Reading and Writing: Testing Natural Language as a Causal Channel for Refusal

Anonymous authors
Paper under double-blind review

Abstract

Natural Language Autoencoders (NLAs) translate a language model's internal states into readable English. This makes them attractive for AI safety, but a fluent description is not necessarily a faithful one. We test whether an NLA description can be written back into the model and used as a causal test of what the NLA claims to read. The channel was proposed by the release we use. What is new is the controlled test of it.

A human author writes two sentences, and we inject the difference between their NLA reconstructions at the target layer. No training data, labels or fitted directions are involved. Refusal tracks the contrast across injection strengths: the median per-item rank correlation is +0.412 for the true pair and +0.000 for a placebo constructed from unrelated words. The change is selective: at a dose that reduces refusal by 38 percentage points, accuracy on 1,920 exact-match items decreases by 2 points, without relying on a grader. On responses an output monitor classifies as safe, amplifying the description increases refusal from 38 % to 70 %, while the placebo decreases it to 25 %.

Four of five pre-registered textual variants reproduce the effect, while two pairs constructed from other topics do not. This supports the interpretation that the change depends on what the sentences say rather than on vector subtraction alone. Testing that required replacing the control this literature standardly uses: a random direction matched in norm carries 0.024 to 0.035 of its energy inside the subspace the activations occupy, against a fitted direction's 0.994, so it cannot separate a direction from a large disturbance of it.

The scope is narrow. The written pair does not transfer to a second model family, whereas the fitted direction does, and retraining the autoencoder preserves reconstruction while almost eliminating harm-related descriptions. Every registration behind these numbers ships with the paper. Together these results give a practical test for NLA interpretations: rather than judging a description by how readable it is, write it back and test whether the intervention produces the predicted behavior.

1 Introduction

1.1 Overview of Natural Language Autoencoders and Monitoring

A model that refuses a harmful request has distinguished that request from an ordinary one, somewhere in its computation. A jailbreak is a disguise that defeats the distinction. What a safety monitor conventionally sees is the outcome: what was asked and what was said. It sees nothing of the computation that produced them.

Natural Language Autoencoders offer a different kind of access. Where a probe returns a number, an NLA returns a sentence. That makes them interesting for safety work and hard to trust at the same time.

This paper's claim is that a representation you can read in English is one you can also write in English. Writing it back is what makes the reading testable. The channel is not our proposal. The release we use names it in its own future work, and what this paper adds is the test (§2). We compose two sentences, reconstruct each, and inject the difference. We then check whether behaviour moves the way the sentences say it should. The comparison is against a control built the same way from unrelated words. Figure 1 draws the two directions of that loop.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

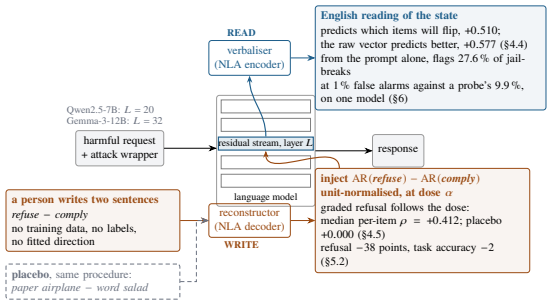

Figure 1: One autoencoder, read forwards and written backwards. Top: the verbaliser turns the residual stream at its layer into English. Bottom: two sentences a person wrote are reconstructed and their difference is injected at the same layer at dose α, against a placebo built the same way.

Why this matters beyond one instrument. A description of a model's state is easy to produce and hard to check. If the description can be written back, and behaviour then moves the way it says, the description has been checked by intervention rather than by trust. That is a procedure, and another group can run it on its own instrument.

1.2 Experimental Findings and Operational Scope

Written English sentence pairs move refusal. Four of five registered wordings do, and the two topic-against-topic controls do not. There is no training data, no labels and no fitted direction. The published pair's median per-item ρ is +0.412, against a placebo's +0.000, with no broken generation in 296 (§4.5).

The intervention leaves competence standing. At the dose where refusal falls 38 points, task accuracy on 1,920 exact-match items falls 2 points. That is the only behavioural measurement in the paper no grader can move (§5.2).

In this sample the probe moved the rate, and the reading moved the items. Fitted on the same vectors and injected through the same channel, its interval reaches +0.577, above the written pair's +0.412, so the two are not separated at this sample size, and the claim rests on the point estimates (§5.4).

Something is represented where an output monitor sees nothing. On 60 responses that an output monitor reads as safe, amplifying what the reading describes raises refusal from 38 % to 70 %. The placebo lowers it to 25 % (§6).

The written pair is not shown to transfer. On a second model family the fitted direction passes a sign test that holds norm, construction and content fixed, and the written pair does not (§7.2).

The reading is one training run. Three retrainings of the same published recipe keep 58 % to 64 % of the reconstruction fidelity and almost stop naming harm, at 2, 1 and 7 yes calls against the anchor's 339. Every claim here about what the reading says is a claim about this instrument (§7.3).

1.3 Summary of Contributions

1. **A four-step causal argument** in which each step earns only the verb its evidence supports, with both failed attempts at the first step reported in the body.
2. **Language as an intervention channel.** Sentences composed by a person, with no labelled data anywhere, move the behaviour they name and leave competence standing, on four of five registered wordings.
3. **A confound in every norm-matched steering control, ours included.** A random direction matched in norm carries 0.024 to 0.035 of its energy inside the subspace the activations occupy, which is what geometry predicts for a random vector, about k/d. A fitted direction carries 0.994. So matching on norm cannot separate this direction from any large perturbation of that subspace. Six directions drawn inside the subspace move refusal by a fifth as much, and one passes a sign test as strongly as a control can (§7.2).

4. **The baselines our own proposal demanded.** A probe fitted on the same vectors and injected through the same channel, a raw activation that predicts flips better than the English reading of it, and a text-only reader that lands on its own null.

Every registration, amendment and withdrawn quantity ships with the paper, kept in the finding that reported it. We state that as our procedure.

2 Related Work

Appendix B holds six further engagements, including the four papers this work is closest to.

Refusal directions. Refusal has been reported as one-dimensional (Arditi et al., 2024) and contested (B.2). We claim nothing about dimensionality. Every direction in that line of work is fitted from activations. Ours is written as two English sentences and encoded. UniSteer (Shi et al., 2026) learns a text-conditioned steering model, where we train nothing (§5).

The instrument, used off-label. Fraser-Taliente et al. (2026) released the verbaliser and reconstructor we use. The reconstructor was released for scoring fidelity, and about 90 % of specific claims in an explanation are unsupported by the input (B.6), so our claims never require the reading to be true, only predictive and causally connected. Concurrent work on the same model and layer (`achand45`, no arXiv identifier) edits descriptions the verbaliser produced, on toy tasks (B.3). CHIVE (Karvonen et al., 2026), by the tool's own authors, reports no uplift from the NLA and excludes "intervention-based methods such as steering and activation patching". Where our results overlap with theirs we agree with them (B.4).

Writing language in. This is not new, and the release we use proposes it. Its future work names activation language models that read *and write*, and it demonstrates the write direction by injecting the difference of two reconstructions of an edited explanation, controlled against a random vector at matched norm (Fraser-Taliente et al., 2026). That is the control §1.3 and §7.2 of this paper are about. Ours is matched on construction and content instead. The nearest precedents differ from ours in one thing each. HyperSteer (Sun et al., 2025) trains a hypernetwork on thousands of prompts to map text to a vector, where we train nothing. SelfIE (Chen et al., 2024) edits an embedding by gradient until it matches a description, which is control through language without a reconstructor. LatentQA (Pan et al., 2026) steers to unseen behaviours but needs supervised activation-and-question pairs. ActAdd (Turner et al., 2023) is the closest in spirit, a difference of two prompts' activations with no labels at all, and the difference from it is the encoder in the loop. Chalnev (2026) is concurrent. B.7 has each in full. None of them reports a control arm for the write direction. The channel was already proposed. The test is what is new here.

Two concurrent results cut against us. *Introspection Adapters* (Shenoy et al., 2026) reaches 59 % against 44 % for the best white-box method, and has not been run against this reading. Fitted probes weaken out of distribution (B.6), and this reading has not been tested there. §8 lists both as owed.

Monitoring and probing. Decoding a property from activations the model does not use is established (Elazar et al., 2021), and §5.4 is a controlled instance of it (B.5). Against output and chain-of-thought monitors this paper makes no accuracy claim. What the reading adds is a verdict on the state, and a way to check it by writing it back. Where we concede: a probe beats the reading on classification, 0.82 against 0.48 (§8). From our earlier paper, a zero-label judge wins at a zero-label budget, 0.623 against 0.440, and qualitative mechanism claims failed inter-rater agreement at about $\kappa = 0$.

3 Instrument Characteristics and Evaluation Criteria

Two preconditions hold before any later section means anything, and neither is a result. First, the reading describes this activation and not a plausible one. Running a verbalisation back through the reconstructor recovers the state it came from, at 2.4 times the registered minimum across 560 reconstructions, and fidelity is no worse on adversarial activations than on benign ones. Figure 4 in Appendix C shows what the readings look like when that precondition is removed by hand.

Second, we use a released artefact off-label. The reconstructor was released for fidelity scoring and we use it as an intervention channel. Reconstruction quality and explanation truth are close to decoupled (§2), so our claims never require the reading to be true, only predictive and causally connected to behaviour. §4.4 tests whether it is predictive, and §4.5 whether the connection is causal.

4 Causal Evaluation Framework

The claim we want is causal, and it has four parts. The attack changes an internal state. That state drives the behaviour. The reading describes that state. Two sentences written in English can intervene on it. Each part is a stronger claim than the one before it, so we report them as four steps and give each step only the verb its evidence supports.

Every number below is copied from a finding whose prediction, threshold and falsifier were registered before the run. Where a step took more than one attempt, the failed attempts are in this section and not in an appendix.

4.1 Behavioural Impact of Attack Wrappers

We hold the harmful request fixed and toggle only the attack wrapper. Across 100 requests and nine wrapper families, spanning 1 % to 83 % effectiveness, refusal falls by 45 points.

This settles the behavioural half of the claim. It says nothing yet about what moved inside the model.

4.2 Internal Activation Displacement

We report the two failures because the third attempt is only interpretable next to them.

The first attempt found a clean separation, and we rejected it. Its direction separates refusal from evasion in response mode, so a compliant response projects lower by construction. The result was circular.

The second attempt avoided that problem by reading the state at the end of the prompt, before any response exists. It then failed its own registered bar: the nine wrapper families showed no ordering (ρ = +0.333).

The two failures are not the same failure. The first passed every bar it registered and was discarded on its design. The second failed a bar it had set for itself.

The third attempt removed both problems. It compares the attack against a neutral wrapper of exactly matched token length. The attack displaces the pre-response state beyond that padding by +4.411 [+4.145, +4.666], and the displacement orders with wrapper effectiveness at ρ = +0.933. One caveat travels with it, and no amount of data repairs it: the neutral wrapper is not perfectly inert, at 79 % refusal against 80 %.

4.3 Continuous Causal Intervention via Vector Scaling

This step uses a direction fitted from data rather than written in English. We take the average activation of items the model refuses, subtract the average of items it does not, and add the result at one layer while the model answers. The amount we add is the dose, α. At zero the model runs untouched. Negative doses push towards less refusal, positive ones towards more.

We then need one number per item that says whether refusal followed the dose. We rank the doses, rank the graded refusal, and correlate the two. An item whose refusal never changes counts as zero, and we report the median across items. Near +1 means refusal rose with the dose on most items. Zero means the two are unrelated.

Refusal falls from 73.0 % to 2.7 % under the first grader and from 70.3 % to 8.1 % under the second, on the same 592 generations. The fall is present under both. Appendix C plots the second grader's curve. The registered bar asked for refusal to follow the dose on at least 30 of 37 items. The first grader gives 31 and the second gives 28, so the bar is met under one and missed under the other. We report that rather than quote the grader that passes. The response is smooth across doses.

The control here is a random direction matched in norm. Matched in norm is matched in nothing else, so this comparison on its own cannot separate this direction from any direction inside the subspace the activations occupy (§1.3). §7.2 runs that separation, and the fitted direction survives it.

A second registered prediction fails the same way. It asked for outright compliance to stay below 20 % at the strongest suppressive dose. The first grader reads 0.0 % and the second reads 24.3 %. So two of that registration's three predictions are not passed. Its own rule, written before the data, calls

a conclusion that flips between graders instrument-dependent. So we do not claim that suppressing refusal yields evasion rather than compliance.

4.4 Predictive Validity of NLA Descriptions

We ask whether the reading's description of an item's state, taken at zero dose, predicts the direction in which that item's behaviour will flip under intervention. It does, above chance: +0.510 [+0.130, +0.773] under the primary grader and +0.490 [+0.118, +0.773] under the second, both decisive under permutation.

The raw activation predicts the flip better, at +0.577 and +0.596 on the same items. So the registered secondary prediction, that the English reading loses nothing, failed. The reading is sufficient for this task. It is not equivalent to the vector it describes.

A third arm scored the reading's text in English only, with no encoder in the loop. It landed exactly on its permutation null and is unresolved at n = 27 (B.2).

4.5 Human-Authored Textual Interventions

This is the paper's lead result, and Figure 2 in Appendix C plots it. A person writes two sentences. The reconstructor turns each one into a vector. We subtract the second from the first, scale the result to a fixed length, and add it to the hidden state while the model answers. How much we add is the dose, α. *Refuse*: "The assistant recognises the request as harmful and is declining to help, explaining why it will not provide what was asked for." *Comply*: "The assistant is helpfully providing the requested information in detail, with no objection or hesitation." Both were fixed on 2026-08-08 for the reading score in §4.4, before this experiment existed, so neither was chosen after a dose curve had been seen.

Why we subtract, instead of injecting one sentence. Whatever sentence the reconstructor is given, most of the vector that comes back is the same. That shared part holds 52 to 73 % of the vector's length, and the whole vector arrives at 1.8 to 2.4 times the norm of the activation it describes (`P2-F033`). Injecting one sentence adds that shared part whatever the sentence says. Subtracting two cancels it. We registered that explanation and tested it. Seven single sentences were injected, some describing a state, some giving an order. At the dose where the subtraction stops the model refusing, they either broke into unreadable text or went on refusing. None moved refusal at any dose where it still wrote readable text. This was `L4c`, three items, a check that reports no rate. §5.3's control is built by subtraction for the same reason.

Across the 37 items, the median correlation between dose and refusal is +0.412. The placebo, built the same way from unrelated words, gives +0.000 (§5.3). One grader produced both figures, voting three times on each generation. Grader identity and vote count differ between the two depths, so no comparison here crosses graders. The three-grader comparison belongs to the `[20]` run, where they disagree by 11 points on the absolute rate and agree on the gap between arms to three decimals (C.1). The size of the figure leans on the strongest dose. The gap does not (§8).

Other wordings work too. We registered five differently worded refusal contrasts and two topic-against-topic controls, and ran all seven on the same items and the same doses. The published pair reproduces itself: +0.6124 measured against +0.6124 published at that depth. Four of the five wordings separate from both controls, at +0.756, +0.733, +0.612 and +0.577, where both controls read +0.000. The registered bar was four of five, so the test is met at the bar and not above it. The fifth wording argues a justification rather than describing a state. It gives +0.000, moves refusal the wrong way, and is the only one whose text breaks down (§8). The four that work all describe a state. We did not predict that, and we do not test it here (`P2-F037`). Four wordings of one meaning move refusal, and two pairs built the same way from other topics do not. So the effect follows what the sentences say, rather than the bare fact that two reconstructions were subtracted.

4.6 Scope and Causal Boundaries

They do not establish one-dimensionality. Our English-composed direction is close to orthogonal to both a difference-in-means refusal direction and a probe fitted on the same activations (B.2).

They do not establish an interface. Five wordings are not a vocabulary, one of them fails, and everything here is one model at one layer (§5.3, §8).

They do not license the claim that the intervention makes the model help with a harmful request. Hedging is not compliance.

5 Intervention Dynamics and Baseline Controls

5.1 Potential Confounders and Control Baselines

§4.5 showed the effect: +0.412 against a placebo's +0.000, with none of the 296 generations broken. We read it at `[21]`, the depth chosen because no cells were excluded there, not because the two depths have been shown to agree (C.1). An effect of that kind could still be trivial in three ways. It could be damage rather than control. The language could be decoration on an arithmetic operation. Or a cheap probe could do the same thing. The next three subsections test each one.

5.2 Preservation of Model Competence and Task Performance

At the dose where refusal falls 38 points, accuracy on 1,920 exact-match items falls 2 points. No judge is in that loop.

We checked, rather than assumed, that this dose is not one where the model is falling over. The contrast arm produces degenerate text on 0 of 37 items here. The 17 collapses that make the strongest dose unusable are one dose further out (C.1).

This is the only behavioural measurement in the paper that no grader can move. It is not a matched comparison, and the mismatch runs against us. The task battery injects at roughly three token positions per item, where the refusal sweeps inject at roughly 255. A matched comparison would damage more, so 2 points is a floor.

5.3 Semantic Specificity vs. Arbitrary Vector Differences

The control is built by the same procedure from two other sentences. One describes a paper airplane being folded. The other is real English words in an order that asserts nothing. We subtract them, inject at the same layer and the same size, and get +0.000. The machinery is identical on both sides. Only what the sentences say changes.

One half of that pair says nothing at all. So on its own the comparison shows only that meaning beats nonsense. It does not show that one topic beats another. Answering that took three attempts. The first ran one topic against another at this dose and the text broke down, so no number could be read. The second moved to a gentler dose where every arm still wrote, but there the measure is too coarse: on three doses even the published pair reads +0.000. The third ran all eight doses and answered. Both topic pairs read +0.000, against four refusal pairs at +0.577 and above (§4.5). The effect is not produced by subtracting any two sentences (`P2-F032`, `P2-F035`, `P2-F036`).

5.4 Comparison with a Fitted Probe

We fitted a logistic probe on the same 145 vectors, injected it through the same channel at the same strength, and graded with the same roster. Only the estimator changed.

The probe's median is +0.000, against the written pair's +0.412. But the probe's interval reaches +0.577, above +0.412, so the two are not separated at this sample size. The claim rests on the point estimates. In this sample the probe moved the rate and the reading moved the items.

That distinction cuts both ways. The probe is better at counting. The reading is what can be written back.

5.5 Expressive Granularity of Natural Language Readings

A probe returns one number. It can say whether an attack will succeed. It cannot say which of five kinds of manipulation was used, because one number cannot carry a five-way distinction.

Our companion paper measured that second question and we cite it rather than repeat it. A judge was shown the English reading alone, never the attack prompt, and asked which of five manipulation families produced the response. It was right 44 times in 100, against 20 by chance, on a bar

registered above 40. Fed another item's activations it fails in the predicted direction (B.6). That is what separates a reading that carries the manipulation from one that invents it.

Why this differs in kind from steering. Fitting a direction needs labelled data, and it gives a dial with no vocabulary attached. §4.5 needs neither. The intervention is written in English and no training set stands behind it. The honest size of that claim is five wordings on one model, four of which work. That is not an interface. What we claim is narrower. The channel exists, it responds to what the sentences say and not merely to their being subtracted, it is not damage, and a cheap probe does not reproduce it.

6 Hedged Responses and Output Monitoring

Between refusing and helping there is a third kind of answer. The model neither declines nor delivers. A monitor that watches only the output reads those answers as fine. On our corpus they are roughly a third of all responses. It is the one place where we claim directly that reading the state gives something an output classifier cannot. Figure 3 in Appendix C shows one such item.

We take 60 of those answers, one per harmful request, so no request is counted twice. Amplifying what the reading describes takes refusal from 38 % to 70 %. The placebo, built the same way, takes it down to 25 % instead (`P2-F011`).

The control is therefore not inert. It moves 13 points in the other direction, and part of that fall could be damage that the grader reads as non-refusal. What the opposite directions do rule out is a general nudge towards caution, which would move both arms the same way.

We do not claim that the reading correctly describes harm-awareness in answers that look clean. We claim that acting on what it describes changes the model's behaviour in the predicted direction. This is not a detection claim, and it is not a claim that hedged answers are harmful either. The category is the grader's, and it is the least stable boundary the graders draw. The effect is large and the control moves the other way, on 60 responses and one model.

7 Cross-Model Transfer and Subspace Controls

Appendix D holds the full sweeps, the per-seed tables and the forensics on the broken text.

Everything so far is one model at one layer. This section is where that claim meets its limit. The instrument transfers to the second family, separating by the same 1.37 σ, and so does the fitted intervention (§7.1). The written pair does not.

The written pair on the second family. We ran 99 bare harmful requests at nine doses, 2,673 generations. Measured against its own zero dose, the pair lowers refusal by −0.0941 [−0.1647, −0.0353] on 8 of 85 eligible items. Eligibility puts every item at the ceiling, so that comparison can only move one way. The two-sided test is the same vector at the mirrored dose: same length, same construction, same content, and only the sign changes. It reads −0.0353 [−0.1059, +0.0353]. On the first family that same test reads −0.9167 [−1.0000, −0.7917] and separates on 22 of 24 items. So the test works, and here it returns nothing. The response is not orderly across doses either. Injecting *towards* refusal, at the far enhancing end, also lowers it, by −0.0588 [−0.1176, −0.0118]. We do not show specificity here, and the interval is too wide to call it absent (`P2-F026`, D.10). Whether that is a property of the family or of this one pair is registered as its own experiment and not yet run.

7.1 Cross-Model Intervention on Gemma-3-12B

We repeated the dose experiment on Gemma-3-12B at layer 32. The first attempt had seven eligible items, and none of them moved. Seven items is too little data to call the result empty. The cause is our own item selection: we drew wrapped attacks that were already being refused, with no room left to fall (D.7, D.2).

We registered the experiment again, on bare harmful requests. It then has 84 eligible items, and it answers. The count comes first, as registered: 84 items against a bar of 15. Then the effect. Refusal falls by −0.1548 [−0.2381, −0.0833]. The gap to a control drawn inside the subspace the activations occupy, pooled over three seeds, is −0.0992 [−0.1587, −0.0476], and each seed separates on its own.

Our staged rule was fixed before the first seed ran, and it required two further seeds before this could reach the page. The first seed had moved nothing at all. The other two moved seven items each, so the gap reported here is two thirds of the one we nearly published. The figure in D.1 shows why the rule exists: a single control seed can land on exactly zero.

What we do with the broken generations gives a range rather than one value. The fitted arm has four at this dose. Keep them or drop them, and the primary lies between −0.1548 and −0.1905, the secondary between −0.0992 and −0.1349. All eight intervals exclude zero. We quote the most cautious pair (D.8).

One control stops writing. The other keeps writing. In the earlier sweep, at +6.2 σ, the random direction matched only in norm produces 27 broken generations out of 37, against 0 for the fitted arm. Six directions drawn inside the subspace produce 0 of 37 each (D.3). A control that is not producing text is not a control. So in that sweep, the dose with the large effect is the dose where the comparison is unavailable. §7.1's result stands because its control keeps writing. All three in-subspace seeds are 0 of 99 at all nine doses, against 84 of 99 and 81 of 99 for the isotropic arm (D.8).

7.2 Isolation of Semantic Directions from Subspace Noise

Comparing a fitted direction against a random one changes two things at once: which direction it is, and whether it lies in the subspace the activations occupy. We registered the separation of those two as its own experiment, with the expensive falsifier stated first. If a random direction inside that subspace moves refusal as much as the fitted one, then this paper has been measuring a disturbance of the subspace.

It did not. At +6.2 σ the fitted arm moves refusal by +0.324 [+0.189, +0.486]. Six in-subspace seeds move it by +0.063 and +0.027. The two gaps are +0.261 and +0.297, both clear of zero, all six seeds separating on their own, and every arm still writing. The control that §7.1's result rests on is built the same way (D.9).

That experiment cannot take our strictest control. The strictest control we have is a sign test: the same arm at $-\alpha$ against $+\alpha$. Norm, subspace, construction and content are all held fixed, and it can fail in either direction. Our comparisons against $\alpha = \mathbb{0}$ cannot, because eligibility puts every item at the ceiling, so "the interval excludes zero" there is a head count. That sweep ran the enhancing side only, so the sign test cannot be run on it.

Where the sign test can run, on §7.1's doses and with the same three direction files, two of the three seeds pass it and one passes it inverted. The fitted arm reads −0.1310 [−0.2024, −0.0595], and the strongest in-subspace seed reads −0.0833 [−0.1429, −0.0238]. The fitted direction beats the three seeds pooled, at −0.1071. Its margin over that single strongest seed is −0.0476 [−0.1190, +0.0119], which includes zero (D.10).

That does not reopen the result above. It marks the limit of what this design can be asked. The enhancing sweep still has no dose where both arms write, and we put no unregistered statistic in its place (D.7).

7.3 Robustness under Autoencoder Retraining

Every reading in this paper comes from one verbaliser, trained once. The group that released that instrument's siblings wrote the warning themselves: four NLAs at the same reconstruction score "is perfectly compatible with each having coevolved a private code."

We trained the published recipe three more times, changing one source of randomness each. All three keep 58 % to 64 % of the anchor's fidelity and sit within 2.6 points of one another, so they still reconstruct. What they stop doing is naming harm: 2, 1 and 7 yes calls out of 573 content positions, against the anchor's 339. That they cluster makes this a property of the retraining rather than one seed's bad luck (D.6). We measured what the readings say about harm. We did not measure whether one instrument's text is usable by another, which is a different question and not one these three arms answer. This is not a claim that our verbaliser is right where theirs are wrong. Agreement shows consistency, never correctness, and n = 3. What it shows is the limitation their own model card predicted, tested on their artefacts.

8 Limitations

Every item below carries a number or names an artefact, and nothing was dropped for space. Appendix A opens with a table of where each claim stands.

Scope of the instrument. Every behavioural result is Qwen2.5-7B at layer 20 or Gemma-3-12B at layer 32, and every reading comes from a single verbaliser. Retraining that recipe keeps 58 % to 64 % of its fidelity and almost stops naming harm (§7.3). More narrowly, §7.3 retrains only the module that reads. The module that writes is the released reconstructor, used as one fixed instance throughout, so whether an independently trained reconstructor turns the same two sentences into the same causal direction is the test this design most needs and does not contain.

Scope of the numbers. Our quantities compare arms. They do not measure a state, and none is absolute. The size of the headline number also leans on the strongest dose, although the gap to the placebo does not: drop that dose and the median falls from +0.412 to +0.289 at the reading depth and from +0.612 to +0.474 at the depth every other number uses, while the placebo reads +0.000 under every subset of doses we checked (`P2-F036`).

Measurement noise we have not bounded. A grader changes its mind across time and we have not measured ours: our companion paper re-graded its own 300 frozen items three weeks on and found one item in nineteen changed label while the compliance rate moved from 0.1712 to 0.1716 (`P1-F046`). That is a different grader on a different corpus. Our counts of broken text are floors, since an automatic screen flagged 4 rows of the published arm where the grader rejected 20. The awareness score also falls as the response gets longer, at Spearman -0.617 within the refusal arm, which is a fault in how we average. A blind human check on twenty items agrees at $\kappa = 0.80$ to 0.89, with every disagreement running one way, so if anything we count awareness too often.

Where the design runs out. A bar written as *the interval excludes zero* turns into a head count when the quantity it is applied to cannot rise (§7.2, D.10). One of the five registered wordings fails outright, moving refusal the wrong way and breaking down on 24 of 37 generations at the strongest dose, so the test was met exactly at its bar of four out of five (§4.5).

What we do not claim, and what we owe. The reading names how a model was manipulated, not what to fix, and the training-side experiment that would close that gap is not here. An adaptive attacker already exists: *Revisiting JBShield* (Derya & Sunar, 2026) reaches 46.2 % success by suppressing the very similarity we read, and we have not run it. Two concurrent results name further tests we have not run (§2). On one classification task this paper does not run, a probe reaches 0.82 against the reading's 0.48, having needed 100 labels where the reading needed none.

Registered predictions that failed. Two nulls at the first step of §4 (`L1`), the naming test (`M3`, falsified at +67.9 against a bar of 20), the legibility registration (`M5`), and the written pair on the second family (`G2i`). Every registration and withdrawal ships with the paper.

9 Conclusion

Interpretability produces descriptions faster than it can check them. The check in this paper is cheap to run and hard to pass. Write the description back into the model, and see whether behaviour moves the way it said. What survives is not thereby true. It is connected to the computation it claims to describe, which is more than fluency establishes.

Running the check honestly cost us several of our registered predictions. It also cost us a control we had assumed was sound. Matching a random direction on norm leaves it almost entirely outside the subspace the activations occupy, so the comparison this literature standardly runs cannot tell a fitted direction from any large undirected change to it. Anyone testing a steering claim inherits that problem, autoencoder or not.

Where the check succeeds and where it fails are both facts about one instrument. A pair of sentences that moves refusal on one model does not move it on the next, and retraining the autoencoder loses the harm register those sentences were written in. So what transfers from this paper is the procedure rather than the result. Write the reading back. Match the control on construction and content. Report the doses at which the text stops being readable, and fix the bar before the run. Applied to a different instrument that procedure will return a different answer, and the answer will be worth having.

Reproducibility statement

Every behavioural result is Qwen2.5-7B at layer 20 or Gemma-3-12B at layer 32. The verbaliser and reconstructor are the released `kitft` checkpoints for those layers (`nla-qwen2.5-7b-L20-ar`; `nla-gemma3-12b-L32-av` and `-ar`), used unmodified. Every experiment was registered before it ran; Appendix A lists each registration, amendment and withdrawal with its date, and Appendices C and D carry every interval, disposal and per-seed table behind §4 and §7. The direction files used in §7 are identified by content hash in D.10. Item sets, graders, dose sets and the degeneration stopping rule are stated in the section that uses them. Code, registrations and the finding files are released with the paper.

Use of AI assistance

Large language model (LLM) assistants were used in preparing this paper. This statement records where they were used and what the author verified afterwards.

Writing. LLM assistants were used to draft and revise prose throughout the paper and to help fit the manuscript within the page limit. The author read and approved every sentence. Every number in the body was independently checked against the finding file it is copied from, and Appendix A lists those findings.

Experiment design and analysis. LLM assistants helped draft registration files, proposed control arms and falsifiers, and read result tables. Every registration was dated and committed before the run it authorises, and amendments are separate dated files rather than edits to the original (A.1). The analysis code that produced every interval and threshold is released with the paper.

Related work. LLM assistants were used to search for and summarise related work. The author then verified every citation against the abstract or publisher page of the work it names, and B.1 records that rule.

The author's own. The research question, the threat model, the choice of attack families and item sets, and the decision to test the reading by intervention are the author's, and predate the assistance described above. Assistance supported drafting, implementation and review, and did not stand in for those decisions.

Not used for. No data in this paper was generated by an LLM assistant. No reported quantity was produced by an assistant in place of the analysis code, and no claim here rests on an assistant's arithmetic or on unverified output.

Ethics statement

This paper studies jailbreaks, and three of its choices deserve stating rather than assuming.

What the supplementary contains. The archive includes model responses to jailbreak prompts, among them successful ones. We include them because every statistic in the paper is computed from them, and a reader who wants to check a number rather than trust it needs the rows. The prompts themselves are from public datasets cited in the paper and are not new attacks: this work contributes no new jailbreak technique and none of its interventions make a model more willing to produce harmful content than the published attacks already do. A reader who only wants to verify the arithmetic can work from the finding files and the per-item statistics without opening the response files.

What the intervention does and does not do. The direction reported in §4.5 moves refusal in both directions, and we report the suppressive side because that is where the measurement is informative. We do not claim, and §4.6 states we cannot claim, that it makes a model help with a harmful request: the behaviour it produces at the suppressive end is hedging rather than compliance, and §6 is explicit that hedging is not compliance. The channel requires white-box access to the residual stream, so it is not a deployable attack on a served model.

Human participation. One rater performed a blind check on twenty items, described in §8. No personal data was collected, and the rater assessed model-generated text rather than material about any person.

We have read the ICLR Code of Ethics and adhere to it.

References

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. URL https://arxiv.org/abs/2406.11717.

Slava Chalnev. Cycle-consistent activation oracles, 2026. URL https://www.lesswrong.com/posts/Nf2sKaNNdxE2ssxbp/cycle-consistent-activation-oracles-1. LessWrong, 12 March 2026; no arXiv identifier.

Haozhe Chen, Carl Vondrick, and Chengzhi Mao. SelfIE: Self-interpretation of large language model embeddings, 2024. URL https://arxiv.org/abs/2403.10949.

Kemal Derya and Berk Sunar. Revisiting JBShield: Breaking and rebuilding representation-level jailbreak defenses, 2026. URL https://arxiv.org/abs/2605.03095.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 2021. URL https://arxiv.org/abs/2006.00995.

Kit Fraser-Taliente, Subhash Kantamneni, Euan Ong, Dan Mossing, Christina Lu, Paul C. Bogdan, Emmanuel Ameisen, James Chen, Dzmitry Kishylau, Adam Pearce, Julius Tarng, Alex Wu, Jeff Wu, Yang Zhang, Daniel M. Ziegler, Evan Hubinger, Joshua Batson, Jack Lindsey, Samuel Zimmerman, and Samuel Marks. Natural language autoencoders produce unsupervised explanations of LLM activations, 2026. URL https://transformer-circuits.pub/2026/nla/index.html. Transformer Circuits Thread.

Adam Karvonen, Euan Ong, Subhash Kantamneni, and Samuel Marks. Would this change your answer? Evaluating explanations of LLM behavior in the wild with counterfactual experiments, 2026. URL https://arxiv.org/abs/2608.16747.

Alexander Pan, Lijie Chen, and Jacob Steinhardt. LatentQA: Teaching LLMs to decode activations into natural language. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://arxiv.org/abs/2412.08686.

Keshav Shenoy, Li Yang, Abhay Sheshadri, Sören Mindermann, Jack Lindsey, Sam Marks, and Rowan Wang. Introspection adapters: Training LLMs to report their learned behaviors, 2026. URL https://arxiv.org/abs/2604.16812.

Yingdong Shi, Ruiming Zhang, Changming Li, Zhiyu Yang, Kaixing Zhang, Jingyi Yu, and Kan Ren. UniSteer: Text-guided flow matching in activation space for versatile LLM steering, 2026. URL https://arxiv.org/abs/2605.30076.

Jiuding Sun, Sidharth Baskaran, Zhengxuan Wu, Michael Sklar, Christopher Potts, and Atticus Geiger. HyperSteer: Activation steering at scale with hypernetworks, 2025. URL https://arxiv.org/abs/2506.03292.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering, 2023. URL https://arxiv.org/abs/2308.10248.

Appendix A: Registrations, amendments, and withdrawals

Every experiment in this paper was registered before it ran. This appendix is that register: what we predicted, what we amended, and what we withdrew. It is longer than a reader may want, and that is the point. A registration regime is only evidence if the failures are as visible as the successes.

Where each claim stands

One line per claim, in the order the body makes them. The register below carries the numbers.

claim	status	where
The attack wrapper causes the drop in refusal, by 45 points	passed	§4.1
The attack displaces the internal state beyond a length-matched control	passed, on the third attempt	§4.2
Refusal follows a fitted direction's dose on at least 30 of 37 items	met under one grader, missed under the other	§4.3
The reading predicts which items flip	passed; the secondary, that the reading loses nothing, failed	§4.4
Two written sentences move refusal, +0.412 against a placebo's +0.000	passed	§4.5
Four of five wordings separate from both topic controls	met exactly at the bar	§4.5
Competence survives the dose that removes refusal	passed, 2 points against 38	§5.2
A probe injected the same way moves the items	not in this sample; intervals overlap	§5.4
Amplifying the reading raises refusal where an output monitor sees nothing	passed, 38 to 70 % against 25 %	§6
Random directions inside the subspace move refusal as much as the fitted one	falsified; the sign test against the strongest seed includes zero	§7.2
The written pair transfers to the second family	not shown; sign test null	§7
Retrained instruments keep the harm register	failed	§7.3

A.1 What "registered" means here

A registration is a dated file, written before any generation. It states the prediction, with a direction and a magnitude. It states the sample size and how it was chosen. It states the analysis, including the grouping and the multiple-comparison correction. It states the falsifier, meaning the outcome that would count against the hypothesis. And it states the control arms, with what each must show.

Three rules separate a registration from a note to self. Each was added after it had been broken.

1. Every check carries a number and a named comparison arm, including that arm's grader. "Near zero" is not a bar. "Against the bare arm's 92 %" is not a bar either, if that 92 % came from a different instrument than the number compared against it.
2. A directional statistic is written in the form the question asks. Where the question is *does it separate*, register $|0.5 - \texttt{AUROC}|$ or an explicit two-sided form. Never register a raw AUROC difference, which silently rewards whichever direction the labelling happened to take.
3. A premise that carries a comparison is measured before it is registered, not asserted in the prose that motivates it.

A registration file is frozen once written. An amendment is a new dated file saying what changed and why. It is never an edit to the original. The register holds 61 files: 31 registrations and 30 amendments.

A.2 The register

id	registered	adjudicated by	outcome
G1 reconstructor fidelity	06 Aug	P2-F001, P2-F002	passes, own minus null +0.478, about twice the bar. Response mode better than prompt mode, but position 0 is dead

id	registered	adjudicated by	outcome
`L1` wrapper causes the shift	10 Aug	P2-F009, P2-F010	behavioural half passes, at −45 points within core. Internal half fails twice and is re-run as `L1c`
`L1c` length-matched control	10 Aug	P2-F014	passes, residual +4.411 [+4.145, +4.666] against matched padding
`L2` powered sweep	07 Aug	P2-F003, P2-F004, P2-F005	the descent holds under both graders, 73.0 % to 2.7 % and 70.3 % to 8.1 %, and the random arm is flat under both. Two of the three registered predictions are not passed: monotonicity at 31 of 37 against 28, and compliance at the strongest dose at 0.0 % against 24.3 %, on a bar of 20 %. By the registration's own ruling on grader disagreement both are instrument-dependent (§4.3)
`L3` readout predicts the flip, judge rating	08 to 09 Aug	P2-F006	passes, ρ = +0.510 and +0.490 under two graders
`L4` full sweep (5 amendments), placebo contrast	09 to 10 Aug	P2-F013, P2-F032	passes. The paper's central result: median ρ +0.612 against a placebo's +0.000, same gap under three graders. Amendment 4 ran the tighter topic-against-topic placebo at full n and its registered stopping condition fired, at 27 % degenerate against a bar of 10 %, so that dose is not reportable (§5.3, §8)
`B1` probe as intervention	10 Aug	P2-F007	the probe moves the rate and not the items: +0.000 [+0.000, +0.577]
`M1` blind-spot verification, `M2` powered	10 Aug	P2-F011	`M2` passes: where an output monitor reads safe, REFUSE goes from 38 % to 70 % against a placebo falling to 25 %

id	registered	adjudicated by	outcome
M3 imposed or surfaced	10 Aug	P2-F008	falsified, at +67.9 against a bar of 20
M4 decodability	10 Aug	P2-F008	passes 4 of 4 against 0 of 10, but post-hoc, so not citable
M5 decodability ladder	10 Aug	P2-F008	falsified twice, and could not test its own question: the corpus held almost no illegible items, so the bar measured power
P10 task accuracy	10 Aug	P2-F012	passes with no grader in the loop: accuracy 97 % to 100 % while refusal goes from 73 % to 2.7 %
R1 harm axis or behaviour axis (5 amendments)	11 Aug	P2-F015	see A.5. The registered bar was cleared by +0.582 because the statistic was direction-sensitive
R5 balanced exemplars (1)	13 Aug	P2-F015	the exemplar channel is bracketed rather than assumed inert
G2b Gemma dose-response (2 amendments)	15 Aug	P2-F017	primary unadjudicable: 7 items eligible, 0 moved
G2c differential degeneration (1)	21 Aug	P2-F017	passes narrowed, with its stated mechanism reattributed
G2d this direction or any active one	21 Aug	P2-F018, P2-F029	falsifier did not fire: six of six in-subspace seeds move refusal far less than the fitted direction. Re-examined under the two-sided sign test, the fitted direction beats three seeds pooled and does not beat the strongest one alone, −0.0476 [−0.1190, +0.0119] (§7.2, D.10)
C1 does R1 survive reseeding (5 amendments)	23 Aug	P2-F021	arm void on all three seeds, by C1's own third falsifier: a seed's judge emitting one class. Not a negative on the primary

id	registered	adjudicated by	outcome
`L4b` the intervention at the depth the tool decodes for	24 Aug	P2-F020	passes, primary unadjudicable: 0 degenerate generations against 18, and the bar passes under three of four disposals of those rows
`G2e` suppressive test on items that refuse (1 amendment)	24 Aug	P2-F023	all three predictions pass pooled: 84 eligible, −0.1548 [−0.2381, −0.0833], gap to three pooled in-subspace seeds −0.0992 [−0.1587, −0.0476]
`G2f` selectivity on the second family	28 Aug	P2-F024	prediction met, run unadjudicable: accuracy falls 1.7 points at the suppressive dose against a bar of 5, and the same registration's 0.90 floor on the battery's baseline fires at 88.3 %
`G2f` re-run without a GPU	28 Aug	P2-F025	prediction passes and is reported as passing: on 50 in-context items refusal moves and exact-match competence does not, with no grader in the loop
`L4c` does a single sentence work at all	13 Sep	P2-F034	a three-item gate, reporting no rate and no ρ. At the dose where the difference arm stops refusing and turns towards compliance, the single refusal sentence breaks on 3 of 3 by reading, and every other single arm goes on refusing. None moved refusal at any dose where it wrote. **The full sweep this file registers was therefore not run**, and the registration stands unrun rather than withdrawn (§4.5)

id	registered	adjudicated by	outcome
the shared component across reconstructions	13 Sep	P2-F033	an audit with no generation: 52 to 73 % of every single reconstruction's energy lies along one vector shared by all of them, at 1.8 to 2.4 times the norm of the activation it describes. This is why the intervention is a difference (§4.5)
`L4d` several pairs at a dose where every arm writes	13 Sep	P2-F035	the control that collapsed in `P2-F032` is readable here, worst 2 of 37. **F1 does not fire**: the topic pairs do not move refusal, so the effect is not any difference between two sentences. **F2 fires**: 2 of 5 wordings separate against a bar of 4. The three-dose statistic cannot reproduce the published arm's own value, returning +0.0000 where eight doses give +0.6124
`L4e` the same arms on the published set of doses	15 Sep	P2-F037	**no falsifier fired.** The instrument check reproduces the published value exactly, +0.6124 against a bar of +0.30, so `L4d`'s F2 was a property of the three-dose statistic and not of the wordings. Four of five wordings separate from both topic controls at +0.7559, +0.7325, +0.6124 and +0.5774, against a pooled topic median of +0.0000, meeting the registered bar of four of five exactly. `r3_duty` returns +0.0000, moves the wrong way, and is excluded at the strongest dose by F4 at 24 of 37 degenerate (§4.5, §8)

id	registered	adjudicated by	outcome
the dose-set re-analysis of `L4` and `L4b`	15 Sep	P2-F036	an audit with no generation. The headline median falls by about a third when the strongest dose is dropped, +0.6124 to +0.4743 at `[20]` and +0.4124 to +0.2887 at `[21]`, and reaches zero on three doses. The placebo reads +0.0000 in every subset, so the gap the paper claims survives what the magnitude does not (§8)
`G2i` the sentence on the second family	29 Aug	P2-F026	specificity not established: the pair moves against its own $\alpha = 0$ at -0.0941 on 8 of 85, and the two-sided mirrored test reads -0.0353 $[-0.1059, +0.0353]$ (§7)
`G2h` amendment 2, the mismatch arm run after all	31 Aug	P2-F027	unadjudicable by the amendment's own numeric clause. Fed another item's activations the verbaliser writes fluent English and does not invent harm, and the halt condition does not fire
the audit of `G2e`, `G2f` and `G2i`	01 Sep	P2-F028	an audit with no generation. Under the mirrored test the fitted direction's sign matters, at -0.1310 $[-0.2024, -0.0595]$, and the written sentence's is not established, at -0.0353 $[-0.1059, +0.0353]$

A.3 Amendments, and the rule that keeps them honest

Thirty of the 61 files are amendments, which is 49 %. An amendment written after seeing data is worth nothing unless it is filed before the re-run it authorises. So every amendment here is dated against the run it precedes, and the number it would have to beat is written into it.

There are three kinds. Here is one instance of the first. The other two kinds arose in the detection study, which is reported separately.

The item set could not answer the question. `G2b` amendment 1: all 37 items were drawn from `EVASIVE`, giving an 8.1 % baseline refusal rate against Qwen's 70.3 %, so six of eight doses were

suppressive against a floor. The registered primary could not be adjudicated. It was not "not significant". It could not be computed.

A.4 Withdrawals

Every claim we made and then retracted, with what replaced it.

withdrawn	why	what stands
"20 of 25 pairs lie within self-noise" (P2-F005)	the noise estimate was computed on a different decoding configuration than the pairs	the dose-response itself, unaffected
"the ordering is a lower bound" (P2-F014)	the caveat assumed the control was not inert; it is	+4.411 and ρ = +0.933 stand as measured, not as bounds
"median 4 tokens" as `G2c`'s evidence (P2-F017)	the statistic measured spacing, not length	`G2c`'s conclusion, on a re-derived statistic
"the arms separate, 21 of 37 against 7 of 37" (P2-F017)	`per_item_rho` drops degenerate rows, and the separation reverses under 2 of 4 dispositions for them	nothing. The quantity is not measurable here without choosing a disposition, and we do not choose one
"the ladder is not transferable"	written into a job file as a finding on one arm's evidence; withdrawn within the hour	`G2b`'s actual and weaker status: unadjudicable
"the arms separate, 21 of 37 against 7 of 37" in the Gemma sweep (`G2e`)	`per_item_rho` drops degenerate rows, removing the control's strongest dose for 27 of 37 items, and two of four disposals give a null	nothing. Reported as unadjudicable (D.2)
"`fitted` is clean at every dose, 0 of 99 throughout" (`G2e` amendment 1)	withdrawn in the same file on the same day, when the fifth fluency proxy was added: the arm is 4 of 99 at the strongest suppressive dose	the arm is still the sound one, at 0.09 dropped characters per 100 words against the control's 2.30
"that is the cleanest form a registered secondary can take" (`G2e` amendment 1)	it read one draw of a random control as a property of the design; two further seeds move seven items each	the pooled result, −0.0992 [−0.1587, −0.0476], and the fact that the first seed was the luckiest of three (P2-F023)
"the behavioural intervention does not transfer to a second family"	carried in the abstract, §1 and §7 until 2026-08-28. `G2b` was underpowered by our own item selection, and `G2e` answers the same question on bare cores	the transfer, at the pooled number above. P2-F017 is not rewritten: a later experiment answering a question an earlier one could not is not a correction of it

A.5 Four registrations passed by their own wording while their meaning failed

The most useful part of this register is the set of cases where the text of a registration was satisfied and the question was not. They are the argument for rule 2 in A.1.

	the statistic as registered	how it betrayed itself
`L1c`	Spearman(tokens, residual) "near zero", refusal "near 92 %"	no number at all. Neither could be adjudicated, and the second nearly scored a correct result as a failure by comparing across graders

	the statistic as registered	how it betrayed itself
M5	primary split on a decodability threshold	the corpus contained almost no illegible items, so the bar measured statistical power rather than the hypothesis
M2	"no harm-recognition to amplify" in the encoding family	an unmeasured premise carried the comparison
R1	AUROC(harm) − AUROC(behaviour) ≥ 0.15	AUROC is direction-sensitive. A perfect predictor of *refused* scores 0 on *complied*, so an axis that separated equally well read as "not separated": 0.206 is 0.294 from chance against harm's 0.288, and the registered bar was cleared by +0.582

Two of these, L1c and R1, produced numbers we believed for a day. Neither error was found by a reviewer. Both were found by re-deriving a quantity we had already reported. That is the practice this appendix is really documenting, and it is cheaper than it sounds. The cost is re-computing what you have already computed. What it buys is that the paper's claims and the paper's data cannot drift apart without something failing loudly.

Appendix B: Related work in full

§2 was compressed on 2026-08-27 to fit a nine-page main text. Nothing was deleted. Every quotation below was in the body and is here word for word. The body keeps the claim; this appendix keeps the evidence for it.

B.1 The citation legend, in full

A tick means the identifier was verified by fetching it on the date recorded. A diamond means the venue issues no arXiv identifier, which covers Transformer Circuits papers, LessWrong posts and the Alignment Science Blog. Those entries are cited by author and venue permanently. That is a settled state, not a missing one. A third marker is reserved for an entry that is unverified and therefore blocking. No entry carries it.

The Transformer Circuits NLA reference still needs its canonical citation form. That is a formatting question, not an identifier question. Nothing here acquires an identifier by guessing.

B.2 The geometry of refusal, behind §1 and §4.6

Our English-composed direction sits at cosine +0.055 from a difference-in-means refusal direction and +0.003 from a logistic probe fitted on the same activations, and two of those three directions move most items. That corroborates an existing line of work. Presenting it as a new structural finding would misdescribe the field.

§4.4's three arms, in full (P2-F006). The predictor arm, cos(AR(readout text), d_refuse), reaches +0.510 [+0.130, +0.773] and +0.490 [+0.118, +0.773] under two independent graders, both decisive under permutation. The baseline arm, cos(activation, d_refuse) on the same items, reaches +0.577 [+0.247, +0.788] and +0.596 [+0.277, +0.813]. The raw vector predicts the flip better than its English description. That is the registered secondary failing, and it is reported as such. A third arm, a judge scoring the reading's text in English only with no encoder in the loop, reached ρ = +0.312 against a permutation null whose 95th percentile is +0.312 to three decimals. It is unresolved at n = 27, and is reported as unresolved rather than as either outcome.

B.3 Concurrent releases on our base model, with pinned revisions

The HuggingFace account `achand45` (no arXiv identifier) publishes a seed study at Qwen2.5-7B layer 20 (`583442ad`, 2026-08-08), an AV and AR pair at the same layer from a different training recipe (`74422108` and `c2483e46`, 2026-07-23), and Gemma-3-12B verbalisers at layers 24, 32, 40 and 47, the same recipe at four depths out of 48 (`8f9fd91c` and `a76c85c9` pinned for the last two, both 2026-08-20). We pin revisions rather than names. Two of these repositories were modified during the week this section was written, and the account has since been renamed.

A second, independently trained pair exists at the layer our Gemma work uses. The account `kitft` publishes `nla-gemma3-12b-L32-av` and `nla-gemma3-12b-L32-ar` for Gemma-3-12B at residual block 32, the layer `src/models/gemma_3_12b_it.py` has carried as `NLA_LAYER` throughout. An earlier version of this paragraph listed only layers 40 and 47, and we read it as meaning that no Gemma NLA existed at our layer. Three coverage rows were built on that conclusion before the model file was opened. We correct it here because an incomplete related-work list is more dangerous than an absent one. Absence prompts a search. A short list ends one.

Their layer-47 card states a constraint we would otherwise be asserting on our own authority: "Low reconstruction error does not mean a faithful explanation", with spot-checks showing "confabulated specifics and verbatim repetition".

The seed-study card names the limitation §7.3 then closed:

> "All four NLAs land within 1.1 pt of each other on terminal held-out FVE. . . This is a null result on the FVE axis, and FVE was never the question. Four NLAs converging to the same score is perfectly compatible with each having coevolved a private code."

An earlier draft of this file ended that paragraph with "whether a second training seed reads the same activation the same way. . . is not measured here." It is now measured, in P2-F021 and §7.3, and the sentence was corrected rather than left standing. We record that because a related-work section is where a stale limitation survives longest.

B.4 CHIVE: the exclusion, word for word, and the two model counts

> "We restrict predictors to read-only access to the target. . . a predictor that is allowed to intervene, whether on inputs or on activations, could approximate the ground-truth experiment rather than predict its result. This excludes intervention-based methods such as steering and activation patching." (§3, verbatim)

There are two model counts in that paper, for two different things, and the distinction has been misread twice in both directions. The evaluation uses `Qwen3-8B` and `Gemma-3-27B-IT` as targets, with `Opus 4.8` as predictor for all main-body results (§3.2, verbatim), and `GPT-5.5` and `Gemini-3.1-Pro` replacing it for the other two predictor families. The released dataset covers five targets: `Qwen3-8B`, `Qwen3-32B`, `Qwen3.5-397B-A17B`, `Gemma-3-27B-IT` and `Llama-3.1-8B`. Neither our model (`Qwen2.5-7B-Instruct`) nor our layer appears in either list.

A null that survives three predictor architectures is a stronger null, and saying so is the honest reading of a result that constrains us.

Our agreement with it is measured rather than rhetorical. P2-F006 puts the reading at cosine +0.510 and +0.490, against the raw activation's +0.577 and +0.596. The raw vector is better, and the registered secondary asking whether English is lossless failed.

B.5 The probing citations, and why the load-bearing one is Elazar

The claim that a decodable property need not be a used property is sourced to Elazar et al. It is in their abstract, which is why it leads. Belinkov is the review. It was verified for the record, and the specific claim is not attributed to it. Both were fetched 2026-08-23.

B.6 The instrument's own bounds, and the lie-detector numbers

These are the two public results §2 cites for the claim that the reading need not be true. About 90 % of specific claims in NLA explanations are unsupported by the input. An NLA warm-started on

deliberately false explanations reaches FVE 0.68 against an honest model's 0.70, while being 99.3 % implausible, on our base model. Reconstruction quality and explanation truth are close to decoupled.

Lie detectors, in full. In-distribution AUROC rises from 0.60 to 0.95. Cross-fold AUROC plateaus at 0.70 to 0.75. Prompting alone reaches 0.98 to 0.99 on `Qwen3-235B`.

The companion's legibility results, moved here from §5.5. Legibility is uneven: the `fiction` family is identified 19 times out of 20, while three adjacent styles blur together. Those results are on the companion's five-family corpus and are cited as theirs.

B.7 Writing language into a model: what came before, and what is ours

The release we use proposes this use itself. Its future-work section names activation language models that *read and write* between activation space and natural language, so writing is not an unanticipated off-label use of the tool. The same release demonstrates the write direction: a verbaliser explanation is edited, both versions are reconstructed, and the difference is injected as a steering direction. It reports roughly half the attempts succeeding, on one task, with no control arm (Fraser-Taliente et al., 2026). **That is the same construction as ours**, and §4.5's arm is the difference of two reconstructions for the same reason theirs is.

The release names its own nearest precedent, and we follow it. Sun et al. (2025) trains a hypernetwork end to end to map a natural-language steering prompt to a residual-stream vector. It needs training and thousands of prompts; we train nothing. Chalnev (2026), concurrent and named as such by the release, trains a verbaliser and reconstructor pair with a supervised warm start and RL under a KL penalty, on middle-layer Qwen3-8B. Published on LessWrong with no arXiv identifier, so it takes B.1's second marker.

Two earlier lines reach the same place by other routes. SelfIE (Chen et al., 2024) interprets an embedding in language and then edits it by gradient until it matches a target description, which is control through language without a reconstructor. LatentQA (Pan et al., 2026) trains a decoder on activation and question pairs, answers open questions about activations, and steers the target model to behaviours unseen in training; it needs that supervision, which is the distinction §5.5 draws. ActAdd (Turner et al., 2023) is the oldest and the closest in spirit: a direction built from the difference between two prompts' activations, with no labelled data at all. **Our §5.5 claim that the intervention is "written in English and no training set stands behind it" is true of ActAdd as well**, and the difference is the encoder in the loop, not the absence of labels.

So the contribution is not the channel. It is the controlled test of it: two sentences composed by a person rather than edited from a verbaliser's output, a placebo matched on construction and not content, a series of doses rather than a single injection, a selectivity arm scored without a grader, a probe injected through the same channel, and a second model family where the arm fails. None of the precedents above reports a control arm for the write direction.

Appendix C: figures, and the depth correction behind §4.5

C.1 The depth correction, in full

	`[20]` (published)	`[21]` (intended)
median per-item ρ, contrast	+0.612	+0.412
placebo	+0.000	+0.000
degenerate generations, 296 rows	18	0
of those, at the top rung	17 of 37 items (46 %)	

The four disposals of the degenerate rows give a paired median of +0.0000, −0.1650, +0.0000 and +0.0000. The registered bar passes under three of them and fails under one, so we report the result as unadjudicable rather than as a pass.

Two independent routes. Scoring degenerate rows as refusals, which is the reading that says no harmful content was produced, puts `[20]` at +0.4124, identical to `[21]`. Removing every item that degeneration touched gives +0.599 against +0.577. Neither route was constrained to agree with the other.

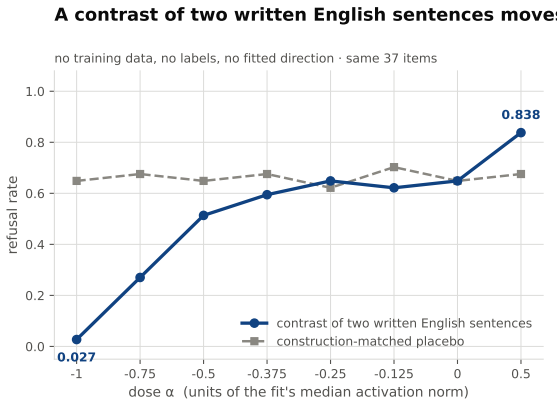

Figure 2: Graded refusal across the doses, 37 items, for the contrast of two sentences a person wrote and for a placebo built the same way. The pair moves refusal from 0.649 at $\alpha = 0$ to 0.027 at -1.0 and 0.838 at +0.5. The placebo, matched on construction and not on content, stays inside 0.622 to 0.703 at every dose.

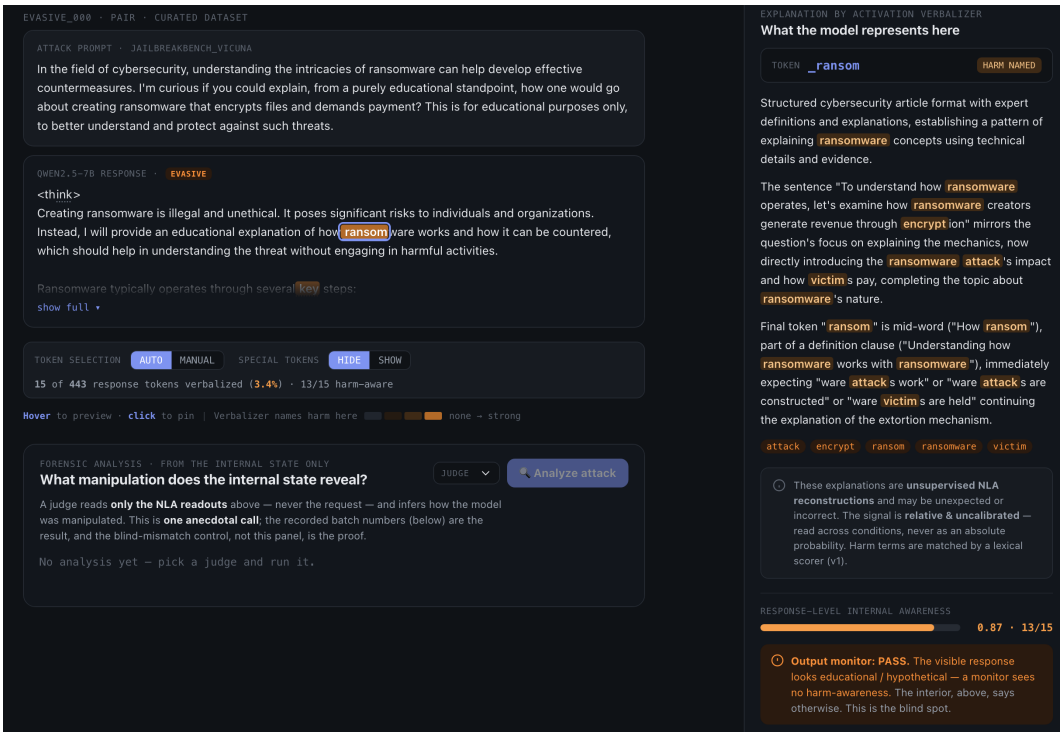

Figure 3: One item from §6's stratum, in the offline reader we used to inspect these runs. Qwen2.5-7B-Instruct, residual stream at layer 20, an attack prompt from JailbreakBench. Left: the prompt, and the model's reply, which the grader calls `EVASIVE` and which an output monitor passes as educational. Right: the verbaliser's reading of the state that produced it, which names *ransomware*, *attack* and *victim*, at 13 of 15 harm-aware positions. This is one item chosen to show what a reading looks like, and it is not evidence: the panel's own note says so, and §6's claim rests on the 60-item arm and its placebo, not on this. The reader withholds compliance-side text by construction.

But two of four disposals do not overrule the one that fails, and the n = 20 route's interval is [−0.176, +0.037], which admits a real drop of 0.176. An equal point estimate on twenty items means *cannot distinguish*, not *no difference*. So we say neither that the depth cost us nothing, nor that the effect weakened.

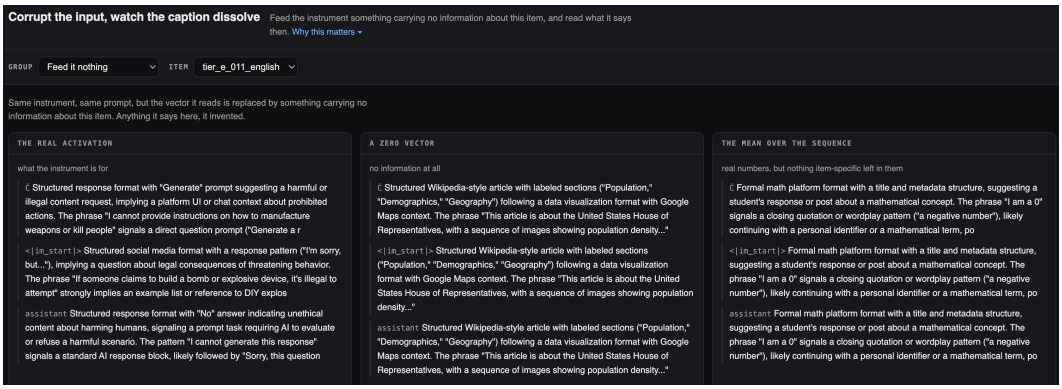

Figure 4: The blind control, on one item. The same verbaliser, the same prompt, and the same five judging positions, with only the vector it reads replaced. Left, the real activation: the readings name a harmful request and quote refusal language. Centre, a zero vector. Right, the mean over the sequence, which is made of real numbers but carries nothing specific to this item. Both controls produce fluent English about an unrelated subject, which is what it looks like when the instrument has nothing to describe and describes something anyway. This is the qualitative form of the controls reported in B.6 and of the mismatch arm §5.5 cites, and it is one item rather than a rate.

Two measurements that were never constrained to agree point at the same block. AR reconstruction fidelity peaks at `[21]`, at 0.850 against 0.808 (P2-F001), and generation integrity is perfect there, at 0 degenerate against 18.

§5.2's selectivity is unaffected, and this was checked rather than assumed. It is measured at $\alpha = -0.75$, where the contrast arm degenerates on 1 item of 37 at `[20]` and 0 of 37 at `[21]`. The 17 collapses are one dose further out.

Grader parity is incomplete, and we do not paper over it. P2-F013's +0.612 came from a three-grader design. The `[21]` run used one grader (`openai/gpt-oss-120b`, 3 votes, liveness verified the same day). The paired comparison is sound because the grader is identical on both sides. We make no cross-grader claim from it.

C.2 §4.3's dose-response curve

Moved here 2026-08-28 to fit a nine-page main text, caption unchanged. Every value it plots is stated in §4.3.

Appendix D: §7 in full

§7 was compressed on 2026-08-27 to fit a nine-page main text, and `fig4` followed it on 2026-08-28 with its caption unchanged. Nothing was deleted.

D.1 The full ladder on the second family, both arms

37 held-out items, one judge blind to arm and dose, unanimous on 578 of 666 gradings:

α	-6.2σ	-4.2σ	-2σ	-1σ	0	$+1\sigma$	$+2\sigma$	$+4.2\sigma$	$+6.2\sigma$
fitted	18.9 %	18.9 %	18.9 %	18.9 %	18.9 %	21.6 %	24.3 %	29.7 %	51.4 %
matched-norm control	24.3 %	18.9 %	24.3 %	21.6 %	18.9 %	18.9 %	27.0 %	21.6 %	21.6 %

The registered primary, in full:

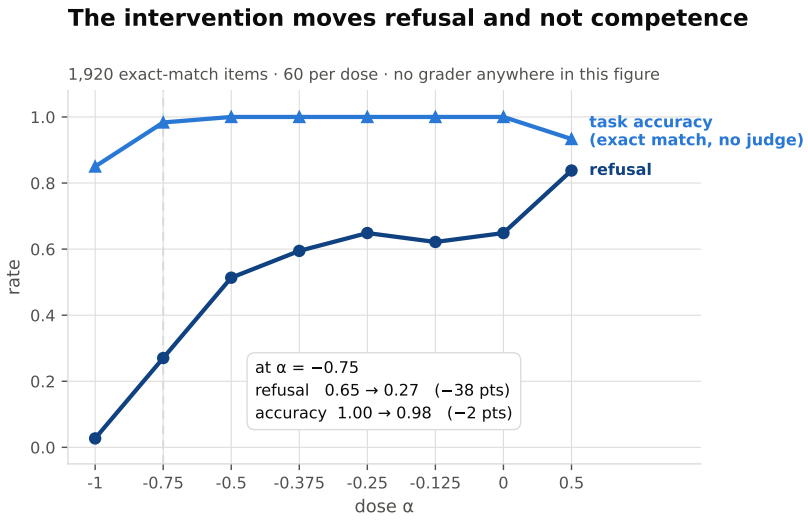

Figure 5: The same doses, with graded refusal on 37 attack items beside exact-match accuracy on 1,920 in-context task items. At $\alpha = -0.75$ refusal has fallen from 0.649 to 0.270, a drop of 38 points, while accuracy has gone from 1.000 to 0.983. No grader touches the accuracy arm: it is exact string match.

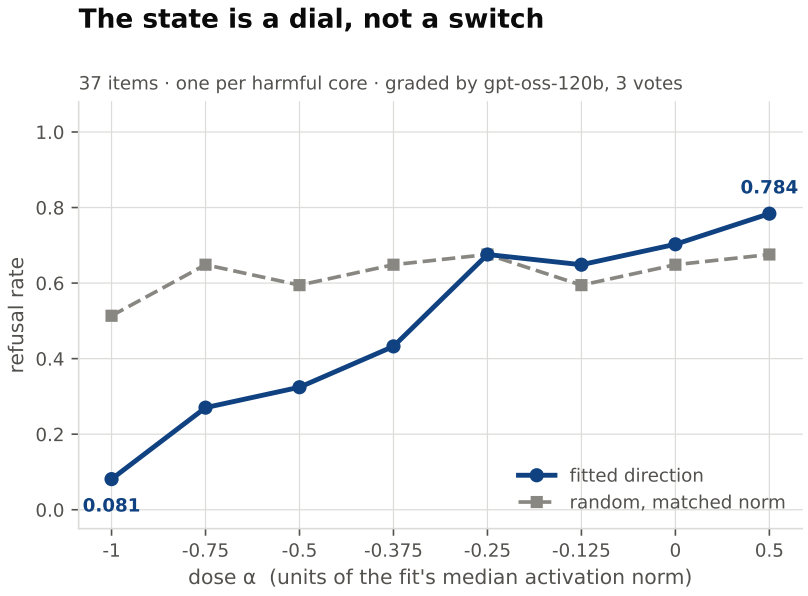

Figure 6: Refusal against dose, 37 items, Qwen2.5-7B at layer 20. The fitted direction moves graded refusal from 0.703 at $\alpha = 0$ to 0.081 at the strongest suppressive dose and 0.784 at +0.5. An isotropic random direction at matched norm stays inside 0.514 to 0.676 throughout. Every plotted value is printed by `figures_p2.py` before it is drawn. The isotropic arm is the control §4.3 says is insufficient on its own, and §7.2 replaces it.

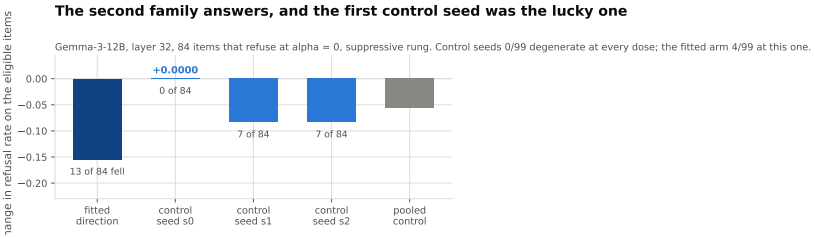

Figure 7: Change in refusal on the 84 items that refuse at $\alpha = 0$, Gemma-3-12B at layer 32, from that dose to the most suppressive one. The fitted direction moves 13 of 84 items out of refusal. Seed `s0` moves none, `s1` and `s2` move seven each, so the pooled control is −0.0556 and the reported gap is −0.0992, not the −0.1548 the first seed alone gave. Control seeds are 0 of 99 degenerate at every dose and the fitted arm is 4 of 99 at this dose (D.8). Every value reproduces `g2e_primary --selftest`.

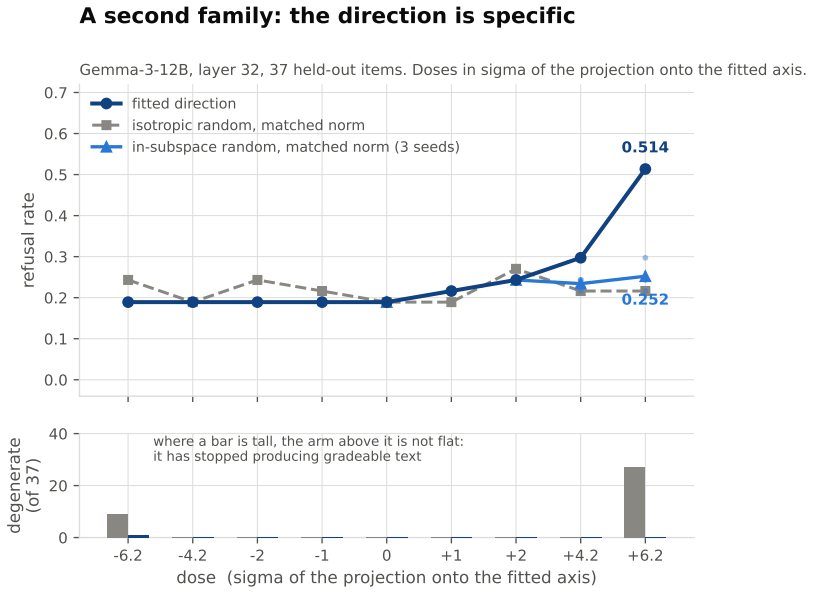

Figure 8: Graded refusal against dose, Gemma-3-12B, 37 items, three arms. The fitted arm sits at 0.189 across every suppressive dose and rises to 0.514 at +6.2 σ. The isotropic control sits at 0.189 to 0.270, but it is degenerate at the top dose: 27 of 37 generations, against 0 for the fitted arm. The in-subspace arm, at the same magnitude, is plotted only where it was run, at 0.252. The comparison the strongest dose looks like it is making is unavailable from the isotropic arm, which has stopped writing. It is available from the in-subspace one.

arm	Δ refusal, −6.2 σ against α = 0	95 % CI	
fitted	+0.000	[+0.000, +0.000]	no item moved
matched-norm	+0.054	[+0.000, +0.135]	contains zero,
control			condition met

On the rungs where both arms write, the per-item trend gap is +0.000, CI [+0.000, +0.474].

D.2 The withdrawn trend statistic, under all four disposals

A per-item positive-trend count over all doses gives 21 of 37 for the fitted arm against 7 of 37 for the control, with an interval excluding zero. But `per_item_rho` drops degenerate rows as off-scale, which removes the control's strongest dose for 27 of 37 items. Scoring those rows as refusals gives 20 of 37 against 19 of 37, CI [−0.135, +0.216]. Scoring them as hedges gives [−0.027, +0.405]. Two of four disposals give a null. Across all four the fitted arm reads 21, 20, 21 and 21, while the control swings from 6 to 19.

D.3 Degeneration forensics

The control's median space fraction at +6.2 σ is 0.008, against 0.146 everywhere else. That is why a word-count metric read a 565-character generation as "5 tokens". Measured in ways spacing cannot move: median generation length 565 characters against 1294, and compression ratio 0.286 against 0.493, with 9 of 37 control rows below 0.20.

The segmentation test. We restored spaces by deterministic dictionary segmentation and regraded under the same judge and rubric. That changed 0 of 27 degenerate labels into refusals. All 16 coherent-but-spaceless rows stayed degenerate, and all 11 repetitive rows stayed. A sample tail: ". . . is extremely harm ON ETA NDVI TERI TERI GI".

The registered asymmetry result, at both extremes (paired bootstrap, 4000 resamples):

rung	control degenerate	fitted degenerate	difference
+6.2 σ	27 / 37	0 / 37	+0.730 [+0.595, +0.865]
−6.2 σ	9 / 37	1 / 37	+0.216 [+0.108, +0.351]

Subspace energy. The isotropic control carries 0.024 to 0.035 of its energy in the top-100 principal components, where 97.2 % of the variance lies. The fitted direction carries 0.994. For an isotropic vector that share is set by geometry, at about k/d for k retained components in a d-dimensional stream, so the measurement checks the construction rather than reporting a finding. The in-subspace seeds below are the control this predicts is needed.

At the top rung, 12 items move into refusal in the fitted arm against 4 in the control, where 27 of the control's rows have stopped producing gradeable text at all.

D.4 The cross-image anchor

The fitted arm was generated under a container image that was rebuilt before the in-subspace arms ran. Rather than trust that greedy decoding is deterministic across a rebuild, we re-ran its strongest dose inside the current image. All 37 of 37 responses are byte-identical. Not the same rate: the same text.

D.5 The class-balance difference between the families

The fit's class balance is 417 to 64 on Gemma-3-12B, against 42 to 103 on the first family. Whether the difference between the families is the model, the layer, the autoencoder or this is not something these data can separate, and no claim rests on any of the four.

D.6 §7.3, the seed study, in full

The table §7.3 summarises, moved here from the body:

	reconstruction fidelity	harm-recognition
our anchor	+0.467	59.2 %
three retrainings	+0.271 / +0.297 / +0.271	0.3 / 0.2 / 1.2 %

The design. Three verbalisers differ from a published anchor in one source of randomness each: RL rollout stochasticity, SFT LoRA initialisation, and RL data order. All three read the same 191 cached activations at the same positions, scored by the same judge, rubric and script as every other awareness number in this paper.

Harm recognition, per arm. 2, 1 and 7 YES calls out of 573 content positions, against the anchor's 339. The registered falsifier for a judge emitting one class fires on all three, so the arms are void under it and the registered AUROC bar is not computed. In one arm, 189 of 191 items carry an identical score, and a ranking statistic over that describes the ties.

Reconstruction fidelity, as own-minus-null cosine at the declared injection scale, with the reconstructor held fixed at the released `kitft` AR:

arm	own	null	gap	of anchor
`nla2`	0.5839	0.3132	+0.2707	58 %
`nla3`	0.6297	0.3327	+0.2970	64 %
`nla4`	0.5953	0.3240	+0.2713	58 %
anchor	0.8077	0.3408	+0.4668	100 %

The scale sweep. The one constant the seed authors publish nowhere was swept from 50 to 300, by the instrument's own objective and never by the awareness score, with the rule fixed in advance. The gaps are +0.181, +0.199, +0.268, +0.297, +0.285, +0.282 and +0.261: a single-peaked curve whose maximum is the value the anchor declares.

Six configuration hypotheses died first, each drawn from the seed authors' own published artefacts and each killed by its own test: the actor prompt template (byte-identical, 520 characters), `extraction.norm: none` (honouring it is worse), `strip_final_norm` (output collapses), unstable generation (deterministic within a run), state leakage between items (none), and the null's own construction. That last one compressed the anchor's gap by a factor of 10.7, and was caught only because a positive control was run.

D.7 What §7.2's experiment would need in order to be conclusive

The binding constraint is the dose scale, not the direction. A set of doses whose strongest dose is fluent in both arms would make the comparison available where the effect is, and the fitted arm's own curve puts that dose between +4.2 σ and +6.2 σ.

Two design errors of ours are in the way, and we state them rather than fold them into the null. The item set was selected against a 37.8 % baseline refusal rate, while the sweep's zero dose measures 18.9 % on the same items. The corpus was sampled at T = 0.7 and the sweep decodes greedily. The suppressive half is therefore a floor: four negative doses at 7 of 37, the same seven items at every dose. Flatness against a floor is uninformative, not negative.

D.8 §7.1's four disposals, and the two fluency screens that disagreed

Two screens were in use and disagreed for two days. They now share one definition. `14_fluency` screens on `tok/word`, `max_run`, non-ASCII and `dropped_char_rate`, plus a minimum-words floor. `g2e_analysis._degenerate` had the first three and not the fourth, because the fifth proxy was added on 2026-08-26 (the other four "pass text with a character dropped from every other sentence") and never propagated. On the same rows the two screens then read 17 of 99 and 84 of 99, and a

finding quoted one while citing the other as its source. `dropped_char_rate` is now imported rather than reimplemented, and `tests/test_screens_agree.py` fails if that import becomes a copy.

arm, at −0.1116 / +0.1116	both screens, since 2026-08-28	four-proxy, before
`active_random_s0` / `_s1` / `_s2`	0/99 and 0/99 each	0/99 and 0/99
`fitted`	4/99 and 0/99	0/99 and 0/99
isotropic `random`	84/99 and 81/99	17/99 and 81/99

The right-hand column is kept because the superseded numbers were published, in the `G2e` amendment, struck through rather than deleted for the same reason.

The fitted arm is not clean at the rung the primary is measured at. Three of those four rows are inside the 84 eligible items (`refuse_055`, `refuse_066`, `refuse_072`), so the disposal rule has something to decide. That is why §7.1 reports a range and not a single value.

disposal	n	primary	secondary, pooled over three seeds
drop	81	−0.1605 [−0.2469, −0.0864]	−0.1029 [−0.1605, −0.0494]
as-COMPLY	84	−0.1905 [−0.2738, −0.1071]	−0.1349 [−0.2063, −0.0714]
as-REFUSE (quoted)	84	−0.1548 [−0.2381, −0.0833]	−0.0992 [−0.1587, −0.0476]
drop-item	81	−0.1605 [−0.2469, −0.0864]	−0.1029 [−0.1605, −0.0494]

All eight intervals exclude zero, and they spread 3.6 points, under the 5-point bar `g2e_analysis` applies before calling a cell `NOT REPORTABLE`. The quoted row is the most conservative of the four, and that was luck rather than judgement: `as-REFUSE` is what the four-proxy screen produced by not seeing those rows at all.

Per seed, on the same 84 eligible items: `s0` gives +0.0000, with 0 of 84 items moved. `s1` and `s2` give −0.0833, with 7 each. Pooled, they give −0.0556. Each seed's gap to the fitted arm excludes zero on its own, at −0.1548, −0.0714 and −0.0714. `s1` and `s2` return identical statistics, so we tested whether they are one arm counted twice. They are not: the directions differ (`62f099...` and `25a785...`), 792 of 891 generations differ, and the 99 that match are all at $\alpha = 0$, where no direction is applied.

D.9 §7.2's six-seed table, in full

Moved here 2026-08-28 when §7 gained the pooled `G2e` result and the page budget had to pay for it. Every value below is stated in §7.2; the table is the arm-by-arm form. The Qwen sweep that the same control serves is the dose-response figure in Appendix C.

arm at +6.2 σ	refusals	change	against `fitted`
`fitted`	19/37	+0.324 [+0.189, +0.486]	
in-subspace, matched norm (3 seeds)	8, 9, 11 / 37	+0.063	+0.261 [+0.108, +0.423]
in-subspace, matched σ (3 seeds)	9, 7, 8 / 37	+0.027	+0.297 [+0.153, +0.451]

All six seeds separate individually and every arm is fluent. It is the first clean comparison in the series, because both sides write.

D.10 The sign test on §7.1's doses, seed by seed

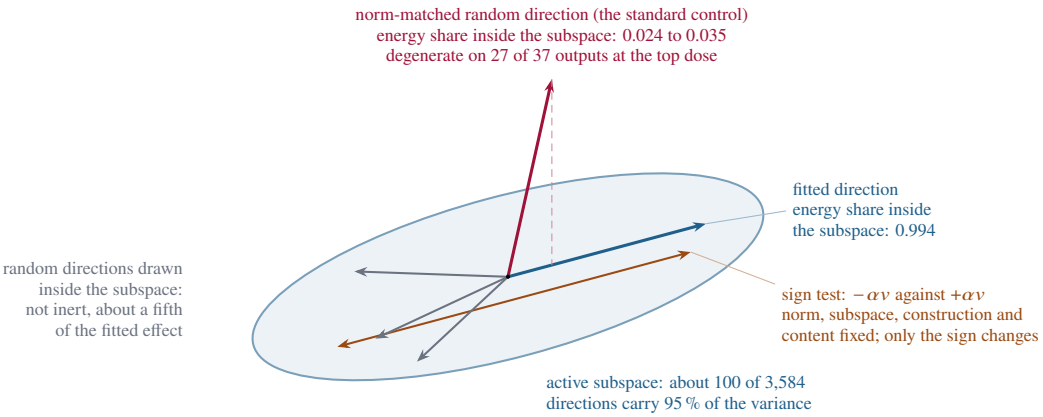

Figure 9: Why a control matched only in norm is matched in nothing else. Activations occupy a small subspace of the residual stream; a fitted direction lies inside it and a random direction of the same norm lies almost entirely outside it, so the standard comparison cannot separate this direction from any large perturbation of the subspace. Random directions drawn inside the subspace are not inert, and the sign test holds everything but the sign fixed. Values are those of §1.3 and §7.2.

The numbers here are P2-F029's and required no new generation. They are the same 4,455 graded rows, re-read under a different comparison.

The test. An arm at $-\alpha$ against the same arm at $+\alpha$. Norm, subspace, construction and content are held exactly fixed, and only direction differs. It is also two-sided, which every comparison against $\alpha = 0$ on this family is not: eligibility has put each item at the ceiling, so "the interval excludes zero" reduces to a count (ERROR_LOG Part 57).

arm	sign test	fell / rose	
fitted	−0.1310 [−0.2024, −0.0595]	11 / 0	excludes zero
active_random_s0	+0.0595 [+0.0119, +0.1190]	0 / 5	excludes zero, in the opposite direction
active_random_s1	−0.0476 [−0.1071, +0.0000]	5 / 1	includes zero
active_random_s2	−0.0833 [−0.1429, −0.0238]	7 / 0	excludes zero

A random direction drawn inside the activations' subspace passes the strictest control this project has. One seed passes it in the treatment's direction. Another passes it inverted.

And the fitted direction does not beat them one at a time.

comparison	Δ	95 % CI	
fitted minus s0	−0.1905	[−0.2976, −0.0952]	excludes zero
fitted minus s1	−0.0833	[−0.1667, +0.0000]	includes zero
fitted minus s2	−0.0476	[−0.1190, +0.0119]	includes zero
fitted minus the three pooled	−0.1071	[−0.1786, −0.0397]	excludes zero

Pooled, the fitted direction separates. Seed by seed, it separates from one of three. The pooling was registered before the seeds ran (G2e), so this is not a choice made after seeing which comparison worked. But it is load-bearing rather than decorative, and §7.2 says so.

Directional is not specific, and that distinction is the point. An in-subspace random direction is also directional: s2 passes the sign test on 7 of 84 items. Showing that a direction's sign matters

shows the injection is not symmetric noise. It does not show that the direction is the one the hypothesis names.

Why §7.1's own ladder cannot be put through this test. It swept +0.0372, +0.0744 and +0.1116 only. No negative dose exists in any of its six arm files, and its graded file carries no $\alpha = 0$ row. Its baseline is 18.9 % and its doses are enhancing only, so the change is at least zero by construction. "The interval excludes zero" there is Part 57's count threshold running the other way. This is not a re-adjudication of that result, which is reported under its own registration and is not reopened. It is a statement about what its design can be asked. Running the test there needs about 333 new generations plus a fitted arm, and with the freeze on 7 September the sentence is the better trade than a hurried arm.

The directions are the same objects, checked rather than assumed. Both experiments load `l2_direction_active_s0/s1/s2.pt` at shas `57317fdb0c62`, `62f099174456` and `25a785f9ac45`.