *The objection:* the agent prompt has a different residual norm, so a fixed α is a different relative perturbation, and the midpoint moves for a pure scale reason. *The measurement:* the norm ratio $\rho = \|r\|_{\text{agent}}/\|r\|_{\text{chat}}$ is 0.86–0.91 with a CI excluding 1 in 5 of 5 cells: the effect is real, and it is the other way round, the agent residual being *smaller*. But $\log \rho$ is negative in 5 of 5, i.e. **sign-homogeneous**, while Λ is two-sided: $\Lambda = +0.190$ on Llama against $\log \rho = -0.118$ there, and negative-on-negative elsewhere. **A scale account predicts one sign for every model; the data show two.**

Baseline alignment. *The objection:* the unsteered residual already has a component along the steering direction, and that component differs by context, so what looks like a potency shift is an offset. *The measurement* (registered before the analysis code existed and before any such cosine had been computed in this project): the cosine c_0 is nonzero in 5 of 5 cells (e.g. -0.328 chat vs. -0.175 agent on Llama) and context-dependent in 5 of 5 (*the mechanism is real, large, and present everywhere we looked*) and the composed scale-plus-translation law it generates is nonetheless rejected in **3 of 5 scorable cells against a rejection threshold of 3 registered in advance**. On the roster's only positive cell it predicts $\Lambda_{\text{pred}} = -0.506$ against a measured $+0.190$ (**the wrong sign**) and dividing the drive out makes the displacement $3.7\times$ *larger* ($+0.696\,[+0.628, +0.776]$). The affine scale factor of Cheng & Kriegeskorte (2026), reported here because no one else has for this context pair, is inadequate at median $R^2 = 0.189$. Two disclosures: our own registered prediction that the law would *undershoot* was refuted (the realized branch is mixed: 2 cells oppose in sign, 2 undershoot, 1 overshoots), and every c_0 here is read at the final prompt token while the deployment operator adds $\alpha \cdot u$ at every position.

Dose transduction. *The objection:* dEC_{50} is a difference of midpoints in *commanded-coefficient* units, and there is no reason a commanded unit buys the same residual displacement in a 188-token ReAct prompt as in a 44-token chat prompt. *The measurement:* the objection is real. Delivered displacement per commanded unit differs by $+19.3\,\%$ (Llama) and $-2.5\,\%$ (Qwen2.5-7B bypass), both intervals excluding 1, *and it runs the wrong way*. Delivery predicts a dEC_{50} of the *opposite* sign on 2 of 3 cells with intervals excluding zero on the wrong side, and explains $+3.0\,\%$ of the shift on the one cell where the sign agrees. Re-expressed in chat-delivered units, all three cells keep sign and CI-exclusion and the only positive cell more than doubles ($+1.013 \rightarrow +2.141\,[+1.883, +2.403]$). **Bound to its design**: three cells over two models, one dose each, delivery and behaviour on *independent* pools so it is not paired, and normal-approximation midpoints.

A fourth account, excluded by construction rather than rejected by measurement. Kang et al. (2026) attribute multi-turn coherence failure to KV-cache contamination. Every agent-context rollout behind every dEC_{50} here runs under mandatory full re-encode at every turn, so cache reuse is absent *by construction*, verified structurally in both agent code paths, a check that can fail. Three scope limits bind. We have *not* measured what cache reuse contributes, only that our result does not depend on it; we do *not* say Kang et al. (2026) are wrong, since the control they name is the one our protocol runs by default; and recompute guarantees turn-uniform re-application, *not* an exclusion of the duality reading, our steering being turn-constant.

7 A REGISTERED PREDICTOR, REFUTED OUT OF SAMPLE

The registration. We frame this as what it is: **we learned the predictor is wrong**. We do not present it as a weak positive. Before the seventh family's sweep existed, our pre-registration staked a signed out-of-sample prediction on Gemma-3-12B: dEC_{50} is *positive*, and its normalised value ranks 1st of 6, above Llama's $+0.0779$; refuted if the fitted dEC_{50} is negative with a CI excluding 0, *or* if its normalised value ranks below Llama's.

The outcome: both clauses fail. Measured $\mathrm{dEC}_{50} = -886.066\,[-945.445, -806.491]$: negative, CI excludes 0. Normalised, -0.2163 against Llama's $+0.0779$: not rank 1, not above Llama. The cell is α^*-GREEN, so no not-applicable escape was available and none was taken. The in-sample Spearman correlations that motivated the feature were $+1.000$ at $N = 4$ and $+1.000$ at $N = 5$; they did not survive one out-of-sample family. And **the map that fit best in-sample is the one that got the sign wrong**: the frozen affine map forecast $+56\,245$ (a flagged extrapolation $6318\times$ outside its training range) where a sign-only theory rule with no fitted parameter forecast negative. The realized affine slope is $a = -0.0316\,[-0.0362, -0.0270]$ against the theory-implied $+1$: the best-fitting map's own coefficients refute its own derivation.

One arm is a clean transfer failure. At $R^2_{\text{pred}} = -312.9$, A_{lin} is $313\times$ worse than predicting the roster mean: a within-context steering-success diagnostic predicts a cross-context shift *worse in squared error than the intercept baseline*. The defensible verb is *does not transfer to* dEC_{50}, never *refuted*: neither Fan et al. (2026) nor Billa (2026) claims a cross-context quantity, so this falsifies nothing they assert. The per-example geometry arm of Fan et al. (2026) is **not scored** here: its classifier needs per-example success labels this roster never produced.

Why $N = 4$ and not 5, and why that is the same problem as §4. Llama-3.1-8B is predictor-*undefined*: its agent-side signed susceptibility sits at its own noise floor ($z = 1.83$ against $z_{\text{crit}} = 1.96$, versus $z = 34.08$ in chat), and Llama is the roster's only admitted positive. That set is therefore entirely negative-signed, the configuration in which a sign-insensitive baseline is hardest to distinguish from a signed one: the selection geometry of §4 and this negative are one problem surfacing twice.

8 Discussion

What it means for anyone steering an agent. A coefficient calibrated in chat does not carry a calibrated *meaning* into an agent scaffold: the vector still works, but the same behavioural fraction now sits at a different dose, in a direction not constant across models. Sweep the curve in the deployment context as well as the calibration one.

An implication for how transfer is measured. Neither estimator family named in §1 holds position on the dose axis fixed. Agarwal (2026) reports a $71.0\,\%$ win rate at the best positive strength per pair against a mean $\Delta = -0.004$ at a fixed strength for the same method, with 4 to 1 admissible strengths per (model, method) by that paper's own usability threshold. **We do not decompose that gap and claim no reading of it**: per-cell variances are unpublished, a maximum over k noisy estimates is upward-biased by selection alone, and the restriction is disclosed there.

The experiment we did not run. The intervention that would establish dEC_{50} as causally *movable* (editing the ReAct format scaffold with task information held fixed) was designed, pre-registered, priced, and **not launched**. Its free forward-pass gate returned zero rankable cells in 6 of 6 (family $\times$ readout-horizon) cells, so the planner refused to launch and no money was spent. That refusal bought two reported things: a content-free placebo edit classifies the blocking gate's two failure modes 6 of 6, separating "reference set deficient" from "real information lost"; and an attainability calculus priced the headline experiment out *before* paying for it. The readout invalidity is a property of this roster's 2024-era models, and we do not generalise it to ReAct scaffolds. On the pre-registered reading this is therefore a *measurement* paper, and we say so rather than reporting the shortfall as a result.

Limitations. Beyond those stated where they arise: **(1)** the positive side rests on one model and a monotone filter biases admission against positives, so the rate is not a population rate (§4); **(2)** that cell's CI-exclusion is rule-dependent (§3); **(3)** two nuisance parameters were chosen on one leg, only the second swept; **(4)** c_0 is read at the final prompt token while the operator acts everywhere, and prompt-block variance understates generation-time variance; **(5)** a dose is not a shared unit, so every cross-family magnitude carries a normalisation (§5); **(6)** the curves are not translates, so dEC_{50} summarises location at one response level and the claim is about sign, not form; **(7)** four control-record errata are named (§2); **(8)** the causal half was registered, priced and not run; and **(9)** the register census bounds name-reachability, not executability (Appendix B.1).

REPRODUCIBILITY STATEMENT

Every number in this paper is transcribed from a committed machine-readable artifact and is recomputed from those artifacts by code released with it. We separate that from two stronger claims we do *not* make. Regenerating the committed artifacts themselves from the released tree is only partly possible, because the fits, control-spine audits and forecasts read rollout records and activation bundles too large to ship; and regenerating the original model outputs is not possible from the release at all, since it requires the inference runs themselves. Appendix E states which of the three each artifact supports. The release carries the fitting and scoring code, the frozen roster with pinned model revisions, the pre-registration register of all 47 registered designs, the guard suite that checks the instruments, and a self-test that runs inside the released tree. Appendix B states which registered branches were scored, which were not, and why.

Two limits are worth stating plainly. First, the scored rollouts behind the headline table are not distributed in full: a majority of their bytes are model generations from a refusal-bypass arm, and we withhold those for the reason given in the ethics statement below. The release instead carries the same rollouts with the generated text removed, and the headline table refits identically from them, including every midpoint, every bootstrap interval, and the heterogeneity and power blocks. Second, the dose grids are model-specific and were frozen per family before any agent-side outcome was observed, so a reader reproducing a cell must use the roster's grid for that cell rather than a common one.

ETHICS STATEMENT

This work measures how much of an activation-steering intervention is required to change behaviour, and one of the two arms it measures is refusal bypass. Studying that arm is what makes the paper's safety claim possible: an intervention calibrated in single-turn chat does not carry its calibration into an agent scaffold, so a control validated in one deployment context can be substantially mis-dosed in another. Reporting that requires measuring the arm where the failure has consequences.

The corresponding risk is that the rollouts contain model outputs produced under a bypass intervention. We therefore withhold those generations from the public release rather than publishing them, and we do so without trading away reproducibility: the released rollouts retain every field the estimators read and drop only the generated text, and the headline table refits identically from them. No new jailbreak method is introduced here, and the steering technique itself is the standard difference-of-means construction already in the literature we cite. All models used are publicly released checkpoints at pinned revisions; no human subjects and no private or personal data are involved.

AI USE STATEMENT

We used generative AI tools extensively in this work, and we state the extent precisely because the disclosure is load-bearing for how the results should be read.

Of the tasks ICLR marks as requiring disclosure, we used AI tools to propose and refine hypotheses, to design research methodology and experiments and to give feedback on them, to implement methods, to clean and reformat data, and to interpret results against pre-registered criteria. We did not use them for translation, and we did not use them to generate synthetic data sets: every number here comes from rollouts of the pinned checkpoints. Support for qualitative and thematic data analysis does not arise; all analysis reported here is quantitative. We did not develop theoretical models with them, did not use them to formulate mathematical claims, did not use them to provide critical ingredients for proving mathematical claims, and did not write proofs with them; the theoretical results we rely on are cited from prior work and are not our own contributions. Of the tasks ICLR marks as recommended for disclosure, we used AI tools for software implementation, literature search and analysis, gap identification, figure generation, drafting and editing of this paper, reference formatting, and title and keyword selection.

Concretely, the experimental programme was run by an autonomous research loop. That loop authored pre-registrations, built the instruments and their tests, launched and monitored the compute,

and scored outcomes against registered branches. The human author set the research direction, chose which questions to pursue and which to abandon, ruled on scope and cost, decided what the paper claims, and reviewed the artifacts and the text.

We have reviewed all AI-assisted work, and the paper's verification apparatus exists because the process was automated. Most studies were registered before their data existed, and the register is published with the paper rather than summarised. It is not a universal claim and we do not make one: the control-spine conformance audit is marked POST-DATA, the causal-attainability derivation is marked post-hoc on frozen artifacts, and the extraction-frame study is a pre-specified analysis of already-public data rather than a blind pre-data registration. Each says so on its own face, and the register records which kind each study is. The guard suite tests the instruments against deliberately planted defects, so a check that cannot fail is ordinarily visible as such. That guarantee is not complete, and we state where it failed. The suite plants each defect against a check's *declared* subject, so where a check's declared subject and its actual one diverge, the planted defect diverges with it. One rail here read a figure caption from that figure's provenance sidecar while its description, its inline comment and its mutant all said it read the caption the paper ships; rail and falsification test were aimed at the same wrong file, and the shipped caption had meanwhile lost the law-exclusion the figure exists to carry. Adversarial review found it and the suite did not. We report it for the same reason we report the refuted forecaster: an apparatus that records its own miss is worth more than one that claims completeness. A census of our own register found registered branches that no code in the tree could emit, and we report the count and each disposition in Appendix B rather than repairing them silently. Our pre-registered forecaster of the estimand was refuted on the first held-out family and is published as refuted. These are not incidental: they are the reason we think results produced this way can be checked by a reader who did not run them.

We take responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI tools.

REFERENCES

Ayushi Agarwal. Cross-architecture steering transfer in language models: A systematic empirical study. *arXiv preprint arXiv:2608.05164*, 2026.

Jayadev Billa. Predicting where steering vectors succeed. *arXiv preprint arXiv:2604.15557*, 2026.

Joschka Braun. Understanding unreliability of steering vectors in language models: Geometric predictors and the limits of linear approximations. *arXiv preprint arXiv:2602.17881*, 2026.

Fan L. Cheng and Nikolaus Kriegeskorte. Decomposing how prompting steers behavior. *arXiv preprint arXiv:2606.03093*, 2026.

Chenrui Fan, Yize Cheng, Ming Li, Soheil Feizi, and Tianyi Zhou. When is your LLM steerable? *arXiv preprint arXiv:2606.11599*, 2026.

Lang Gao, Jinghui Zhang, Wei Liu, Fengxian Ji, Chenxi Wang, Zirui Song, Akash Ghosh, Youssef Mohamed, Preslav Nakov, and Xiuying Chen. The cylindrical representation hypothesis for language model steering. *arXiv preprint arXiv:2605.01844*, 2026.

Diancheng Kang, Zheyuan Liu, Ningshan Ma, Yue Huang, Zhaoxuan Tan, and Meng Jiang. Prompt-activation duality: Improving activation steering via attention-level interventions. *arXiv preprint arXiv:2605.10664*, 2026.

Jiaqian Li, Yanshu Li, and Kuan-Hao Huang. Steering vector fields for context-aware inference-time control in large language models. *arXiv preprint arXiv:2602.01654*, 2026.

Lucas Pinto. Present but rescaled: Chat-to-agent transfer of additive activation steering. *arXiv preprint arXiv:2607.09156v1*, 2026.

Julian Skifstad, Xinyue Annie Yang, and Glen Chou. Local linearity of LLMs enables activation steering via model-based linear optimal control. *arXiv preprint arXiv:2604.19018*, 2026.

Magamed Taimeskhanov, Samuel Vaiter, and Damien Garreau. Towards understanding steering strength. *arXiv preprint arXiv:2602.02712*, 2026.

Daniel Tan, David Chanin, Aengus Lynch, Brooks Paige, Dimitrios Kanoulas, Adrià Garriga-Alonso, and Robert Kirk. Analysing the generalisation and reliability of steering vectors. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, 2024. doi: 10.52202/079017-4417. Preprint: arXiv:2407.12404.

A ERRATA TO THE CONTROL RECORD

An audit of our own control spine returns four record-level errata. We state them rather than silently repairing them, because each is a case where a control's *name* asserted more than the field it was computed on.

1. **Matched-norm random directions** enter as a per-dose diagnostic and as the α^* separability gate, rather than as a correction of the fitted response. Every midpoint in Table 2 is a baseline-corrected (induce) or absolute (bypass) fit; **none is a random-corrected fit**, and no line of the estimator subtracts the random arm from the response the curve is fitted to. Applied to the estimand as a reported sensitivity (same estimator, seed, B and items) the correction leaves sign, CI-exclusion and status unchanged in 5 of 8 cells, and we name the three it moves. Under the plain-reading `mean` aggregation `OLMo-2-7B:bypass` strengthens from -5.497 to -8.274, *away* from zero; under the conservative `maxabs` band the same cell is censored, which would cost the law its fifth family. `Qwen2.5-3B:induce` and the α^*-RED `Starling-7B:bypass` change sign inside intervals that span zero in the primary, and under `maxabs` alone `Qwen2.5-3B` reaches $+1.730\,[+0.328, +2.985]$, a CI excluding zero. We read `maxabs` as a gate and never as an estimator (a maximum over five draws is a positively biased estimate of a random direction's effect) but the registered primary is the uncorrected fit either way, and that is the erratum.
2. **The "length-controlled" combined metric is length-*correlated***, not length-controlling. It holds on the field it names (the combined binary is not the judge-raw score that carried the length artifact) but the name asserts a property the construction does not deliver, and a residual length–response correlation remains.
3. **One pairing control holds on the raw record and not on the set the estimator averages.** Item pairing is exact in the logged rollouts; the coherence-gated subset the estimator averages over can drop an item in one context and not the other at a single dose. The same gate also acts on a whole *dose*: each context's grid is derived downstream from the rollouts that survived scoring, so a dose at which every rollout in one context has collapsed leaves that context's grid with no field recording that it was ever swept, while the other context retains it and the two legs of a paired midpoint difference are fitted over different ranges. We re-score the realised support of every cell in the canonical fit on four rails: both contexts fitted on the same dose set; a finite bootstrap sd at every fitted dose: the weighted fit floors a *non-finite* sd to the same minimum it uses for a saturated one, which hands the least-determined dose the greatest weight; no swept dose removed from a fitted grid by the scoring filter; and the ≥ 5 doses spanning $\geq 4\times$ rule re-checked on the grid that was fitted rather than the one that was planned. All four rails pass on all 6 law-included cells and all 12 legs, at a minimum scored n of 26 per leg. Both conditions fire in exactly 1 cell, the already-quarantined α^*-RED `OLMo-2-7B:induce`: at dose 43 the chat context has 0 of 200 rollouts coherent and loses the dose, the agent context keeps 2 and retains it, and that retained point's undefined bootstrap sd is floored to the minimum.
4. **The chat gate ran on one clause of two.** Its registered form is a conjunction (a ≥ 25 pt swing *and* beating matched-norm random); the executed check evaluated the swing clause alone. The random comparison is made at the α^* stage, but as a *cross-context* separation against the random band rather than as the gate's per-family clause, so the registered conjunction was never enforced in either place.

None of the four moves a fitted number, a CI, or a verdict; they are errata to the record of how a control was realised. Each is stated with what would falsify it in the full limitations register (§H).

B PRE-REGISTRATION AND AUDIT APPARATUS

Every paid experiment in this project carries a dated pre-registration committed before the run it governs, with its live/die branches (including its null and its adverse branches) written before any number exists. Three machine-scored registers keep that discipline honest after the fact rather than by vigilance: a *pre-registration ledger* linking each registered rule to the verdict token some artifact actually emitted; a *methods register* that fails closed if a registered design's procedure goes unstated in the full methods appendix or in the companion document its row names; and a *limitations register* that fails closed if a threat frozen before the data is absent from the full limitations register. Both scored texts are carried here (§D, §E, §H). A whole-manuscript consistency audit mechanically rechecks that every dEC_{50} asserted in prose is a value the canonical fit contains and that every "CI excludes zero" claim is true of the interval printed beside it. Two registered verdicts reported in this paper (the predictor stop of §7 and the launch refusal of §8) are branches those registrations wrote in advance and that we would have preferred not to fire.

B.1 TWO AUDITS OF THE APPARATUS, AND WHAT THEY RETURNED

Both registers below were run on this repository and both returned defects. We report them at the size the finding had, because a falsifiability claim that is never itself measured is exactly the failure this paper is otherwise about.

Branch-emitter census. For each registered branch label, is there a string in non-guard, non-registry code that the tree can produce and that equals it: as a literal, an f-string with substantial literal evidence, a run-time prefix rewrite, or a marker written by a shell launcher? Scale: all 299 registered decision rules, 641 branch mentions, 569 distinct labels, against every string constant the analysis code in this repository can produce. Mentions resolve as `EMITTED_IN_CODE` 521, `NAME_UNRESOLVABLE` 79, `ONLY_IN_REGISTRY` 29, `DERIVED_IN_CODE` 7, `COMPOSED_IN_CODE` 3, `ONLY_IN_GUARD` 1, `ONLY_IN_SHELL` 1. The 79 unresolvable mentions are the 41 labels that are ordinary words (`PASS`, `NULL`, `GREEN`, `MIXED`, ...); they are counted, listed and excluded from every verdict rather than resolved in the direction that flatters the tree.

Three limits are declared. The census measures static *name-reachability*, not executability: a label found here is emittable in principle, and it does not prove the emitting line is reachable from an entry point on the realized inputs. The error therefore runs in the safe direction: a label *not* found cannot have been emitted by name at all. And nothing here says whether a registered prediction was any good; only whether the register could have recorded its failure.

What it found, and the disposition of what remains. Twelve branches belonged to a rule whose study had landed and had no emitter. Four were the real defect: both branches of two registered rules on the cache-reuse study, an α^* recomputation and a grid diagnostic. Both estimators were then written, both are refits of records that already exist, and both shipped the branch their pre-registration had named in advance. The remaining eight each carry a disposition naming the class it expects, so writing a missing estimator turns the entry stale and forces its deletion: an explanation may not outlive the thing it explains.

disposition	n	why it is not an owed estimator
`STAGE_NOT_REACHED`	3	Stage 1 was gated on a clean Stage 0; Stage 0 failed two of three bounds and Stage 1 was never launched, so the labels describe an experiment that does not exist in this tree
`RULE_NOT_EVALUABLE`	3	the rule shipped its not-evaluable or withdrawn branch, so the measurement that would select between the substantive branches was never built
`GATE_LABEL_NOT_A_VERDICT`	2	a named precondition, not a branch the scorer selects between

A further 20 branches belong to studies registered and not yet run, where no estimator is owed yet. We state the number because it is the difference between *we will run it* and *we have not written it either*, and it is part of what running those studies would cost.

Validating the detector before quoting its negatives. A branch that is the realized verdict of a rule scored in a committed artifact was demonstrably emitted, and the ground truth is re-read *from that artifact* rather than from the registry field being audited. There are 170 demonstrably-emitted branches; the scanner must find an emitter for every one. Unwaived misses: 0. Waivers: 0. Stale waivers: 0. This is not ceremony. The first f-string rule required only one literal part, so a two-variable template matched nearly every namespaced label and, in one run, all 40 then-unemittable branches turned emittable and the census reported a clean tree. The repair was a stricter detector, not a waiver; the waiver table exists, is subject to a stale-waiver rule, and is empty.

Methods register. A second register takes the mechanical set of pre-registration documents as its denominator and scores the merged Methods document against the tree: a `MAIN` row must name a section of that document that exists *and whose own body cites the pre-registration*, so a row cannot pass by being written down. Measured when the re-cut was made, against a corpus of 44 pre-registrations, the frozen Methods draft cited 10 of 44 registered designs by filename (13 by a looser stem reading) and the merged draft cited 5 (14 by stem): 31 registered designs (including every deflationary measurement this paper calls its hardest-to-dismiss result) had no procedural description anywhere a reviewer reads. The pre-registration ledger could not see it, because it asks whether every registered *rule* has a recorded *verdict*, one step later in the reading order than *where is the procedure that produced this number*. The register now reports 47 of 47, 0 defects, disposed as 25 stated in the full methods appendix, 18 in the named companion documents, 2 infrastructure and 2 registered-but-not-run.

The α^*-trimmed horizontality secondary, per cell. Re-running the horizontality test on grids cut at the cell's single paired α^*, applied to both contexts, gives `HORIZONTAL_VALIDATED` again at median $H = 0.964$, but 4 of 5 rather than 5 of 5. Writing $\Delta\mathrm{AIC}_{R2}$ for the model-comparison margin (not a coefficient of determination), full grid $\rightarrow$ α^*-trimmed: `Qwen2.5-7B:induce` $+347.87 \rightarrow +347.87\,[264.5, 431.2]$ and `Qwen2.5-3B` $+9.80 \rightarrow +9.80\,[1.4, 22.2]$, both unchanged because neither carries an α^*-excluded dose; `Llama-3.1-8B` $+87.54 \rightarrow +38.45\,[16.6, 70.7]$ on five of seven doses; `Qwen2.5-7B:bypass` $+24.81 \rightarrow +0.17\,[-4.9, +4.9]$ on six of seven; and `OLMo-2-7B:bypass` $+46.37 \rightarrow -11.87\,[-33.4, +6.6]$ on five of seven. The two zero-loss cells act as a control: neither moves at all, which is what attributing the change to lost grid support requires. Among the cells that do lose doses the ordering is *not* simply how many go, since `Qwen2.5-7B:bypass` loses one dose and collapses further than `Llama-3.1-8B` losing two. `OLMo-2-7B:bypass`'s point estimate does reverse the preference; what we rely on is that its interval spans zero, so the trimmed grid does not *establish* a reversal, rather than that it shows none. We report it because the primary fit deliberately retains doses at and above α^*, so a reader is owed the version that does not.

C Related work, in full

The body concedes only what its guardrails require, in the space a 9-page limit leaves. This section is the full positioning. Its organising claim is that **nothing here measures a ratio across two deployment contexts for a fixed model and a fixed direction, and yet every measurement primitive this paper uses is owned by someone else.** The job below is therefore to concede every primitive to its owner at exactly the right altitude; under-conceding is the failure mode a reviewer punishes.

Why this section is in the appendix rather than the main text. The 9-page body ends flush with the bottom of its last page: it has no free line. Restoring this section to the main text would cost a section of the identification argument, which is the contribution the body exists to carry. The concessions the guardrails require (the slope-estimator credit, the kinship between T and a ratio of Braun's slopes S, the Billa altitude, the SVF phrase) are each discharged *in the body* at their first mention; what is here is the complete version, not the load-bearing part.

C.1 The steerability slope and relative steerability

Owned: the propensity-curve-and-slope machinery; steerability s; relative steerability s_{rel}; and deployment-context manipulation at readout. We concede all four.

Tan et al. (2024) add $\lambda \cdot v_L$ at the last token position of layer L, read propensity as the logit difference between correct and incorrect multiple-choice answer tokens, and define **steerability** s as the slope of a least-squares line fit to the mean scores at each multiplier λ_i over the grid $\Lambda = \{-1.5, -1.0, -0.5, 0, 0.5, 1.0, 1.5\}$. **Relative steerability** $s_{\text{rel}}(v_A, D_B, \Lambda) = s(v_A, D_B, \Lambda)/s(v_B, D_B, \Lambda)$ is the ratio of two such slopes measured in the *same* application context. They also vary the deployment context at readout, extracting under one prompt setting and applying under another.

Three concessions follow and we make all three explicitly. **(i)** The slope estimator is theirs; where §C.4 places T alongside a ratio of Braun's slopes, Braun borrows S from here, and the credit for the primitive belongs at this citation. **(ii)** T is a member of the relative-steerability family (a ratio of two effect summaries formed to describe transfer) so it was not a new object, and this paper says so at the first mention of T. **(iii)** "Steering vectors change effectiveness when the deployment context changes" is not ours to claim as new.

What is ours is the *axis*, and it is a design difference rather than a correction. A slope taken over $|\lambda| \leq 1.5$ is not a location parameter: it does not separate a change in attainable effect from a change in the dose that attains it. Two further separations are worth stating. Their context shift injects instructions that name the steered trait, so the steerability change is entangled with a shifted baseline propensity on the same axis, whereas a ReAct scaffold is behaviour-orthogonal to refusal. And their grid carries three positive multipliers gated at the top, which fails this paper's registered dose rule by a wide margin: even with their raw per-λ data a midpoint could not be fitted to it.

The obvious reinterpretation is dead, and we do not make it. We do **not** write that their conclusions are a multiplier-window artifact. They ran that ablation themselves and report that the overall trends in steerability remain highly consistent across multiplier ranges. Making the claim anyway would require arguing that their own registered robustness check is inadequate, using data we do not have. Their readout compounds it: a logit difference selected partly for linearity, and a horizontal shift of a locally linear curve does not change its slope.

One inherited overstatement we retire rather than repeat. A secondary source summarises this work as consistently finding lower effect sizes out of distribution. That is not what they report: their headline is that steering vectors generalise reasonably well, and their principal scatter falls on both sides of $x = y$ ($\rho = 0.891$ on Llama-2-7b, 0.694 on Qwen-1.5-14b). Those are not to be confused with their weaker panels (0.535/0.341; 0.769 in-distribution and 0.586 out; -0.26/-0.46). We cite the weaker, correct version.

C.2 The shape of the dose–response curve

Owned: the shape of a single steering dose–response curve, and the three results that constrain any susceptibility readout. Our estimand is the cross-context difference in the one quantity they derive for a single context and never compare across contexts.

Taimeskhanov et al. (2026) give the closest theoretical account of what a steering dose–response curve is, and three of their results are load-bearing here. Their **Thm 3.6** makes the concept-level response a $\tanh$ up to a context-dependent horizontal reparametrization ν_j, in a simplified model: they derive ν_j in closed form (their App. B.4) for a single context, and never form a cross-context difference, ratio or comparison of it. Their **Thm 3.8** makes the first-order term of $\Delta\mathrm{CE}(\alpha)$ exactly zero, for the training-set cross-entropy and under that theorem's assumptions. *We* read it as suggesting that a susceptibility defined as a first derivative of any similarly sign-symmetric aggregate will vanish likewise; that step is ours, not theirs, and our readouts are signed and directional because of it. Their **Prop 4.1** makes LayerNorm force prompt-independence at large $|\alpha|$, so the contexts converge toward a shared limit and the dynamic range that identifies any cross-context object collapses: a normalisation artifact. It drives the *raw* logits together; it does not by itself pin a baseline-subtracted ratio to exactly 1, since the contexts differ at $\alpha = 0$. We therefore determine a single paired-context α^* and report where the comparison sits relative to it. The fitted *midpoints* lie strictly inside α^* in both contexts; the swept grids do not, since the primary fit deliberately retains doses at and above the ceiling and reports α^* post hoc. The registered α^*-trimmed secondary is reported in the body, and it is weaker than the primary.

The delta is one word: they derive *one* curve's shape and its reparametrization analytically in a simplified model; we estimate the difference in that curve's location **between two deployment contexts**, empirically, with a CI. Their ν_j is a term inside our dependent variable, not identical to it (§H.2).

C.3 Per-example prediction of steering success within one context

*Owned: early-decoding residual geometry, and cheap forward-pass diagnostics, as predictors of steering outcome within a single context. Both are cited in the introduction. One is run as a baseline; the per-example geometry arm is registered and **not scored**, for the reason the Results give. This is where altitude discipline matters most.*

Fan et al. (2026) predict, within one fixed chat context, which (prompt, concept, strength) triples steer successfully, from early-decoding residual geometry; their Steering Geometry feature group is fully specified and is reused here as baseline P1. Billa (2026) predicts *where* (at which layer, for which family) a steering vector will succeed, from training-free forward-pass diagnostics: a logit-lens argmax accuracy A_{lin} and a perturbation sensitivity λ.

We take both as genuine prior art on their own question. Both of Billa's diagnostics are run as registered refutation arms. Fan et al.'s per-example Steering Geometry group, baseline P1, is registered and **not scored**: its classifier needs per-example success labels this roster never produced. The anchors we score them against are read from the published papers: A_{lin} predicts steering effectiveness at $\rho = +0.86$ to $+0.91$ *cross-model* across five models and layer selection at $\rho = +0.63$ to $+0.92$; the distinct pooled within-model figure is $\rho = +0.777$ at $n = 130$, and we never attribute the cross-model range to a single model. λ correlates with steering effectiveness at $\rho = -0.706$ pooled, reaching -0.990 on one family, over 26 layers and 5 families: high λ meaning fragile, meaning steers poorly. A_{lin} is a training-free top-1 argmax accuracy of the logit lens with no probe fitted, and we reproduce it exactly, applied through the model's own final norm with a frozen refusal/compliance token family, since refusal-steering has no native next token.

Altitude. We never write, at this paper's own altitude, that a cheap forward-pass diagnostic predicts steering effectiveness: that sentence is owned at this citation. Our forward-pass measurement forecasts a *cross-context horizontal shift*, a quantity none of these methods forms. Our predictor is kept provably distinct from A_{lin} (a soft expected loading of the *steered* softmax under a continuous signed loading, against hard argmax-membership of the *unsteered* logit lens) and the distinction is enforced by a unit test rather than asserted, because refuting a diagnostic with a relabelled copy of itself would be a circular refutation and would deserve to be caught.

And the arm we called cheap and decisive returned against us. The registered stop rule fired: one of these two published within-context diagnostics recovers dEC_{50} on this set at the roster's power (§7). The defensible verb for a transferred within-context feature is *does not transfer to* dEC_{50}, never *refuted*: neither incumbent claims a cross-context quantity, so a failure here falsifies nothing either of them asserts.

C.4 Steerability as a fitted dose–response slope

Owned: the steerability score S, a fitted dose–response slope. We concede the kinship up front, and decline to overstate it as an identity.

Braun (2026) defines a steerability score S as a fitted slope of the dose–response curve, borrowed from §C.1, to whom the credit for the primitive belongs. The prior measurement's per-model coupling T is the same kind of object as $S_{\text{agent}}/S_{\text{chat}}$, and the body says so in the first paragraph that mentions T. It is not the same quantity, and we do not claim it is: S is a least-squares slope fitted across a multiplier grid, whereas T is a ratio of effect summaries read at an operating point, so the two agree only where the response is proportional and linear. That is precisely the condition this paper's dose–response fits reject, which is why the distinction is worth keeping rather than rounding off. Our estimand is deliberately not a slope ratio: a slope ratio is dose-dependent and dies against a ceiling, whereas a horizontal displacement is estimable where a vertical gain ratio is not. Conceding the kinship is what lets the switch be argued as a repair rather than a rebrand; overstating it as an

identity would concede more than is true, and would sit badly beside our own finding that a ratio read at one operating point does not identify potency and efficacy.

His geometric covariates, stated as the position actually stands. The two published predictors of steerability there are a directional-agreement covariate (the mean cosine between each per-example activation difference and the resulting steering vector) and the discriminability index d' of positive and negative activations projected onto the difference-of-means line. Both are properties of the extraction pool and of the direction; because this design holds the direction fixed while the context varies, neither is a within-cell explanation of dEC_{50}, and they can only explain between-model variation in it. That test is pre-registered rather than asserted: both covariates, in both deployment frames, on the pinned roster, scored against dEC_{50} under the same frozen leave-one-model-out rule the other arms use. **At the time of writing that measurement has not been run, and this paper makes no claim to have conditioned on those covariates.** We note in advance that the roster supports $n = 4$ models and that no rank correlation at $n = 4$ can reach conventional significance (the exact two-sided permutation floor is 0.0833) so the conditioning is roster-limited by construction. That is a limit we state, not a result we defer.

C.5 "Context-dependent steering direction"

Owned: the phrase. We concede it by name. Li et al. (2026) introduced the phrase "context-dependent steering direction", and their Fig. 5 is a per-example slope. We concede the phrase explicitly and mark the difference in object: theirs is a per-example directional quantity in one setting, ours an aggregate cross-context *location* of a fixed direction carried into a second deployment format. We do not reuse their phrase for our estimand.

C.6 Layerwise-Jacobian and subspace transfer instruments

Skifstad et al. (2026) own a layerwise-Jacobian instrument and a subspace transfer metric, and measure within a single deployment format, so the cross-context ratio reported here is identically 1 in their setup and there is no two-context object to preempt. We pre-empt the obvious appendix question directly rather than wait to be asked: the contribution is the cross-context comparison their instrument is not run across, not a better within-context instrument.

C.7 Cross-architecture steering transfer

A complementary axis. Agarwal (2026) varies the target **model** with a translated concept direction over all 20 ordered pairs of five models, so the unit of analysis is the directed model pair. This paper varies the deployment **context** with model and vector held fixed, so the unit is the context pair. The axes are orthogonal. Two boundaries keep the paragraph from becoming a fight.

Altitude of claim. That headline is a **win rate**: whether an effect occurs, at the best positive strength per pair: $71.0\,\%$ against $68.0\,\%$ native, $65.7\,\%$ baseline and $67.3\,\%$ universal. Ours is a signed magnitude and direction of a location shift. A win rate cannot be confused with a displacement, so there is no overlap to concede.

Statistical honesty, at our own bar. We cite the cross-model advantage as that paper's own data reports it: directional but not significant, paired $t(4) = 1.85$, $p = 0.14$, with the abstract itself calling the gap non-significant. Presenting it as an established effect would violate the standard we hold ourselves to, and the honest version *strengthens* the estimand boundary: a binary quantity its own $N = 5$ leaves underpowered is precisely the small-N regime a leave-one-family-out design is built to survive. That work runs a 9-point strength sweep and plots raw strength–response curves but fits no dose–response location, which is the accurate sentence and the one we write.

C.8 The Cylindrical Representation Hypothesis

Not a novelty threat (single context, per-sample unit, never forms a cross-context object) but a validity threat, and handled, an asset. Gao et al. (2026) hold that steering sensitivity is controlled sample-specifically, so the per-sample quantity we aggregate has high variance by construction. Two facts from the full text turn that into support for the design.

First, **the variance is real and geometrically unpredictable at the sample level**. That paper reports no quantitative variance statistic, and its own third implication test finds difference-vector similarity does *not* predict steering behaviour (Pearson -0.034, $p > 0.05$; a robustness check on Qwen2.5-14B gives 0.019, $p = 0.297$). So the per-sample susceptibility variance the design must be powered against is not importable from that paper, and we estimate it empirically from our own chat-only pilots; the family-level contrast is powered against *that*, never against cell count.

Second, it supplies **a published mechanism for why the aggregate is the tractable unit**. If per-sample geometry does not predict per-sample outcome even within one context, a per-example predictor is in trouble by that paper's own result, and the stable object is an aggregate, which is what dEC_{50} is. We cite it as corroborating mechanism for the predicted failure of the per-example baseline, and carefully: it shows per-example geometry does not predict per-example outcome within one context, while the arms concern per-example predictors of a cross-context quantity. This is evidence about mechanism and does not constitute proof.

C.9 The predecessor measurement

The systematic chat-to-agent transfer measurement this paper reinterprets (Pinto, 2026) established that the chat-extracted direction survives into ReAct deployment at near-full residual strength while the behavioural coupling T is reset per model, and (to its credit) publicly refuted its own pre-registered baseline-shift predictor out of sample. We cite that refutation as prior art and as a credibility asset. This paper is not the extended version of it: it changes the estimand from a dose-dependent gain ratio to a horizontal location shift, motivated directly by the theory of §C.2, establishes that the ratio does not identify potency and efficacy on grids built to support a midpoint, and reports a registered forecast of the location shift that was refuted out of sample. Its published numbers are reinterpreted under the new estimand; its claim is not extended.

D Methods, in full

The body states the estimand and the scope conditions in one section. This is the complete protocol. Every procedure below is fixed in a dated pre-registration committed before the run it governs; the register in §E names which one, and is machine-scored against this text rather than trusted.

D.1 The estimand

Why not the gain ratio. The prior measurement on this question reports a per-model behavioural coupling $T = \Delta_{\text{agent}}/\Delta_{\text{chat}}$: the ratio of steering-induced behaviour change in the ReAct-agent context to that in the chat context, read at a matched operating point. T is not wrong. It is a well-defined functional of two curves and the prior paper stated the operating point at which it was read. The claim made against it is narrower and it is an **identification** claim: *a scalar read at one operating point does not determine the pair (potency shift, coherent efficacy)*. We establish it by reconstruction rather than by argument. The prior estimator is re-implemented in both of its published forms: per-dose effect, single-dose ratio, AUC-over-doses ratio, and the sub-saturation filter at margin 0.05: copied from its own source, and re-run on grids swept to their plateaus. Every per-dose effect recomputed reproduces the committed fit's own record exactly (maximum absolute difference 0.0, 7 of 7 cells), so what the reconstruction shows is a property of the estimator and not of a re-implementation.

Two of *our own* registered predictions were refuted in that section and both constrain what the paper may say downstream. On grids built to the dose rule, T is computable in every cell (34 of 48 cell$\times$dose pairs are defined and no cell yields none) so the prior paper's three families that "yield no coupling estimate" are a *grid* property and not a model property. And the identification level set contains zero in only 2 of 3 scorable cells against a registered threshold of 3.

The estimand and its poolable companion. $\mathrm{dEC}_{50} = m_{\text{agent}} - m_{\text{chat}}$, where m is the fitted 4PL midpoint of the registered combined-binary response in each deployment context, in commanded-coefficient units. Its scale-free companion is $\Lambda = \log m_{\text{agent}} - \log m_{\text{chat}}$. Theory motivates each property of that choice and we say which result motivates which, and where the step from their result to our design is ours rather than theirs: Thm 3.6's reparametrization acts on a curve location,

which is why we estimate one; Thm 3.8's vanishing first-order term is proved for the training-set cross-entropy, and it is *our* reading, not theirs, that a sign-symmetric first-derivative susceptibility will vanish likewise, so we make the readouts signed and directional, which a midpoint difference is; Prop 4.1 makes any cross-context object lose identifiability at large dose, and a midpoint (unlike a top-dose ratio) can be checked against that ceiling, and is.

Two dose axes are not commensurable across model families, which is why dEC_{50} is reported per cell and Λ carries the cross-family panel: Λ is invariant to a rescaling of the dose axis where dEC_{50} scales with it. Λ is **not new to this project and is not claimed as new**: the same log ratio was computed weeks earlier, behind a stricter gate, by a different instrument, and the later instrument reproduces it to the printed digit in 5 of 5 cells. The independent reproduction is reported as a reproduction.

What the estimand is not claimed to be. Two registered tests bound the claim from opposite sides and both are reported whichever way they come out. A shape-invariance test, registered before the fit, compares a horizontal reparametrization against a vertical rescaling *at equal model complexity* and reports ΔAIC intervals per cell. And a translation test registers $\Delta_q = \mathrm{dEC}_{0.75} - \mathrm{dEC}_{0.25}$, which is zero for a pure translation: the paper is bound by its outcome, and because that outcome rejects, no sentence in this paper may call the two curves translates of one another. The same registration carries the upper-plateau comparison. That comparison establishes only that the fitted asymptotes agree; it does not establish equal achievable effect, and our own therapeutic-window audit says why it cannot: on a refusal-bypass cell the asymptote is the floor the refusal rate is driven to, so the contrast is blind to how much refusal there was to remove.

D.2 Contexts, the control spine, and the metric

The contrast is between two deployment contexts related by a **matched-information template ladder**: **C0**, the bare chat turn, and **C3**, the same task wrapped in a ReAct scaffold with deterministic dummy tools of fixed names, signatures and return values.

The match is exact on the task and asymmetric on one span, and we state both. Audited on the committed model-facing prompts across all seven probe cells (six distinct models) $\times$ 200 items $\times$ both contexts, the task instruction is **byte-identical**, occurs exactly once in each context, and the rendered prompts are byte-equal outside the system span: 1400 of 1400 item checks, each rebuilt from its pinned tokenizer revision. The C0$\rightarrow$C3 token delta is a per-cell constant identical for every item ($+144$ tokens on six cells, $+158$ on the seventh), so the wrapper is exactly additive. What is *not* matched is one span: C0 carries a compliance-valenced directive that C3 does not carry at all, so the contrast bundles the ReAct frame with the removal of a compliance instruction. That non-frame component is the first rung of the predecessor's frame ladder, whose marginal CI spans zero against a format-wrapper rung whose CI excludes it; we report it as an upper **bound** and draw no licence from it, because it was measured on T at a single dose on one family and arm. Prompt *length* cannot be matched (the agent prompt is 3.3–5.1$\times$ longer by construction) which is why the length control acts on the *output*.

The spine as executed, with its errata carried rather than buried. Appendix A states the four record-level errata; here is where each one bites.

- **A: matched-norm random directions, ≥ 5 per cell**, swept at every dose in both contexts on the same item identifiers. They enter as the per-dose reported diagnostic $\Delta(\text{real}) - \max|\Delta(\text{random})|$ and as the α^* separability screen; the fitted response is **not** the random-corrected one, and applying the correction as a sensitivity leaves sign, CI-exclusion and status invariant in 5 of 8 cells. Control A is a per-dose diagnostic and a gate, never a correction of the fitted response.
- **F: KV recomputation at every turn**, verified by execution against the pinned generator source with five predicates each shown able to fail. It guarantees turn-uniform, position-uniform *re-application*; it is *not* what excludes the prompt–activation duality account, because our steering is turn-constant and a correct cache would be numerically equivalent. §D.6 registers the missing half as a measurement. **The guarantee is bounded by how often a rollout actually reaches the turns it covers**, and the registered rule is that it may

not be quoted without that bound: at the agent midpoint the proven five-turn binding lower bound is 0.729 on Llama-3.1-8B and 0.639 on `OLMo-2-7B:bypass`, but 0.040 on the weakest α^*-GREEN cell. The roster weakest is `OLMo-2-7B:induce` at 0.000, and exact turn depth is not recorded.

- **M: the combined binary** (refused *and* coherent), never judge-raw. The inherited name *length-controlled* does not hold on the property it asserts: the correction the judge applies is length-correlated as registered but **one-directional**, so it strengthens rather than removes the metric's own length association. A second, estimand-specific limit is measured rather than argued: a context-differential one-directional correction displaces the two midpoints by different amounts: *the same signature as* dEC_{50} *itself*. The correction is therefore transported to act identically in both contexts and the frozen fitter re-run on the transported records, gated by two identity checks. Sign is stable 5 of 5 and CI exclusion 4 of 5, and the one cell that moves moves *toward* exclusion. No canonical number moves, and the registration forbids reporting a transported value as the paper's dEC_{50} under every branch.
- **G: the chat gate**, a swing of at least 25 points per family per arm. The clause *"and beats matched random"* **never ran**: admission gates on the bootstrap CI lower bound of the chat swing alone, proven by mutation, and the generator's own two-clause field has zero consumers. Applying the missing clause now removes exactly one cell, which one is metric-dependent, and **no sign-carrier is removed under either rule**. A gate failure is a **drop after diagnosis**, never a re-run at a bigger coefficient: the inherited lesson is that a baseline refusal of 0.200 caps the achievable swing below the gate, and no coefficient clears a floor.
- **N and BS: greedy decoding, 200 items per cell with identical item identifiers, and a** $B = 10\,000$ **item-paired bootstrap.** Greedy is verified on all $137\,600$ canonical rollout records. The *fitted* set is smaller and dose-dependent, because coherence-gated rollouts are dropped before the fit: the resampled anchor is 195–200 and the per-dose paired count falls as low as 0. And the pairing, while genuinely implemented, is **inert for this estimand**: dEC_{50} is a difference of fitted *locations*, which does not inherit the cancellation a difference of means at a matched dose does.

Arm selection is pre-committed per family from precheck baselines, choosing the arm (induce (raising refusal) or bypass (lowering it)) with dynamic range in *both* contexts. Where a family has range on one arm only, that arm is frozen in advance. **Statistical honesty:** a sign claim requires the bootstrap CI to exclude zero; we never claim significance where an interval spans the null, and a null is reported as a null.

D.3 Dose grids, the validity ceiling, and admission

Fitting a midpoint to a narrow grid is the single most expensive way to be wrong here, and the inherited grids demonstrate it: the predecessor's coupling grids cluster just under saturation, two points about ten percent apart, and cannot support a midpoint fit at all.

The dose rule: at least five doses spanning at least $4\times$, bracketing the fitted midpoint, in *both* contexts, or the cell is not fitted. No full grid is committed blind: a cheap chat-only pilot locates the transition first, and the paid grid is built from a frozen width prior around it.

The opposite-sign pair, and the comparison set it was drawn from. The overlapping-gain pair reported in §3 is an *existence* result and we give its denominator: the registered scan compares every negative cell×dose against every positive one, 80 pairs in all, of which 9 have overlapping gain intervals and 5 have both intervals tight. Every one of the 9 uses the sole positive cell, but their negative sides span three distinct cells (OLMo-2-7B bypass, Qwen2.5-7B bypass, Qwen2.5-7B induce), so the search ranges over negative cells as well as over doses. The claim the pair supports is that a gain can coincide across opposite potency shifts, not that it commonly does.

What that rule costs, censused over the predecessor's entire raw corpus rather than argued from one configuration per family. The paragraph above was originally read off the coupling configurations alone, which is a sample of the corpus and not the corpus. We have since streamed all 79 of the predecessor's rollout files end to end — $349\,843$ records — and scored every grid in them against the rule: 0 unparseable lines, 0 records whose deployment context the declared map

could not classify, and 0 disagreements between the dose named in a run's setting label and the injected coefficient's magnitude. The corpus holds 73 (model, arm, layer) cells, of which 25 carry a dose grid in *both* deployment contexts. **Exactly one of the** 25 **meets the rule in both, and it is the cell whose midpoints the predecessor already published** (Qwen2.5-7B, induce, $L16$), passing inside that configuration's own file — the only file in the corpus that passes alone. Of the 22 two-context pairs outside the three configurations that already carry a published midpoint, **none** does. The instrument recovers the one known-good cell and nothing else, on thresholds fixed before the corpus was read. This is the answer to *why pay for new sweeps instead of refitting what is already published*: refitting was not an alternative that existed. The roster is small because the corpus is, not because the search stopped.

The honest qualifier is a single near miss, and it does not become a free cell. Llama-3.1-8B on the *bypass* arm carries 5 chat doses spanning $8\times$ and 4 agent doses spanning $4\times$: it fails on the counting clause, in the agent context only, by one dose. That is the pair we would most want, because this paper carries Llama on the induce arm alone, where it is the roster's only positive cell, so a bypass cell would be a second arm on the positive anchor. It is not one. The response rate is 0.72 at dose 4 and 0.000 at 8, 12, 16 and 32 in *both* contexts, so the whole transition lies inside $(4, 8)$ and no swept dose lies inside it; a fifth agent dose anywhere outside that interval satisfies the clause and carries no information about the midpoint. The pair also exists only as a union of seven files whose item universes range from 40 to 200, and no single file passes. We report two further readings rather than choosing the flattering one. The rule as stated is **stricter than two of the three predecessor configurations that do publish a midpoint** (3 doses over $4\times$; 5 doses over $2\times$), which makes a null under it easier to obtain than under the exemplars' own shape; relaxing it to that weakest endorsed shape (≥ 3 doses spanning $\geq 4\times$) admits exactly this near miss and no other pair. Both readings are descriptive: neither gates anything here, and neither would have licensed a fit.

The α^* validity ceiling operationalises Prop 4.1. A cell is admitted only if both fitted midpoints sit below the binding α^*. That ceiling is behavioural, not logit-level, and it is computed with the two contexts *merged*. Writing $\Delta S(\alpha)$ for the real direction's cross-context success-rate separation minus the median separation of matched-norm random directions, α^* is the first dose *above the peak of* ΔS at which $\Delta S \leq 0$. The high-side anchor is load-bearing rather than cosmetic: two ascending curves with different midpoints also separate little at low dose, and that ramp is not the merged limit this ceiling exists to catch. A within-context saturation candidate is computed and reported and does *not* bind. **Three cells have no located ceiling**: for Gemma-3-12B induce, Qwen2.5-3B induce and Qwen2.5-7B induce, ΔS never returns to ≤ 0 above its peak anywhere on the swept grid, so α^* is undefined for them and every swept dose is in-regime. Their GREEN therefore records *no ceiling reached on this grid*, a weaker statement than the GREEN of a cell whose midpoints were checked against a located ceiling, and one bounded by the grid we swept rather than by the model. We name the ceiling this way rather than as a prompt-dependence regime, because the estimator reads rates and never reads logits. That screen sets the in-regime dose set and each cell's verdict. A within-context diagnostic is computed per context and reported alongside it. That diagnostic fires at the lowest dose where *either* the signed real effect stops exceeding the matched-norm random band *or* the fitted upper asymptote is reached, and the second branch is not decorative: `OLMo-2-7B`'s chat leg has a positive real-minus-random margin at every dose and its diagnostic still fires, at 30, on saturation alone. It does not bind. Llama is α^*-GREEN at $\alpha^* = 9$ with that diagnostic reading 1. α^*-RED cells are **printed and excluded from every law claim**, and from the heterogeneity specification this paper reports (model unit, α^*-GREEN only). They are law-excluded rather than merely flagged, and they bind no frozen hyper-parameter. **Registered aggregates that do contain them, which we do not claim otherwise**: the fitter's `power_crh` block keeps one representative per *fitter label* (the cell with the smallest standard error). There are seven such labels, which is neither the eight cells nor the six distinct base models: Qwen2.5-7B's two arms carry different labels and are never collapsed, and the artifact records this as `k_is_neither_unit`. Both α^*-RED representatives survive that reduction, and the dEC_{50} forest figure's title reports the block, labelled on its face; and the canonical verdict block counts every fitted cell without an α^* filter, so RED `Starling-7B:bypass` raises `n_families_fitted`, RED `OLMo-2-7B:induce` enters `pos_sign_cells`, and the two-sidedness and phase flags are unfiltered. We read no law from either. The GREEN-only two-sidedness test is reported separately and passes on its own, which is the version the law rests on. This is a weaker position than the registration's blanket exclusion and is stated rather than papered over. The exception is diagnostic and labelled as such: the dEC_{50}

forest figure's title reports the heterogeneity block *as the fitter emitted it*, which includes them, and says so on its face so it cannot be read as the registered figure.

Is the fitted shift resolved by the grid, and does its sign need the 4PL? A registered resolution check requires the displacement to be larger than the grid can hide, and a fit-free reading of the same sign means the headline does not rest on a parametric form. The estimator fits the full swept grid and reports α^* post hoc; it does *not* apply the stricter dose-exclusion sentence that two of our own pre-registrations also register. Re-run under that rule (same estimator, seed, B and items) the roster's only positive cell keeps its sign and loses CI exclusion. The registered primary does not move, because the full-grid rule was in force before any paid data existed, but the number that would change a reader's mind is reported in the results and not only in the limitations.

The random arm's item universe is itself a registered choice, because the separability screen reads $\max|\Delta(\text{random})|$ and a narrower universe makes the screen easier to pass. The sensitivity is reported beside the α^* analysis rather than folded into it.

The grid that was swept is not the grid that was fitted, so the rule is re-checked on the realized support. Each context's dose vector is derived downstream from the rollouts that survived scoring, and a rollout is scored only if its metric field is in $\{0, 1\}$ *and* it was coherent. A dose at which every rollout in one context has collapsed therefore produces no key and leaves that context's grid with nothing recording that it was swept, which the launch predicate cannot see, because it is a property of the realized data. The weight has the same shape: the weighted 4PL floors a bootstrap sd below a minimum so that a *saturated* dose cannot take infinite weight, and the same line floors a *non-finite* sd, so a dose whose uncertainty could not be computed is given the smallest uncertainty in the vector. We therefore re-score the realized support of every cell in the canonical fit on four rails, over law-included cells only: **(D1)** both contexts are fitted on the same dose set; **(D2)** every fitted dose has a finite bootstrap sd; **(D3)** no swept dose is removed from a fitted grid by the scoring filter; **(D4)** the dose rule above holds on the grid that was fitted rather than the grid that was planned. All four pass on all six law-included cells and all twelve legs, at a minimum scored n of 26 per leg.

The one cell at which either condition fires is `OLMo-2-7B:induce`, which is α^*-RED and already excluded from every law claim, and both fire there: at dose 43 the chat context has 0 of 200 rollouts coherent and the dose leaves its grid, the agent context keeps 2 and retains it, and that retained point's undefined bootstrap sd is floored to the minimum. The cell is discussed as an artifact in the main text; what these rails add is the two-directional statement that the mechanism reaches no cell the law uses.

D.4 Estimation, inference, and the analytic multiverse

The fit. A four-parameter logistic in each context, with m its midpoint. Bootstrap $B = 10\,000$, item-paired, seed fixed in advance. The per-dose bootstrap standard deviation weights the fit, which is why B is not a cost dial: lowering it moves the fitted location rather than only widening the interval.

Is the interval calibrated? A registered simulation calibrates the reported dEC_{50} interval against known ground truth on the realized grid geometry, with the repair (recomputing the per-dose weights *within* each resample) registered **conditionally** before any replicate was drawn, and applied to all law-eligible cells or to none. Point estimates move by less than 5×10^{-4} and no cell's CI-exclusion status changes. Prereg §6.1 requires *both* intervals published, the committed one dated rather than deleted, so both are printed here. Writing committed $\rightarrow$ repaired: `Llama-3.1-8B:induce` $+1.013$, $[+0.777, +1.273] \rightarrow [+0.760, +1.295]$; `OLMo-2-7B:bypass` -5.497, $[-7.851, -3.175] \rightarrow [-7.687, -3.045]$; `Qwen2.5-3B:induce` -0.237, $[-1.303, +0.818] \rightarrow [-1.358, +0.818]$; `Qwen2.5-7B:bypass` -10.855, $[-15.814, -4.614] \rightarrow [-15.790, -4.253]$; and `Qwen2.5-7B:induce` -12.368, $[-13.726, -10.876] \rightarrow [-13.915, -11.017]$. Four cells clear zero before and after, the one that never did still does not, and Table 2 prints the committed interval because that is the one the registration dates. Three cells carry no repaired interval, for two different reasons, and the distinction is stated rather than smoothed: two are α^*-RED and outside the repair's frozen roster, while one is law-eligible but *post-dates* that roster, which was frozen before the seventh family existed. Re-running the repair to admit it would re-open a roster frozen before any replicate was drawn, so the gap is named rather than closed by amendment.

Power, at the unit the paper reports. Because steering sensitivity is controlled sample-specifically (§C.8), the family-level contrast is powered against **per-sample susceptibility variance** rather than against cell count, and the per-sample variance is reported here rather than only asserted, with its mean beside it as the registration asks: across the sixteen cell-context estimates the variance spans 0.130 to 0.278 and the mean spans -0.285 to $+0.809$. The scorer previously stored only the variance, so an earlier version of this paper reported the range and stated that no mean was available; the scorer now records both, and the refit that produced them is bit-identical to the committed artifact in every other field. **The unit of analysis is the model, never the cell** (two arms of one model are one model) and every roster count in the paper is stated at that unit.

The selection bound, signed and measured. Rather than hedge about the roster's admission geometry we measure it. All three agent-side measurability gates are monotone in absolute dose, so the admission window reaches less far above zero than below it, and the design under-observes exactly the sign that is scarce. The observed positive rate is therefore reported as **biased downward by the admission geometry** (not as a bound on a population rate, which five cells drawn under no sampling frame cannot support) with the share attributable to our own width prior separated out ($63.0\,\%$ of the pooled total).

The analytic multiverse. A registered specification curve varies response level, link function, fit window and metric, and reports the **sign agreement rate** across specifications. The defensible claim is about the *sign*: $\mathrm{dEC}_{0.50}$ varies by up to $2.99\times$ across links and the AIC-preferred link is our registered 4PL in only 1 of 5 cells: a prediction our own pre-registration made *against* our registered choice, and it held.

Coherence is fitted as an outcome and never used to exclude a cell. The coherent margin $M = m_{\mathrm{coh}} - m$ is registered as a second dose–response dimension: where the behavioural midpoint sits relative to the dose at which coherence half-collapses. This is what makes a fitted midpoint readable as behaviour rather than as decoding failure, and it is the instrument that identifies the one cell this paper quarantines.

D.5 Roster, and how the seventh family was chosen

The backbone roster is inherited (continuity is what makes this a reinterpretation of published numbers rather than a cold start) and every family passes the chat gate on a pre-committed arm or is dropped before it consumes a verdict.

The seventh family was selected under a registered protocol that exists to defeat one specific objection: choosing a family *because* our forecast scores it well and then validating the forecast on it. The registration therefore fixes, before the run: the predictions of **all three** forecasting arms for the chosen family, so the new datapoint is a fair three-way test and not a showcase; selection on **predictor-definedness** as well as on predicted sign, so a family predicted to sit at its own noise floor is not bought; and the **full candidate ranking**, with each arm's forecast for each candidate and the reason the winner won, so the selection is auditable. A candidate that failed definedness was dropped under that rule rather than run in hope.

D.6 Deflationary accounts, measured before generation

Four mechanisms could produce a context-dependent horizontal displacement with no reparametrization at all. Three are measured and rejected; the fourth is excluded by construction, and the difference in logical status is stated in the first sentence of its subsection rather than left to the reader. Each of the three is measurable from forward passes alone, before any generation, and each was registered before the quantity it measures existed anywhere in this project.

Residual-norm rescaling. *The account:* the agent prompt has a larger residual norm, so a fixed α is a smaller relative perturbation, so the midpoint moves: a pure scale effect. *The measurement:* the norm ratio $\rho = \|r\|_{\mathrm{agent}}/\|r\|_{\mathrm{chat}}$ is estimated per cell with a CI and its *sign pattern* compared against Λ's. The registered falsification is structural rather than quantitative: a scale account predicts **one sign for every model**, so a two-sided Λ refutes it whatever the magnitudes.

Baseline alignment. *The account, and it is the sharper one:* the norm law is the $c_0 = 0$ special case of the general drive $\cos\theta_v(g) = (c_0 + g)/\sqrt{1 + 2c_0 g + g^2}$ with $g = \alpha/\|r_L\|$. If the unsteered residual is already partly aligned with the steering direction and that alignment differs by context, the composed scale-plus-translation account generates a displacement with no reparametrization. *The measurement*, committed before any cosine between a residual and a steering direction had been computed once in this project, and CPU-only: four registered questions (is c_0 nonzero, does it differ by context, does the *composed law* reproduce the measured Λ, and does the affine form in the literature fit) each with its rejection threshold registered in advance. The composed-law question carries a *count* threshold, and its verdict is quoted as "k of n scorable cells reject, against a rejection threshold registered in advance", never as a bare ratio implying unanimity.

Scope, shipped in the artifact's own field. Every c_0 above is read at the **final prompt token**, while the deployment operator adds $\alpha \cdot u$ at **every** position; a prompt-block readout also understates generation-time variance. A registered completion removes the first half of that limit by measuring c_0 over the whole steered position set, block-partitioned into the instruction span the two contexts share verbatim and the scaffold span only C3 has. That partition makes the instruction-block contrast a context effect on a *fixed token set*: the matched-information control applied to a geometric readout. A free preflight established that all six roster cells partition cleanly and that the instruction block is token-identical across contexts in every one.

Dose transduction. *The account:* dEC_{50} is a difference of midpoints in **commanded-coefficient** units, and there is no reason a commanded unit buys the same residual displacement in a long ReAct prompt as in a short chat prompt. *The measurement:* the delivered displacement per commanded unit is measured per context with an interval, the objection is scored on its own terms, and the location is then **re-expressed in chat-delivered units** by the algebraic correction $R\,m_3 - m_0$. No curve is re-fitted, and we do not claim one is. The registered reading requires all three: whether the axes are commensurable, how much of dEC_{50} delivery explains and *with what sign*, and whether sign and CI-exclusion survive the correction. **Scope, binding on every quotation of this result**: $n = 3$ cells over two models, one behaviour family, one dose per cell; R comes from a 120-item projection replay while m_0, m_3 and dEC_{50} come from independent behavioural sweeps, so it is not a paired comparison; and m_0, m_3 are normal approximations.

KV-cache reuse: excluded by construction, and then registered as a measurement. This account has a different logical status from the three above and the paper says so in its first sentence: those are mechanisms we measured and rejected, this one is a mechanism our design never admits. Kang et al. (2026) attribute multi-turn coherence failure to KV-cache contamination: each generated token writes steered states into the cache, so later tokens attend to an increasing set of perturbed states. If a ReAct rollout reused its cache across turns, the agent response could differ for a reason unrelated to a dose–response reparametrization.

The design: every agent-context rollout behind every dEC_{50} in this paper runs under **mandatory full re-encode at every turn**, verified structurally in both agent-context paths. In the single-item loop each turn rebuilds the prompt string and passes a list of *strings*, with no parameter through which a cache could be threaded, plus a hard assertion naming the control; in the batched path, batching is across items within a turn and recompute across turns within an item, and the two are orthogonal. The check *can fail* (adding a cache parameter to either call site flips it) which is what makes it a verification rather than a restatement.

Three scope limits bind, and all three are stated: we have **not** measured what cache reuse contributes, only that our result does not depend on it; we do **not** say that account is wrong, only that the mechanism is untested there and inapplicable here; and recompute guarantees *re-application*, not duality exclusion, because under turn-constant steering a *correct* cache would be numerically equivalent. §G describes the registered arm that converts the first limit into a measurement.

D.7 Frame, extraction, and injection-site controls

The objection, and it is mechanical: the direction is chat-native, so of course it behaves differently under a ReAct scaffold: you are measuring how badly a chat-fitted probe is mis-specified for the agent context and calling the mis-specification a reparametrization.

The **extraction-frame control** answers it on already-existing data. The predecessor project ran a pre-registered, clean-room-verified reciprocal-extraction matrix in which the same instructions were also rendered inside the ReAct frame, and this paper reads that matrix on its own estimand for the first time. Four questions are registered (sign invariance across frames, magnitude stability against a registered ratio bound, whether the *agent-native* direction *widens* the displacement, and whether the two arms agree) with the fourth's unfavourable branch written as unfavourable. That registration discloses in full that the analyst had already seen a point-estimate preview, so it is a **pre-specified analysis of already-public data, not a blind pre-data registration**, and it is described that way wherever it is cited.

The injection layer is the other nuisance parameter of a cross-context estimand that was chosen on one of the two legs, and a registered sweep bounds what that choice buys. The two incidents (extraction frame and injection layer) are named in the limitations as a **class**, not as two unrelated events. Four further frame controls are registered and reported in their own companion documents: the predecessor's five-rung matched-information frame ladder read on the horizontal estimand, the same ladder on the refusal arm, the frame-state ablation and conditional-transport stages, and the injection-*site* confound.

D.8 The forecast and the two refutation arms

The forecast predicts dEC_{50} from a pre-generation directional susceptibility measurement, **leave-one-family-out, with no per-model fitted parameter**. Because Thm 3.8 zeroes the first-order term of the training-set cross-entropy, and because we read that as extending to sign-symmetric aggregates generally, the susceptibility is signed and directional by construction; a symmetric readout would on that reading be identically zero and is used nowhere in this paper.

The two refutation arms run before anything downstream depends on the forecast, because they are cheap, forward-pass only, and decisive. A published within-context perturbation sensitivity, and a published logit-lens accuracy, are each fit per context and each scored as a predictor of dEC_{50} on the same leave-one-family-out protocol, against a naive roster-mean baseline. The published *per-example* geometry classifier is registered and **not scored**: it needs per-example success labels this roster never produced. The registered stop rule is explicit and it fired: *if a published within-context diagnostic also recovers* dEC_{50} *on this set, the forecast contributes nothing that is separable at this roster size*. That is reported rather than routed around. An identification gate registered with the arms bounds when a leave-one-out affine map on a small N is identified against a collinear competitor at all.

The out-of-sample test is the one that settles it. A signed prediction for a family whose dEC_{50} did not yet exist (sign, normalised rank, and both refutation clauses) was committed before that family's sweep ran. Both clauses failed. We report the refutation as the result, we do not describe the forecast as a validated predictor, and we take no not-applicable escape that the registered branches did not authorise. **No entrant is promoted after the fact**: a five-entrant gauge board registered before the landing is scored as registered, its own reading is gauge-dependent, and the post-hoc winner is not promoted.

D.9 Register-then-run, and what the record can and cannot re-run

Every paid experiment has a design document and a pre-registration with live/die criteria committed before launch, and the corpus of those registrations is the denominator of two instruments. One asks whether every registered *rule* has a recorded *verdict* and where; §E asks whether every registered *design* has a stated *procedure* and where. Both fail closed: a rule answered nowhere is a defect rather than a silent pass, and so is a design described nowhere.

Guards are run as a suite, with the exceptions named. Each control's verification lives in an executable guard, and every guard is required to be able to **fail**: verified by mutation, not by review. Until late in the project no session had run them as a suite; the first full-suite run returned four red where the operating log tracked two, and none of the four named the defect it actually had. The suite now carries a declared environment per guard and a third status, *not runnable here*, because an absent environment is neither a pass nor a failure. **We do not claim a uniformly green suite. We claim a suite that has been run as a suite, with its exceptions named.**

Data availability, measured rather than asserted. The reproducibility census is re-measured against the tree a reviewer actually receives rather than against the working tree, which holds tens of gigabytes of rollout records and activation bundles that the ignore file excludes. Fitted dEC_{50} values, the control-spine audits and the forecasts are computed from data that is **not** in the code repository. **We do not claim that the repository reproduces the results**; we state what it does regenerate and name what it does not. Run-time provenance is a separate and weaker record than pinning, and is reported as such: a pinned revision records what was *requested*, and only a capture taken while the machine is alive records what was *loaded*.

E THE STUDY REGISTER: EVERY REGISTERED DESIGN, AND WHERE ITS PROCEDURE IS STATED

This table is the denominator's counterpart and it is **scored rather than taken on trust**. An audit takes the mechanical set of pre-registration documents released with the paper as its denominator; a registration missing from this table is a defect, a row naming a section that does not exist (or a section that does not cite it) is a defect, and a row deferring to a document that has since appeared is a defect the moment it appears, so the deferral clears itself rather than ageing into cover. Each key below names one pre-registration document, released with the paper. Dispositions: **main** (procedure in §D); **companion** (procedure in the named companion document, which must cite its own pre-registration); **infra.** (operational design with no reviewer-facing protocol); **not run** (registered, never run, reported as such, §G).

The register reports **47 of 47** with **0** defects, disposed as 25 stated in the full methods appendix, 18 in the named companion documents, 2 infrastructure and 2 registered-but-not-run. It was at **10 of 44** when it was first measured, which is the number Appendix B reports and the reason the instrument exists.

Table 3: The study register: all 47 registered designs, where each is reported, and what it fixes.

registration	where	section	what it fixes
`week1_dose_grid`	main	§D.2	the two-context sweep, the control spine and the arm-selection rule
`scalar_gain_nonidentification`	main	§D.1	the reconstruction of the prior estimator on our grids
`scale_commensurability`	main	§D.1	the identity gate under which the log ratio may be pooled
`potency_identification`	main	§D.1	translation test and upper-plateau comparison
`shape_invariance`	main	§D.1	horizontal against vertical at equal model complexity
`alpha_star`	main	§D.3	the Prop 4.1 validity ceiling and the separability screen
`dose_resolution`	main	§D.3	grid resolution, and the sign without the 4PL
`random_universe`	main	§D.3	the item universe of the matched-norm random arm
`estimator_calibration`	main	§D.4	is the reported interval calibrated on the realized geometry
`estimator_repair`	main	§D.4	the conditional within-resample weight repair
`power_analysis`	main	§D.4	powered against per-sample variance, at the model unit
`selection_correction`	main	§D.4	the admission geometry, measured and signed
`spec_curve`	main	§D.4	the analytic multiverse and the sign-agreement rate
`coherent_efficacy`	main	§D.4	coherence fitted as an outcome, never used as a filter
`sixth_family_selection`	main	§D.5	three-arm registration, definedness screen, full ranking

Table 3: The study register *(continued)*.

registration	where	section	what it fixes
norm_rescaling	main	§D.6	the residual-norm account and its sign-homogeneity test
baseline_alignment	main	§D.6	the composed drive law and its registered rejection count
baseline_alignment_allposition	main	§D.6	the all-position completion and the block partition
dose_transduction	main	§D.6	commanded against delivered dose, and the re-expressed law
cache_reuse	main	§D.6	the reuse leg paired against the committed recompute leg
cache_reuse_policy	main	§D.6	resolving what "cache reuse" means before the pod
extraction_frame	main	§D.7	the reciprocal-extraction matrix read on this estimand
layer_sensitivity	main	§D.7	what the one-leg choice of injection layer buys
predictor	main	§D.8	the signed directional-susceptibility forecast, LOFO
refutation_arms	main	§D.8	the two published baselines and the registered stop rule
judge_transport	companion	judge_transport_appendix	the context-differential judge correction, transported and re-fitted under two identity checks
positive_arm_replication	companion	positive_arm_replication_appendix	the positive arm re-fitted on a second, independent dose grid from the parent repository
format_ladder_potency	companion	format_ladder_potency_results	the matched-information format ladder read as a horizontal displacement rather than a gain
refusal_ladder	companion	refusal_ladder_results	the same matched-information ladder run on the refusal arm of the roster
frame_scaffold	companion	frame_scaffold_results	the parent's five-rung frame ladder, re-read on the horizontal estimand rather than on the gain
frame_transport	companion	frame_transport_stage0_results	frame-state ablation stages and the conditional transport arm that was gated on them
phase_dose	companion	phase_dose_results	the injection-site (phase) confound, measured against the horizontal displacement
protocol_integrity	companion	coherent_efficacy_results	agent protocol integrity treated as a context-exclusive response dimension of its own
generated_block_readout	companion	generated_block_readout_results	the steering axis read over the model's own generated tokens rather than over the prompt
therapeutic_window	companion	therapeutic_window_results	the deliverable dose-response, and this project's first efficacy-level rather than location estimand
baseline_offset	companion	baseline_offset_results	whether the parent paper's own refuted baseline-shift predictor recovers the horizontal estimand
gemma3_scaffold_gate	companion	gemma3_scaffold_gate_results	the recoverability gate and reader validity, measured on a fourth family before spending
gemma2_fold_control	companion	gemma2_fold_control_results	breaking the chat-template fold confound with a 2024-era model that folds the system turn

Table 3: The study register *(continued)*.

registration	where	section	what it fixes
qwen3_generation_control	companion	qwen3_generation_control_results	the empty cell of the generation-by-format design, and the first test of format capture there
q2_internlm_feasibility	companion	q2_internlm_feasibility_results	whether the salvage cell that the gain ratio loses is fundable at all before committing to it
phase_b_sharding	infra.	—	how the paid sweep is sharded across pods and co-tenants; a scheduling design with no estimand of its own
react_batching_revalidation	infra.	—	re-validation of batched ReAct against the already-accepted noise floor; a harness authorisation, not a reported measurement
scaffold_intervention	not run	§G	the causal scaffold edit, registered in full and never run: its own attainability analysis showed the registered verdict was unreachable at the affordable design, and the manuscript reports it as the experiment we did not run
braun_covariates	not run	§G	Braun's geometric covariates against the horizontal estimand, registered with both covariates in both contexts and not yet measured; Limitations states the gap rather than the prereg implying a result
roster_currency_rescore	companion	roster_currency_rescore_results	the deflationary-accounts re-score that admitted the post-freeze α^*-GREEN cell as a secondary panel; a record-integrity repair with no estimand of its own, reported because it moved one verdict (NR−Q6) and discharged threat 4.37
gemma_ladder	companion	gemma_ladder_results	the parent's Gemma C0/C2/C3 format ladder read on the horizontal estimand at both dose scales, localising the displacement to the format rung on a second family
parent_grid_census	companion	parent_grid_census_results	the complete dose-grid census of the parent's raw corpus against the fittability rule: whether any unfitted two-context cell exists to refit, answered over all 349,843 records

F ADDITIONAL FIGURES AND TABLES

The body carries one figure, because at nine pages a second float costs a paragraph of the identification argument. The rest of the registered figure set is here. Every panel is rendered from a committed fit artifact by a driver whose consumed fields were declared in a figure plan frozen before the fit, and every null render (what the panel would look like if the result had gone the other way) was written into that plan in advance. Two figures of the frozen set are **withdrawn rather than nulled**, and §F.1 says why, because drawing a pre-written null for an experiment that never ran would be a fabricated measurement.

What the scorecard earns, and what it does not. Our forecast dominates the recovering baseline on every point estimate: mean absolute error 0.28 and predictive R^2 0.9989, both **at gauge** G0_RAW ($\kappa = 0$) and at no other, against 2.96 and 0.908 for the incumbent; every gauge the registered audit calls defensible puts R^2_{pred} below zero, sign 4 of 4 against 3 of 4, and their error CIs **overlap heavily** ($[0.23, 4.86]$ against $[0.65, 7.16]$), so the two are not separable at $N = 4$ under the identification gate registered with the arms. It is the best available forecaster on these data, and the design lacks the power to certify it against the incumbent. It is not described as a validated predictor. The A_{lin} arm's predictive R^2 of -312.88 ($313\times$ worse than predicting the roster mean) means a within-context

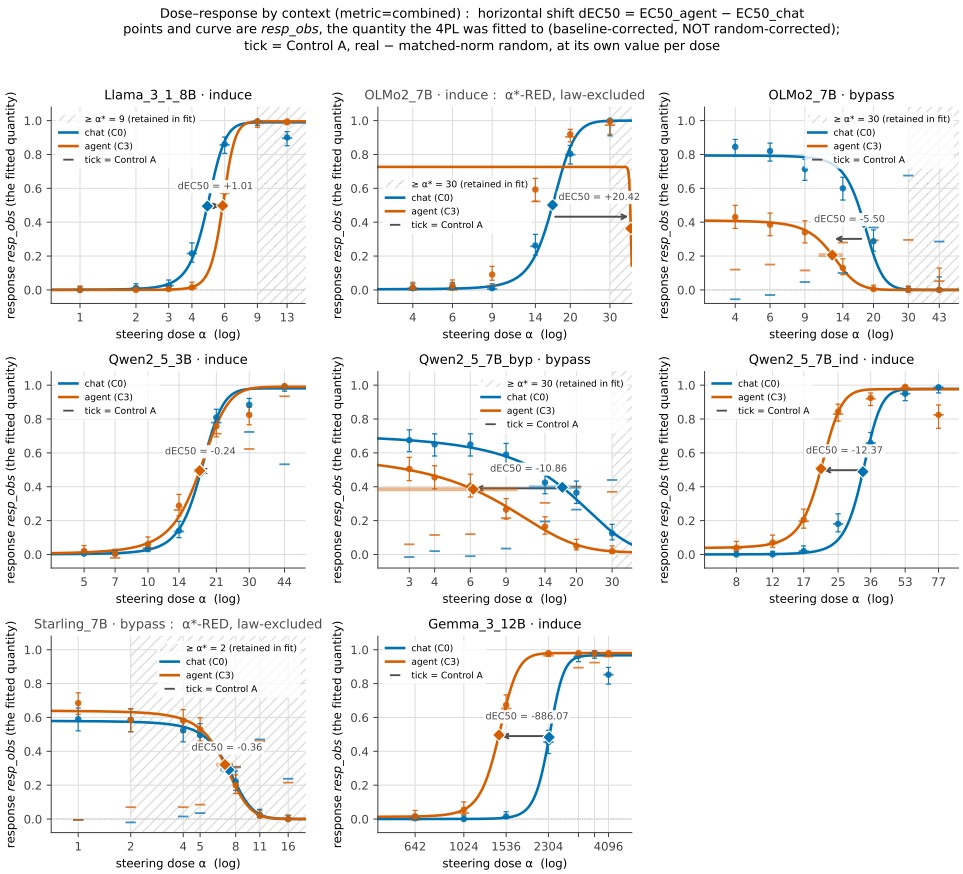

Figure 2: **The law, cell by cell.** One panel per fitted cell; each overlays the chat and agent observed responses, their fitted 4PL curves, midpoint markers with CIs, and a hatched region for doses at or above the validity ceiling α^*. **Points and curve are the same quantity**: the response the 4PL was fitted to, which is baseline-corrected and, as errata item 1 records, *not* random-corrected. Control A is drawn per dose as a separate tick at its own value, so the matched-norm random comparison is visible without being mistaken for the fitted series. Those doses are *retained* in the fit: the primary estimator is the full swept grid and α^* is reported post hoc (§D.3), so the hatching marks a scope condition, not an exclusion. The registered null render for this panel is overlapping curves with coincident midpoint bands: a family with no shift would be reported, not dropped.

Table 4: **The refutation scorecard.** Leave-one-family-out mean absolute error with its bootstrap CI, predictive R^2, and sign accuracy, for the naive intercept-only baseline, the two published within-context diagnostics, and our forecast. **The registered PASS condition was that the published baselines are refuted *and* ours recovers; the realized verdict is that a published baseline recovers**, so the condition is not met and this is reported as the headline negative. Read sign accuracy with care: the head-to-head families are all negative-signed, so a constant-sign rule scores 4 of 4 by construction: which is exactly what the naive roster-mean baseline does at $R^2_{\text{pred}} = 0.000$.

predictor	MAE (LOFO)	MAE 95 % CI	R^2_{pred}	sign	arm verdict
naive (roster mean)	9.39	[8.06, 10.84]	0.000	4/4	—
per-example geometry	not scored on this set				
A_{lin}	68.83	[58.76, 78.86]	$\mathbf{-312.88}$	4/4	does not transfer
perturbation sensitivity λ	2.96	[0.65, 7.16]	0.908	3/4	**recovers**
ours (gauge `G0_RAW`, $\kappa = 0$)	**0.28**	[0.23, 4.86]	**0.9989**	4/4	—

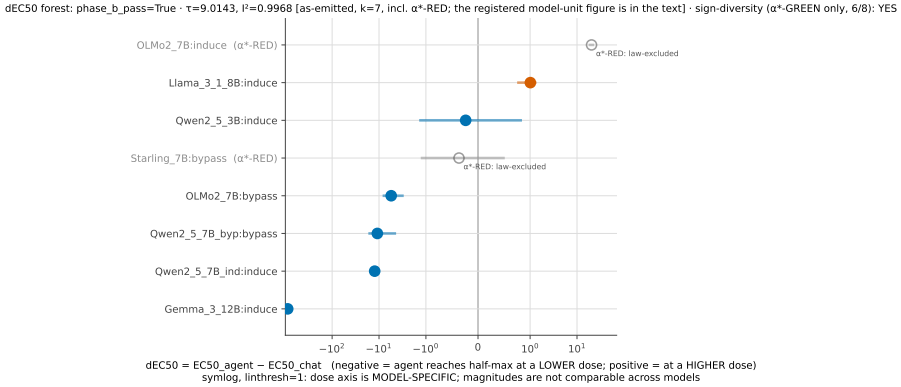

Figure 3: **Forest of dEC_{50}, and the heterogeneity the aggregate rests on.** One row per fitted cell, point and item-paired bootstrap CI, ordered by value, with the null line always in frame. The axis is symmetric-log rather than linear: on a linear axis one cell's model-specific dose units compress every other row to a point, which would be a presentational choice doing the work of a result. α^*-RED cells are drawn in a subordinate style and tagged law-excluded: they stay *in frame*, because dropping them would be the mirror-image dishonesty, but they cannot be misread as evidence, and the sign-diversity annotation is computed over GREEN cells only. The legend deliberately does not say "amplifier" or "attenuator": those are gain words, and dEC_{50} is a location.

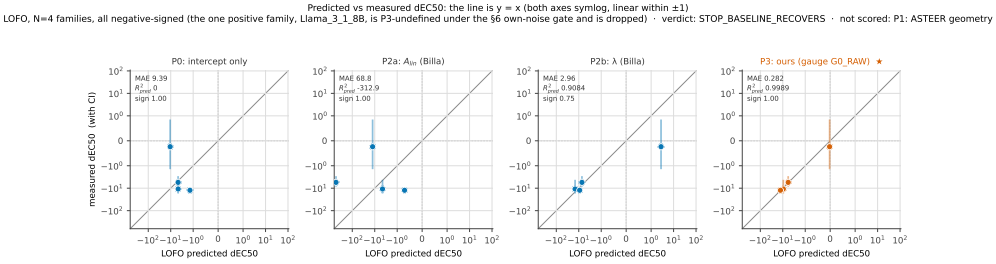

Figure 4: **Leave-one-family-out forecast against measured dEC_{50}, one facet per predictor. Our facet's printed MAE and R^2_{pred} are at gauge `G0_RAW` ($\kappa = 0$) and hold at no other**: every gauge the registered audit calls defensible puts R^2_{pred} below zero, and the slope's sign moves with the gauge. The facet set is the four predictors that were actually scored; the per-example geometry baseline was never scored on this set and the figure says so rather than drawing an empty panel. **The danger render is drawn as faithfully as the favourable one**: a baseline hugging the diagonal is the registered stop outcome, and that is what one facet shows. The A_{lin} facet's caption labels its published correlation as the cross-model, within-context figure it is, and states that the diagnostic was built to predict within-context success rather than to locate a cross-context displacement (§C.3).

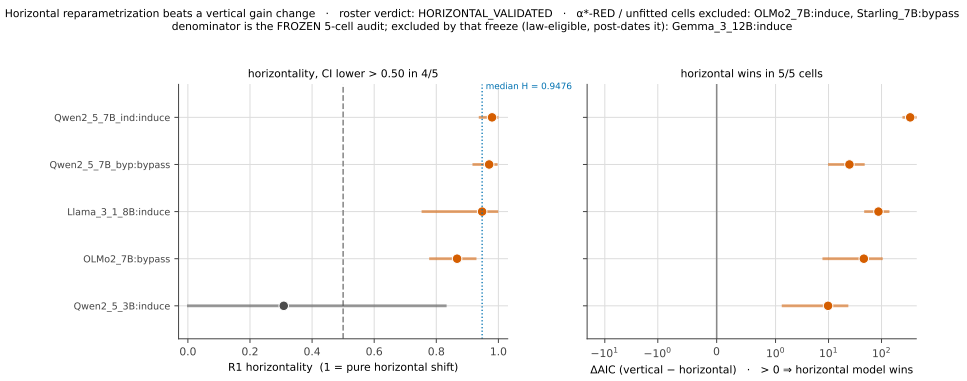

Figure 5: **Horizontal against vertical, at equal model complexity.** *(left)* per cell of the **frozen five-cell audit**, the horizontality statistic with its bootstrap CI against the registered threshold of 0.50, with the median annotated. *(right)* the same five cells, ΔAIC of the vertical model minus the horizontal one with its CI, null line always in frame; positive means the horizontal model wins. **The denominator is five, not the roster's six α^*-GREEN cells**: `Gemma_3_12B:induce` is law-eligible and α^*-GREEN but landed after this audit was frozen, so it is named on the figure as excluded-by-freeze rather than left to be inferred from a short list. Law-excluded and unfitted cells are not drawn (the verdict is defined over GREEN cells) but their exclusion is likewise named with the cell keys. A CI crossing either reference line would be drawn crossing it; nothing about the layout presumes the pass.

steering-success feature carries *anti-information* about a cross-context horizontal shift. It *does not transfer* to dEC_{50}; it does not mean the published within-context result is wrong.

F.1 The two figures that are withdrawn rather than nulled

The frozen figure plan registered two figures for the causal half of the paper (the measured movement of dEC_{50} under a scaffold edit, and the forecast against that movement) each with a pre-written null render. **Neither may be drawn, and neither may be drawn on its null branch either.** A null render communicates a measured null; these experiments produced no measurement at all, so drawing "all CIs cross zero" would assert a measurement that does not exist. They are replaced by two figures that draw what was actually measured: why the experiment was not launched.

G Registered, and not run

Three registrations in §E carry no result. All three are reported here rather than deleted, because a registration that vanishes when its experiment does not happen is exactly what the register-then-run discipline exists to prevent. **They are not the same kind of no-result**, and the difference is stated rather than levelled: two were never launched, while the KV cache-reuse arm was run in full and *halted at its own registered gate*, which is a measurement about the arm and not an absence of one.

G.1 The causal scaffold intervention

The intervention that would establish dEC_{50} as causally *movable* (editing the ReAct format scaffold with task information held fixed) was designed, pre-registered, priced at about $205 of agent-context generation, and **not launched**. Its free forward-pass gate returned zero rankable cells in 6 of 6 (family $\times$ readout-horizon) cells, so the pre-registered planner refused to launch. Budget and statistical power were both available; the missing thing was a readout that could rank the edits.

The causal half alone was scored separately, and it is also unattainable. The registered headline splits into a causal claim and a forecast claim, and one registered branch for the realized refutation outcome was to run *the causal half only*, which uses no forecast. Its marginals are an in-calibration placebo, at least one substantive keep, and a powered gate, on at least two families.

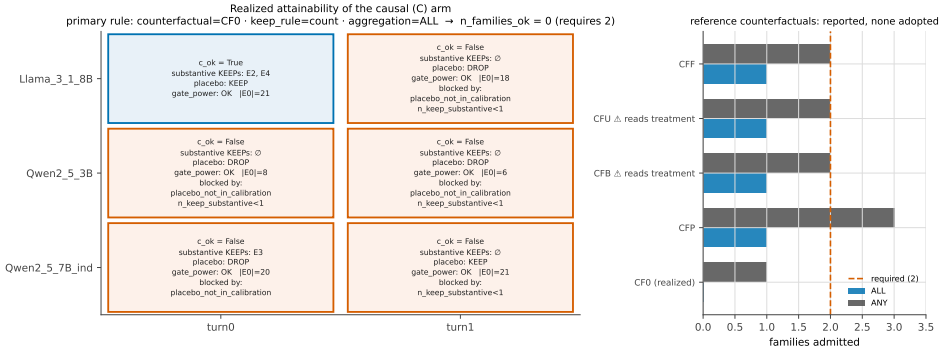

Figure 6: **The attainability matrix: why the causal arm was not scorable.** The six realized (family, readout-horizon) cells, each reporting the recoverability gate's substantive keeps, whether the content-free placebo edit was itself kept (the in-calibration condition) the gate's power flag, and the resulting cell verdict, beside four reference counterfactuals. The derivation is **post-hoc**: it was written after all six cells and the scenario matrix existed, and its own record says so. Read it as an exhaustive sensitivity analysis over a frozen roster, not as a pre-data prediction. Encoding is categorical: nothing here is signed. The power flag is printed in every cell so a reader can see the block is not a power problem. Counterfactuals that read treatment information are marked as such and drawn visually subordinate; the realized, frozen decision is the only one drawn as primary, and no counterfactual column is presented as an alternative result.

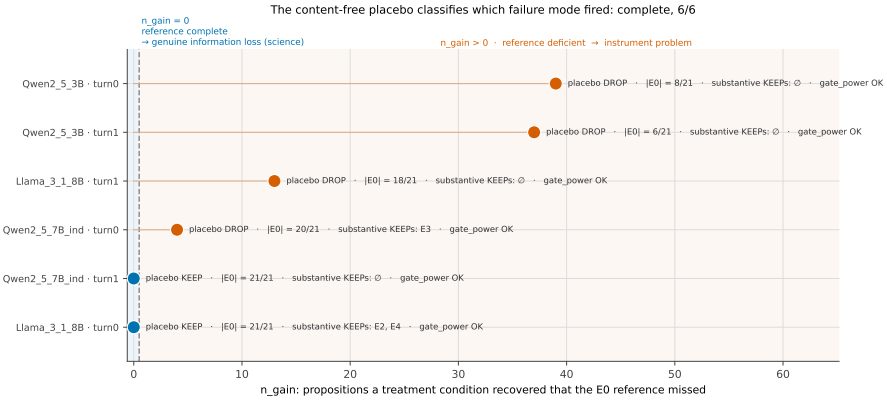

Figure 7: **The placebo separates the gate's two failure modes on all six realized cells.** The same six cells on two axes: the number of propositions the treatment conditions recovered that the reference read missed, against the placebo's own gate decision. The separation is complete in 6 of 6 cells, which names which failure mode fired in each: *reference deficiency*, an instrument problem, against *genuine information loss*, a science result. The registered null render was an incomplete separation, drawn in the off-diagonal quadrant with a caption saying the placebo is not a complete diagnostic; the layout does not assume the result. The claim is a *diagnostic* for which failure mode fired, never that the design was validated.

Under the registered reference and rule, no aggregation reaches two families. Four reference repairs were scored, every one reported and none adopted: the repair is **necessary** (without it no aggregation reaches two families) and **not sufficient** at the conservative rule, where every repair and both keep rules return exactly one family. A reading of the gate restricted to the first readout horizon would make the causal half attainable today on two families, unanimously across all repair and rule combinations; **that reading is refused by name** in the registration's own analysis, because choosing the horizon after seeing which horizon admits is the post-hoc selection the registration was written to prevent.

What the refusal bought, and it is not nothing. Two things, and both are reported. A content-free placebo edit classifies the blocking gate's two failure modes in 6 of 6 cells, separating "the reference set is deficient" from "real information was lost" (Fig. 7), which is what makes this a diagnosis rather than a null. And an attainability calculus priced a headline experiment out *before* rather than after paying for it. The two near-miss cells fail for *opposite* reasons, and neither is a power problem.

Three threats retained verbatim and un-discharged so that a future launch inherits them rather than re-deriving them: the matched-information invariant is the crux and can only be audited, not proven; the mechanism is a hypothesis and the experiment tests the prediction rather than the mechanism; and holding the chat-context midpoint fixed by construction is a single point of failure, because an edit that perturbs the chat leg at all moves the difference for a reason the design cannot see. A fourth constraint is not a retained threat but a measured property of the edit space, and a future launch inherits it too: ReAct scaffold edits are *parser-bound*, since removing a structural marker changes the judged metric, so the edit space is smaller than the design assumed.

The scoping statement, and its limit. The readout invalidity is a property of the 2024-era models on this roster; the current-cohort cell behaves differently. That is what makes this a scoping statement rather than a dead end. It is also, precisely, a claim about a cohort of which we have one member. The one format-scaffold manipulation this paper *can* show is weaker than the registered one: three doses, a different behaviour, and point estimates that were not blind. It is reported as a demonstration that a scaffold edit moves the response at all, never as the registered intervention and never as evidence for the forecast direction.

G.2 THE KV CACHE-REUSE ARM

§D.6 states the scope limit that our design excludes cache reuse by construction and has not measured what cache reuse contributes. A registered arm converts that limit into a measurement: the same agent leg on one cell under cache *reuse*, against a recompute leg, with all outcome branches (including the null) registered before any number exists. The registration proposed pairing the reuse leg against the *committed* recompute leg; what actually ran generated *both* legs fresh under an identical loop, and the paragraph below reports the realized design rather than the proposed one.

Because "cache reuse" is ambiguous between several implementable policies, and an ambiguous term silently resolved after the data is exactly the failure register-then-run exists to prevent, a second registration resolves the term to a named policy set **before** any machine is rented, and that resolution is itself committed as a registration. **The arm then ran** (both legs newly generated under the identical loop, $2 \times 7{,}400$ rollouts on one cell) and **halted at clause 1 of its registered Q_0 gate**: the two legs' item-id sets are 199 and 200 and the gate requires array equality, so the primary questions ship as not-scored rather than as a cache-irrelevance finding the registration forbids reporting from a failed gate. The manuscript's sentence therefore remains *"absent by construction"* and nothing stronger: not because the arm was skipped, but because it was run and its gate was honoured. Two registered branches on this study had no emitter when the study landed; both estimators were then written, both are refits of records that already exist, and both shipped the branch their pre-registration had named in advance (Appendix B.1).

G.3 BRAUN'S GEOMETRIC COVARIATES

§C.4 states the position: both covariates, in both deployment frames, on the pinned roster, scored against dEC_{50} under the same frozen leave-one-model-out rule: registered, and not measured. The

limitations section states the gap rather than the registration implying a result. The roster supports $n = 4$ models and the exact two-sided permutation floor at $n = 4$ is 0.0833, so the conditioning is roster-limited by construction whenever it is run.

H THE LIMITATIONS REGISTER, IN FULL

H.1 WHY THIS SECTION, OF ALL SECTIONS, NEEDS A MACHINE TO SCORE IT

The threat list opens with the reason, written before any paid number existed: *a limitations section written after the results is where a paper quietly drops whichever weaknesses its data happened to expose. A reviewer cannot tell an honest limitation from a self-serving omission after the fact; the only defence is to have enumerated the threats before knowing which ones the data would let us dodge.*

That defence was half-built. The enumeration happened (21 headings committed before the paid stage) but nothing checked the enumeration against the paper. The corpus has since grown from 21 headings to **66**, and the pre-register manuscript named **five** of them. A reader had no way to tell whether the other 61 were absent because they were subsumed, because they were moot, or because they were inconvenient.

So the register in §H.12 has the corpus as its denominator and this section as its numerator. Every heading in the frozen threat list gets exactly one row and exactly one disposition. A row may not pass by being written down: the section it points at must exist and must carry that threat's marker in its own body, so a pointer that goes nowhere is a defect rather than a skip. A threat present in the frozen cut and dropped from the paper is not an ordinary defect but a named failure: the exact failure the preamble above warned about.

Two honest limits of the instrument itself. Its denominator is our own threat list, so a threat nobody in this project ever wrote down is invisible to it: it measures completeness *relative to what we already knew*, which is a real but bounded claim, and it is not a substitute for adversarial review. And a threat can be discharged badly: the register checks that a threat is *stated* where the row says it is; it does not and cannot check that the statement is adequate.

One disposition deserves its own warning. *Did not materialize* is the disposition most open to abuse, because "the data let us dodge it" is exactly what a self-serving omission sounds like when written down honestly. It is therefore the only disposition that requires an artifact pointer in its note, it is confined to §H.9, and that section is written to say what the non-materialization *cost* rather than to bank it. There is exactly one such row, and it records a shortfall.

H.2 CONSTRUCT VALIDITY: WHAT dEC_{50} READS, AND WHAT IT DOES NOT

The location is not a pure readout of the reparametrization: it reads the reparametrization together with the baseline offset, which is an interpretive bound and the reason the paper's guardrails forbid claiming the contexts differ *only* horizontally. The 4PL can be misspecified, and one cell admitted a descending fit. Coherence gating changes the denominator the estimator averages over, which is paired with the coherence-as-response finding rather than treated as a nuisance. Equal upper plateaus mean the vertical half of the reparametrization is *unconstrained*, not shown absent. Coherence is fitted as a response throughout, and the reported flag is blind to its own collapse. And the deliverable maximum is right-censored in four of ten context-cells, so the efficacy contrast names its scale rather than being quoted scale-free.

H.3 IDENTIFICATION OF THE FITTED LOCATION: THE GRID, THE INTERVAL, AND THE CEILING

Two limitations of the positive anchor, stated here rather than only in the register. **(a)** The positive arm's displacement is read on a sub-grid: its *sign* survives without a curve model, but its magnitude is not resolved by these data, and no magnitude for it should be quoted as settled. **(b)** The interval on that anchor is *anticonservative*, which is why a repair was registered separately rather than folded into the number. Both intervals are printed in §D.4: the committed one, which

the registration dates and Table 2 carries, and the repaired one beside it, so a reader sees both rather than a silently improved figure.

The α^* ceiling is an estimate with its own uncertainty, and two registered families carry a null value for it. The primary fit runs the full swept grid rather than the stricter registered dose exclusion, and both readings are printed. The finer grid that resolves the positive arm fails the span half of the identifiability rule, so the two grids are not pooled. And the visual stress test in the body is a law-excluded cell: its second summary is not a corrected estimate, and its screen is post-hoc.

H.4 The control spine as realized: four errata, and what would falsify each

Appendix A states the four errata and §D.2 states where each bites. The register carries, in addition: scaffold integrity collapses with dose and α^* does not screen it, so the top of the grid is the least trustworthy part of it; the judged object changes kind with dose; the judge artifact and the estimand share a signature, which is measured and separated rather than assumed away; and the one family with a valid gate readout is the one whose chat template folds the system role structurally. Each erratum is stated with what would falsify it.

H.5 Nuisance parameters, and the surfaces the readouts are taken on

The readout surface, and where it reverses. Every residual readout we *report a result from* is taken on the prompt, and the one study that read the alternative surface is reported here rather than folded into those results. In the one cell where the comparison is movable, the frame difference *reverses sign* on the model's own generated tokens. We therefore make no claim about the generated-token readout and scope the prompt-side result to the prompt.

The same commanded coefficient is injected at more sites in the agent context; matching the injection phase set across the two contexts leaves the displacement intact and sign-identical, while the agent-only sites do add potency, so a minority of it is dose bookkeeping and the majority is not. The carrier cannot be separated from the delimiters, so no residual-stream mechanism is claimed. The localizing ladder runs on the leg that is *not* information-matched and on the movable metric, and the finest frame ladder is single-dose with an inherited ranking and a random band undefined in 17 of 20 cells.

H.6 What the deflationary rejections license, and what they do not

The norm null is rejected against one normalization model, on a turn-0 surface, under an assumption that cannot be verified from this data. dEC_{50} is not a translation along the steering direction, and the neighbouring predictor does not transfer to it. The delivered-unit correction changes no sign, and the transport model that would have explained the shift fails. Braun's covariates were claimed before being computed, and the erratum is recorded rather than silently corrected. Every deflationary count is scored on a five-cell panel frozen before the roster's sixth GREEN cell existed; the relative-dose commensurability verdict is a five-cell result whose spread the sixth cell takes from 2.68 to 21.74, flipping it. A post-hoc plateau-gap diagnostic correlated with the *absolute* dose axis; scale-free, it never held: on seven cells or on eight. And two consumers of a shared loader are invisible to the currency check because its artifact set is keyed to a declaration; both are registered deferrals.

H.7 The forecast: refuted out of sample, and the ways it could have been right for the wrong reason

Parameter-free means no *per-model* parameter, and the global choices are named rather than hidden. The forecast shares its assumptions with the theory it is derived from, and the realized slope refutes its own derivation. The arms did not fail benignly: the registered verdict is that a published baseline recovers, and the forecast is never called validated. The readout is a point on a trajectory, the registered gauge board reads gauge-dependent, and no entrant is promoted post hoc.

H.8 THE CAUSAL CLAIM: THREATS RETAINED UN-DISCHARGED, BECAUSE THE EXPERIMENT WAS NOT RUN

§G.1 carries this group in full: three threats retained verbatim and un-discharged, three that materialized as measurements and are reported as such, and one weaker manipulation that is reported as a demonstration and never as the registered intervention.

H.9 THE ONE THREAT THAT DID NOT MATERIALIZE, AND WHAT ITS NON-MATERIALIZATION COST

Left-censored salvage bounds are directional rather than two-sided: where a fitted midpoint is only *bounded*, dEC_{50} is a one-sided directional bound and weaker evidence than a two-sided CI, and the frozen posture was that the salvage claim would never rest on a bound alone. **It did not arise**: no cell in this paper is reported as a censored bound, so no headline rests on one.

And that is not good news. The reason no bound was reported is that the salvage arm produced no salvage cell at all. The salvage question (whether a horizontal displacement is estimable in families where a vertical gain ratio dies against a ceiling) was called, before the data, the cheapest thing that could have made this paper, and **the shipped fit does not answer it**. The paid stage passes on two-sidedness alone. The threat is moot because the arm it threatened is empty, which is a strictly worse outcome than the threat materializing, and it is recorded as such rather than counted as a clean sheet.

H.10 STATISTICAL THREATS, EXTERNAL VALIDITY, AND THEORY-DEPENDENCE

Multiplicity across families, arms and edits: exploratory verdicts are labelled as such and the law rests on registered primaries. Greedy decoding gives item-resampling intervals with no decoding variance, which is a scope statement rather than a defect. Per-sample variance is high by construction under the cylindrical hypothesis, so the contrast is powered against variance rather than cell count (§D.4). The boundary of the claim across steering type, behaviour, roster, scaffold, decoding and language is stated as a scope condition. And theory buys the motivation and the α^* scope condition: the difference of fitted midpoints survives the theory failing.

H.11 OPERATIONAL AND REPRODUCIBILITY THREATS, EACH ONE MEASURED

The guard suite had never been run as a suite, and a later audit found seven guards invariant to their inputs. Most committed artifacts do not regenerate from the tree a reviewer receives, and every document that quotes one names the tree it was measured in. The registered reading rules had no working call site until a prereg-to-verdict ledger enforced them. One registered rule had no recorded verdict, and it was the α^* control's own live/die criterion. Cells are pinned to checkpoints and one of the eight records what it loaded, so run-time provenance is absent for the other seven. Each of these is printed at the size the finding had, unflattering.

H.12 THE THREAT REGISTER

Every heading in the frozen threat list, with its disposition and the section that carries it. *container* rows are group headings whose children carry the content; *scope* rows are boundary statements rather than defects; *moot (unrun)* rows are the causal-claim threats retained verbatim against a future launch.

Table 5: The threat register: every heading in the frozen threat list, its disposition, and the section that carries it.

threat	disposition	where	note
1	CONTAINER	§H.2	container heading for the construct-validity group; its four children carry the threat content
1.1	APPENDIX	§H.2	the location reads the reparametrization together with the baseline offset; interpretive bound and guardrail G5
1.2	APPENDIX	§H.2	4PL misspecification, its frozen mitigations, and the one cell where a descending fit was admitted

Table 5: The threat register *(continued)*.

threat	disposition	where	note
1.3	APPENDIX	§H.2	coherence gating changes the denominator; paired with the coherence-as-response finding
1.4	DID_NOT_MAT.	§H.9	no censored bound is reported in a committed artifact (bound_direction null 8 of 8) because Q2_salvage_pass is false and Q2_hits is empty; recorded as a shortfall
2	CONTAINER	§H.8	container heading for the causal-claim group; framing only, children carry the threats
2.1	MOOT_UNRUN	§H.8	matched information can only be audited; retained verbatim per its own registration so a future launch inherits it
2.2	MOOT_UNRUN	§H.8	the experiment tests the prediction and not the mechanism; retained against its own registration
2.3	MOOT_UNRUN	§H.8	EC50_chat held fixed is a single point of failure; retained against its own registration
2.4	APPENDIX	§H.8	control and instrument want different readout points and the overlap is empty; this is the measurement that priced the launch out
2.5	APPENDIX	§H.8	the two failures are marginal rather than joint and no horizon repairs them; cost was not the binding constraint
2.6	APPENDIX	§H.8	the causal half alone fails on a different marginal, and the placebo classifies the gate 6 of 6
3	CONTAINER	§H.7	container heading for the predictor group; its four children carry the threat content
3.1	APPENDIX	§H.7	parameter-free means no per-model parameter and the global choices are named
3.2	APPENDIX	§H.7	the predictor shares its assumptions with the theory, and the realized slope refutes its own derivation
3.3	APPENDIX	§H.7	the arms did not fail benignly; the registered verdict is STOP_BASELINE_RECOVERS and the predictor is never called validated
3.4	APPENDIX	§H.7	the readout is a point on a trajectory, the gauge board reads GAUGE_DEPENDENT, and no entrant is promoted post hoc
4	APPENDIX	§H.3	the alpha-star ceiling is an estimate with its own uncertainty, and two registered families carry a null value
4.1	APPENDIX	§H.3	the primary fit runs the full swept grid rather than the stricter registered dose exclusion; both readings printed
4.2	APPENDIX	§H.4	Control A enters as a diagnostic and as the admission gate, never as a correction of the fitted response
4.3	APPENDIX	§H.4	Control B matches the task and not one span of the system prompt; a bounded mismatch we name
4.4	APPENDIX	§H.4	scaffold integrity collapses with dose and alpha-star does not screen it; the grid top is least trustworthy
4.5	APPENDIX	§H.4	Control M is length-correlated rather than length-controlling, and the judged object changes kind with dose
4.6	APPENDIX	§H.4	Control G ran on one clause of two; the beats-random clause was read by nothing
4.7	APPENDIX	§H.4	Control N holds on the rollout record and not the fitted set, and the pairing is inert for a location estimand
4.8	APPENDIX	§H.11	the guard suite had never been run as a suite, and a later audit found seven guards invariant to their inputs
4.9	APPENDIX	§H.11	most committed artifacts do not regenerate from git archive HEAD; every quoting document names the tree
4.10	APPENDIX	§H.11	the registered reading rules had no working call site; now enforced by the prereg-to-verdict ledger
4.11	APPENDIX	§H.11	one registered rule had no recorded verdict and it was the alpha-star control's own live-die
4.12	APPENDIX	§H.4	the judge artifact and the estimand share a signature; measured and separated rather than assumed away
4.13	APPENDIX	§H.3	the positive arm is a sub-grid displacement whose sign survives without a curve model and whose magnitude is unresolved
4.14	APPENDIX	§H.3	the interval is anticonservative on the positive anchor and armed a separately registered repair rather than a number change

Table 5: The threat register *(continued)*.

threat	disposition	where	note
4.15	APPENDIX	§H.3	the finer grid that resolves the positive arm fails the span half of the identifiability rule; the two are not pooled
4.16	APPENDIX	§H.11	cells are pinned to checkpoints and 1 of 8 records what it loaded; run-time provenance is absent for seven of the eight fitted cells
4.17	BODY	§8	a dose is not a shared unit, so cross-family magnitudes carry a normalization and the log-ratio is the poolable companion
4.18	BODY	§8	the curves are not translates; PI_Q2_TRANSLATION_REJECTED, and dEC50 is a location summary at one response level
4.19	APPENDIX	§H.2	equal plateaus mean the vertical half of the reparametrization is unconstrained rather than shown absent
4.20	APPENDIX	§H.2	coherence is a response not a filter, and the reported flag is blind to its own collapse
4.21	APPENDIX	§H.6	the gain is dose-indexed and two of our own registered predictions about it were refuted in Results
4.22	APPENDIX	§H.2	the deliverable maximum is right-censored in four of ten context-cells and the efficacy contrast names its scale
4.23	APPENDIX	§H.6	the norm null is rejected against one normalization model on a turn-0 surface with an unverifiable assumption
4.24	APPENDIX	§H.4	the one family with a valid gate readout is the one whose chat template folds the system role structurally
4.25	BODY	§8	the injection layer is one per family and was chosen on the chat leg for the Qwen anchor; still unswept
4.26	BODY	§8	the direction is chat-extracted and unvaried in the reported fits; swept only on already-public reciprocal data
4.27	APPENDIX	§H.5	the same commanded alpha is injected at more sites in the agent context; matching the phase set leaves the displacement intact and a minority of it is dose bookkeeping
4.28	APPENDIX	§H.6	dEC50 is not a translation along the steering direction, and the neighbouring predictor does not transfer to it
4.29	APPENDIX	§H.6	Braun's covariates were claimed before being computed; the erratum is recorded rather than silently corrected
4.30	APPENDIX	§H.8	the showable scaffold manipulation is three doses in another behaviour with non-blind point estimates
4.31	APPENDIX	§H.5	the carrier cannot be separated from the delimiters, so no residual-stream mechanism is claimed
4.32	APPENDIX	§H.6	the delivered-unit correction changes no sign and the transport model that would have explained the shift fails
4.33	APPENDIX	§H.5	every residual readout is on the prompt and the frame difference reverses on generated tokens in the one movable cell
4.34	APPENDIX	§H.5	the localizing ladder runs on the leg that is not information-matched and on the movable metric
4.35	APPENDIX	§H.5	the finest frame ladder is single-dose with an inherited ranking and a random band undefined in 17 of 20 cells
4.36	APPENDIX	§H.3	the visual stress test is a law-excluded cell; its second summary is not a corrected estimate and the screen is post-hoc
4.37	APPENDIX	§H.6	every deflationary-accounts count is scored on a five-cell panel frozen before the roster's sixth α^*-GREEN cell existed
4.38	APPENDIX	§H.6	relative-dose commensurability (NR-Q6) is a five-cell result; the sixth α^*-GREEN cell takes the spread from 2.68 to 21.74 and flips the verdict
4.39	APPENDIX	§H.6	the post-hoc plateau-gap diagnostic correlated with the absolute dose axis; scale-free it never held, on seven cells or eight
4.40	APPENDIX	§H.6	roster_currency's artifact set is keyed to a declaration, so two consumers of the shared loader are invisible to it; both are registered deferrals
4.41	APPENDIX	§H.6	a deliberately pinned historical record scored as stale; the exemption that fixes it is measured on four clauses, never listed
5	CONTAINER	§H.10	container heading for the statistical group; its three children carry the threat content
5.1	APPENDIX	§H.10	multiplicity across families arms and edits; exploratory verdicts are labelled and the law rests on registered primaries
5.2	SCOPE	§H.10	greedy decoding gives item-resampling intervals with no decoding variance; a scope statement rather than a defect

Table 5: The threat register *(continued)*.

threat	disposition	where	note
5.3	APPENDIX	§H.10	per-sample variance is high by construction under CRH, so the contrast is powered against variance not cell count
6	SCOPE	§H.10	the boundary of the claim across steering type behaviour roster scaffold decoding and language
7	SCOPE	§H.10	theory buys the motivation and the alpha-star scope condition; the difference of fitted midpoints survives the theory failing
8	APPENDIX	§H.11	operational disciplines were measured rather than asserted, and the measurements are printed unflattering

What a reader should take from this register. Not that the paper is unusually fragile: it is that the enumeration was frozen before the data and then scored against the manuscript by a machine, so the list a reviewer sees is the list that existed before we knew which threats the data would let us dodge. The instrument's denominator is our own threat list, which is the bound stated in §H.1 and is not a substitute for adversarial review.