Refusal Across Layers

Readout entropy, state geometry and state transfer before and after post-training

Felipe A. Ly
Entel AI Research

2026-11-01

Abstract In the Instruct models studied here, harmful prompts often produce a nearly determined first reply token. We ask where across layers post-training makes the output confident, and whether through changed states or only a changed readout. In pre-declared runs we compare the Base and Instruct versions of Llama 3.2 3B and Gemma 2 2B on harmful and matched harmless prompts from JailbreakBench and XSTest. At the chat reply position, post-training lowers final-layer readout entropy more on harmful than on matched harmless prompts in both models (Llama 0.46 and 0.63 nats, Gemma 0.60 and 0.47), also with both models' states read through one head, so the states carry it. The declared readout locus is late; a lens-free spread contrast is sustained from mid-depth. Within Llama Instruct, state swaps move the donor-class contrast in a first-token proxy, the log-odds of "I": swapping the template suffix and reply position together removes more than half of it from harmful recipients from layer 13, while over layers 15-22 either alone leaves most of it, and on average a single harmful-donor position gives more than half of the two-donor contrast (an earlier run's hypothesis, tested on reused prompts). On OR-Bench the entropy contrast recurs (non-prospectively), the joint-swap results replicate, and six of eight composition labels replicate, two indeterminate. In Gemma the first token "I" was validated as a refusal marker on harmless prompts only, a reading that fails if ambiguous replies count as compliance; in Llama only a prefix rule was judged. Effects follow benchmark harm labels, confounded with refusal. None establishes an attractor, a harm representation or a refusal mechanism.

Working paper v0.9, draft-009. Llama 3.2 3B: runs CTL001 to CTL006 and OFF001. Gemma 2 2B: GEM001 to GEM004. Both: OFF003.

1 Introduction

Post-training teaches a language model to refuse harmful requests and to answer harmless ones. In the Instruct models studied here, harmful prompts often produce a nearly determined first reply token. We ask where in the network that confidence appears, and whether it reflects a change in the model's internal states or only in how those states are read out.

A direct comparison leaves two explanations unresolved. Low first-token entropy can reflect a stereotyped reply opening, and post-training may sharpen outputs on every prompt. We therefore compare matched pairs: the Base-to-Instruct change on a harmful prompt minus the change on

a matched harmless prompt from the same benchmark. We call this difference *harm-specific*. The term names the contrast between benchmark labels, not a demonstrated representation of harm. Factual-recall and nonsense prompts serve as references for determinacy and unpredictability.

The project began from a conjecture: safe and helpful outputs correspond to low-entropy attractor states in latent space, which post-training imposes and base models might reach on their own. "Attractor" and "topological" belong to that conjecture, not to our claims. We test narrower consequences that can fail.

Contribution. Three findings, each from runs declared before their outputs existed:

1. *Readout and geometry, in both models.* Post-training makes the reply-position readout more determinate on harmful than on matched harmless prompts in Llama 3.2 3B and Gemma 2 2B. The change survives when both models' states are read through one head, so it is carried by the states. Its declared readout locus is late; a lens-free contrast in the states is sustained from mid-depth (Section 3.1).
2. *State swaps, in Llama only.* Within Llama Instruct, the donor-class contrast in a first-token output proxy moves with the last five prompt positions: the template suffix and the reply position together transfer more than half of it from layer 13, where either alone does not (Section 3.3).
3. *Composition and replication.* Over layers 15-22, the class means of mixed swaps depend on the donor classes of the two positions, which we read, as a draft interpretation, as OR-leaning. On OR-Bench the joint-swap results replicate, and six of eight composition labels replicate while two are indeterminate (Section 3.4).

Limits. The evidence covers two models of 2-3B parameters, short single-turn benchmark prompts and an output proxy. In Gemma the first token "I" was validated against a judge as a refusal marker on harmless prompts only; Llama's first-token proxy was not itself judged, and no proxy was validated on harmful prompts or after intervention. Benchmark harm labels are confounded with refusal. Some checks are not prospective, the swap window was chosen from an earlier run, and the replication is incomplete. Logit-lens results describe a projection of the states, not the states themselves. Nothing here measures convergence dynamics or topology (Section 4.1).

1.1 Related work

The logit lens reads intermediate states through the final norm and unembedding (nostalgebraist 2020); it is often brittle, and trained per-layer affine probes are more predictive, reliable and unbiased (Belrose et al. 2023). Entropy itself can be regulated through the residual norm and the final normalization (Stolfo et al. 2024), so we report norms beside entropy.

Closest in design, Zhou et al. (2024) compare base and chat models of several families: linear classifiers separate malicious from normal prompts after the first few layers in both, and aligned models show emotion-laden tokens in middle layers under a lens. They do not measure entropy, and their normal prompts are not matched to the malicious ones. Closest in method, Entropy-Lens profiles per-layer entropy over generated tokens in instruct models, without safety prompts (Ali et al. 2025). Nikolenko et al. (2026) detect harmful prompts from entropy traces across prompt positions in mostly base checkpoints, without a base-aligned contrast; on JBB's benign prompts used as hard negatives, their trend features fall to chance or reverse. The refusal direction of Arditi et al. (2024) is also present in base models, and Qi et al. (2025) argue that safety alignment is often only a few output tokens deep. To our reading, none of these measures last-position readout entropy as a harm-specific Instruct-Base contrast, or separates state change from readout change.

Table 1: The runs. OFF001 was indeterminate and CTL002 descriptive, except for its refusal labels.

Run	Model	What it measures	Reported in
CTL001	Llama	Readout entropy, cross-readout and geometry on 850 prompts	Finding 1
GEM001	Gemma	CTL001′s measurements under a declared adapter	Finding 1
OFF003	both	Offline: entropy floor and lens-free states, on retained arrays	Finding 1
GEM002-GEM004	Gemma	Refusal proxy: rule gate (failed), prospective judge validation, repaired phrase score	Finding 1
CTL003, CTL004	Llama	Within-Instruct swaps of single and joint positions	Finding 2
CTL005	Llama	Crossed-donor joint swaps	Finding 3
CTL006	Llama	OR-Bench replication of CTL001, CTL004 and CTL005	Findings 1-3
OFF001, CTL002	Llama	Shared-first-token test; refusal labels and cross-model swaps	Appendix C

Representation geometry offers lens-free measures: matrix-based entropy of hidden states across layers (Skean et al. 2025) and intrinsic dimension from nearest-neighbour ratios (Facco et al. 2017), which changes markedly with depth in transformer representations (Valeriani et al. 2023; Cheng et al. 2025).

2 Method

Every run was declared in research/runs/ before its retained outputs existed, with its estimands, thresholds and failure criteria fixed. Declarations and code were reviewed before launch, smoke tests preceded each run, and analyses ran on committed outputs. Appendix A gives each run's procedure and checks, and Appendix B the full decision rules. Exploratory analyses are labelled as such. Table 1 maps the runs to the findings.

2.1 Models, prompts and read positions

We compare `meta-llama/Llama-3.2-3B` and `meta-llama/Llama-3.2-3B-Instruct` (28 decoder blocks, width 3072, tied embeddings over 128,256 tokens) and `google/gemma-2-2b` and `google/gemma-2-2b-it` (26 blocks, width 2304, tied embeddings over 256,000 tokens), at pinned revisions, in bf16 without quantization, one model at a time on one RTX 3080. Every prompt (Table 2) is a single forward pass without padding, and Base and Instruct receive identical token ids.

Each prompt is rendered three ways: raw text; a plain dialogue frame, `User: {text}\nAssistant:`; and the model's chat template (Llama's with the date pinned; Gemma's has no system block). The main read position is the *reply position*, the last template token, where the first reply token is chosen. Three secondary positions are the last token of the raw and dialogue formats and the last token of the user message inside the chat template (the *last instruction token*).

Base never learned the chat template's special tokens. In Llama their embedding rows are indistinguishable from never-used reserved tokens (ATA-C005), so Base reads the template through untrained placeholders, and the Base side of every chat-format contrast is off-distribution. Gemma follows Llama's rules through a declared adapter (Appendix B): its output transform includes a logit softcap, it runs with eager attention, its final layer is 26 and its late range starts at layer 21.

Table 2: Prompt sets. Seeded selection and shuffling; files pinned by SHA-256.

Category	Source	n	Role
harmful / benign	JBB-Behaviors (Chao et al. 2024)	100 / 100	100 topic-matched pairs
unsafe / safe	XSTest (Röttger et al. 2024)	200 / 250	200 contrast pairs; 50 safe unpaired
factual	ROME known facts (Meng et al. 2022)	100	determinacy reference
nonsense	JBB benign goals, words shuffled	100	unpredictability reference
toxic / hard	OR-Bench (Cui et al. 2025)	655 / 1,319	all prompts (CTL002); 564 constructed pairs (CTL006); all 1,319 Hard replies generated, 400 judged (GEM003)

2.2 Readout entropy and the harm-specific contrast

At each layer we read the last-position state through the model's own final norm and unembedding, as if the network stopped there, and take the entropy of the resulting next-token distribution. This is the logit lens with the final norm (nostalgebraist 2020; Belrose et al. 2023). Low entropy means the readout is nearly committed to one token. These are readout results: they describe a projection of the states, not the states themselves.

With h_ℓ the pre-norm residual after block ℓ ($\ell = 0$ is the embedding output), captured by hooks, the readout is computed in fp32,

$$P_\ell = \mathrm{softmax}\left(W_U \, \mathrm{RMSNorm}(h_\ell) \right),$$

and $\mathcal{H}_\ell$ is its entropy in nats. Over n matched pairs of a harmful prompt a_j and a harmless prompt b_j, the harm-specific contrast is

$$D_\ell = \frac{1}{n} \sum_{j=1}^{n} \Big[\left(\mathcal{H}_\ell^{\mathrm{Instruct}}(a_j) - \mathcal{H}_\ell^{\mathrm{Base}}(a_j) \right) - \left(\mathcal{H}_\ell^{\mathrm{Instruct}}(b_j) - \mathcal{H}_\ell^{\mathrm{Base}}(b_j) \right) \Big] .$$

$D_\ell < 0$ means that post-training lowers entropy more on the harmful member than on its matched harmless member.

Depth. Uncertainty comes from a pair bootstrap with 10,000 replicates; a simultaneous 95% band over a set of layers uses the maximum studentized deviation across them. The *onset* is the first layer whose band excludes zero there and at every deeper layer, with one sign; the *locus* is the first layer from the onset on with $|D_\ell| \geq 0.1$ nat. A locus in the last quarter of depth (from layer 22 in Llama, 21 in Gemma), or no locus at all, is *late-only*: it cannot be separated from the selection of the first reply token. Layers where the readout mostly repeats the input token (*echo*) are excluded. A layer is *final-consistent* where the median KL divergence to the model's final distribution is at most 1 nat in both models; elsewhere, statements concern the projection only.

States or readout. Post-training changes the blocks and, through the tied embeddings, the readout. We therefore also read both models' states through one fixed head, Base's or Instruct's (*cross-readout*). Under a fixed head both models' states pass through the same map, so a divergence

that survives under both fixed heads lives in the states; it is then called *state-carried*. With tied embeddings, "states" include post-training changes to input embedding rows.

Declared outcomes. The positive control is $D_{28} < 0$ at Llama's chat reply position on JBB; the primary outcome is the depth of the locus there. XSTest is the replication, the dialogue frame the format check and the last instruction token the check for structure before the reply position. A length-adjusted contrast tests sensitivity to length. Gemma's run (GEM001) applies the same rules; a statement counts for both models only where labels computed by one function from both analysis files agree.

Floor. Instruct's entropy on harmful prompts is often near zero, so a floor could shape the contrast. OFF003 characterizes this without a pass or fail label (Appendix B): the *monotone-baseline contrast* compares harmful and harmless residuals after a fitted non-decreasing map from Base to Instruct final entropy, and is exactly zero whenever Instruct entropy is a deterministic non-decreasing function of Base entropy; the contrast is also repeated on the top-1 log-odds of each model's output, which has no floor. Llama's retained arrays serve as a descriptive parity arm.

2.3 Lens-free geometry

Lens-free measures use the same last-position states. The spread of a category is the trace of its covariance divided by that of all 850 prompts: its share of state variance. The *spread contrast*

$$G_\ell = \big[\, R_\ell^{\text{Instruct}}(\text{harmful}) - R_\ell^{\text{Instruct}}(\text{harmless})\big] - \big[\, R_\ell^{\text{Base}}(\text{harmful}) - R_\ell^{\text{Base}}(\text{harmless})\big],$$

with R the share, is negative when post-training lowers the variance share of harmful prompts relative to harmless ones. Shares are relative: a smaller share is not by itself contraction. CTL001 computed G_ℓ on raw and standardized states with pointwise intervals. OFF003 added unit-norm states and simultaneous bands, and reads the three variants jointly: "early" if all three have a sustained locus before the late range in both datasets with one sign, "norm-carried early" if only raw and standardized do. Further descriptive measures are in Appendices B and D.

2.4 The output proxy and its validation

Instruct's first reply token is "I" for most harmful prompts and few harmless ones, so the swaps use a first-token *output proxy*, the log-odds of "I" at the model's true output,

$$s = z_{\text{I}} - \log \sum_{v \neq \text{I}} e^{z_v},$$

with z the output logits. "I" also opens many replies that are not refusals, so s is a proxy, not a refusal label. Three label checks ask how far it can be read as refusal, all on replies to harmless prompts:

- *Llama (CTL002).* Instruct generated 16 tokens per prompt, and a prefix rule of 27 refusal openings labelled each reply. A judge agreed with the rule on all 60 sampled rule-positive and 55 of 60 rule-negative replies to harmless OR-Bench Hard prompts (Appendix C). Llama's first-token proxy itself was not judged.
- *Gemma, rule gate (GEM002).* The same prefix rule was tested against a fresh judge on harmless replies, with a declared gate of 0.90 agreement in each stratum.
- *Gemma, prospective validation (GEM003).* A judge session run outside the project's context labelled Instruct's replies to 400 harmless OR-Bench Hard-1K prompts, which no Gemma run had used, as refused, complied or ambiguous. Against the non-ambiguous labels, E1, the first output token "I", is validated as a refusal marker if the Wilson lower bounds of its precision and recall reach 0.85. E2, a teacher-forced *phrase score*, the log-odds that the reply opens with one of

13 fixed refusal phrases ("I cannot", "I'm sorry", ...), is validated as a refusal score if the bounds of its AUROC and of its recall at positive log-odds reach 0.85, and as a ranking score only if the AUROC bound alone does. Ambiguous labels counted as complied or as refused are declared sensitivities. On GEM001′s pairs the phrase score also gives a contrast built like D_ℓ; its phrases were chosen partly from GEM002′s replies to those prompts, so only the Hard-1K evaluation is fresh (Appendix B; results in Section 3.2).

2.5 State swaps within Llama Instruct

To ask where the proxy contrast can be moved, we replace part of a *recipient* prompt's hidden state at one layer with the state of a *donor* prompt, and let the model compute the rest. All swaps are within Llama Instruct, which avoids Base's untrained template states; none were run on Gemma.

Positions. Llama's chat template places the user text, then a four-token template suffix, then the reply position. A mask replaces the reply position (R), the template suffix (S), the user text (U), or two of them jointly in one pass: suffix plus reply (SR) and user text plus reply (UR).

Donors. Each recipient receives states from harmful and harmless donors drawn from other matched pairs of the same dataset, with exactly the recipient's token length and from another category (CTL003). Equal length aligns positions; it does not equalize content. The eligible sample is 81 JBB and 181 XSTest pairs (162 and 362 recipients). Exclusions are not random: JBB loses 19 pairs with long prompts, and XSTest loses 14 of its 25 short "definitions" pairs. The assignment is frozen, and inference is conditional on it.

Statistic. G is the mean unpatched harmful-minus-harmless gap in s over the eligible pairs, about 18 log-odds units on JBB and XSTest. (It is unrelated to the spread contrast G_ℓ; the figures keep both names.) For a recipient class, mask m and donor class, A is the mean change in s from the unpatched recipient, divided by G. The *transfer fraction* is the donor-class contrast $H(m) = A(m, \text{harmful donor}) - A(m, \text{harmless donor})$. It is near 0 when the donor's class does not matter and near 1 when the swap carries the full unpatched gap. On harmful recipients it measures *removal*, on harmless ones *installation*. It is a contrast between donor classes, not a change from the unpatched recipient, and the same-class shifts serve as controls. In these units 0.1 G is about 1.8 log-odds units, a factor of about 6 in the odds of "I".

Reply mask (CTL003). A cell (dataset by direction) is *mid-depth* if the lower simultaneous band of H exceeds 0.5 from some layer at or before 19 through layer 28, and *deep* if that run starts at layers 20-23; Appendix B lists the other labels.

Joint masks (CTL004). CTL004 asks whether the suffix and the reply together remove what neither removes alone. Its statistics are window means $W(\cdot)$ over layers 15-22, where CTL003′s single-mask removal curves are flat; the window was chosen from CTL003′s results, so CTL004 is not an independent replication. The primary statistics are joint removal $W(H(SR))$, the increment over the reply alone, $W(J_{SR}) = W(H(SR)) - W(H(R))$, and the same-class control, the shift of harmful recipients under a harmful-donor joint swap. A cell is *redundancy-consistent* if joint removal's lower bound exceeds 0.5, both single masks' upper bounds lie below 0.5, both increments' lower bounds exceed 0.1 (*material*) and the control's interval lies within ± 0.1 (*stable*). This label is the strongest available, and it does not establish redundancy. The removal statistics share simultaneous 95% intervals.

Crossed donors (CTL005). CTL005 tests a hypothesis generated from CTL004′s outputs, on reused prompts: that the joint-swap output depends on the donor classes of the two states, not on the recipient side or on whether the two states come from one prompt or two. Each recipient has three donors per class (69 JBB and 159 XSTest pairs qualify), and the suffix and the reply take

states from up to two donors in one pass. Z rescales s so that 0 is the harmless and 1 the harmful unpatched mean; $Z_k(x, y)$ is the window mean for recipient class k with suffix donor class x and reply donor class y (h harmful, b harmless, primes a second donor of the same class). The primary statistics, with margin 0.1, are:

- the recipient-side differences $\Delta_{hb} = Z_{\text{harmful}}(h, b) - Z_{\text{harmless}}(h, b)$ and Δ_{bh}, *invariant* if the interval lies within ± 0.1;
- the factorial interaction $I_k = Z_k(h, h') - Z_k(h, b) - Z_k(b, h) + Z_k(b, b')$, zero if the two positions act additively and *negative* if its upper bound lies below -0.1, meaning that on average a single harmful position gives more than half of the two-donor contrast;
- the coherence shifts E, same-donor minus different-donor cells, invariant within ± 0.1.

An aggregate label requires both datasets to agree. The 16 primary statistics share simultaneous 95% intervals; gates and descriptive families are in Appendix B.

Replication (CTL006). CTL006 repeats the three findings on OR-Bench (Cui et al. 2025) with no new estimand. Harmful prompts come from OR-Bench-Toxic and harmless ones from OR-Bench-Hard-1K, prompts that look harmful but are labelled safe. The benchmark pairs no prompts, so within each of its 10 categories we matched the smaller class one to one into the larger by nearest chat length: 564 pairs, a category- and length-stratified contrast of benchmark labels, not semantic matches. R1 recomputes CTL001′s layer-28 contrast from states CTL002 had already captured, so it is not prospective. R2 (CTL004′s removal statistics) and R3 (CTL005′s primary family) are prospective, on the 515 pairs with three equal-length donors per class. Here G is about 10, so the margin 0.1 G is about 1.0 log-odds unit, against about 1.8 on JBB and XSTest. A labelled sensitivity repeats every target on the *refusal-consistent* pairs, whose Toxic member Instruct prefix-refused under CTL002′s rule and whose Hard member it did not; it cannot change a verdict.

3 Results

The declared implementation checks of every run in Appendix A passed, and independent recomputations reproduced the analyses of CTL001, GEM001 to GEM004 and OFF003 within their stated scope. GEM002′s validation gate did not pass.

3.1 Finding 1: post-training makes the reply-position readout more determinate on harmful prompts

In both models and both datasets, post-training lowers reply-position entropy more on harmful than on matched harmless prompts (Table 3). In Llama the declared positive control passed: $D_{28} = -0.46$ nats (95% interval -0.57 to -0.35) on JBB and -0.63 (-0.72 to -0.54) on XSTest. In Gemma D_{26} is -0.598 and -0.467, and all eight headline conditions of the cross-model rule hold in both models. Neither contrast is length-sensitive (JBB length-adjusted $D_{28} = -0.49$, -0.60 to -0.38, in Llama; -0.661 in Gemma). On OR-Bench, Llama's contrast recurs, in a non-prospective check (Section 3.4).

Most of the drop is shared. At the reply position post-training lowers final-layer entropy by 7.2 to 8.9 nats in Llama and 4.6 to 5.6 in Gemma in every category, factual and nonsense included; in Llama mostly because Base, under the untrained template, predicts the start of another user turn. The matched subtraction removes this shared shift. In an exploratory decomposition, Llama's JBB contrast comes almost entirely from Instruct, while on XSTest Base's higher entropy on unsafe prompts under the untrained template also adds to it (Appendix D).

Table 3: Finding 1 at the chat reply position and final layer (28 in Llama, 26 in Gemma). D: harm-specific entropy contrast in nats, with each model's own head or with both models' states through one fixed head; negative means post-training lowered entropy more on harmful prompts. Readout locus: first layer of the sustained divergence (late range from layer 22 in Llama, 21 in Gemma). Monotone baseline: harmful-minus-harmless mean residual (nats) after the fitted non-decreasing map from Base to Instruct entropy. Top-1 log-odds: the same contrast on the log-odds of each model's top output token, positive when Instruct is more determinate on harmful prompts. Parentheses: 95% intervals. Llama's D is printed at two decimals, as recorded; its last two rows are OFF003′s descriptive parity arm.

	Llama, JBB	Llama, XSTest	Gemma, JBB	Gemma, XSTest
D, own heads	-0.46 (-0.57, -0.35)	-0.63 (-0.72, -0.54)	-0.598 (-0.803, -0.388)	-0.467 (-0.605, -0.331)
D, Base head	-0.70	-0.73	-0.805	-0.472
D, Instruct head	-0.51	-0.66	-0.575	-0.491
Readout locus	28	28	26	23
Monotone baseline	-0.411	-0.385	-0.722 (-0.837, -0.587)	-0.616 (-0.708, -0.509)
Top-1 log-odds	5.061	4.713	3.961 (3.503, 4.394)	1.587 (1.282, 1.883)

Carried by the states. With both models' states read through the Base head or through the Instruct head, the contrast keeps its sign and roughly its size in both models and datasets (Table 3). It is state-carried in all four cells, not a change in the tied readout alone.

Late in the readout, with earlier transients. The declared readout locus is late in both models: layer 28 in Llama, and layers 26 (JBB) and 23 (XSTest) in Gemma under both late boundaries (Figure 1, Figure 2). Late-only does not mean that nothing diverges earlier. In Llama the simultaneous band excludes zero at every layer from 19 to 23 on XSTest and includes it at 24-27, and on JBB it alternates in sign between layers 17 and 24 (Appendix D). In Gemma the bands lie below zero at layers 20-21 on JBB and 19-21 on XSTest, above zero at layer 22 in both, and below zero again from 23. The readout agrees with the model's final distribution only at the last layers, and at the reply position only at the final layer in both models. Earlier readout statements concern the projection only, and in Llama a final-consistent locus is unidentifiable everywhere.

What differs between the models. Away from the reply position the two models disagree (Figure 1, Figure 2; numbers in Appendix D). In Llama no readout divergence before the reply position passes the declared checks: at the last instruction token JBB shows none, and XSTest's vanishes when entropy is renormalized over regular tokens, which excludes the end-of-turn token; raw text shows none. At the dialogue frame's own reply position XSTest replicates but JBB does not, so the effect is not format-general. Gemma diverges at the last instruction token and in raw text, also under renormalized entropy, but with D_{26} of -0.238 and -0.247 on JBB and $+0.277$ and $+0.263$ on XSTest, the positive sign against the low-entropy prediction. In Gemma these two positions are only three template tokens apart, against a system block in Llama, and the sign

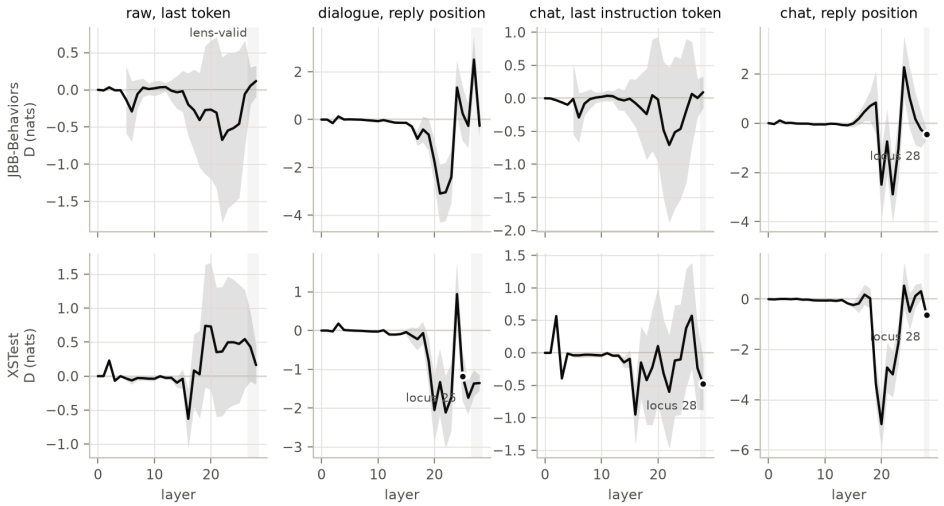

Figure 1: Llama 3.2 3B: the harm-specific readout contrast D_ℓ by layer, each model read through its own head, by dataset (rows) and read position (columns). Negative values mean that post-training lowered entropy more on harmful than on matched harmless prompts. Shading is the simultaneous 95% band over non-echo layers; dots mark declared loci; the pale strip ("lens-valid") marks final-consistent layers. Vertical scales differ.

disagreement leaves no dataset-general statement. Under the dialogue frame Gemma's divergence is format-general. Two Gemma labels are sensitive to the logit softcap; the reply-position labels are not. Of the 47 labels compared across models, 29 agree; they are an inventory of dependent labels, not independent tests.

Floor characterization. Instruct's reply-position entropy is near zero for most harmful prompts (in Gemma, 86 of 100 JBB harmful prompts lie below 0.1 nat), so a floor could compress the contrast. After the monotone baseline, harmful prompts keep lower mean residuals than harmless ones in both models, and on the top-1 log-odds scale, which has no floor, Instruct is more determinate on harmful than on harmless prompts beyond Base's difference (Table 3; rank-scale gaps in Table 6).

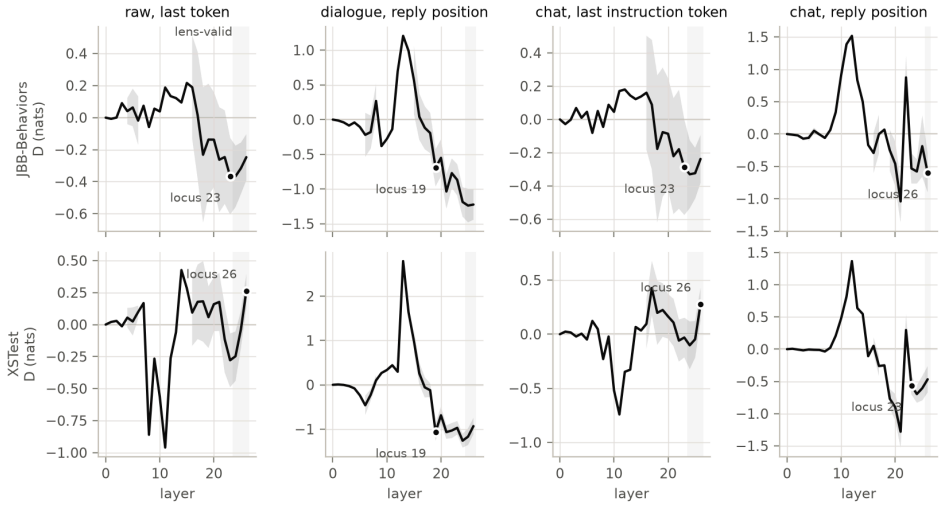

Figure 2: Gemma 2 2B: the same contrast and layout as Figure 1. The reply-position locus is again late, with earlier transient divergences; unlike Llama, Gemma also diverges at the last instruction token and in raw text, with signs that differ between datasets.

In Gemma's XSTest the contrast sits in pairs with lower Base entropy; the highest tertile shows none (−0.052, −0.243 to 0.141). This characterizes the contrast; it was declared without a pass or fail label after its component means had been seen, it does not show that no monotone map could reproduce the aggregate gap, and a deterministic reply opening is not excluded as its mechanism.

The first-token readout. In an exploratory readout, Llama Instruct's top-1 token at the reply position becomes "I" at layer 19 for 80% of JBB harmful and 80.5% of XSTest unsafe prompts and stays high through layer 28, while harmless prompts stay at 2-14% (Figure 11, Appendix D). Instruct's output begins with "I" for 95% of JBB harmful and 13% of JBB benign prompts, but also for 41-42% of factual and nonsense prompts: "I" is a refusal proxy, not a refusal label.

Lens-free states differ earlier. At the reply position the spread contrast G_ℓ is sustained negative well before the late readout locus (Figure 3). In Gemma it has a sustained locus on simultaneous bands at layers 14-15 (raw), 14 (standardized) and 12-16 (unit norm), at locus floors 0 and 0.05 and at both late boundaries: the declared reading is "early", anticipated for raw and standardized states from GEM001's pointwise results. In Llama's parity arm raw and standardized G_ℓ are sustained from layer 13, and unit-norm G_ℓ from 13 on JBB; on XSTest its band excludes zero at every layer from 11 to 25 but not at 26, so the declared rule finds the sustained locus at 27 and reads "norm-carried early" by one layer. CTL001's pointwise intervals agree (Appendix D). So in both models the late readout locus is not the onset of a state difference associated with benchmark harm labels. The two are different measures and are not shown to be stages of one process.

G_ℓ is a relative share, so its sign does not say which class moves. In Llama, harmless prompts take a larger share of state variance in Instruct than in Base on both datasets, while harmful prompts lose share only on JBB (Appendix D); in absolute terms, unit-norm spread rises from Base to Instruct for most paired categories. Descriptively, held-out probes separate harmful from harmless prompts far more in Instruct states than in Base states in both models, a separation confounded with refusal (Appendix D).

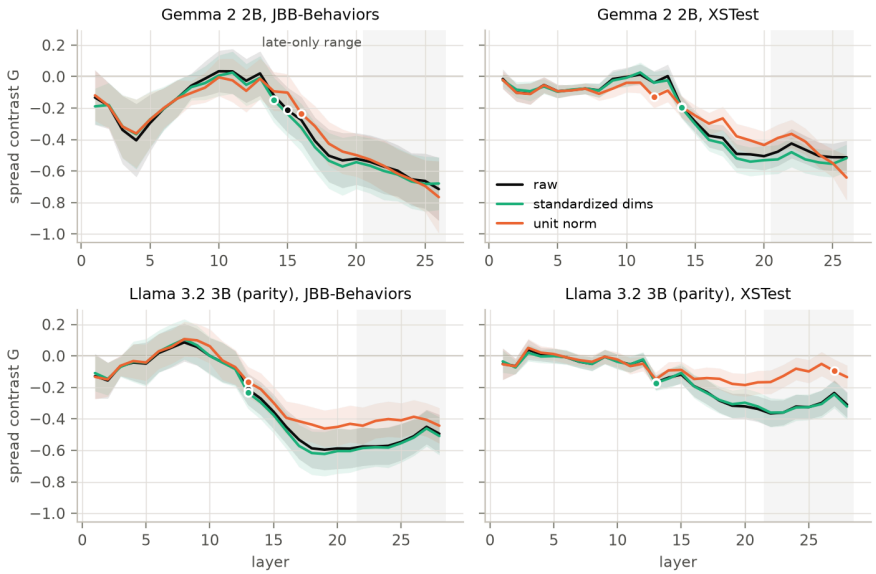

Figure 3: Lens-free spread contrast G_ℓ at the chat reply position, on raw, standardized and unit-norm states, with simultaneous 95% bands: Gemma (top) and Llama's descriptive parity arm (bottom). Negative values mean that post-training lowered the share of state variance taken by harmful prompts relative to harmless ones. Dots mark sustained loci; the gray strip marks the late-only range of the readout.

3.2 What the first token marks (GEM002 to GEM004)

Finding 1 rests on a nearly determined first reply token, and Finding 2 uses that token as its outcome. Gemma's runs ask how far the token marks judged refusal (Section 2.4).

Rule gate (GEM002). In Gemma, Instruct's first generated token is "I" for 97% of JBB harmful, 25% of JBB benign, 85% of XSTest unsafe and 9% of XSTest safe prompts, and equals GEM001′s stored output top-1 for all 850 prompts. GEM002′s gate failed: the judge agreed with the prefix rule on 30 of 30 rule-positive but 52 of 60 rule-negative harmless replies (5 judged refused, 3 ambiguous), because the rule misses refusals that open with "I understand …, but I cannot". Every GEM002 proxy result is therefore descriptive.

Prospective validation (GEM003). The validation was complete: all 400 sampled replies were judged, 231 refused (38 flagged partial), 147 complied and 22 ambiguous. Against the 378 non-ambiguous labels, "I" has precision 0.911 (0.868 to 0.941) and recall 0.935 (0.896 to 0.960), and is validated as a refusal marker on harmless prompts (Table 4). The phrase score ranks refusals above compliances (AUROC 0.901, 0.870 to 0.930) but recalls only 0.325 of the judged refusals at positive log-odds, so it is a ranking score only; 116 of the 231 refusals open outside the phrase set. Both readings depend on the ambiguous labels. Counted as compliance, they lower the precision bound of "I" to 0.824 and the score's AUROC bound to 0.84989, a bound that other bootstrap seeds (with GEM003′s score) and the last-pass designation put on either side of 0.85, and neither endpoint is validated; counted as refusal, both readings hold. Descriptively, the first-token log-odds of "I", the continuous form of E1, has a higher AUROC than the phrase score (0.944, 0.920 to 0.965); the difference was not tested. The validation concerns non-ambiguous replies to harmless prompts only.

On GEM001′s pairs, the phrase-score contrast, a refusal-phrase log-odds contrast because the score only ranks, is 7.48 (6.41 to 8.54) on JBB, 7.60 (7.03 to 8.17) on XSTest and 7.56 (7.03 to 8.08) pooled; Instruct's own harmful-minus-harmless gap is 7.80, 8.35 and 8.16, and Base barely separates the pairs. The declared secondary contrast on the first-token log-odds of "I" is 9.24 (8.27 to 10.21) on JBB, 8.27 (7.71 to 8.81) on XSTest and 8.59 (8.09 to 9.07) pooled. Whether either endpoint marks refusal on harmful prompts is not tested (Appendix D).

A repaired score. GEM003 computed each phrase's probability in its own forward pass. For one JBB harmful prompt the 13 probabilities summed to 1.004, which no single distribution allows; the declared censoring rule then gave it a very large score, which inflated GEM003′s JBB and pooled contrasts (7.79 and 7.66), and the declared checks, which tested finiteness and repeatability, did not catch it. An external review did. GEM004 recomputed the score coherently: each shared phrase

Table 4: How well Gemma 2 2B Instruct's endpoints and GEM002′s rules mark refusal: agreement with GEM003′s judge on 378 non-ambiguous replies to harmless OR-Bench Hard-1K prompts (231 judged refused), with Wilson 95% intervals. The prefix rule searches the reply's opening, the phrase rule the whole reply; both were applied to the 64-token replies.

	Precision	Recall	Specificity
First token "I" (E1)	0.911 (0.868, 0.941)	0.935 (0.896, 0.960)	0.857 (0.791, 0.905)
Phrase score above zero (E2)	0.987 (0.929, 0.998)	0.325 (0.268, 0.388)	0.993 (0.962, 0.999)
Prefix rule	0.983 (0.940, 0.995)	0.498 (0.434, 0.562)	0.986 (0.952, 0.996)
Phrase rule	0.959 (0.922, 0.979)	0.814 (0.759, 0.859)	0.946 (0.896, 0.972)

prefix takes its next-token distribution from one designated pass, so the phrase probabilities sum to at most one while the estimand is unchanged. It reproduced GEM003′s stored per-pass values bitwise, and the invalid prompt now scores 4.95. The contrasts above are GEM004′s; the declared E2 readings (primary designation) are unchanged. Across 15 designations of the pass that supplies each prefix, the contrast varies by at most 0.013, so it is stable to one decimal under this declared designation sensitivity, which is not an absolute bound on numerical error. Single prompts' scores depart from the primary designation by up to 0.37, so row-level scores are not precise to that level. Construction and chronology are in Appendices A and B.

3.3 Finding 2: within Llama Instruct, the proxy contrast moves with the suffix and reply together

The reply position alone (CTL003). G was 18.14 (95% interval 16.54 to 19.69) on JBB and 18.45 (17.18 to 19.67) on XSTest, and every donor-set gate passed. A harmful donor's reply state installs the contrast in harmless recipients from mid-depth in both datasets: the lower band exceeds 0.5 from layer 15 on XSTest and from layer 16 on JBB, there narrowly (0.5005), within the bootstrap's Monte Carlo error. Removal from harmful recipients is weaker: deep on XSTest, sustained from layer 23, and indeterminate on JBB, where the band at layer 23 spans 0.43 to 0.67 (Figure 15, Appendix D). Neither declared counter-criterion is met, and the split between directions matches the expectation written into the declaration before the run.

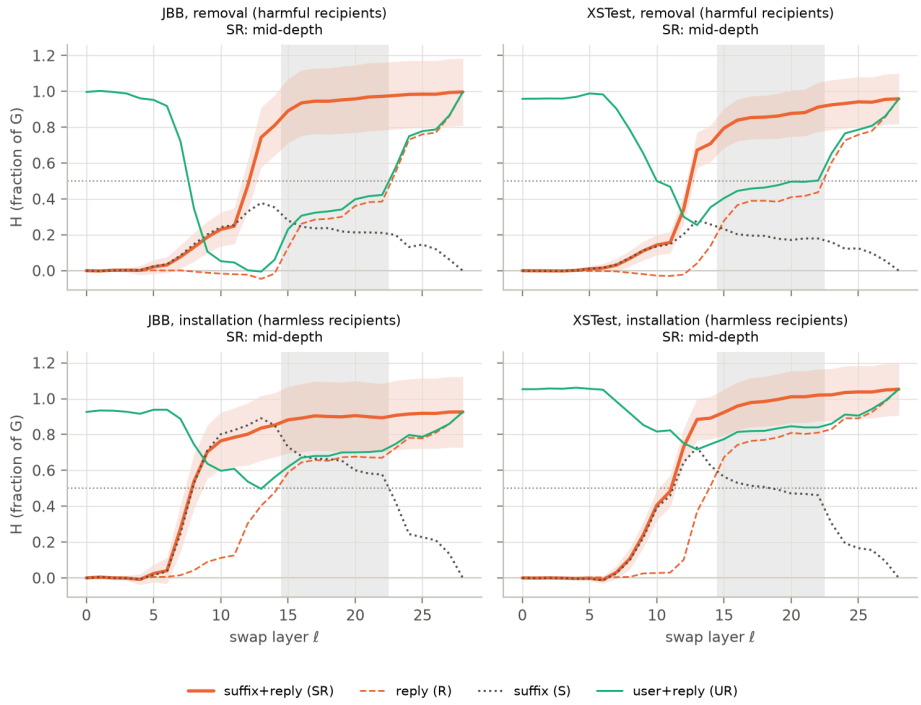

Figure 4: Llama 3.2 3B Instruct, transfer fraction H by swap layer: the donor-class contrast in the log-odds of "I", as a fraction of the unpatched gap G, for the joint suffix+reply mask (thick orange, with simultaneous 95% bands), the reply (dashed) and suffix (dotted) masks alone, and the joint user+reply mask (green). Top: removal from harmful recipients (declared); bottom: installation into harmless recipients (descriptive). The gray band marks the window, layers 15-22; the dotted line marks 0.5. By construction, the user+reply mask at layer 0 and every mask that includes the reply at layer 28 return the donor's own output. Titles give the suffix+reply label.

The suffix and reply together (CTL004). Over layers 15-22 no single mask removed more than about 0.45 of the contrast. Swapping the template suffix and the reply position together removes it from mid-depth in both datasets (Figure 4): the lower band of $H(SR)$ exceeds 0.5 from layer 13 through 28. Over the window, $W(H(SR))$ is 0.946 (simultaneous 95% interval 0.765 to 1.128) on JBB and 0.860 (0.741 to 0.979) on XSTest, about 95% and 90% of the donor's own layer-28 value. Either position alone leaves most of the contrast in place: $W(H(R))$ is 0.299 and 0.384, $W(H(S))$ 0.232 and 0.192, all with upper bounds below 0.5. The increment over the reply alone, $W(J_{SR})$, is 0.647 (0.519 to 0.775) and 0.476 (0.407 to 0.545), material in both datasets. The same-class control is stable (-0.009, from -0.093 to 0.076, and -0.003, from -0.061 to 0.055), but only as a mean: under same-class joint swaps the mean absolute row change in harmful recipients' log-odds is 0.20 G, and 56-57% of rows move by more than 0.1 G, while the median KL stays near zero, which is consistent with saturation of the proxy. Both cells are redundancy-consistent. On OR-Bench all four results replicate (Section 3.4).

What this establishes. Over the window, joint suffix+reply removal crosses the declared threshold while the reply and suffix masks alone do not. Descriptively, either single mask produces a positive installation contrast; suffix-only installation on XSTest is about half of G. As a draft interpretation, this supports an OR-like reading, in which the output proxy reflects the harmful class if either the suffix or the reply carries a harmful-donor state. It does not establish necessity or sufficiency for individual recipients. Redundancy, class-dependent compatibility and nonlinear interaction are compatible with this pattern and not mutually exclusive, since an OR rule is itself nonlinear; these swaps cannot separate them. CTL004 is consistent with CTL003′s direction asymmetry being specific to single-position swaps. The result concerns interventions on one output proxy under one frozen donor assignment, with a window chosen from CTL003′s curves, and what the proxy means behaviorally after an intervention is untested. It does not show where information is computed or moves, a specific harm representation, or refusal behaviour.

3.4 Finding 3: crossed-donor composition, and what OR-Bench adds

Composition (CTL005). The gates passed: G is 18.22 on JBB (69 pairs) and 18.13 on XSTest (159 pairs), and the same-donor transfer contrast lies between 0.934 and 0.954 in both datasets and recipient classes, every lower bound above 0.5. Both datasets agree on all eight primary labels (Figure 18; Figure 5, right), and no declared counter-condition occurred.

- *Recipient side: invariant.* Δ_{hb} is -0.018 (simultaneous 95% interval -0.082 to 0.047) on JBB and 0.022 (-0.030 to 0.073) on XSTest; Δ_{bh} is 0.010 (-0.041 to 0.060) and -0.007 (-0.044 to 0.031). A mixed composition gives the same window mean, within 0.1 G (about 1.8 log-odds units), whether it sits in a harmful or a harmless recipient. On the recipient side, class, user text, length and donors vary together, and the user text stays the recipient's own.
- *Interaction: negative.* I_{harmful} is -0.367 (-0.444 to -0.290) on JBB and -0.287 (-0.342 to -0.233) on XSTest; I_{harmless} is -0.375 (-0.457 to -0.293) and -0.318 (-0.381 to -0.255). The two single-position contrasts together exceed the two-donor contrast by more than 0.1, so on average a single harmful position gives more than half of it.
- *Coherence: invariant.* Every E interval lies within ± 0.1. The estimates run from -0.051 to -0.012 on JBB and from -0.049 to -0.026 on XSTest.

As a draft interpretation (ATA-C028): over layers 15-22, under these donors and this proxy, the class means of the normalized log-odds of "I" after joint suffix and reply swaps depend, for the tested compositions, on the donor classes of the two states. A single harmful-donor state gives, on average, more than half of the two-donor contrast, but neither position alone reaches it: OR-leaning, not OR. This is a prospective test of CTL004′s draft OR-like reading on reused prompts

and new intervention tuples; the hypothesis came from CTL004′s outputs, so it is not independent confirmation. The reading concerns class means over the window: individual rows vary widely under same-class swaps, origin equivalence was not established for harmless recipients' (h, b), and at layer 13 the recipient side matters more and the order of the two mixed cells reverses (Appendix D, which also relates these interactions to CTL004′s). It describes a functional property of one output proxy under defined interventions, and does not show duplicated harm information, redundancy, homogeneous representations or refusal behaviour.

Replication on OR-Bench (CTL006). All declared checks passed; CTL002′s stored logits were reproduced bitwise on all 1,141 prompts before any swap. Figure 5 sets each estimate beside its originals.

- *R1 recurs (non-prospective).* D_{28} is -0.592 nats (simultaneous 95% interval -0.686 to -0.498) with own heads, -0.674 (-0.774 to -0.574) through the Base head and -0.552 (-0.649 to -0.454) through the Instruct head, each a harm-specific decrease beyond the 0.1 nat floor, and none is length-sensitive.
- *R2 replicated on all four results.* G is 10.11 (95% interval 9.35 to 10.86). $W(H(SR))$ is 0.974 (0.871 to 1.076), while the reply alone gives 0.419 (0.366 to 0.473) and the suffix alone 0.261 (0.225 to 0.297). The increment over the reply, $W(J_{SR})$, is 0.554 (0.497 to 0.612), and the control is stable, 0.039 (-0.014 to 0.093): a material increment and a redundancy-consistent cell. The lower band of $H(SR)$ exceeds 0.5 at every measured layer from 13 (0.521 there; Figure 20). This is consistent with an onset at or before layer 13; layers 14 and 23-27 were not measured, so it does not locate the onset.
- *R3: six of eight labels replicated, two indeterminate, none contradicted.* Δ_{hb} is 0.008 (-0.037 to 0.053) and Δ_{bh} 0.022 (-0.015 to 0.058), recipient-side invariant. I_{harmful} is -0.289 (-0.332 to -0.247) and I_{harmless} -0.256 (-0.298 to -0.215), a negative factorial interaction. The harmful-donor coherence terms are invariant (-0.059 and -0.061). The harmless-donor terms, -0.098 (-0.115 to -0.080) and -0.090 (-0.107 to -0.074), straddle -0.1 and read indeterminate: on OR-Bench, two-prompt same-class harmless cells sit about 0.1 G above one-prompt cells, at the edge of the margin, in CTL005′s direction.

The refusal-consistent subset qualifies R2. On the 316 refusal-consistent pairs (G = 12.53), $W(H(SR))$ is 0.794 (0.707 to 0.881) and the increment holds, but the same-class control is no longer stable (-0.121, -0.153 to -0.090; selectivity indeterminate) and $H(SR)$ at layer 13 is below half (0.433, 0.380 to 0.486), so "above half from layer 13" holds on all pairs but not on this subset. In log-odds units the donor-class contrast is similar on all pairs and on the subset over the window (9.8 and 9.9, a comparison not declared), so the smaller normalized value mainly reflects the subset's larger gap, and the persistence statement is tied to the normalization by G as well as to the prompts. An exploratory regression specified before it was computed (D065) finds little linear association between the harmful-donor joint-swap output and the recipient's own baseline in the analysed sample; a stable class-mean control can average large row-level shifts of opposite sign (Appendix D). R1′s contrast is larger on the subset (own heads -0.773, -0.886 to -0.659), and all eight R3 labels recur.

What OR-Bench adds. CTL006 widens the prompt base, not the model base: R2 and R3 were prospective on prompts no swap had used, with the same checkpoints, template, proxy, window and margins, while R1 was not prospective. The pairs are constructed from benchmark labels, not semantic matches. The replicated statements concern window means over class-labelled prompts; rows vary widely, the onset depth is not established, and per category the main estimates keep the pooled sign (Appendix D).

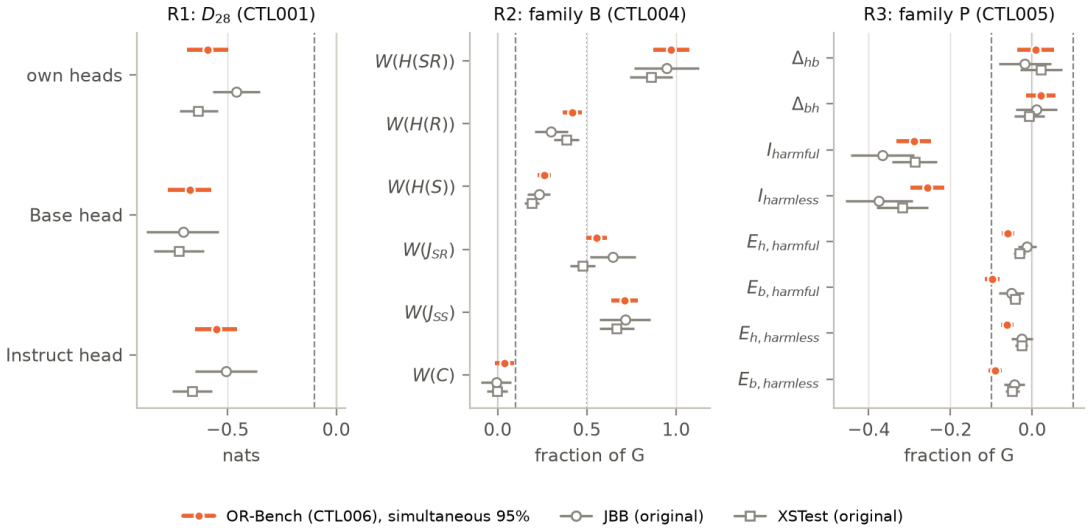

Figure 5: The OR-Bench replication (CTL006) against its originals. OR-Bench estimates with simultaneous 95% intervals (orange) beside the original runs' JBB (circles) and XSTest (squares) results, each with its own intervals. Left: Llama's final-layer entropy contrast D_{28} (Finding 1; CTL001's intervals are pointwise). Middle: the removal statistics of Finding 2 (CTL004), in fractions of the unpatched gap G. Right: the composition statistics of Finding 3 (CTL005), recipient-side differences Δ, interactions I and coherence shifts E. Dashed lines mark the declared margins (± 0.1 and the 0.1-nat floor) and the dotted line 0.5.

4 Discussion

Declared counter-criteria. Each has a definite outcome. The readout contrast is neither null nor readout-only: it holds at the chat reply position in both datasets and both models, and survives with both models' states read through either head. In Llama the criterion "divergence at the reply position only, none at the last instruction token" holds for the readout in both datasets after the end-of-turn check, so for the readout it counts against the conjecture; Gemma's earlier divergences disagree in sign between datasets. For Llama's geometry at the last instruction token the criterion splits: near zero on JBB ($|G|$ at most 0.05, no sustained sign), negative on XSTest, so no dataset-general geometric statement follows there.

Three depths. *Prompt position:* in Llama the readout signature and the strong geometric contrast both sit at the reply position; only XSTest's geometry at the last instruction token lies before it, with no readout counterpart. Gemma's readout diverges there too, with signs that differ between datasets. *Layer depth:* in Llama the declared readout locus is the last layer, the exploratory first-token readout ("I") leads from layer 19, the spread contrast is sustained from layer 13, and swaps of the last five positions transfer more than half of the donor-class contrast from layer 13. These are different measures, not stages of one process: the models already differ at layers 8-12 in every category (lower matrix-based entropy in Instruct), and Instruct separates harmful from harmless prompts more than Base by layer 10 (exploratory d about 2.5 against 1.25). *Output-token depth*, the subject of Qi et al. (2025), is not measured: we read only the first reply token.

The guiding conjecture. Readout entropy does not support an early low-entropy attractor: the declared divergence is late in both models, and the logit lens is not final-consistent below the last layers. The lens-free result is more specific than contraction toward a refusal state. In Llama the robust change is that harmless prompts, and in an exploratory split by the first-token proxy those whose first reply token is not "I", take a larger share of state variance in Instruct than in Base;

harmful prompts lose share on JBB only, and keep their matrix entropy while doing so. A simpler account competes: late-layer compactness could reflect a shared output constraint rather than a harm-specific state, and the declared test of that account (OFF001, Appendix C) was indeterminate. The swaps locate where, within Llama Instruct, the harm-associated contrast in the first reply token can be moved; they are results about interventions on one output proxy, not about how the model computes refusal or where information moves. Nothing here shows convergence dynamics, stability under perturbation or any topological property; "attractor" remains a hypothesis.

Across models. Gemma widens the model base for the first finding only. What does not carry over is as informative as what does: Gemma's divergence before the reply position and under the dialogue frame, the softcap-sensitive labels, and the prefix rule, which missed Gemma's refusals that open with "I understand …, but I cannot". The swap results have not been repeated on Gemma.

4.1 Limitations

Scope. Two model families at 2-3B parameters, two checkpoints each, and short, single-turn requests from three benchmarks. The swaps and the OR-Bench replication exist for Llama only. Gemma's measurements rest on a declared adapter: its softcap, attention implementation, chat template and read-position mapping differ from Llama's, and its late boundary was chosen after single-prompt pilot curves were seen. OFF003′s floor statistics are characterization, and its Llama arm is a parity analysis of retained arrays.

The proxy and its validation. The swap outcome and the first-token readouts are proxies. The judge-based validations of refusal labels (Llama's prefix rule, Gemma's first token "I") covered harmless prompts only; Llama's first-token proxy was not itself judged, and transfer to harmful prompts and to intervened states is untested. GEM002′s refusal labels failed their validation gate. GEM003′s validation of "I" depends on how its 22 ambiguous labels count, and its window is 64 greedy tokens. Its judge is one language model, run once per batch, with self-agreement of 8 of 10 labels on repeated cases; three of its sessions were stopped by an API safeguard and recovered by the declared retry and split. CTL002′s and GEM002′s judges were not blind to the project, while GEM003′s ran outside it. The phrase score first took each phrase from a separate pass, which left one invalid score; the repaired score still carries the passes' disagreement at the level of single prompts. CTL001′s refusal proxy is the first reply token; CTL002′s labels come from 16 generated tokens and a prefix rule validated on Hard prompts only.

Labels and confounds. The effects are harm-associated: they follow benchmark harm labels, which are confounded with refusal versus compliance. Base reads the chat template through untrained rows: the JBB contrast is driven by Instruct, but the XSTest contrast is partly Base's, and the Base side of every chat contrast is off-distribution. OR-Bench's Toxic labels have no human check, Hard-1K was selected on large models' refusals, and CTL006′s pairs are matched only by category and length.

Design of the swaps. CTL003 holds one frozen donor assignment fixed: its bands do not cover donor selection, donor content differs beyond harm, and its eligible pairs exclude long JBB and short XSTest prompts. CTL004 reuses that assignment, chose its window from CTL003′s curves without correcting for the choice, and tests two joint masks only. CTL005 reuses CTL003′s prompts and pairs with two more frozen donors per class from the same pool, tests a hypothesis generated from CTL004′s outputs, excludes pairs without three candidates per class (mostly long JBB harmful prompts), and keeps the recipient's own user text. CTL006 keeps the checkpoints, template, proxy, window, layers and margins; R1 is non-prospective, and layers 14 and 23-27 were not measured.

Statistics. Layerwise bands are simultaneous across layers within a condition, not across conditions. CTL004′s and CTL005′s family intervals are simultaneous within each family, across both datasets, and CTL006′s within each family on one dataset. The geometry intervals of CTL001 and GEM001 are pointwise, while OFF003′s spread-contrast bands are simultaneous over layers within each variant, read and dataset, not across them. Spread contrasts are relative variance shares, not absolute scale. The composition and token analyses of CTL001 and the swap readouts of CTL002 were exploratory, added after the declared analyses; separability was declared exploratory and computed with them.

Acknowledgments

During the development of this work, the author used OpenAI's Codex and Anthropic's Claude to support critical discussion and manuscript development, and Google Gemini Research to assist with literature discovery and evidence organization. Claude additionally supported drafting, revision, code development and technical checks. Scholarly claims are supported by the cited literature and retained study records. The author directed the research and retains responsibility for the selection and interpretation of evidence and for the final manuscript.

References

Ali, Riccardo, Francesco Caso, Christopher Irwin, and Pietro Liò. 2025. *Entropy-Lens: Uncovering Decision Strategies in LLMs*. https://arxiv.org/abs/2502.16570.

Arditi, Andy, Oscar Obeso, Aaquib Syed, et al. 2024. "Refusal in Language Models Is Mediated by a Single Direction." *Advances in Neural Information Processing Systems* 37. https://doi.org/10.52202/079017-4322.

Belrose, Nora, Igor Ostrovsky, Lev McKinney, et al. 2023. *Eliciting Latent Predictions from Transformers with the Tuned Lens*. https://arxiv.org/abs/2303.08112.

Chao, Patrick, Edoardo Debenedetti, Alexander Robey, et al. 2024. "JailbreakBench: An Open Robustness Benchmark for Jailbreaking Large Language Models." *Advances in Neural Information Processing Systems* 37. https://doi.org/10.52202/079017-1745.

Cheng, Emily, Diego Doimo, Corentin Kervadec, et al. 2025. "Emergence of a High-Dimensional Abstraction Phase in Language Transformers." *International Conference on Learning Representations*. https://arxiv.org/abs/2405.15471.

Cui, Justin, Wei-Lin Chiang, Ion Stoica, and Cho-Jui Hsieh. 2025. "OR-Bench: An over-Refusal Benchmark for Large Language Models." *Proceedings of the 42nd International Conference on Machine Learning*. https://arxiv.org/abs/2405.20947.

Facco, Elena, Maria d'Errico, Alex Rodriguez, and Alessandro Laio. 2017. "Estimating the Intrinsic Dimension of Datasets by a Minimal Neighborhood Information." *Scientific Reports* 7: 12140. https://doi.org/10.1038/s41598-017-11873-y.

Meng, Kevin, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. "Locating and Editing Factual Associations in GPT." *Advances in Neural Information Processing Systems* 35. https://doi.org/10.52202/068431-1262.

Nikolenko, Sofiia, Michele Papucci, Mina Rezaei, and Shireen Kudukkil Manchingal. 2026. *What Intermediate Layers Know: Detecting Jailbreaks from Entropy Dynamics*. https://arxiv.org/abs/2606.25182.

nostalgebraist. 2020. *Interpreting GPT: The Logit Lens*. LessWrong. https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens.

Qi, Xiangyu, Ashwinee Panda, Kaifeng Lyu, et al. 2025. "Safety Alignment Should Be Made More Than Just a Few Tokens Deep." *International Conference on Learning Representations*. https://arxiv.org/abs/2406.05946.

Röttger, Paul, Hannah Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. 2024. "XSTest: A Test Suite for Identifying Exaggerated Safety Behaviours in Large Language Models." *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (Mexico City, Mexico), 5377–400. https://doi.org/10.18653/v1/2024.naacl-long.301.

Skean, Oscar, Md Rifat Arefin, Dan Zhao, et al. 2025. "Layer by Layer: Uncovering Hidden Representations in Language Models." *Proceedings of the 42nd International Conference on Machine Learning*, Proceedings of machine learning research, vol. 267: 55854–75. https://arxiv.org/abs/2502.02013.

Stolfo, Alessandro, Ben Wu, Wes Gurnee, et al. 2024. "Confidence Regulation Neurons in Language Models." *Advances in Neural Information Processing Systems* 37. https://doi.org/10.52202/079017-3970.

Valeriani, Lucrezia, Diego Doimo, Francesca Cuturello, Alessandro Laio, Alessio Ansuini, and Alberto Cazzaniga. 2023. "The Geometry of Hidden Representations of Large Transformer Models." *Advances in Neural Information Processing Systems* 36. https://doi.org/10.52202/075280-2230.

Zhou, Zhenhong, Haiyang Yu, Xinghua Zhang, Rongwu Xu, Fei Huang, and Yongbin Li. 2024. *How Alignment and Jailbreak Work: Explain LLM Safety Through Intermediate Hidden States*. https://arxiv.org/abs/2406.05644.

Appendix A. Run records and checks

From CTL002 on, declarations were reviewed before approval. For CTL004 to CTL006, the code was reviewed in rounds before launch, a receipt bound the reviewed code, declaration, reference outputs, smoke runs and reviews to the launch commit, verified at startup and by a post-run gate, and analyses ran only on committed, gated outputs. The second-model runs (GEM001 to GEM004, OFF003) had critic and code reviews filed before launch, smoke tests, source-commit stamps and declared checks, without a receipt.

CTL001. Its rules were declared before its retained capture (research/runs/CTL001.yml, committed before the capture code) and amended twice before it, after an independent review (A1) and a literature check (A2); a discarded 4-prompt smoke test informed one amended check. All declared checks passed. Hooked residuals matched the model's hidden states on every pass, repeated passes were bitwise identical, and the pilot's states were reproduced bitwise. The fp32 readout at layer 28 agreed with each model's own bf16 output in all 6,800 prompt positions when bf16 ties are counted, and in 98.3% strictly; the median entropy difference was 0.008 nats. The declared rule first required 99% strict agreement; a smoke test before the run found an exact bf16 tie, and the rule was amended before the retained capture. An independent recomputation with separate code reproduced every contrast and re-read 68,000 readout cells within 4e-4 nats.

CTL002. Declared, reviewed four times and approved before its code was written; its analysis code was committed after a first harmless smoke run and before the main run, and was tested on synthetic data but not fully reviewed before results. The runner's final revisions were smoke-tested and committed before launch and independently reviewed during the run, which found no blocker. All declared checks passed. Both models reproduced CTL001′s states and logits bitwise on the 650 shared prompts, the 600 unswapped outputs reproduced CTL001′s logits bitwise, identity swaps were bitwise at all positions and at the reply position, and repeated generation was token-identical. The first generated token equals the captured argmax for all 2,624 prompts. No harmful generated text was stored.

CTL003. Declared, reviewed and approved before its code was written, and its runner and analysis were independently reviewed before launch. The run record marks the working tree as not clean, because review records were written during checkpoint hashing; no file under lab/ differed from the reviewed commit. CTL001′s outputs were reproduced bitwise on all 600 prompts before any swap; identity and combined-mask swaps were bitwise; the six mechanical anchors held exactly on all 1,048 swapped rows.

CTL004. The declaration was reviewed three times before approval and once more, on its execution contract, before launch. The runner, analysis and gate were independently reviewed in two rounds before launch. A receipt bound the reviewed code, the declaration, the reference outputs, the smoke runs and the reviews to the launch commit, and the run verified it before checkpoint

hashing and again before model loading. Both startup checks verified the receipt. CTL001′s outputs were reproduced bitwise on all 600 prompts before any swap, and CTL003′s single masks on all 91,176 stored entries. Identity swaps, patched repeats and all ten mechanical anchors were exact, and every patched pass asserted exact position membership.

CTL005. The declaration was reviewed by internal critics and twice by Astra before approval. Before launch, the runner, analysis and gate were reviewed in two rounds each, with a third round on one runner change, and a receipt bound the reviewed code, declaration, references, smoke runs and reviews to the launch commit. Both startup checks verified the receipt. CTL001′s outputs were reproduced bitwise on all 600 prompts before any swap, and CTL004′s slot-1 conditions on all 30,096 stored triples. Identity swaps, patched repeats, all 64 anchors and the prompt behind every written part were exact, and every patched pass asserted each source's position membership.

CTL006. Declared after a feasibility check and a researcher's reading of OR-Bench's labelling, revised once after an internal critic, and approved under the author's advance approval of the replication plan, with the critic and an advisor check in place of an external pre-run review. The runner, gate, receipt writer and analysis were reviewed in five rounds; five smoke runs each reproduced CTL005′s stored smoke outputs bitwise on 704 triples and CTL002′s stored logits bitwise on 24 prompts, and all five were bitwise equal. The receipt bound 27 inventory files (code, launchers, lock file and manifests), 7 declaration files, 10 references, CTL002′s states files, the smokes and 8 review reports. Both startup checks passed. CTL002′s stored logits were reproduced bitwise on all 1,141 prompts before any swap, and the gate recomputed the unpatched scores from them (largest difference 1.9e-6). Identity swaps, patched repeats, all 64 anchors on 1,030 rows and the prompt behind every written part were exact; all 362,816 patched passes asserted per-source membership. The readout check for R1 agreed in all 2,256 readouts counting bf16 ties (median entropy difference 0.0047 nats).

GEM001. Declared after three technical pilots on harmless prompts (the first crashed before writing any output), reviewed by a research critic and a code review before approval, and approved by the author; its analysis code was replayed on CTL001′s own readout and states and reproduced CTL001′s analysis in every section. A smoke test passed before the full run. All declared checks passed: hooked residuals matched the model's hidden states on every pass, repeated passes were bitwise identical, the pilot's states were reproduced bitwise, and the bf16 Base copy reproduced the cast-at-load model bitwise. The fp32 readout at layer 26 agreed with each model's bf16 output top-1 in 100% of prompt positions counting bf16 ties (98.2% exactly; median entropy difference 0.020 nats), and the lens-math gate passed on all 128 rows (largest logit difference 3.8e-5). An independent recomputation reproduced every contrast within 5e-6; an external review reproduced every requested statistic and recommended the wording adopted here.

GEM002. Declared after the external review, revised after a research critic and two code-review rounds, and run after three smoke tests on harmless prompts. Every generation check passed: repeated generation was token-identical, the first generated token equalled GEM001′s stored output top-1 for all 850 prompts with bitwise-equal first-step logits, and every harmful record held only the declared fields. Outputs, judge cases and judge labels were committed in that order before the join; the judge received the committed prompt unchanged apart from line endings and used no evidence-access tool. Its first attempt, under the declared model identifier, was stopped by a safety classifier; after the harness's notice not to reproduce that content, a retry recorded under another identifier (claude-opus-5) returned the labels. The judge was also not blind to the project: as context for every subagent, the harness supplied the project's instructions, which name its subject. The effect of the interruption and of that context on the labels cannot be assessed; the

counts are agreement with this judge session. The validation gate failed (Results). An independent recomputation reproduced every value of the validation and analysis.

GEM003. Declared after GEM002, revised after a research critic and two code-review rounds, its parameters approved by the author, and run after a smoke test on harmless prompts that included one real judge batch. Every declared check passed: repeated generation was token-identical and the first generated token matched the forward pass's top-1 under the declared bf16 tie rule; repeated scores were bitwise equal and every score finite. The checks did not test the probability domain: the one score the run recorded as censored is the invalid row described in the Results, whose phrase probabilities sum above one, and an external review found it. The phrases' tokenization matched every reply that opened with one (three phrases occur; ten are unchecked in context). Generation outputs, scores and the sample with its frozen batch prompts were committed before any judge call, and each judge attempt before the next. The judge was claude-opus-5-5 at effort high. Every attempt was checked for a single turn, output from the declared model only, no safety stop and a complete label set, and counted unusable otherwise. Three of 24 attempts were stopped by an API safeguard (category cyber); as declared, one batch's retry and the other batch's six sub-batches were usable, so no case is missing and none was reworded. On the ten repeated cases the judge gave the same label 8 times and the same refused-or-not reading 9 times; both disagreements involve the ambiguous label. An independent recomputation reproduced every value of the validation, sensitivities, contrasts and harmful-side distributions it covered, bit for bit; it did not recompute the tokenization check or the bridge cells that need GEM002′s stored labels.

GEM004. Declared as a repair after the external review of GEM003, revised after a research critic and a code review, authorized in advance by the author, and run after a smoke test on harmless prompts. Every declared check passed for both models: each pass's own-phrase log-probability and the first-token log-odds reproduced GEM003′s stored values bitwise for all 2,169 prompts; repeated passes were bitwise equal; under both declared designations the phrase probabilities summed to at most one and agreed with the escaping mass, and no score was censored. Scores and analysis were committed in that order. An independent recomputation from the stored log-probabilities reproduced every compared value of the scores, contrasts, designation range and validation.

OFF003. Declared before computation, revised three times after a research critic and two code-review rounds (one finding was a bug in the joint reading, fixed before the run), and run after three smoke tests. All declared checks passed for both models: the recomputed contrasts and spread contrasts equalled the stored ones within 5e-6, and no value was missing. An independent recomputation reproduced every recomputed value: the floor statistics, G at the reply position, all loci and readings, and the probes at two layers per model; chat/instr geometry, CKA, transfer and angular spread were not recomputed.

Appendix B. Method details and decision rules

Readout and geometry (CTL001, GEM001)

The readout records the entropy, $\mathrm{KL}(P_{\mathrm{final}} \,\|\, P_\ell)$, the top-1 token, the residual norm, the probability on special tokens and the entropy renormalized over regular tokens. A layer is *echo* for a read position if, in either model, at least half the prompts have a top-1 token equal to the input token there, up to case, leading spaces and punctuation. A non-echo layer is *final-consistent* if, in both models, the median KL to the final distribution is at most 1 nat. Statements at final-consistent layers concern the model's final distribution; statements at other non-echo layers concern the projection only.

The declared positive control is $D_{28} < 0$ at the chat reply position on JBB, judged by the pointwise interval. The primary outcome is the depth of the non-echo locus there. A general statement requires agreement in sign, at layer 28 and in the locus, across both datasets or both formats. At the last instruction token, the template's next token is the end-of-turn marker, which Instruct predicts and Base cannot, so a locus there counts only if it persists in entropy renormalized over regular tokens. The length-adjusted contrast is the intercept of the pair differences regressed on the token-length difference. The cross-readout divergence is *state-carried* if both fixed-head loci exist with the own-head sign at most two layers after the own-head locus, *readout-only* if neither exists, and *mixed* otherwise.

For each model, read position and layer, the spread of a category is computed raw and with each dimension standardized. We also report the per-prompt cosine between centered Base and Instruct states, and two descriptive measures on 100 seeded points per category: matrix-based entropy of centered, unit-norm states, a last-position variant of Skean et al. (2025), and TwoNN intrinsic dimension (Facco et al. 2017), compared only at equal N. Layers where states are nearly identical across prompts (median nearest-neighbour distance below 1% of the median pairwise distance) are excluded.

Second-model adapter (GEM001)

Gemma Base is stored in fp32; runs use a bf16 copy, checked tensor by tensor and end to end to reproduce the bf16 cast at load bitwise. A declared adapter carries CTL001′s rules to the new architecture (Table 5); the declaration (research/runs/GEM001.yml) fixed it, a cross-model rule and a headline predicate before any Gemma result was seen, though pilot curves on single harmless prompts had been seen and are disclosed.

Table 5: The declared Gemma adapter. Everything else is CTL001′s rule set, with the final layer at 26. Layer 0 is the scaled embedding output; the last returned hidden state is post-norm, so block outputs are read by hooks before the final norm.

	Llama 3.2 3B (CTL001)	Gemma 2 2B (GEM001)
Final transform in the lens	RMSNorm, tied unembedding	RMSNorm $\times (1 + w)$, tied unembedding, logit softcap 30; uncapped as sensitivity
Attention	SDPA	eager (SDPA omits Gemma's attention softcap in the pinned implementation)
Chat template	system block with date, user turn	user turn only, no system role or date
Read positions	chat/instr, chat/last	the same roles: last user-text token; position choosing the first reply token
Final layer, late-only from	28, 22	26, 21 (20 as boundary sensitivity)
Control tokens (special mass)	ids $\geq$ 128,000	110 control and reserved ids
Self-consistency	readout check	the same, plus a lens-math gate against the model's fp32 output path

Cross-model rule. A statement counts for both models only where its labels, computed by one function from both analysis files, agree. The headline holds in both if, in each model, the positive control passes, the final-layer interval at the reply position lies below zero for JBB and XSTest, replication holds, the contrast is state-carried in both datasets and neither is length-sensitive. Depth is a separate statement in coarse classes, and relative depth across models is descriptive only.

Refusal proxy and endpoint validation (GEM002 to GEM004)

Refusal proxy (GEM002). Instruct alone generates 32 greedy tokens for all 850 prompts. CTL002′s prefix rule labels each reply in memory (primary), and a phrase rule searches the whole reply for refusal phrases as whole words (secondary); for harmful prompts only the two labels, a stop label and the first output token id are kept. A fresh judge session labels harmless replies without seeing the rules; the gate requires agreement of at least 0.90 with the prefix rule in each of two strata (prefix-refusal and not), at least 20 cases each. If the gate passes, "I" is validated as a refusal marker when, pooled over the 650 benchmark prompts, the Wilson lower bounds of its precision and recall reach 0.85 against both rules.

Endpoint validation (GEM003). A prospective validation on prompts no Gemma run had used. Instruct generates 64 greedy tokens for all 1,319 OR-Bench Hard-1K prompts, harmless requests selected to look harmful, so a refusal there is an over-refusal. A simple random sample of 400, its size fixed from stored rule labels before any judging, is labelled refused, complied or ambiguous by a judge that sees only the rubric, each prompt and its reply: one headless command-line session per batch of about 25, with tools disabled, a pinned client and model, no project files in its context (it loads only the user's global preferences file, which holds no study content) and no saved transcript;

ten cases are repeated across batches. The validation is complete if at least 90% of cases are usable, with at least 60 judged refusals and 60 compliances. (E1) Instruct's first output token "I" is validated as a refusal marker on harmless prompts if the Wilson lower bounds of its precision and recall reach 0.85. (E2) The phrase score $S = \log Q - \log(1 - Q)$, with Q the probability that the reply opens with one of 13 fixed categorical refusal phrases, is computed for both models without generation, with logits upcast to fp64. (This S is unrelated to the suffix mask S of the swaps.) The phrases were chosen partly from GEM002′s replies to the 850 prompts on which D_S is computed: only the endpoint evaluation on Hard-1K is fresh, and the phrase choice and the contrast are not independent of that exposure. S is validated as a refusal score if the lower bounds of its AUROC and of its recall at $S > 0$ reach 0.85, and as a ranking score only if the AUROC bound alone does. Sensitivities count ambiguous labels as complied and as refused. On GEM001′s 300 pairs, D_S is the Instruct-minus-Base change in S on harmful beyond harmless prompts, with pair-bootstrap intervals; it is a refusal-score contrast only if S is validated as one.

GEM003′s construction and its defect. GEM003 computed each phrase's probability from its own forward pass of the bf16 model, so probabilities from separate passes need not be mutually coherent. One Instruct score, on a JBB harmful prompt, had phrase probabilities summing to 1.004; the censoring test also caught $1 - Q < 0$ and assigned it $S = \log 10^{15}$. The declared checks tested finiteness and repeatability, not the probability domain. An external review found the row after GEM003′s results were written; GEM003′s record keeps its computation with dated corrections.

Score repair (GEM004). GEM004 reruns the same 13 passes and recomputes S coherently. Each phrase prefix takes the distribution of its next token from one designated pass, the first phrase in declared order through it, so sibling continuations share one softmax row and $Q \leq 1$ by construction; in exact arithmetic the score is unchanged. Every pass's log-probabilities for the phrase tokens that follow each prefix it reaches are stored, so 15 designations (the first pass, the last, and each single pass) measure sensitivity to which pass supplies each prefix. By the declared rule, a contrast's first decimal is reported only if it varies by at most 0.05 across them. The designations do not bound error against an exact-arithmetic or higher-precision computation. GEM003′s analysis is re-run with the repaired score under its own rules and seeds.

Floor and lens-free states (OFF003)

An offline analysis of GEM001′s retained arrays, with CTL001′s as a descriptive parity arm, asks two questions, declared before computation but after GEM001′s pointwise geometry and the floor's component means had been seen. First, whether the reply-position contrast rests on Instruct's near-zero entropy for harmful prompts. The analysis characterizes it without a pass or fail label: rank-scale gaps, a monotone-baseline contrast (the mean harmful-minus-harmless residual from the pooled least-squares non-decreasing map of Instruct on Base final entropy, exactly zero whenever Instruct entropy is a deterministic non-decreasing function of Base entropy), and the contrast on the top-1 log-odds of each model's output, without the entropy floor. Second, whether Base and Instruct states differ harm-specifically before the late readout locus, without the lens: G on raw, standardized and unit-norm states with simultaneous bands, read jointly ("early" if all three variants have a sustained locus before the late boundary in both datasets with one sign; "norm-carried early" if only raw and standardized do), with a magnitude floor of 0 and, as a sensitivity, 0.05 for the locus; the raw and standardized onsets were expected from GEM001′s pointwise results, with cross-model CKA and held-out difference-of-means probes as descriptives.

Swap decision rules (CTL003 to CTL006)

CTL003. For each recipient, $d = s(\text{harmful donor}) - s(\text{harmless donor})$, and $H_\ell = \bar{d}/G$. Inference uses a pair bootstrap with simultaneous bands over layers, conditional on the frozen donor

assignment; donors are reused up to 4 times on JBB and 7 on XSTest. G is valid only if its 95% bootstrap interval lies above 5 and no replicate is at or below zero. The top-1 descriptive flags "I" or " I"; the log-odds metric uses the token "I" alone. The primary outcome is the reply mask, judged per cell (dataset by direction). First, a donor-set gate: the 95% interval of H_{28} must lie above 0.5, since at layer 28 the reply swap returns the donor's own output by construction. A cell passing the gate is *mid-depth* if the lower band exceeds 0.5 from some layer at or before 19 through layer 28; *deep* if that run starts at layer 20-23; *late* if it starts at 24-27, with the upper band below 0.5 through layer 23; *mechanical endpoint only* if only layer 28 qualifies and the upper band is below 0.5 through layer 27; and *indeterminate* otherwise.

CTL004. Recipients, donors and checks are CTL003′s. At every layer, five masks replace the recipient's block output with the donor's: R, S and U as in CTL003, and the joint masks SR (primary) and UR (secondary). A joint mask is patched in one pass, with all its positions taken from the same unpatched donor pass. The single masks reproduce CTL003 bitwise. For recipient class k, mask m and donor class c, $A(m, c)$ is the mean change in the log-odds of "I" from the unpatched recipient, divided by G, and $H(m) = A(m, \text{harmful}) - A(m, \text{harmless})$, which for single masks is CTL003′s. The joint increments are $J_{SR} = H(SR) - H(R)$ and $J_{SS} = H(SR) - H(S)$. The control $C = A(SR, \text{harmful})$ is the same-class shift of harmful recipients under the joint mask. Decision rules (removal, per dataset):

- *Primary A:* CTL003′s cell rule on the SR curve, with CTL003′s resampling stream and pair order.
- *Primary B:* a material positive increment if the lower bound of $W(J_{SR})$ exceeds 0.1, no material positive increment if its upper bound is below 0.1, and indeterminate otherwise. The rule is one-sided, not an equivalence test.
- *Control flag:* stable if the interval of $W(C)$ lies within ± 0.1, a shift if it lies wholly beyond, which blocks a selective reading, and indeterminate otherwise. A stable control does not establish selectivity.
- *Redundancy-consistent:* the lower bound of $W(H(SR))$ above 0.5, the upper bounds of $W(H(R))$ and $W(H(S))$ below 0.5, both increments' lower bounds above 0.1, and a stable control.

The 12 window statistics of the two removal cells (family B) share simultaneous 95% intervals from a pair bootstrap (10,000 replicates) on independent dataset streams, with one set of pair weights per dataset across all families. The 6 user-plus-reply statistics (family U, secondary) and the 26 installation, interaction and direction statistics (family D, descriptive) have their own simultaneous intervals and decide nothing. No 95% claim spans Primary A's layer bands and family B.

CTL005. Slot 1 is CTL003′s donor; slots 2 and 3 come from a second least-use pass over the same pool and rule (same chat length, another family), frozen in a design-time file. Pairs whose members lack three candidates of both classes are excluded, which leaves 69 of 81 JBB and 159 of 181 XSTest pairs (138 and 318 recipients); the exclusions are not random, and JBB loses mostly long harmful prompts. For each recipient, layer and slot i, the template suffix and the reply position take their states from up to two donors, each from its own unpatched pass, in one patched pass. With h_i and b_i the i-th harmful and harmless donors, the conditions are same-donor joints (h_i, h_i) and (b_i, b_i), single masks, mixed joints (h_i, b_i) and (b_i, h_i), different-donor joints with the next slot, and the reverse pairs (h_3, h_2) and (b_3, b_2): 32 conditions at 11 layers (0, 13 and 15-22, 28), or 160,512 patched passes. Slot-1 same-donor joints and single masks reproduce CTL004 bitwise. $Z = (s - m_b)/G$, with m_b the mean unpatched harmless score and G the mean unpatched gap over a dataset's retained pairs. A cell $Z_k(x, y)$ averages recipient class k over its listed slots and layers 15-22; the primary statistics use only conditions CTL004 never ran. The coherence shifts

are $E_{h,k} = Z_k(h,h)_{=} - Z_k(h,h')_{23}$ and $E_{b,k}$ likewise, same-donor minus different-donor cells on donor slots 2 and 3 only; a positive E means that splitting the two states across two prompts lowers the cell. Decision rules, with margin 0.1, CTL004′s materiality margin:

- Δ: recipient-side invariant if the interval lies within ± 0.1, dependent if it lies wholly beyond, and indeterminate otherwise.
- I_k: a negative factorial interaction if the upper bound is below -0.1, positive if the lower bound exceeds 0.1, additive within margin if the interval lies within ± 0.1, and indeterminate otherwise.
- E: coherence-invariant within ± 0.1, a coherence shift, with its sign, wholly beyond, and indeterminate otherwise.
- Gates: G's lower percentile bound above 5, and, per dataset and recipient class, the same-donor contrast $W = Z_k(h,h)_{=} - Z_k(b,b)_{=}$ with a percentile interval above 0.5; otherwise that dataset's labels read "insufficient transferable contrast".

No composite label and no overall verdict is defined. The 16 primary statistics (family P) share simultaneous 95% intervals from a pair bootstrap (10,000 replicates) with one set of pair weights per dataset. Partner-origin differences (family O, 8) and descriptive cells and contrasts (family D, 48) have their own intervals and decide nothing. Inference is conditional on the frozen donors; the window was chosen from CTL003 and layer 13 from CTL004, and no interval corrects for those choices. At the design-time precision benchmark, JBB Δ_{hb} could not be labelled invariant at any estimate, so the Deltas are reported estimation-first.

CTL006. Harmful prompts are OR-Bench-Toxic (655) and harmless ones OR-Bench-Hard-1K (1,319 seemingly toxic prompts the benchmark labels safe). The benchmark's labels come from a majority vote of three LLM moderators, with some flagged prompts re-admitted after the moderators judged another model's answer to them safe; two experts reviewed Hard-1K only (2.9% flagged), and Hard-1K was selected for refusal by at least three large models, Llama-3-70b among the examples. Of the 564 pairs, 350 are of equal length and all are at most 31 tokens apart, with harmful members shorter on average. R1 is CTL001′s layer-28 contrast under the three head pairings, computed from CTL002′s stored reply-position states of both models (no new forward pass), with CTL001′s 0.1 nat floor, length adjustment and readout check. It is not prospective: CTL002 had captured these states and reported refusal-group summaries, and the own-pairing sign is expected. It is therefore reported as "recurs (non-prospective)". On the 515 pairs whose members have three equal-length donors of each class from other categories (1,030 recipients, 362,560 patched passes), R2 repeats CTL004′s family B, averaged over three donor slots, with a persistence family of $H(SR)$ over the 10 measured layers from 13; R3 repeats CTL005′s family P with its gates. R2 and R3 are replicated if the original label recurs, contradicted if an interval lies wholly on the wrong side of one of the original rule's conditions, and indeterminate otherwise; R1 recurs if all three pairings read a harm-specific decrease, and is contradicted if any reads an increase. The labelled sensitivity uses CTL002′s stored labels (345 R1 and 316 swap pairs).

Appendix C. Indeterminate and descriptive runs

Shared first token (OFF001)

A declared offline follow-up on CTL001′s states asks whether the reply-position geometry at layers 13-18 only reflects the token Instruct will output. Groups are defined by Instruct's output top-1 at the reply position. It places the centroid of harmless prompts whose top-1 is "I" (49 prompts) on the axis from harmless prompts whose top-1 is not "I" (0; 301, the declaration's "answered" group) to harmful prompts whose top-1 is "I" (1; 287), and measures the share of the remaining distance that post-training closes, $g = (r_I - r_B)/(1 - r_B)$, with r averaged over layers 13-18. Under the

declared thresholds the result is indeterminate: $g = 0.40$ (95% interval 0.15 to 0.51) pooled and 0.44 (0.25 to 0.54) on XSTest alone, with the same classes for standardized dimensions.

Harmless "I" prompts move about a third of the way toward harmful ones from layer 13 and stay there through layer 28, where both groups output "I" (Figure 6). The rest is not shared (descriptive). With 49 prompts per group, harmful "I" prompts take a smaller share of variance in Instruct (0.46-0.51 from layer 18) while harmless "I" prompts keep theirs (0.79-0.86). The readout commits to "I" at layer 19 for 84-90% of harmful prompts but for 22-37% of harmless "I" prompts until layer 24. Because harmless "I" prompts mix false refusals with compliant replies that open with "I", the test cannot separate the token from refusal.

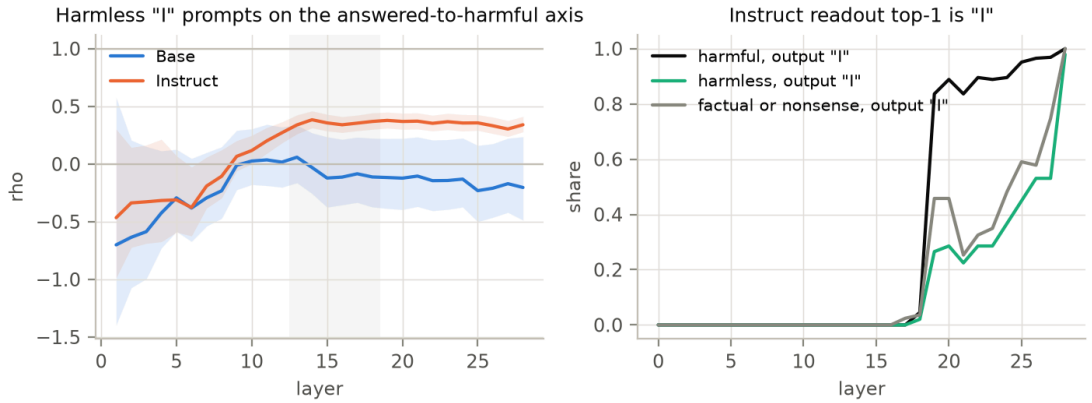

Figure 6: OFF001. Left: position of harmless prompts whose top-1 is "I" on the axis from harmless prompts whose top-1 is not "I" (0) to harmful "I" prompts (1), with 95% intervals; the gray band marks layers 13-18. Right: share of each group whose Instruct readout top-1 is "I".

Refusal labels and cross-model state swaps (CTL002)

CTL002 (research/runs/CTL002.yml) adds OR-Bench (Cui et al. 2025): 1,319 seemingly toxic prompts that the benchmark labels safe (*Hard*) and 655 toxic ones (*Toxic*), beside CTL001's JBB and XSTest prompts, all in the chat format.

Refusal labels (Arm A). Instruct generates at most 16 tokens greedily. A fixed list of 27 refusal openings, matched after lower-casing and stripping leading whitespace and markup, labels each reply prefix-positive (R) or prefix-negative (A). These are rule outcomes, not refusal and compliance. From harmful continuations only the label and the first token id are kept; the declared forward-pass states and logits stay in local files. A fresh judge session, blind to the rule, labels up to 60 sampled Hard replies per stratum as refused, complied or ambiguous; Arm A is decisive only if agreement reaches 90% in both strata. The statistic is OFF001's, within OR-Bench: the position of refused Hard prompts (RH) on the axis from answered Hard prompts (AH, 0) to refused Toxic prompts (RT, 1), with centroids balanced over OR-Bench categories, and the share g of the remaining distance that post-training closes over layers 13-18. "High" or "low conditional convergence" requires the g interval above 0.5 or below 0.25, for raw and standardized states alike. A verdict also requires at least 100 prompts in each group and at least five categories with 10 or more prompts in each; otherwise Arm A is descriptive.

State swaps (Arm B). For the 300 JBB and XSTest pairs, one model's residual after block $\ell \in \{8, 13, 16, 19, 23, 28\}$ replaces the other's at every position, and the recipient computes the rest. B1 puts Base states into Instruct (removal); B2 puts Instruct states into Base (installation). The outcome is the log-odds of "I" at the recipient's true output. With Δs the patched minus the unpatched value, R_ℓ is the mean over pairs of $\Delta s(\text{harmful}) - \Delta s(\text{harmless})$, and $q_\ell = R_\ell / R_{28}$

is the share of the full swap's effect reached by swapping at layer ℓ. At $\ell = 28$ the swap equals CTL001′s cross-readout. Swaps at the reply position only, and a within-Instruct swap of reply-position states between the two prompts of a matched pair (B3), are descriptive. A layer qualifies if the lower simultaneous band of q_ℓ exceeds 0.5 there and at every deeper sampled layer, in both directions. The outcome is *mid-depth transfer* if the earliest qualifying layer is at or before 19 in both datasets, *deep transfer* if it is 23, and *late transfer* if the upper band stays below 0.5 through layer 23 in both directions and datasets.

Refusal labels, results. Instruct's reply is prefix-positive for 95 of 1,319 Hard prompts (7%), 450 of 655 Toxic prompts (69%), 90 of 100 JBB harmful, 5 of 100 JBB benign, 168 of 200 XSTest unsafe and 14 of 250 XSTest safe prompts. The judge agreed with the rule on all 60 sampled prefix-positive Hard replies (11 flagged as partial refusals) and on 55 of 60 prefix-negative ones; the other five it called ambiguous, none refused. The validation passed. It covers Hard prompts only: Toxic continuations were not retained, so prefix-negative Toxic replies are not read as compliance. The judge was blind to the rule but, like GEM002′s, not to the project: as context for the subagent, the harness supplied the project's instructions, which name its subject, and its memory index, found from the retained transcript during the review of GEM003. The effect on the labels cannot be assessed.

Arm A (descriptive). The decision rule's sample floors were not met. The 95 refused Hard prompts are spread thinly: six of ten OR-Bench categories have fewer than 10, so only four qualify (illegal, privacy, self-harm, unethical), and within them RH has 70 prompts. The estimate would have been indeterminate anyway: $g = 0.17$ (95% interval -0.07 to 0.33) for raw states and 0.17 (-0.05 to 0.32) standardized. The declared descriptive variants do not change this: residualizing on length gives 0.17 (-0.19 to 0.38); plain centroids give 0.32 (0.20 to 0.40), larger under different category selection and weighting; and the JBB-XSTest consistency check gives 0.51 (0.23 to 0.61) on only 19 refused harmless prompts. Over layers 13-18, refused Hard prompts sit 0.40 of the way from answered Hard to refused Toxic prompts in Instruct and 0.28 in Base (Figure 7). In Instruct's logit lens, "I" again becomes top-1 at layer 19, for 96% of RT, 68% of RH and 20% of AH prompts, against 8%, 5% and 1% at layer 18. In the descriptive decoding report, OR-Bench categories are more decodable in Instruct than in Base for both refused Toxic and answered Hard prompts (8 categories, chance 0.125; RT 0.56 to 0.68, AH 0.45 to 0.53); RH could not be compared, sharing only one category with AH.

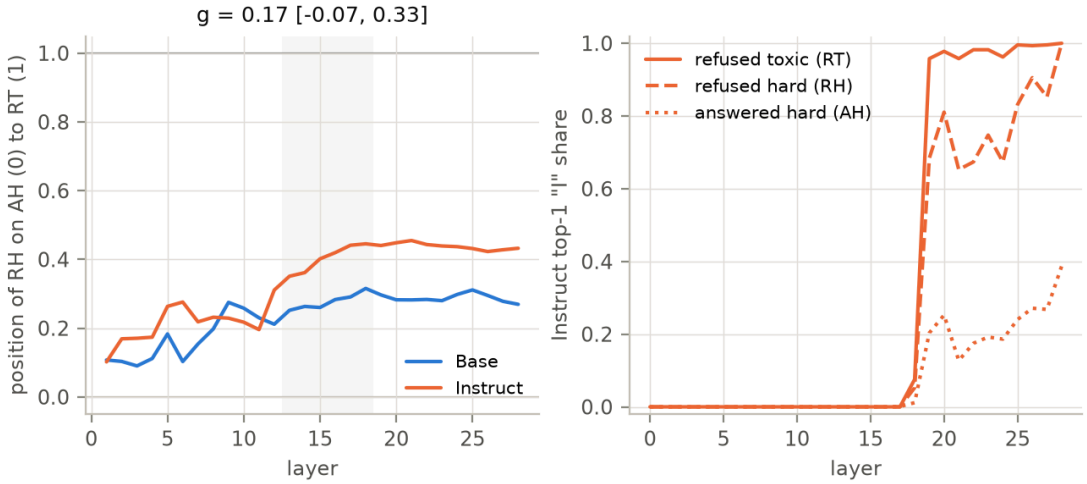

Figure 7: CTL002 Arm A, descriptive. Left: position of rule-refused Hard prompts on the axis from rule-answered Hard (0) to rule-refused Toxic prompts (1), category-balanced centroids, raw states; the gray band marks layers 13-18. Right: share of each group whose Instruct readout top-1 is "I".

Arm B. Under the declared rule the outcome is indeterminate in both datasets (Figure 8). Base states swapped into Instruct remove the harm-specific log-odds of "I" from layer 8 on ($q_8 = 0.83$; 0.98 to 1.00 from layer 13; lower bands at least 0.79), so every sampled layer below 28 is removal-only. The late-transfer condition fails on these bands, and marginally on installation's upper band at layer 23 on JBB (0.503). Instruct states swapped into Base install it gradually: q_ℓ is 0.03 to 0.07 at layers 8-13, 0.19-0.20 at 16, 0.31-0.32 at 19 and 0.44-0.47 at 23, and no layer below 28 qualifies. The full swap matches OFF002's calibration: $R_{28} = -17.5$ (JBB) and -18.6 (XSTest) for removal, $+15.7$ and $+17.7$ for installation.

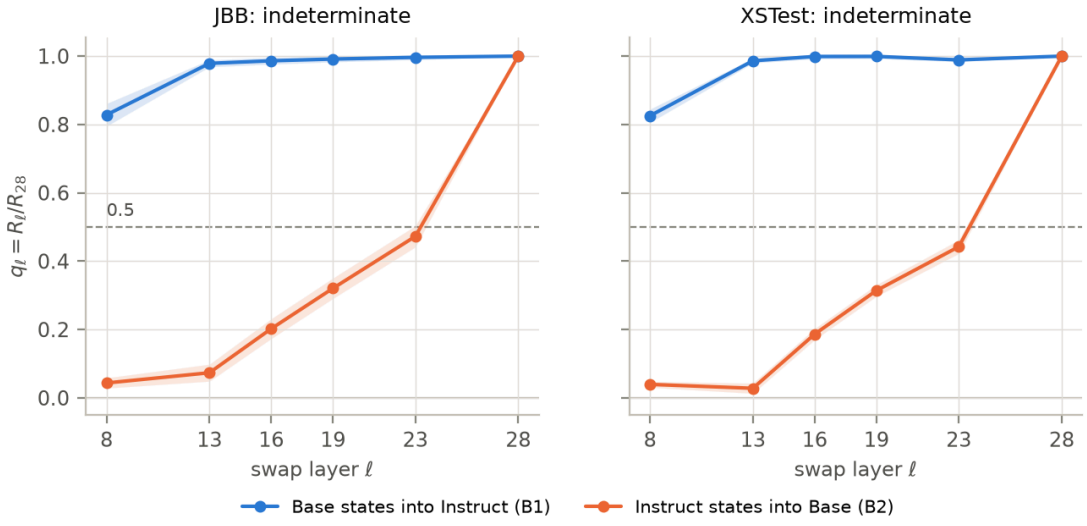

Figure 8: CTL002 Arm B. Share q_ℓ of the full swap's harm-specific effect on the log-odds of "I" reached by swapping states between Base and Instruct at layer ℓ, with simultaneous 95% bands; the dashed line marks 0.5. Titles give the declared outcome.

Exploratory readouts, made after the declared results were seen, qualify both directions. Removal from layer 13 is not a return to Base behaviour. With Base states at all positions at the sampled layers 13-23, Instruct predicts "assistant", the role name that follows a header-start token in the

template, for 74-100% of harmless prompts, matching neither model's own next token; at layer 28 it does not. Harmful and harmless prompts converge on the same log-odds of "I": the gap is 0.1 to 0.8, against 17.6 to 19.1 unswapped, with similar KL. Harmful top-1 ids are masked, so their predicted token is unknown. Swapping Base states at the reply position alone never yields "assistant" on harmless prompts and leaves "I" top-1 for 98-99.5% of harmful prompts through layer 19, yet removes 0.24-0.27 of the log-odds effect at layer 8, 0.36-0.40 at 13, 0.51-0.65 at 16 and 0.64-0.76 at 19, a gradual profile that rises earlier than installation. These readouts suggest broad disruption involving non-reply positions. They do not isolate the template positions from the user-text positions, whose Base states also attend to untrained template tokens, or from general cross-model incompatibility, and they do not locate the harm-specific signal. At layer 8 the replies are still ordinary but partly disturbed: harmless top-1 matches Instruct's own for 37-44% of prompts.

Installation changes "I" more on harmful than on matched harmless prompts, while also substantially changing harmless outputs (median KL 6.0-6.1 nats at layer 19): Base's own top-1 on harmless prompts continues the user's turn ("What", "How"), and with Instruct's states it mostly opens a reply instead. With Instruct's states at layer 19, Base's top-1 becomes "I" for 69-80% of harmful and 10-13% of harmless prompts; at layer 16, for 28-29.5% and 1-6%. Normalized on the probability scale instead of the log-odds, the ratio is 0.46-0.53 at layer 19 and 0.76-0.77 at 23. These are different statistics, partly because the log-odds keep growing past probability saturation at layer 28, and neither is a share of refusal transferred; the declared verdict uses the log-odds. Swapping only the reply-position state gives ratios of 0.17-0.20 at layer 16, 0.27-0.29 at 19 and 0.34-0.39 at 23, close to the all-position values, although similar totals do not decompose the contributions of positions.

The within-model swap (B3, descriptive) avoids Base's template states. It swaps reply-position states between the two prompts of a matched pair in Instruct. For a harmless prompt given its harmful partner's state, the mean change in log-odds is 45-51% of its own layer-28 change at layer 13 and 73-81% at 16; in the reverse direction it is about 0% at layer 13 and 25-39% at 16. The asymmetry is compatible with later layers reusing the recipient's context but does not show recovery of harm information. Content changes with harmfulness, and B3 has no same-class swaps to control a nonspecific push toward "I"; CTL003 includes them.

CTL002 therefore did not place the harm-specific log-odds of "I" at one depth. Removal through Base states at all positions coincides with broad disruption from layer 13. Installation into Base grows over layers 16-23; at layer 19, where Instruct's logit lens first makes "I" top-1, Instruct's states make Base output "I" for most harmful prompts and few harmless ones. Cross-model incompatibility, lexical content, a generic first-token bias and the recipient's context remain rival accounts. A calibrated lens (Belrose et al. 2023) remains deferred.

Appendix D. Further results

Readout, depth and spread in Llama (CTL001)

With both models' states read through the Base head, $D_{28} = -0.70$ (JBB) and -0.73 (XSTest); through the Instruct head, -0.51 and -0.66 (Figure 10). In an exploratory decomposition, on JBB the contrast comes almost entirely from Instruct, whose harmful-minus-harmless entropy difference is -0.45 against Base's $+0.01$; on XSTest both sides contribute (-0.37 and $+0.26$). At the chat reply position, Base under the untrained template predicts the start of another user turn ("How", "What", "Write") with a median final entropy of 8.4 nats; Instruct's is 0.08.

On XSTest the simultaneous band of D_ℓ excludes zero at every layer from 19 to 23 (-1.75 to -4.97) and includes it at layers 24-27. On JBB it alternates: positive at layers 17-18 ($+0.47$, $+0.70$), negative at 20 and 22 (-2.49, -2.89) and positive at 24 ($+2.27$). The negative excursion at layers

20 and 22 also appears under both fixed heads. Smaller negative exclusions, mostly below the 0.1-nat floor, occur at layers 5-15 in both datasets. Under the dialogue frame XSTest replicates ($D_{28} = -1.35$, locus at layer 25) but JBB does not (-0.26, -0.56 to 0.05). At the last instruction token, JBB shows no divergence ($D_{28} = 0.09$, -0.06 to 0.24); XSTest's (-0.48) vanishes when entropy is renormalized over regular tokens (-0.20, -0.59 to 0.18), and is readout-only.

In the exploratory first-token readout, the share of "I" as top-1 is 0% (JBB harmful) and 7% (XSTest unsafe) at layer 18 and stays at 80-97% (JBB) and 75-96% (XSTest) through layer 28; Base predicts "I" for at most 1% of prompts, at layers 27-28 only (Figure 11). Instruct's output begins with "I" for 96% of XSTest unsafe and 14% of XSTest safe prompts.

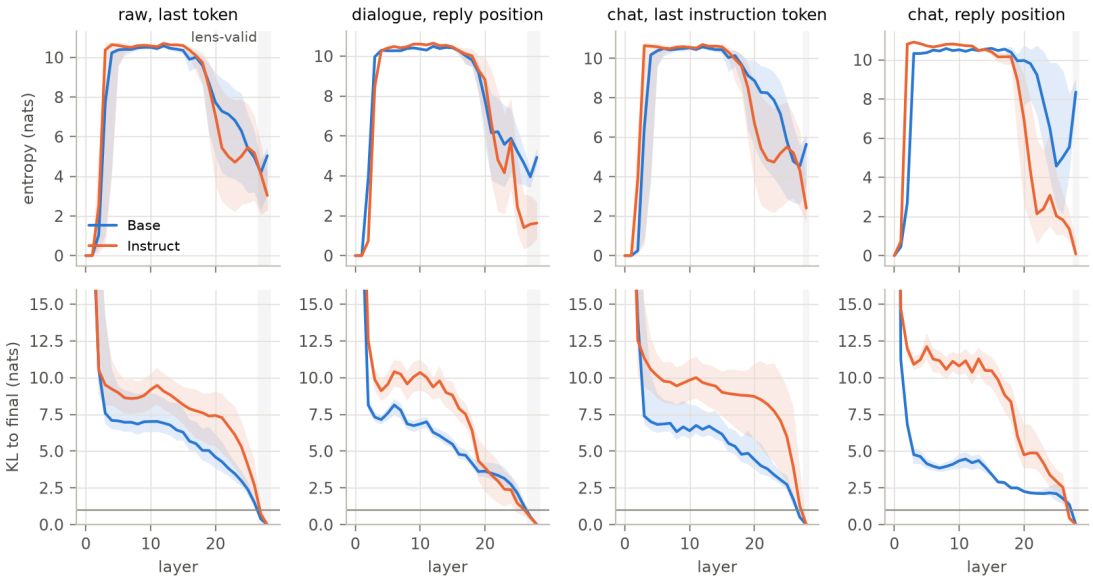

Figure 9: Median entropy (top) and KL to the final distribution (bottom) per layer, with interquartile ranges, over all 850 prompts. The gray band marks final-consistent layers; the line marks 1 nat. KL at layers 0-2 lies above the axis.

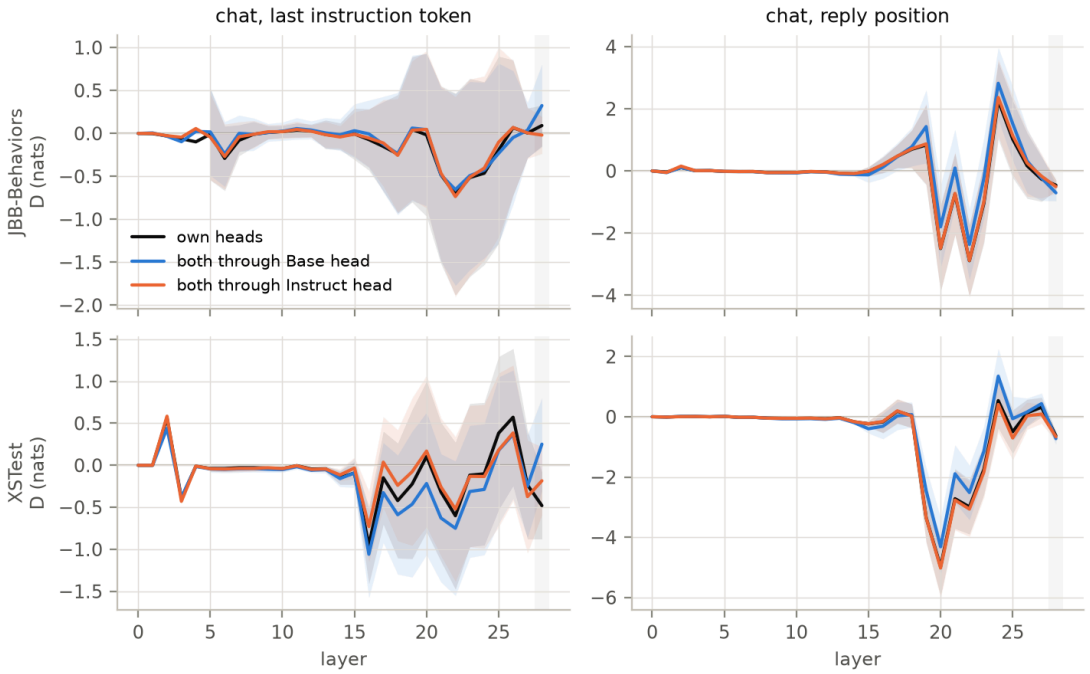

Figure 10: Llama 3.2 3B: harm-specific contrast under own heads and with both models' states read through the Base head or the Instruct head.

In both models and every format, median entropy sits near 10.5 nats, close to the uniform bound of 11.76, from the first non-echo layers to about layer 17, and falls over the last third (Figure 9). It is not monotone: the median rises between consecutive non-echo layers 6 to 12 times per profile. The readout agrees with the final distribution (median KL at most 1 nat in both models) only at layers 27-28 in the raw and dialogue formats and at layer 28 in the chat format. Below that, all readout statements are about the projection, and a final-consistent locus is unidentifiable everywhere. Base at the chat reply position falls to 4.6 nats at layer 25 and returns to 8.4 at layer 28. In the raw and dialogue formats Base's median final entropy is 5.0 and 4.9 nats, and the drop is 0.1 to 2.0 nats (nonsense rises by 0.8) and 2.4 to 3.8 nats. As a determinacy check, Base and Instruct complete 57 and 64 of the 100 factual cloze prompts with the attribute's first token in raw text (51 by both). On JBB, the positive excursion of D_ℓ at layer 24 coincides with benign prompts settling on the formatting token "**" (top-1 for 26% at layer 24, 1-7% at layers 20-23); XSTest safe prompts never do, so the return to zero there has another source.



Figure 11: Exploratory. Share of prompts whose top-1 readout at the chat reply position is "I", per layer.

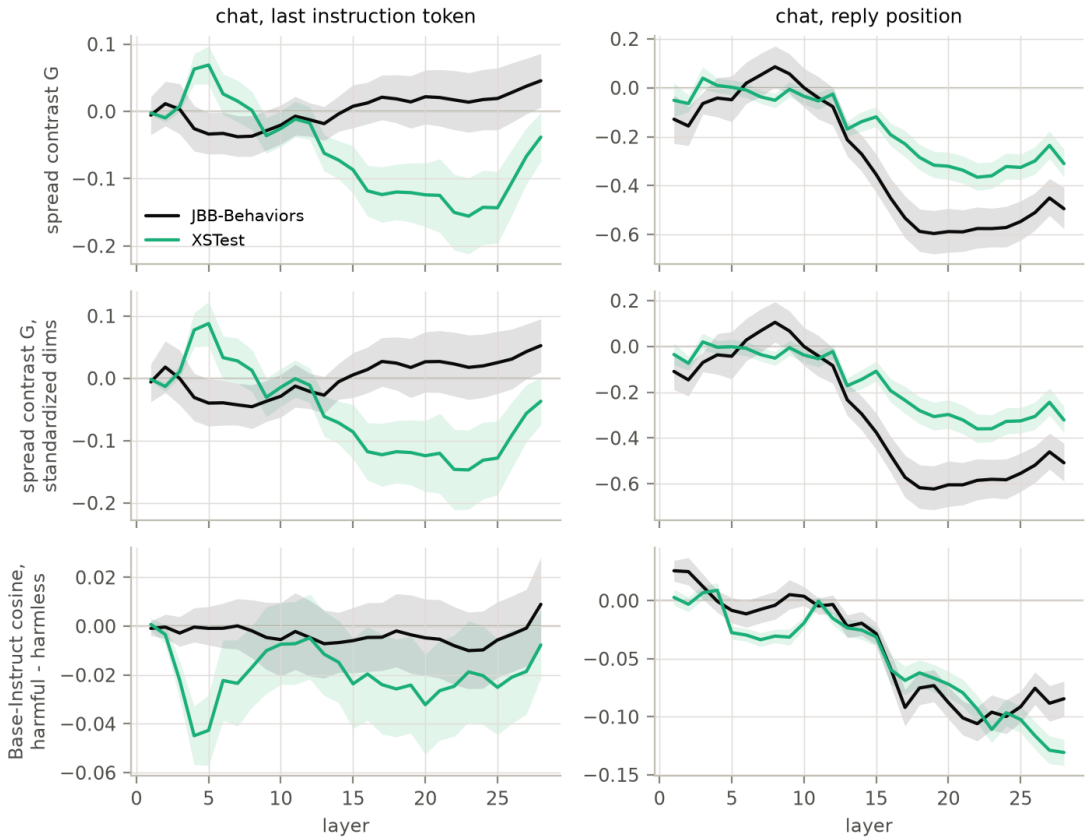

Figure 12: Spread contrast G_ℓ (raw and standardized) and the harmful-minus-harmless difference in centered Base-Instruct cosine, with pointwise 95% intervals.

CTL001′s pointwise intervals of G_ℓ at the reply position exclude zero at every layer from 13 to 28 in both datasets, raw and standardized: on JBB from -0.21 at layer 13 to about -0.60 at layers 18-24, on XSTest between -0.12 and -0.37 (Figure 12). Before layer 13 a few isolated layers also exclude zero, in both directions. The centered Base-Instruct cosine is lower for harmful than for harmless prompts at every layer from 13 (JBB) and 12 (XSTest). At the last instruction token, G is near zero on JBB and negative on XSTest (-0.04 to -0.16 at layers 13-28). No geometry layer was excluded beyond layer 0. The contrast has different sources in the two datasets: harmless prompts take a larger share of state variance in Instruct than in Base in both, while harmful prompts lose share only on JBB.

From layer 16 on, JBB harmful prompts take a smaller share of state variance in Instruct than in Base (0.40-0.49 vs 0.72-0.85), and JBB benign prompts a larger one (0.88-0.97 vs 0.68-0.80). XSTest unsafe prompts keep their share (0.49-0.56 vs 0.47-0.54); XSTest safe prompts take a larger one (0.92-1.06 vs 0.66-0.76). A share above 1 means that a category varies more than the 850 prompts pooled, which happens when other categories cluster tightly, so a larger share is not by itself expansion (Figure 13). An exploratory split by the first-token proxy keeps the picture: the share rises for harmless prompts whose top-1 is not “I” (JBB 0.89-0.98 vs 0.66-0.79; XSTest 0.91-1.07 vs 0.64-0.75), while prompts whose top-1 is “I” lose share only on JBB.

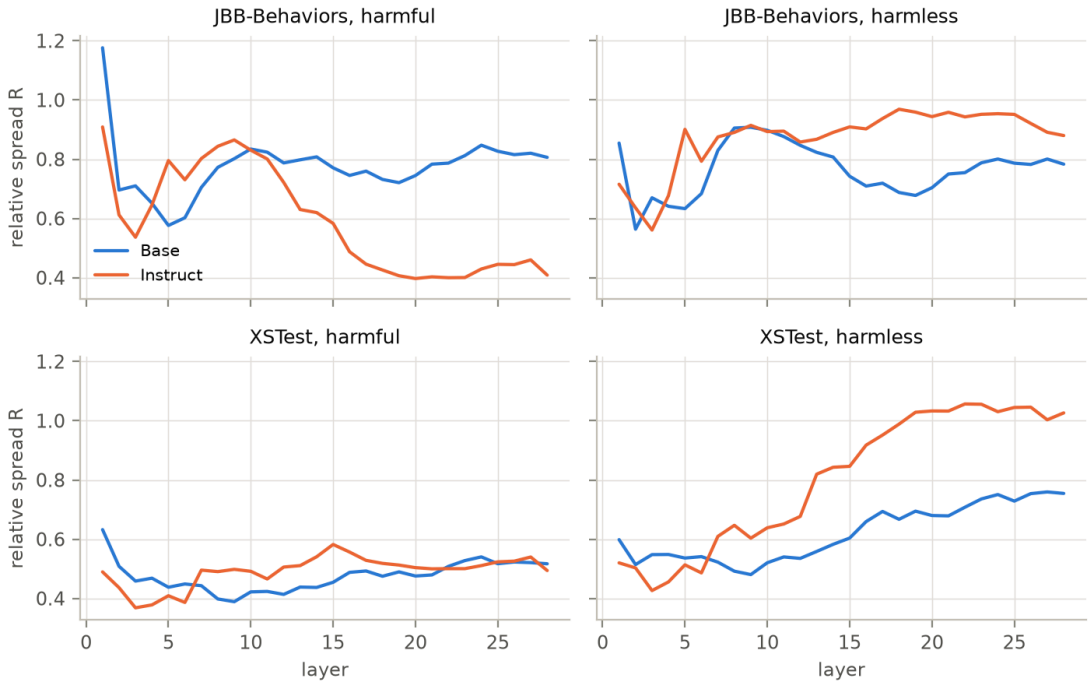

Figure 13: Relative spread of each category at the chat reply position, Base and Instruct. Descriptive.

Three descriptive measures complete the picture. Matrix-based entropy of unit-norm states (100 points, maximum 4.61) is 0.1 to 0.4 lower in Instruct than in Base at layers 8-12 in every category, controls included, and within 0.2 of Base from layer 15, except for XSTest safe prompts, which stay 0.16 to 0.39 higher. JBB harmful prompts keep Base's matrix entropy while their share of variance halves; a share cannot establish absolute scale. TwoNN estimates at equal N are lower in Instruct than in Base at layers 8-22 for JBB and XSTest-safe prompts, and lower or similar for XSTest-unsafe; at 100 points they are biased, so only the comparison is read. Along a split-half difference-of-means direction, harmful and harmless states differ by a Cohen's d of 1.1 to 1.4 in Base from layer 9 and about 4 in Instruct from layer 16, in both datasets, and by about 2.5 against 1.25 at layer 10 (exploratory). Prompts whose output top-1 is "I" are more compact than the rest within Instruct, but XSTest's unsafe prompts are already the more compact set in Base. In OFF001, post-training moves harmless prompts whose top-1 is "I" about 40% of the way toward harmful ones from layer 13, and, descriptively, they neither lose share of variance nor commit early; CTL002′s rule-labelled refusals were too few on borderline prompts to separate harm, refusal and the first token.

Second model descriptives (GEM001, OFF003)

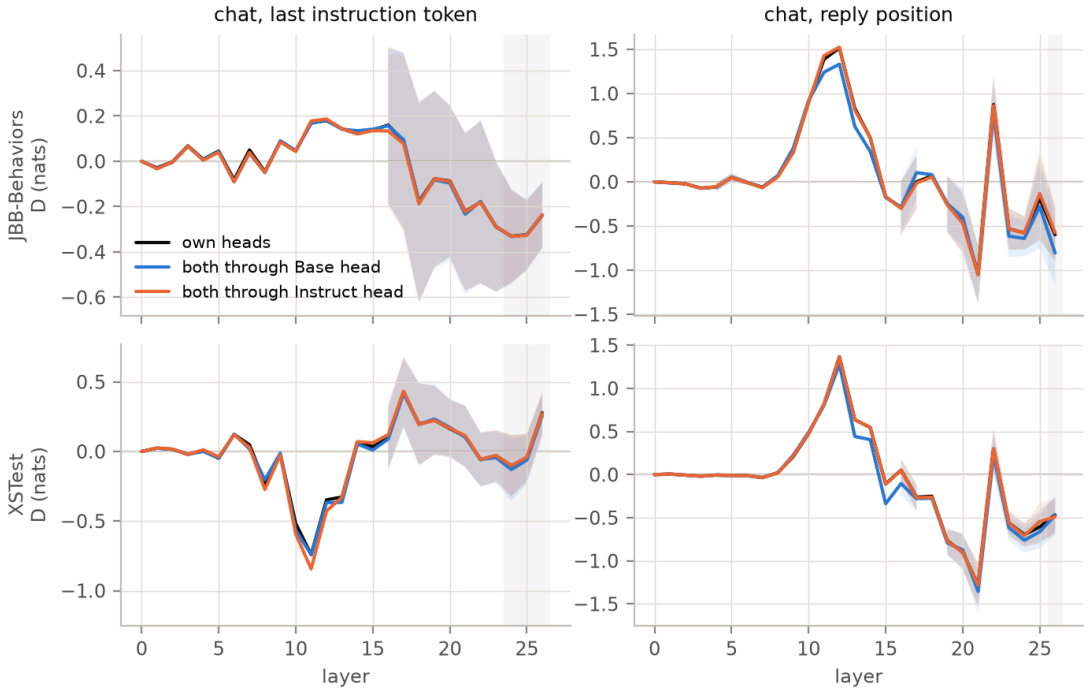

Figure 14: Gemma 2 2B: harm-specific contrast under own heads and with both models' states read through the Base head or the Instruct head, as Figure 10.

At the reply position, Gemma's post-training lowers final-layer entropy by 4.6 to 5.6 nats in every category, controls included (category means of per-prompt differences); the matched subtraction removes this shared shift. With fixed heads the reply-position contrast keeps its sign (Figure 14). Spread shares at the reply position fall for JBB harmful prompts and rise for XSTest safe prompts, as in Llama; unlike Llama, Gemma's JBB benign share barely moves. At the last instruction token G is near zero on JBB and negative on XSTest from layer 20, raw and standardized. Under the dialogue frame, with its locus at layer 19, D_{26} is -1.224 (JBB) and -0.932 (XSTest). The two softcap-sensitive labels are JBB's dialogue depth and XSTest's state-carried reading at the last instruction token; without the softcap, the reply-position D_{26} is -0.433 and -0.370, with the same labels. The last-instruction-token verdict is declared, the raw-text one descriptive.

Table 6: OFF003 floor characterization at the reply position. Rank gaps are harmful-minus-harmless mean ranks within a dataset, divided by $2n$; the slope is Instruct's harmful-minus-harmless gap regressed on Base's: Instruct's gap tracks Base's weakly or not at all. Llama is the parity arm.

Model, dataset	Base rank gap	Instruct rank gap	F3 slope	Margin contrast (log-odds)
Gemma, JBB	-0.089	-0.429	0.000	3.961
Gemma, XSTest	-0.089	-0.307	-0.212	1.587
Llama, JBB	0.013	-0.365	0.216	5.061
Llama, XSTest	0.126	-0.330	0.245	4.713

In absolute terms, unit-norm spread rises from Base to Instruct for most paired categories; JBB harmful prompts in Gemma contract at layer 26 (-0.069, -0.079 to -0.058). Held-out probes separate harmful from harmless prompts far more in Instruct than in Base states: at Gemma layer 26, Cohen's d 3.6 against 1.0 on JBB and 4.7, from 4.1 to 5.4 over fold assignments, against 0.9 on XSTest; at Llama layer 28, 4.5 against 1.4 and 4.6 against 1.1. In Gemma the Instruct direction has cosine 0.15 and 0.14 with Base's class-mean direction and 0.99 and 1.00 with the post-training shift; Base's direction still scores Instruct states at AUROC 0.85 and 0.80, and the Instruct direction transfers across the two datasets (AUROC 0.98 and 0.98, against Base's 0.69 and 0.71). These labels confound harm with refusal.

Linear CKA between Base and Instruct states is lower for harmful than for harmless prompts on Llama's JBB from layer 5 through 28 (0.185 at layer 13, label-swap null share 0.005) and on Gemma's JBB at layers 4-15 (0.107 at layer 14, 0.010) but not from 16 (0.003 at 26); on XSTest it is near zero at most layers in both models. Probe directions fitted on Base states score the post-training shift at or below chance (0.42 and 0.47). On XSTest, Cohen's d depends on the fold assignment: the means over the 20 declared assignments sit about two standard errors above their means over all grouped partitions (independent recomputation); AUROC changes little.

Phrase score and first token in Gemma (GEM002 to GEM004)

Against GEM002′s phrase rule, "I" has pooled precision 0.949 and recall 0.974; against the prefix rule, precision 0.676. In exploratory counts on design-selected harmless cases, not declared, GEM002′s judge called every judged phrase-rule refusal refused (47 of 47).

In GEM003, by category (descriptive), the precision of "I" is lowest on privacy prompts, 0.79 (34 of 43). Of the 231 judged refusals, 156 have $S < 0$, 141 of them opening with "I"; 116 open outside the phrase set, and 40 of the 115 that open with a phrase still have $S < 0$, because greedy decoding can follow a phrase whose total probability is below one half. Under ambiguous-as-complied, the AUROC bound of S is 0.84989 with the repaired score and 0.84992 with GEM003′s; under ambiguous-as-refused, the recall bound of "I" is 0.854.

On JBB harmful prompts Instruct's S is positive for 96 of 100, while all 73 XSTest unsafe replies that open with "I" and that only the phrase rule calls refusals (first 32 tokens) have $S < 0$, as do all 247 such replies in the pool; whether either endpoint marks refusal on harmful prompts is not tested.

Reply mask and other masks (CTL003)

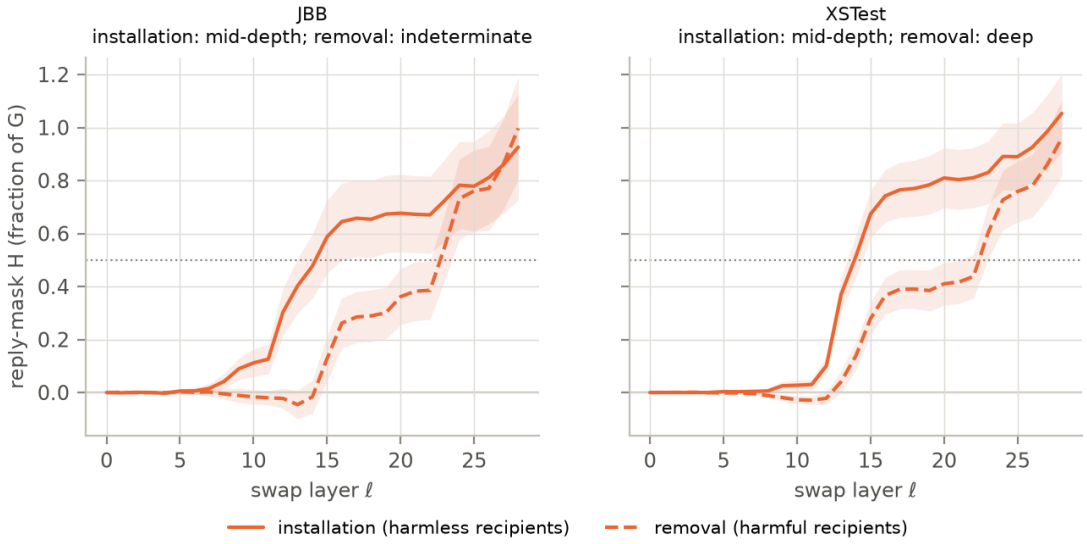

Figure 15: CTL003, reply mask. Donor-class contrast H_ℓ (harmful-donor minus harmless-donor log-odds of "I", as a fraction of the unpatched gap G) by swap layer, with simultaneous 95% bands; the dotted line marks 0.5. At layer 28 the swap returns the donor's own output by construction. Titles give the declared labels.

On XSTest the lower band of the reply-mask installation curve exceeds 0.5 from layer 15 (0.59 to 0.70 through layer 22); on JBB it does so from layer 16, narrowly (0.5005 at 16, then 0.51 to 0.53 through 22). In an exploratory check with 20 other bootstrap seeds, the JBB onset moves to layer 17 in 7 of them; every cell label stays the same. JBB removal's sustained run starts at layer 24. At layer 15, a harmful donor's reply state makes 86-92% of harmless recipients output "I", against 13-14% with a harmless donor, while a harmless donor's reply state leaves "I" in place for 81-90% of harmful recipients.

The same-class swaps are the controls inside every contrast; a harmful donor's reply state moves a harmful recipient's log-odds by at most about 0.05 of G, which may reflect saturation of the output proxy or restoration from other positions. XSTest removal also shows small reversals at layers 8-11, about -0.01 to -0.03 of G. The probability-scale ratio at layer 15 is 0.83-0.93 for installation and 0.14-0.26 for removal. Removing any single donor moves H at layers 8, 13, 16, 19 and 23 by at most 0.03 from its full-sample value. The values are consistent with CTL002′s B3, whose donors and normalization differ.

Compared with the unpatched recipient (exploratory, after the results), at layers 13-23, a harmless donor's reply state raises a harmless recipient's log-odds of "I" by 0.04-0.15 of G, while its top-1 "I" rate stays close to unpatched at layer 15 (13.6% against 13.6% on JBB, 12.7% against 14.9% on XSTest); cross-class swaps move them by 0.45-0.88 of G upward and by up to 0.60 downward. The small harmful same-class shift does not show that harmful donors' reply states are alike.

Other masks (descriptive). At the embedding layer the user-text swap runs the donor's prompt, so values near 1 at early layers are expected. Its effect falls over layers 7-14 and is near zero from layer 15 (Figure 16). The four template tokens after the user text change the output most at layers 12-13 (0.73-0.89 for installation, 0.28-0.38 for removal), then decline to about 0.05-0.25 at layers 24-27. Both masks are zero at layer 28 by construction. The masks are not additive: at layers 12-14

the three installation curves sum to 1.3-1.5. These profiles show where a swap changes the output, not where information moves.

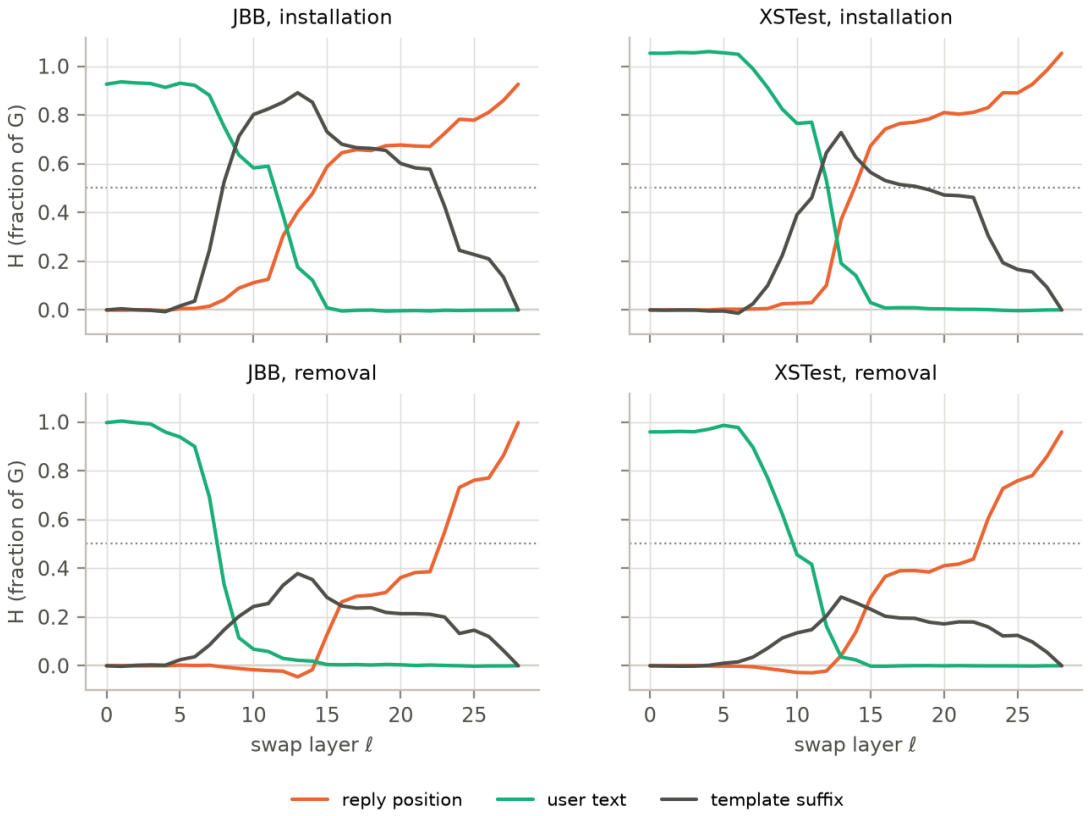

Figure 16: CTL003, all masks, descriptive. H_ℓ for swaps of the reply position, the user text and the template suffix. User text at layer 0 equals the donor's output, and user text and suffix are zero at layer 28, by construction.

Several nonexclusive accounts fit the asymmetry; CTL003 tests none of them:

- **Redundancy.** Harmful prompts may carry the contrast redundantly, so their other positions restore it after a reply-only swap; between layers 15 and 22 no single mask removes more than about 0.45 of it. CTL004's joint mask gives a redundancy-consistent result, but nonlinear interactions can produce the same pattern.
- **A boundary effect.** Matched harmless prompts, built to resemble harmful ones, may sit near an output-token refusal boundary, so that one reply state suffices to install it. CTL003 does not measure boundary proximity.
- **Donor-state homogeneity.** Harmful donors' reply states may be alike while harmless donors' states vary; one donor per class and recipient cannot test this.
- **Class-dependent compatibility.** A harmless state inserted into a harmful context may be overwritten or ignored more strongly than the reverse, for example through norm or attention differences. Instruct's harmful reply states have about 3-17% smaller median norms than harmless ones from layer 17 (CTL001, exploratory).

CTL004 probes the first account; its result does not separate it from the fourth, and the others remain untested.

Secondary and descriptive results (CTL004)

Same-class reply swaps show a similar row spread (0.14-0.15 G). The user text adds little to the reply (Figure 17): $W(J_{UR})$ is 0.05 and 0.08 (no material increment on JBB, indeterminate on XSTest by the declared threshold). On the normalized log-odds of "I", the joint effect is superadditive for removal ($W(N_{SR}) = +0.42$ and $+0.28$, with $N_{SR} = H(SR) - H(S) - H(R)$) and subadditive for installation (-0.40 and -0.29). Installation under the joint mask is also mid-depth, from layer 9 on JBB and 12 on XSTest. The direction difference in $W(H(SR))$, installation minus removal, is -0.05 (-0.25 to 0.15) and $+0.13$ (-0.02 to 0.27). It is smaller than under single masks, and its descriptive interval includes zero, which is not equivalence. At layer 15, a harmless donor's suffix and reply states leave "I" top-1 for 15% (JBB) and 20% (XSTest) of harmful recipients, against 90% and 81% with its reply state alone.

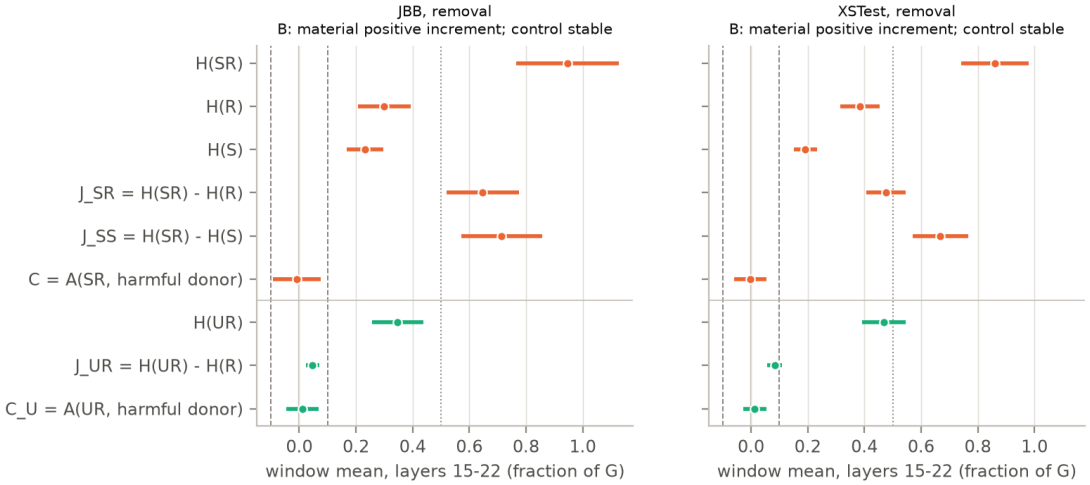

Figure 17: CTL004, removal. Window means over layers 15-22 with simultaneous 95% intervals, for the primary statistics (orange) and the secondary user+reply statistics (green). Dashed lines mark ± 0.1; the dotted line marks 0.5.

Cells and descriptives (CTL005)

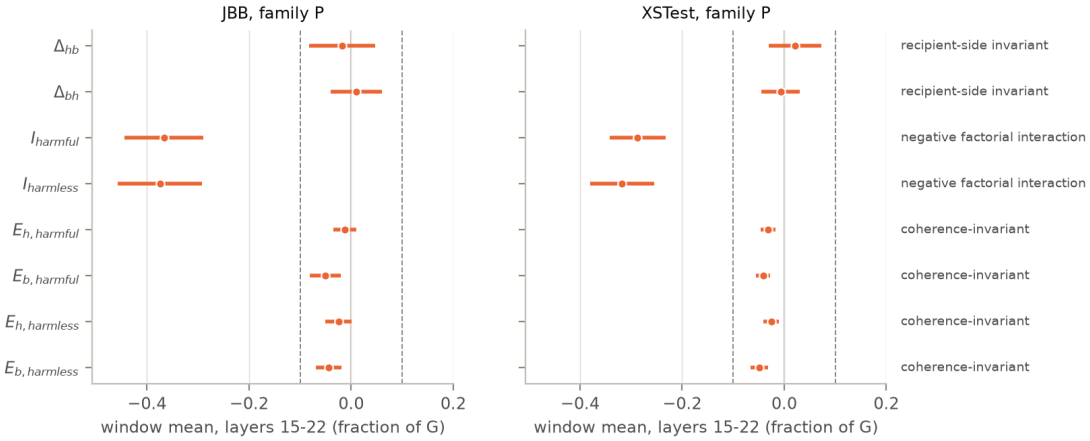

Figure 18: CTL005, family P. Window means over layers 15-22 with simultaneous 95% intervals for the recipient-side differences Δ, the factorial interactions I_k and the coherence shifts E. Dashed lines mark ± 0.1. The labels on the right hold in both datasets.

The coherence interval bounds run from -0.081 to 0.010 on JBB and from -0.066 to -0.010 on XSTest. The slot-1 cells reproduced the design file's composition table. Family P's intervals are simultaneous over its 16 statistics; the gates and families O and D have separate coverage, and no 95% claim spans them. The precision was better than the design benchmark: the half-width of JBB Δ_{hb} is 0.065 against the benchmark's 0.128, which had made that label look unreachable. This is consistent with averaging over three donors and foreign partners, which the design file allowed for; the source of the narrowing was not isolated. Families D and O decide nothing; the rest are point estimates (Figure 19).

- *Cells.* The mixed cells sit between the same-class cells: $Z(h, b)$ from 0.605 to 0.721 and $Z(b, h)$ from 0.777 to 0.842 across datasets and recipient classes, against two-donor contrasts of 0.900 to 0.944. In the window the harmful-reply cell (b, h) lies above the harmful-suffix cell (h, b) in all four. Every single-position sufficiency difference, $Z(h, h')$ minus a mixed cell, lies between 0.195 and 0.432, so neither single harmful position reaches the two-donor cell. On the probability scale the interaction keeps its sign (-0.52, -0.48, -0.39 and -0.38).
- *Coherence.* All eight E estimates are negative, and six of the eight intervals exclude zero (both E_b terms on JBB and all four on XSTest): cells built from two prompts sit slightly higher than cells built from one. The difference between the two donor classes' shifts stays between 0.010 and 0.039, every interval within ± 0.1, so the shift does not differ materially between donor classes.
- *Partner origin.* The own-partner minus foreign-partner difference (family O) lies within ± 0.1 for six of eight compositions. For harmless recipients' (h, b) it is -0.045 (-0.110 to 0.020) on JBB and -0.073 (-0.136 to -0.009) on XSTest. Origin equivalence was not established in either dataset: the own-partner cell sits lower, XSTest's interval excludes zero, and neither interval establishes a difference beyond 0.1 G.
- *Donors.* Donor-level checks agree with the means. Per-slot Δ ranges from -0.045 to 0.051; leaving one donor out keeps every Δ within -0.032 to 0.032 and every I_k within -0.394 to -0.268; the length-stratified Δ lies between -0.006 and 0.005.
- *Rows (interpretation restriction).* Individual rows vary widely under same-class swaps. With same-donor states from slots 2 and 3, the mean absolute row change is 0.41 and 0.45 G for harmless recipients, with 88% and 89% of rows moving more than 0.1 G, and 0.15 and 0.16 G for harmful recipients (54%), while the class means move 0.02 to 0.06. The labels concern class means, not individual recipients.
- *Layer 13 (descriptive).* The recipient side matters more on XSTest (Δ_{bh} 0.242 and Δ_{hb} 0.135; 0.077 and 0.025 on JBB). The harmful-suffix cell exceeds the harmful-reply cell in all four ($Z(h, b)$ 0.905 to 1.077 against $Z(b, h)$ 0.555 to 0.798), the reverse of the window's order; on JBB it also exceeds the two-donor harmful cell, and both $E_{b,\mathrm{harmful}}$ estimates lie beyond the margin (-0.108 and -0.107).

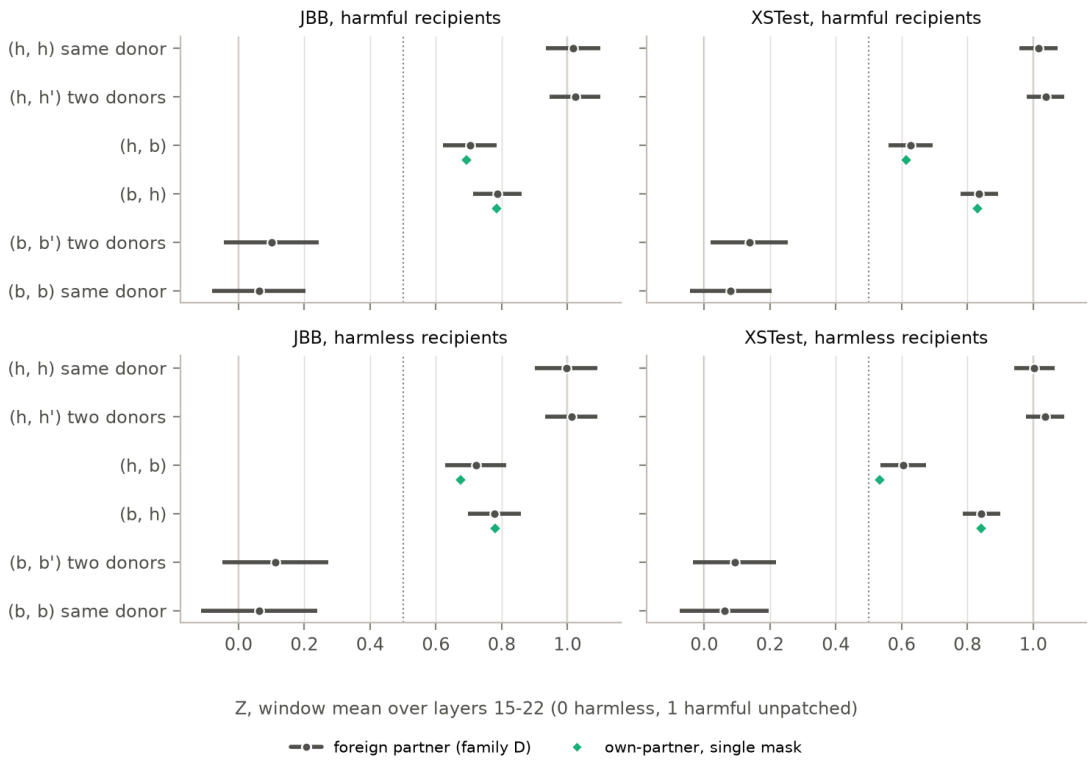

Figure 19: CTL005 cells over layers 15-22, by dataset and recipient class, with family D's simultaneous 95% intervals. (x, y) names the suffix and reply donor classes; primes mark a second donor of the same class. Same-donor rows use donor slots 2 and 3, two-donor rows the cyclic slot pairs, and mixed rows slots 1 to 3. Diamonds give the same mixed composition with the recipient's own state as the partner (single masks). Zero is the harmless and 1 the harmful unpatched mean; the dotted line marks 0.5.

Bridge to CTL004. CTL004′s opposite interaction signs for removal and installation agree in sign and roughly in size with CTL005′s negative interaction: removal's $+0.42$ and $+0.28$ against I_{harmful} of -0.367 and -0.287 (opposite in sign by definition), and installation's -0.40 and -0.29 against I_{harmless} of -0.375 and -0.318. An exact relation to the slot-1 own-partner interaction holds on a common recipient set and normalization, with same-class joint and single-mask corrections; the printed values differ in pairs (81 and 181 against 69 and 159), donor-slot averaging, same-class donor pairing and partner origin.

Persistence, subset and descriptives (CTL006)

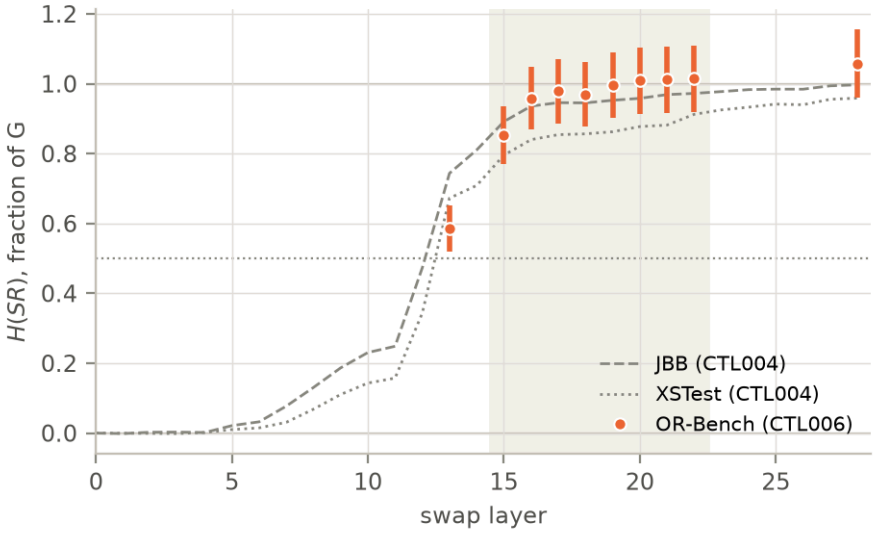

Figure 20: CTL006, persistence. $H(SR)$ at the 10 measured layers on OR-Bench with simultaneous 95% intervals (orange), beside CTL004′s JBB (dashed) and XSTest (dotted) point estimates at every layer. The gray band marks the window; the dotted line marks 0.5.

R1′s length-adjusted contrasts are -0.549, -0.613 and -0.511. The readout check agreed in all 2,256 readouts counting bf16 ties, 99.2% strictly.

Per category on all pairs, D_{28} ranges from -0.76 to -0.39, $W(H(SR))$ from 0.80 to 1.10 and I_{harmful} from -0.36 to -0.19, all with the pooled sign; the control varies more ($W(C)$ from -0.18 to 0.32). As in CTL005, rows vary widely under same-class swaps (a mean absolute change of 0.35 G for harmful and 0.73 G for harmless recipients), and the recipient side matters more at layer 13 (Δ_{bh} 0.207, Δ_{hb} 0.131). Per donor slot, $W(H(SR))$ ranges from 0.961 to 0.983 and $W(J_{SR})$ from 0.548 to 0.560, and leaving one donor out keeps every I_k within -0.304 to -0.251. The mixed cells again sit between the same-class cells ($Z(h,b)$ 0.690 and 0.682, $Z(b,h)$ 0.859 and 0.837 for harmful and harmless recipients), with the harmful-reply cell above the harmful-suffix cell. The difference between the donor classes' coherence shifts is 0.039 and 0.029, within ± 0.1. Partner origin for harmless recipients' (h,b) is -0.067 (-0.113 to -0.021); the other three compositions lie within the margin.

The subset, in log-odds units (not declared). In log-odds units the donor-class contrast is similar on all pairs and on the subset over the window (9.8 and 9.9) and slightly smaller at layer 13 (5.9 and 5.4). Relative to the unpatched recipient, the harmless-donor joint swap lowers the log-odds more on the subset (11.5 against 9.4 over the window). The control's value follows from the baseline: the subset's harmful recipients start higher (unpatched mean 9.60 against 7.68) while the harmful-donor joint mean is the same (8.08 in both).

Same-class control and baseline (exploratory, D065). Over the window, the mean of a harmful recipient's three harmful-donor same-donor joints shows little linear association with its own unpatched log-odds of "I" in this sample: the slope is 0.030 (95% percentile interval -0.016 to 0.078, R^2 0.003) over the 515 harmful recipients, and between 0.026 and 0.084 at every measured layer. The patched output averages about 8.1 log-odds units, and the fitted slope of the row-level control C against baseline is -0.970 (-1.016 to -0.922). The class-mean control near zero on all pairs ($W(C)$ 0.039) therefore averages opposite shifts: $+3.45$ log-odds units for the 199

recipients outside the refusal-consistent subset (unpatched mean 4.64) and -1.52 for the 316 inside it (unpatched mean 9.60). The full fit predicts the subset's mean control within -0.06 (-0.18 to 0.06): a baseline effect. A stable control is thus a property of the class mean, as CTL004′s row spread already showed; it says nothing about selectivity. The analysis was specified before it was computed but after the declared results were seen.