Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

When Does a Readout Reflect Computation? Dose-Response Interventions in Continuous Thought Models

ChaeWoo Son

soncwnetp@gmail.com

Independent Researcher https://orcid.org/0009-0009-6108-2373

Research Article

Keywords: interpretability, mechanistic interpretability, latent reasoning, chain of continuous thought, causal intervention, logit lens

Posted Date: August 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10666101/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

When Does a Readout Reflect Computation?

Dose–Response Interventions in Continuous Thought Models

ChaeWoo Son
Independent Researcher
soncwnetp@gmail.com

Abstract

Chain of Continuous Thought (Coconut) performs computation in a continuous latent space, which makes its intermediate steps difficult to monitor directly. Work seeking to interpret it relies on projecting latent states into vocabulary space, and treats a change in that readout as evidence that the underlying computation has changed—an inference that has not itself been characterized.

We measure it. Using the J-lens readout as a training signal, we plant bridge entities in latent states to obtain a testbed in which the readout is controlled, then apply a dose–response protocol that fixes the magnitude of each intervention. Bridges are planted without loss of accuracy (100% answer accuracy, 0.999 readout probability). We decompose the causal intervention delta into the component the lens detects and the component it does not, and control with a random direction matched on both norm and lens-invisibility.

Across the two relation types used for intervention, three seeds and a 2B model, the readout and the computation have different thresholds along the same direction. The readout flips at relative perturbation magnitudes of 0.13–0.16, but the answer does not move there; the answer threshold is 2.03× higher on one relation and is never reached before the model breaks on the other. Conversely, along a direction to which the readout does not respond once in 1,800 trials, 96–97% of answers are transplanted. The random direction matched on magnitude and invisibility produced no readout change, no answer change and no breakage in the same 1,800 trials.

This implies that minimum-norm readout-targeted intervention stops at the readout threshold and therefore measures below the causal threshold—at half of it on the one relation where that threshold is reached at all. Our own initial experiment produced a false null of 0/90 in this trap, with a stopping magnitude of 0.119 against a readout threshold of 0.125. Readout-based interpretability claims should be reported as dose–response curves rather than at a single magnitude.

1 Introduction

Chain of Continuous Thought (Coconut) performs reasoning in a continuous latent space rather than in language, feeding the model's last hidden state back as the next input embedding instead of decoding it into a token [3]. The design carries expressivity advantages, but because no intermediate step is emitted in natural language, what the model is computing cannot be observed directly.

The standard instrument for observing it is the *readout*: a projection of the hidden state into vocabulary space [4]. The original Coconut paper presented a case study decoding continuous thoughts into tokens, reporting that the first thought corresponds to the first intermediate variable of the calculation [3], and subsequent work has systematized this approach, recovering gold reasoning

traces from a majority of correctly predicted instances [2]. Such results are interpreted under an inference of the form: **if the readout changed, the computation changed.**

This inference has been examined before. The Tuned Lens identified the residual-stream directions most influential on the lens output and showed that ablating them from the model's hidden states disproportionately affects the model's output as well [1]. But such verification is carried out *at a single intervention magnitude*. If the readout and the computation have different sensitivities to the same direction, a measurement at one magnitude cannot reveal the difference.

This work measures that sensitivity as a function of magnitude. Doing so requires two things: a model whose readout is known to point at a specific token, and a procedure that controls the magnitude and the direction of an intervention independently.

For the first, we use the J-lens readout as a training signal to plant the bridge entity in the continuous thought. The answer loss is left unmodified; the result is a readout that places over 0.95 probability mass on the bridge while answer accuracy is retained (§3). This model is not a claim of the paper but the apparatus that makes the measurement possible.

For the second, we fix the relative magnitude of the intervention at α and split the causal delta—the perturbation that transplants the answer in full—orthogonally into the component the lens detects and the component it does not. The two components sum exactly to the original delta, so no magnitude is manufactured. As a control we use a random direction matched on both norm and lens-invisibility (§4.1).

The readout and the computation turn out to have different thresholds along the same direction. The readout flips on every pair at relative magnitudes of 0.13–0.16, yet the answer does not move at that point. The answer threshold is 2.03× the readout threshold on one relation, and on the other it is never reached before the model breaks. The error runs in the opposite direction as well: along the component the lens does not detect, 96–97% of answers are transplanted while the readout neither points at the donor nor departs from the recipient in 1,800 trials. The matched random direction has no effect in the same trials.

This structure explains how a standard intervention test can mislead. A minimum-norm intervention that targets the readout is designed to stop where the readout flips, and on lmk that point is half the causal threshold; on typeA the causal threshold is never reached at all. Our own initial experiment fell into exactly this trap and produced a false null of 0/90; its stopping magnitude (0.119) coincides with the readout threshold (0.125) (§4.4). The bias is not specific to the J-lens or to this task. Any procedure that adopts readout change as its stopping condition measures systematically below the causal threshold, in proportion to the readout's greater sensitivity.

The threshold gap is not a product of J-lens supervision: it persists at 1.74× in the unsupervised condition (Appendix A).

Contributions.

1. A **dose–response intervention protocol** that separates the sensitivity of the readout from that of the computation: fixed magnitude, orthogonal decomposition of the causal delta, and a random control matched on both norm and lens-invisibility.
2. **Quantification of a two-sided dissociation** across the two relations used for intervention, three seeds and a 2B model. The magnitude at which the computation responds varies with the task in a way the readout threshold does not account for, and the computation can be transplanted along a channel to which the readout does not respond at any magnitude.

3. A **diagnosis of minimum-norm readout-targeted intervention**, showing that the procedure produces false nulls by construction, with our own pre-registered test as the worked example.
4. A **controlled testbed** enabling the above: J-lens supervision aligns the readout without loss of accuracy and requires no additional human annotation of bridge entities.

2 Background

2.1 Continuous thoughts in Coconut

Coconut treats the last hidden state as a representation of the reasoning state and feeds it back directly as the next input embedding rather than decoding it into a language token [3]. We refer to this state as a *continuous thought*.

Concretely, starting from the prompt embeddings E, for $t = 0, \ldots, k-1$ the model computes

$$c_t = \mathrm{LM}\big(E \oplus c_0 \oplus \cdots \oplus c_{t-1}\big)^{(\text{final layer})}_{\text{last}} \tag{1}$$

and appends c_t to the sequence. After k iterations the answer is read from a **separate final forward pass**. Any perturbation applied to a continuous thought therefore reaches the answer only through the feedback path.

Training proceeds by curriculum: early stages express intermediate reasoning explicitly in text, and successive stages replace explicit steps with continuous thoughts. We use $k = 2$ throughout.

The interpretability problem. Continuous thoughts are never decoded into vocabulary and so cannot be read directly. This is the monitorability problem of latent reasoning models generally, and it is the object of this work.

2.2 Readouts

Projecting a hidden state into vocabulary space to obtain a distribution is what we call a *readout*. In its simplest form (the logit lens; 4) the hidden state is multiplied directly by the unembedding matrix; refined variants first apply a transformation learned per layer [1].

We use a Jacobian lens. Writing $h \in \mathbb{R}^d$ for the last-position residual at layer L and J_L for that layer's Jacobian, the readout distribution is

$$p(h) = \mathrm{softmax}\big(\mathrm{unembed}(J_L h)\big), \tag{2}$$

and the *readout token* is $\arg\max_v [p(h)]_v$.

Two properties matter for what follows.

Nonlinearity. $J_L h$ is linear in h, but unembed applies RMSNorm internally, so $h \mapsto$ logits is not. An intervention that moves the readout toward a particular token therefore cannot be constructed as a linear inner product (§4.1), and the readout direction is state-dependent and must be recomputed at each step.

Freezing. The lens is fitted to the base model in advance and never updated thereafter. The readout in this work is thus a **fixed probe**, not one that co-adapts with the model being trained. This is deliberate: were the probe to move with the model, a change in the measured threshold could not be attributed to the model rather than to the probe. The limitation this incurs is discussed in §4.5.

2.3 From readout to computation

Interpretation via readouts rests on an inference of the following form.

> If, after some intervention, the readout changes from b to b', then the intermediate representation the model uses has changed correspondingly.

For this to hold, the component the readout reads must correspond to the component the model uses. Prior work has verified this correspondence [1], but such verification is typically conducted at a single intervention magnitude. This work measures the correspondence **as a function of intervention magnitude** (§4).

2.4 Notation

Symbol	Meaning
x, b, y	head entity, bridge entity, answer
h	last-position residual at layer L
J_L	Jacobian lens at layer L (frozen)
k	number of continuous thoughts ($k = 2$)
α	*relative magnitude* of an intervention, $\|\delta\|/\|h\|$
u	normalized gradient of the lens contrast (§4.1)
d	causal delta, $h_{\text{don}} - h_{\text{rec}}$ (§4.1)
recipient / donor	the item receiving an intervention / the item it is drawn from

3 A Controlled Testbed

The intervention experiments of §4 require a model in which what the readout points at is under control. This section constructs that model and verifies that the preconditions of the intervention experiment hold. The training procedure presented here is not a claim of this paper but the apparatus that makes the measurement possible; claims about the relationship between readout and computation are made in §4.

3.1 Task and data

Two-hop factual recall. Each item consists of a head entity x, a bridge entity b, and an answer y. The model must produce y from x; b never appears in the prompt. We use three relation types:

- **typeA**: capital → country → continent (22 chains, 3 distinct answers)
- **lang**: capital → country → language (18 chains, 18 distinct answers)
- **lmk**: landmark → country → capital (15 chains, 13 distinct answers)

Which relation is used where. The three are not interchangeable, and we use different pairs of them for the two purposes of this paper.

Testbed validation (§3.2–§3.4) uses typeA and lang, which differ sharply in answer-space size. Emitting the modal answer as a constant yields 50% accuracy on typeA but only 6% on lang. lang therefore confirms that readout alignment is not explained by a collapse of the answer space.

Intervention experiments (§4) use typeA and lmk. lang is excluded here for a specific reason: in 17 of its 18 chains the bridge and the answer share an orthographic prefix (Germany → German), so an intervention that transplants the answer cannot be distinguished from one that transplants a surface string. This makes lang informative about generalization but unsuitable as a causal testbed. lmk replaces it because its bridges are least reachable by surface association among the three.

We report no held-out numbers for lmk in §3.2–§3.4. Under the bridge-group split described below, lmk is the only relation whose split changes (two of its bridges appear under two head entities each), and we did not re-run the held-out evaluation after that change. This does not affect §4, whose intervention pairs are drawn from training items in all cases.

Single-token constraint. We retain only chains whose bridge and answer both tokenize to a single token. This is required in order to define the readout as a top-1 token and to read the answer at a single position; extension to multi-token answers is left open (§4.5).

Splitting. We split by **bridge entity rather than by item**: all chains sharing a bridge are assigned to the same side. This prevents the same (bridge, answer) pair from appearing on both sides via different head entities, so that a held-out claim is not in fact a claim about an already-supervised fact. On typeA and lang every bridge is unique, so the split coincides with an item-level split; on lmk two bridges appear under two head entities each, and the two splits differ. Immediately after splitting we assert that no (bridge, answer) pair is shared between the two sets; the run aborts if the check fails.

Limits of scale. Each relation has 15–22 chains with 4–6 held out. This scale sharply limits the precision of any accuracy claim (§3.3); we note in §4 that the sample for the intervention experiments is the number of intervention trials, not the number of chains. We attempted to extend to naturally occurring multi-hop data but encountered a tension: the facts a small model knows are largely reachable through surface associations, while facts without such shortcuts are not known to the model. This trade-off is treated as a limitation in §4.5.

3.2 Supervision aligns the readout

Metric. To measure bridge alignment in the continuous thought we use

$$p_{\text{bridge}} = \left[\text{softmax}(\text{unembed}(J_L h_0))\right]_b, \tag{3}$$

where h_0 is the last-position residual at layer L at the *first* latent step and b is the gold bridge token. This is the probability mass the readout distribution assigns to the bridge, which is a different quantity from whether the bridge is top-1. The two move together in strongly aligned regimes but are not identical, and the intervention experiments of §4 use the top-1 readout. We report p_{bridge} here because it is the exact exponential transform of the J-lens loss used in training, and thus reports precisely the quantity the supervision targets.

Objective. We add a J-lens term to Coconut's answer loss:

$$\mathcal{L} = \mathcal{L}_{\text{ans}} + \lambda\, \mathcal{L}_{\text{J-lens}}, \qquad \mathcal{L}_{\text{J-lens}} = \text{CE}\big(\text{unembed}(J_L h_t),\, b\big), \tag{4}$$

computed at each latent step t and averaged. **The answer loss is left unmodified.** Setting $\lambda = 0$ is identical to standard Coconut training at the level of the computation graph, and we use it as our baseline.

Supervision layer. L is selected automatically at roughly 80% of model depth, clamped to the range of layers to which the lens was fitted. For the 2B model this gives $L = 19$ of 24 layers.

Results. Table 1 reports held-out p_{bridge}.

Table 1: Held-out p_{bridge} by condition (mean ± s.e. over 3 seeds).

Relation	Answers	Baseline (λ=0)	Supervised (λ=1)
typeA	3	0.766 ± 0.010	0.952 ± 0.003
lang	18	0.015 ± 0.011	0.975 ± 0.002

The two baselines differ sharply. On typeA the readout is already substantially aligned with the bridge without any supervision; on lang it is essentially unaligned. Supervision raises both above 0.95. The increase is $t(2) = 85.9$, $p = 0.0001$ on lang and $t(2) = 17.8$, $p = 0.0014$ on typeA.

The baseline difference shows that **the natural alignment of the readout is task-dependent.** Where the answer space is small, the continuous thought stays near the bridge; where it has 18 members, it does not. The intervention experiments of §4 are run after supervision has brought the readout into an aligned state, so that what the readout points at is known in every condition.

On the statistics. The t values above are computed over 3 seeds and carry only 2 degrees of freedom. Read against a normal distribution the same statistics would yield $p < 10^{-70}$, a precision the sample size does not support. Furthermore, since the split is fixed, the across-seed standard error contains no item-sampling variance. We use seed-level tests only as statements about whether an effect exceeds seed noise, and make no generalization claims from them.

3.3 Accuracy is retained

Measurement. For each held-out item we run $k = 2$ continuous thoughts and read the answer token from the final forward pass over the whole sequence. Accuracy is the fraction matching the gold answer.

Results. Held-out accuracy was 100% in every run, across both relations, all conditions and all 3 seeds (Table 2).

What these numbers do and do not support. With 5–6 held-out items, observing all correct excludes only true accuracies below roughly 57–61% (Wilson interval). The strong statement "there is no loss of accuracy" is therefore *not* supported by this data. What is supported is that **supervision did not degrade accuracy to an observable degree**, which suffices to confirm that the intervention experiments of §4 are not conducted on a model that fails to produce answers.

Table 2: Held-out accuracy and the Wilson interval it supports.

Relation	Held-out size	Baseline	Supervised	95% CI on true accuracy
typeA	6	6/6	6/6	[61.0%, 100%]
lang	5	5/5	5/5	[56.6%, 100%]

The role accuracy must play as a precondition for §4 is the following: the intervention experiment must be able to distinguish the recipient's answer from the donor's, and the model must produce the recipient's answer absent intervention. This condition is verified on the full training set used in §4 (all 15–22 chains per relation correct), a larger sample than the held-out set.

A structural reason. The answer loss is unmodified by the addition of the J-lens term (Eq. 4). That supervision does not damage accuracy is therefore not merely an observation but a direct consequence of the loss construction. Larger λ could put the two objectives in competition; no such competition was observed at $\lambda = 1$. We did not test λ sensitivity (§4.5).

3.4 The model can supply its own bridge labels

The supervision of §3.2 requires a gold bridge. This section examines whether that label must come from a human.

Procedure. Before training, the base model is given a bridge-eliciting prompt (e.g. "*Fact: {capital} is the capital city of the country of*"). The top-1 J-lens readout at the final position is taken as a candidate label and accepted if its probability exceeds τ. Chains whose labels are not accepted are trained with the answer loss alone, without a J-lens term.

At label-generation time the LoRA parameters are forced to zero, guaranteeing that labels come from the untrained base model. We assert at runtime that the label prompt does not contain the gold bridge; the run aborts if the check fails.

Results. Labels were accepted for every item on both relations, and every accepted label matched the gold bridge (Table 3). The self-labelled and gold-labelled conditions are consequently identical to three decimal places. We used $\tau = 0.7$; coverage and precision both remained at 100% throughout $\tau \in [0.5, 0.9]$.

Table 3: Self-generated bridge labels versus gold labels.

Relation	Coverage	Precision	Self-label p_{bridge}	Gold-label p_{bridge}
typeA	100%	100%	0.952	0.952
lang	100%	100%	0.975	0.975

The scope of this result. The procedure removes **the bridge annotation required by the J-lens term.** It does not make the entire training curriculum bridge-free. Coconut's explicit-reasoning stage (zero continuous thoughts) teacher-forces the bridge as text, and the warm-up stage together with mixed revisits accounts for roughly 40% of all training steps. Because this stage is part of the standard Coconut curriculum and applies *identically to the baseline condition*, the difference

between the two conditions is obtained without additional annotation. We therefore describe the procedure as showing that **readout alignment requires no additional human annotation**, not that the readout can be aligned with no bridge annotation at all.

Replacing the explicit-reasoning stage with self-labels as well is implemented but not evaluated in this work.

4 Readout and Computation Come Apart

4.1 Intervention protocol

Lens direction. Let rec and don denote the bridge tokens of the recipient and donor items. Define the contrast scalar and its normalized gradient

$$s(h) = \big[\text{unembed}(J_L h)\big]_{\text{don}} - \big[\text{unembed}(J_L h)\big]_{\text{rec}}, \qquad u(h) = \frac{\nabla_h s(h)}{\|\nabla_h s(h)\|}. \tag{5}$$

u is the unit direction along which the readout tilts fastest toward the donor. The choice of this contrast is not arbitrary: it is exactly the objective ascended by the standard intervention we examine in §4.4, so both sections operate on the same notion of direction. Since s is nonlinear in h (§2.2), u is state-dependent and is recomputed at each latent step.

Fixing the magnitude. For a unit direction $\hat{v}$ and scalar α we apply

$$\delta = \alpha \cdot \|h\| \cdot \hat{v}, \qquad \text{so that} \quad \|\delta\|/\|h\| = \alpha. \tag{6}$$

We use relative rather than absolute magnitude because $\|h\|$ varies across layers, positions, items and latent steps; a fixed absolute norm is dominant at one site and negligible at another.

δ is recomputed at every latent step against that step's h. At step t=1 the input already contains the continuous thought altered at t=0, so h differs across conditions. Fixing δ once lets the relative magnitude drift apart across trajectories—we measured a spread of 0.25–0.80 across conditions when doing so. Per-step recomputation holds the relative magnitude at exactly α on every trajectory.

Interventions are applied through a forward hook on the output of layer $L-1$, at the last position only, inside each latent step. After the latent loop terminates, the answer is read from a **separate final forward pass with the intervention disabled** (§2.1). The perturbation therefore reaches the answer only through the altered continuous thought.

Decomposing the causal delta. For the recipient state h_{rec} and the donor state h_{don} at the same step, let

$$d = h_{\text{don}} - h_{\text{rec}}, \qquad d_{\parallel} = (d \cdot u)\, u, \qquad d_{\perp} = d - d_{\parallel}. \tag{7}$$

Applying d in full transplants the answer completely (§4.2). $d_{\parallel}$ is the component the lens contrast detects; $d_{\perp}$ is the component it does not. Because

$$d_{\parallel} + d_{\perp} = d, \qquad d_{\perp} \cdot u = 0, \qquad \|d\|^2 = \|d_{\parallel}\|^2 + \|d_{\perp}\|^2, \tag{8}$$

the decomposition **neither creates nor discards magnitude.** The two components are an orthogonal split of a causal delta, not an artificially amplified perturbation.

Matched random control. The control direction is built as

$$r \sim \mathcal{N}(0, I), \qquad r \leftarrow r - (r \cdot u)\, u, \qquad r \leftarrow \frac{\|d_\perp\|}{\|r\|}\, r. \tag{9}$$

The projection removes any lens-visible component; the rescaling matches the magnitude of $d_\perp$. The result is a direction that agrees with $d_\perp$ on *both* invariants—norm and lens-invisibility—and differs only in direction. If $d_\perp$ has an effect and r does not, the difference is attributable to direction rather than magnitude. At $d = 2048$ the expected cosine between a random vector and u is $d^{-1/2} \approx 0.022$, so the projection removes little; making it explicit turns an approximate match into an exact one.

Measured outcomes. Each trial records four quantities: **readout→donor** ($\arg\max p(h + \delta)$ is the donor bridge), **readout=recipient** (unchanged from the recipient bridge), **answer→donor**, and **broken** (the answer is neither the donor's nor the recipient's). The first two are recorded separately because failing to point at the donor and failing to move at all are different facts; this distinction determines the strength of the claim in §4.3.

Note that the direction is defined narrowly—orthogonality is with respect to a single contrast—while the observation is broad, over the full top-1 vocabulary. A component orthogonal to u that nonetheless perturbs the readout in some other way would still register in this measurement.

Samples and grid. We sweep relative magnitude over the grid

$$\alpha \in \{0.10,\ 0.15,\ 0.20,\ 0.22,\ 0.24,\ 0.28,\ 0.32,\ 0.40,\ 0.50,\ 0.60\}.$$

For each relation we sample 30 (recipient, donor) pairs differing in both bridge and answer, giving 90 trials per cell across 3 seeds. Pairs are drawn from the training items, so measurements in this section are in-sample; generalization claims are made in §3, and this section concerns mechanism.

Threshold estimation. We estimate the 50% crossing of each curve by linear interpolation between adjacent grid points,

$$\alpha^* = \alpha_0 + (\alpha_1 - \alpha_0)\, \frac{0.5 - f_0}{f_1 - f_0}. \tag{10}$$

When a curve does not reach 50%, we report it as not reached rather than extrapolating a threshold. The answer curve for typeA is such a case (§4.2); manufacturing a threshold there would be the single most misleading number this procedure could produce.

Procedural verification. We confirmed in every run that the tensor edited by the hook is the same object as hidden_states[L] (maximum absolute difference 0.00e+00). The three decomposition identities, the two matching conditions of the control, and numerical stability in the degenerate case where d is parallel to u were verified over 200 random vectors. Interventions against a fixed checkpoint are fully deterministic: 19 grid cells shared between two separate runs agreed exactly. Training, by contrast, is not exactly reproducible on our hardware (losses differ across runs at identical seed); threshold estimates were nonetheless stable across the three models so obtained (standard error of the readout threshold across seeds: 0.0005).

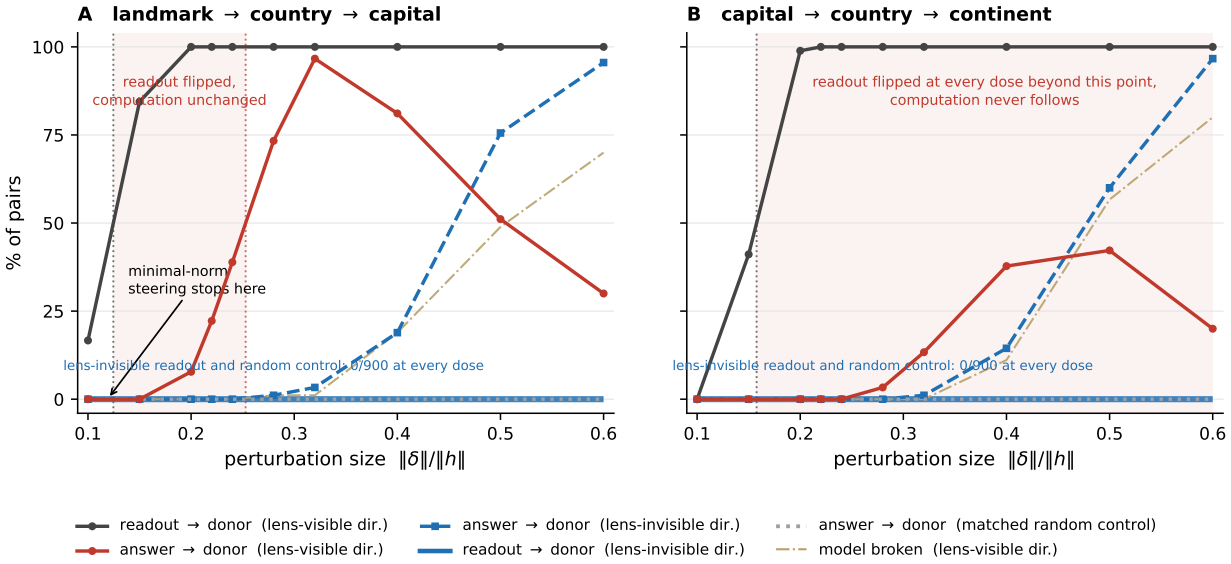


Figure 1: The readout and the computation have different thresholds. Dose–response along the lens-visible direction u and the lens-invisible direction $d_\perp$, on (A) landmark → country → capital and (B) capital → country → continent. Three seeds, 90 pairs per cell. Dotted vertical lines mark the 50% crossings of the readout and answer curves along u; the shaded band is the interval in which the readout has flipped but the computation has not. The lens-invisible readout (solid blue, at zero) and the matched random control (dotted grey, at zero) overlap: neither responds at any magnitude (0/900 each). The decline of the answer curve beyond its peak is a transition into breakage, not a weakening of the effect (dash-dot). The arrow marks the magnitude at which minimum-norm readout-targeted steering stops (§4.4).

4.2 The same direction, two thresholds

Lens-visible direction. Pushing along u moves both the readout and the answer toward the donor, but the two begin to move at different relative magnitudes. On lmk the 50% threshold is 0.125 for the readout and 0.253 for the answer, a **ratio of** 2.03. At $\alpha = 0.20$, where the readout points at the donor for every pair, only 8% of answers have flipped. Both thresholds are stable across seeds; the readout threshold in particular has a standard error of 0.0005 and is effectively constant (answer threshold: 0.0089). The seed-wise mean of the ratio is 2.02 (s.e. 0.077), and the $t(2)$ 95% confidence interval $[1.69, 2.35]$ excludes 1.

The answer threshold varies sharply with the task. On typeA the answer **never reaches majority flipping at any magnitude.** The peak is 42% at $\alpha = 0.50$, and 2 of 3 seeds stay below 50% across the entire grid. Per the convention of §4.1, we report this curve as not reached rather than extrapolating a threshold.

On the same relation the readout threshold is 0.158, close to the 0.125 observed on lmk. Two relations do not license a claim that the readout threshold is task-invariant. What they do establish is that **variation in the readout threshold does not account for variation in the answer threshold.** The readout thresholds differ by 0.033 across the two relations, while the magnitude at which the answer responds is twice the readout threshold on one relation and unreachable before breakage dominates on the other. The sensitivity of the readout does not license an estimate of the sensitivity of the computation.

This asymmetry is the source of the problem examined in §4.4.

We do not report a pooled ratio. The figure "about 2×" obtained by pooling the two relations holds only on lmk, and the task dependence is itself part of the result. All thresholds below are reported per relation (Table 4).

Table 4: Thresholds by relation (relative magnitude, 50% crossing, pooled over 3 seeds).

	Readout (lens-visible)	Answer (lens-visible)	Ratio	Answer (lens-invisible)
lmk	0.125	0.253	2.03	0.455
typeA	0.158	not reached (peak 42%)	—	0.478

At excessive magnitudes, breakage dominates. The answer-flip rate is not monotone beyond the threshold. On lmk it peaks at 97% ($\alpha = 0.32$) and falls to 30% at $\alpha = 0.60$, while over the same interval the breakage rate rises from 1% to 70%. The two rates sum to 100% throughout the supra-threshold region, and the recipient's answer survives in 0% of trials. The decline is therefore a transition into breakage, not a weakening of the effect. This occurs only along the lens-visible direction: neither $d_\perp$ nor the matched random control produces any breakage at $\alpha = 0.60$ (0/900 each).

The usable operating window of this direction is consequently narrow. On lmk, complete transplantation without breakage occurs only between roughly $\alpha = 0.30$ and 0.35.

A check on typeA's answer space. Because typeA has only three answers (continents), a broken output could coincide with the donor's answer by chance and inflate the flip rate. We

checked this in two ways. First, if the perturbation were a constant artefact always producing the modal answer, the expected coincidence rate under our pair-sampling scheme would be 41%, whereas the observed transplantation rate for $d_\perp$ is 97%. Second, this bias inflates flip rates, yet lens-visible flipping on typeA is *lower* than on lmk. The failure of the answer to reach threshold on typeA is therefore not explained by this artefact, and the measurement on this relation is conservative.

4.3 A channel the lens does not register

By construction, $d_\perp$ is orthogonal to the gradient of the lens contrast, while the quantity we observe is the full top-1 readout token (§4.1).

The answer moves and the readout does not. Pushing along $d_\perp$ up to $\alpha = 0.60$ transplants the answer from the recipient's to the donor's (96% on lmk, 97% on typeA; 50% thresholds of 0.455 and 0.478 respectively). Under the same interventions, the readout pointed at the donor bridge in **0 of 900 trials** on each relation (95% upper bound 0.43%), and the top-1 readout departed from the recipient bridge in **0 of 900 trials** as well (900/900 retained; 95% lower bound 99.6% for the true rate). Breakage likewise occurred in 0/900. The lens does not misidentify the donor on this channel; it **registers no change at all** (Table 5).

This statement is unambiguous. Across every direction and magnitude we applied, the top-1 readout was either the donor bridge or the recipient bridge; departures to a third token were absent except at a single grid point in a transition region. "The readout did not change" therefore cannot be confused with "the readout may have pointed at something else."

Magnitude alone does not explain it. The control direction r has the same norm as $d_\perp$ and is equally orthogonal to the lens contrast, differing only in direction (§4.1). It produced no answer flips, no readout changes and no breakage in 900 trials on each relation. Moreover, owing to trajectory divergence, r was applied at a *larger* relative magnitude than $d_\perp$ in the paired-arm measurement (0.80 vs 0.47) and still had no effect. The effect of $d_\perp$ is attributable to direction, not magnitude.

Table 5: Lens-invisible direction and matched control (pooled over 3 seeds, 900 trials per cell).

Direction	Answer→donor	Readout→donor	Readout=recipient	Broken
$d_\perp$ (lmk)	96%	0/900	900/900	0/900
$d_\perp$ (typeA)	97%	0/900	900/900	0/900
Matched random	0/900	0/900	900/900	0/900

How much of the causal delta is removed. A reader may object that if $d_\perp$ constitutes nearly all of d, then transplanting the answer along it is close to applying the full causal delta, and the result is trivial. We measured the norm split directly on a saved checkpoint (lmk, 30 pairs, latent step t=0, where the Pythagorean identity holds exactly):

$$\frac{\|d_\parallel\|}{\|d\|} = 0.329 \pm 0.035, \qquad \frac{\|d_\perp\|}{\|d\|} = 0.944 \pm 0.013, \qquad \cos(d, u) = +0.329. \tag{11}$$

The objection is partly correct on its face: $d_\perp$ is 94% of d by norm and 89% by energy. We state that plainly rather than presenting the decomposition as an even split. Three facts nonetheless close the objection.

First, the removed component is not inert. $d_\parallel$ is 33% of d by norm, and applied alone at the magnitude it actually has, it flips 50% of answers (§4.4). What the decomposition removes is a component that is itself causally effective, not a rounding error.

Second, u is not an arbitrary direction with respect to d. A random direction in $d = 2048$ dimensions has $|\cos| \approx 0.022$; the measured alignment is 15× that. The lens direction captures a real and specific part of the causal delta.

Third, and decisively, magnitude alone does not produce the effect. The control r is also orthogonal to u and carries the same norm as $d_\perp$, yet produces nothing in 900 trials per relation. If "$d_\perp$ is most of d" were sufficient to explain the transplantation, a direction matched on both norm and lens-invisibility would do something. It does not.

At t=1 the ratio rises to 0.411 ± 0.037 as trajectories diverge; we report the t=0 value, where the split is exact. The residual departure of $\|d_\parallel\|^2 + \|d_\perp\|^2$ from $\|d\|^2$ is 0.13%, consistent with bfloat16 rounding.

What this section does not show. The magnitude at which $d_\perp$ moves the answer (0.455–0.478) is larger than the lens-visible threshold (0.125–0.158). We do not claim that this channel is more efficient—only that it is **invisible to the lens.** The orthogonality of $d_\perp$ is moreover defined against a single contrast; a readout measured by rank or by distributional distance could yield a different picture, though any such change large enough to alter the top-1 token would have registered here.

4.4 Why minimum-norm interventions miss this

The design of the standard test. The natural way to ask whether a direction is causal is to push the state along it and see whether the output changes. This requires deciding how far to push, and the usual answer is: **until the readout changes.** Once the readout points at the donor, the intervention has landed, and pushing further only perturbs the state unnecessarily. Our own initial intervention followed this design. It ascended the readout contrast by gradient and stopped the moment the top-1 token became the donor bridge, with landing confirmed by the fraction of pairs whose readout pointed at the donor.

What we registered in advance. Before running the experiment we fixed the decision rule: an answer-flip rate exceeding the baseline by $z > 3$ would indicate a causal relationship, $z \leq 2$ would indicate none, and intermediate values would call for additional seeds. If the landing check failed, the intervention strength was to be raised and the run repeated.

The result. On lmk the intervention landed on all 90 pairs (readout→donor 100%), and the answer flipped in **0 of 90** (95% Wilson upper bound 4.1%). The difference from the baseline condition was $z = -1.00$, falling squarely in the NULL region of the pre-registered rule. We interpreted this as showing that a planted bridge is represented but not used in computation, and prepared to report it in that form.

That interpretation was wrong. Decomposing the causal delta and applying only its lens-visible component, at the magnitude that component actually has, flips 50% of answers. Sweeping

magnitude with the direction held fixed places the answer threshold at 0.253 on lmk (§4.2). It is the same direction. Only the magnitude differs.

The reason for the miss is quantified exactly. The relative magnitude produced by the stopping rule was 0.119. The 50% threshold of the readout is 0.125 (standard error across seeds 0.0005). The two are effectively identical—as they must be, since the stopping rule is *defined* as the point at which the readout flips. And that point is **half** the answer threshold (ratio 2.03). This test does not happen to push too gently; it is **designed to stop before the answer moves.** On typeA the situation is more extreme still: the answer does not reach majority flipping at any magnitude before breakage dominates.

The general form of the trap. The core problem is the use of **the readout as the success criterion for the intervention.** We were testing whether the readout is a reliable indicator while using that same readout to decide when we had intervened enough. This circularity disables the test in exact proportion to the readout's greater sensitivity. That the landing check reported 100% seemed reassuring; in fact it was the mechanism guaranteeing that we stayed below threshold.

This structure is not specific to the J-lens or to this task. Any procedure that targets a readout and adopts readout change as its stopping condition inherits the same bias. So long as the readout is more sensitive than the computation, the minimum-norm solution lies systematically below the causal threshold.

Prescription. Readout-based causal claims should be reported as a function of magnitude rather than at a single magnitude. At minimum, one should report (i) the magnitude at which the readout responds, (ii) the magnitude at which the output responds, and (iii) a random control matched on magnitude and on the relevant invariants. Without these three, a null result cannot distinguish "the direction is not causal" from "we did not push hard enough." Our own pre-registered test failed to make that distinction.

4.5 What these results do not show

Full residual replacement shows only that the site is sufficient. Transplanting the donor's entire state into the corresponding position at layer L flips the answer on every pair. This establishes that the state at this site determines the answer, and thereby rules out the explanation that the intervention site lies downstream of answer formation. However, the donor's state **already contains the donor's answer**, so this transplantation is not evidence that bridge-to-answer computation occurs at this point. This section establishes exactly two things: that the state at layer L determines the answer, and that the component of that state read by the lens is not the component that determines it. Where the computation from bridge to answer takes place is not addressed by this experiment.

The lens direction is one contrast. u is the gradient of the donor-minus-recipient logit difference, not the readout distribution as a whole. A component orthogonal to this contrast could in principle alter the readout in some other way. Any such change large enough to move the top-1 token would have appeared in our measurement, and along $d_\perp$ the top-1 readout was pinned to the recipient bridge throughout (§4.3); we nonetheless cannot exclude changes in rank or distributional

shape. A readout measured by KL divergence or by the donor token's rank could yield different conclusions.

Withdrawal of an ablation-based claim. An earlier diagnostic in this work measured the fraction of items whose answer survives after the bridge is removed from the readout, and found 100% on both relations (0/15 and 0/22 broken; 95% upper bounds on the true breakage rate of 20.4% and 14.9%). We interpreted this as showing that the trained model does not use the bridge, and further, that Coconut training induces a shortcut.

That interpretation cannot be sustained. The ablation in question descends the lens logits and therefore operates in the same space identified as problematic in §4.4—and the lens-visible direction does flip answers at sufficient magnitude (§4.2). What the measurement supports is only that **removing the lens-readable component of the bridge leaves the answer unchanged.** It supports neither the claim that the model does not use the bridge nor any claim about shortcut formation. We report it as a statement about the readout and do not use it as a claim about computation.

Is the dissociation an artefact of supervision? We trained the model so that a frozen lens would read the bridge, so a direction that the lens reads but the computation does not use could plausibly be a by-product of that optimisation. To examine this we ran the same measurement without J-lens supervision ($\lambda = 0$, 3 seeds). The readout threshold was 0.119, effectively identical to the supervised condition (0.125), and the answer threshold was 0.207, **preserving a gap of** 1.74× (standard error across seeds 0.051; 95% confidence interval $[1.54, 1.97]$). The lens-invisible direction likewise transplanted 50% of answers at $\alpha = 0.40$ while the readout pointed at the donor in 0/900 and remained the recipient in 98% of trials, and the matched random control produced no answer flips and no breakage. Breakage along $d_\perp$ occurred in 1 of 900 trials (0.1%).

The threshold gap is therefore not created by J-lens supervision. Its size does differ between conditions, being larger under supervision (2.02 vs 1.74; Welch $t = 2.84$, $df = 3.4$, $p = 0.056$): supervision appears to widen the dissociation without generating it. Supervision also sharpens the readout, which is why top-1 stability at large magnitudes differs between the two conditions (§4.3): without supervision the readout remains the recipient's up to $\alpha = 0.40$ but falls to 82% at $\alpha = 0.60$. See Appendix A.

Frozen lens. The readout is a fixed probe, fitted to the base model and never updated (§2.2). We did not test whether the threshold gap survives refitting the lens to the trained model. If the gap disappeared under a refitted lens, the results of this section would narrow to a statement about probe staleness.

Scope. All measurements were made at a single supervision layer (layer 19 of a 2B model, roughly 80% depth), on one model, two synthetic relations, and $k = 2$ latent steps. Intervention pairs are drawn from training items and are therefore in-sample, and all answers are single tokens. Thresholds are linear-interpolation estimates on a 10-point grid and do not carry standard errors from a parametric fit. How the threshold gap varies across layers, scales, naturally occurring multi-hop data and multi-token answers remains open.

5 Related Work

Latent reasoning models. Coconut treats the last hidden state as a representation of the reasoning state and feeds it back as the next input embedding rather than decoding it into words, enabling reasoning in an unrestricted latent space rather than in language space [3]. Subsequent work has proposed similar architectures, including CODI [5]. Because these models emit no intermediate step in natural language, **their monitorability is structurally limited.** We study Coconut, planting a bridge entity in its continuous thoughts to obtain a condition in which the readout is controlled, and then run interventions.

Interpretation via vocabulary projection. Projecting hidden states onto the unembedding matrix to read them as a vocabulary distribution—the logit lens [4]—is the de facto standard instrument for interpreting latent reasoning. The Tuned Lens refines this by training an affine probe for each block of a frozen pretrained model, making it possible to decode every hidden state into a distribution over the vocabulary [1]. The J-lens used here belongs to the same family.

Importantly, the Tuned Lens made the correspondence between readout and computation an object of verification in its own right. It introduced causal basis extraction to locate the residual-stream directions with the highest influence on the lens output, then ablated those directions in the corresponding model hidden states and found that such features tend to be disproportionately influential on the model output. Our work shows that this correspondence **depends on intervention magnitude.** The magnitude at which the readout responds and the magnitude at which the computation responds differ by roughly a factor of two along the same direction, and the computation can be transplanted along a direction to which the readout does not respond at any magnitude. A correspondence observed at a single magnitude does not reveal this structure.

Interpreting latent reasoning models. The original Coconut paper presented a case study decoding continuous thoughts into vocabulary, reporting that the first thought decodes to tokens corresponding to the first intermediate variables of the calculation [3]. The practice of treating the readout as evidence about computation is thus present from the outset of this line of work.

Dilgren and Wiegreffe [2] systematized it. They show that when latent reasoning tokens are necessary, gold reasoning traces can be decoded from 65–93% of correctly predicted instances, and conclude that current latent reasoning models largely encode interpretable processes, with interpretability itself serving as a signal of prediction correctness.

Our results do not contradict this conclusion but constrain its scope. Their verification procedure constructs counterfactual prompts and checks whether the vocabulary projection at the relevant step changes to its expected new result—that is, **it verifies a readout change with a readout.** That a reasoning trace can be recovered from the readout and that the readout corresponds to the components the model actually uses are distinct claims. We show that the magnitude interval over which the latter holds is narrow, and that outside it errors occur in both directions.

They also note that vocabulary projection reveals only single-token concepts and omits multi-token concepts and latent-space directions not well aligned with vocabulary space, leaving tools for such directions to future work. Our $d_\perp$ channel is exactly an instance of such a direction: along it, 96–97% of answers are transplanted while the readout does not respond once in 1,800 trials.

Position of this work. Prior work asks what can be read from the readout. We ask whether what is read corresponds to the computation, **as a function of intervention magnitude.** To that

end we construct a testbed in which the readout is controlled (§3) and propose a protocol combining fixed-magnitude interventions, orthogonal decomposition of the causal delta, and a doubly matched random control (§4.1).

6 Conclusion

We measured, as a function of intervention magnitude, how far the readout of a model reasoning in continuous latent space reflects its computation.

Two things were required. The first is a model whose readout is known to point at a specific token: using the J-lens readout as a training signal, we aligned the bridge entity in the continuous thought without loss of accuracy. The second is a protocol that controls magnitude and direction independently: we fix the relative magnitude, split the causal delta orthogonally into lens-visible and lens-invisible components, and control with a random direction matched on both norm and lens-invisibility.

The result is a two-sided dissociation. The readout points at the donor on every pair at relative magnitudes of 0.125–0.158, yet the answer does not move there. The magnitude at which the answer responds is 2.03× the readout threshold on one relation and is never reached before breakage dominates on the other. Conversely, along the direction the lens does not detect, 96–97% of answers are transplanted while the readout neither points at the donor nor departs from the recipient in 1,800 trials. A random direction matched on norm and invisibility produced no answer changes, no readout changes and no breakage in the same trials. The gap persists at 1.74× without J-lens supervision.

The methodological consequence concerns single-magnitude interventions. A minimum-norm intervention targeting the readout is designed to stop where the readout flips, which on lmk is half the causal threshold and on typeA is below a threshold the answer never reaches. Our own pre-registered initial test produced a false null of 0/90 in this trap, with a stopping magnitude of 0.119 against a readout threshold of 0.125. The bias is not specific to a particular lens or task: a procedure that adopts readout change as its stopping condition measures systematically below the causal threshold, in proportion to the readout's greater sensitivity.

Readout-based causal claims should therefore be reported as a function of magnitude rather than at a single magnitude. Without the magnitude at which the readout responds, the magnitude at which the output responds, and a random control matched on the relevant invariants, a null result cannot distinguish "the direction is not causal" from "we did not push hard enough."

What remains. Our measurements were made at a single supervision layer, on one model, over two synthetic relations. How the threshold gap varies across layers, scales and naturally occurring data is open. The lens-invisible direction is moreover defined against a single contrast, and a readout measured by rank or distributional distance could yield different conclusions.

Within that scope, however, what we have established is clear. **The sensitivity of the readout is a property of the readout, and the sensitivity of the computation cannot be read off from it.** Work asking what can be read from a readout must also ask at what magnitude what is read corresponds to computation.

References

[1] Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2023.

[2] Connor Dilgren and Sarah Wiegreffe. Are latent reasoning models easily interpretable? *arXiv preprint arXiv:2604.04902*, 2026.

[3] Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. Training large language models to reason in a continuous latent space. In *Second Conference on Language Modeling (COLM)*, 2025. arXiv:2412.06769.

[4] nostalgebraist. Interpreting GPT: The logit lens. LessWrong, 2020. URL https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens.

[5] Zhenyi Shen, Hanqi Yan, Linhai Zhang, Zhanghao Hu, Yali Du, and Yulan He. CODI: Compressing chain-of-thought into continuous space via self-distillation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 677–693, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.36. arXiv:2502.21074.

A The threshold gap without J-lens supervision

We trained the model so that a frozen lens would read the bridge. A reviewer may reasonably ask whether the readout/computation gap is a by-product of that optimisation: a direction that the lens reads but the computation does not use is, on its face, what one might expect from pushing a model to satisfy a fixed probe. This appendix reports the same dose–response measurement without J-lens supervision.

Table 6: Supervised versus unsupervised conditions on lmk (3 seeds, 900 trials per cell).

	Readout	Answer	Ratio	Answer ($d_\perp$)	$d_\perp$ readout	$d_\perp$ broken
$\lambda = 1$	0.125	0.253	2.03	0.455	0/900	0/900
$\lambda = 0$	0.119	0.207	1.74	0.400	0/900	1/900

The readout threshold is essentially unchanged by supervision (0.119 vs 0.125), and the gap between readout and answer thresholds persists at 1.74× (standard error across seeds 0.051; 95% confidence interval $[1.54, 1.97]$, excluding 1). The lens-invisible channel is likewise present without supervision: at $\alpha = 0.40$ it transplants 50% of answers while the readout points at the donor in 0/900 trials and remains the recipient in 98%. The matched random control produces no answer flips and no breakage.

The gap is therefore not created by the supervision. Its magnitude does differ between conditions, being larger when supervised (2.02 vs 1.74; Welch $t = 2.84$, $df = 3.4$, $p = 0.056$), which suggests that supervision widens the dissociation without generating it.

One difference between conditions should be stated plainly. Under supervision the top-1 readout along $d_\perp$ remains the recipient bridge in 900/900 trials; without supervision this falls to 875/900 (97.2%), declining to 82% at $\alpha = 0.60$. This is expected: supervision is what makes the readout

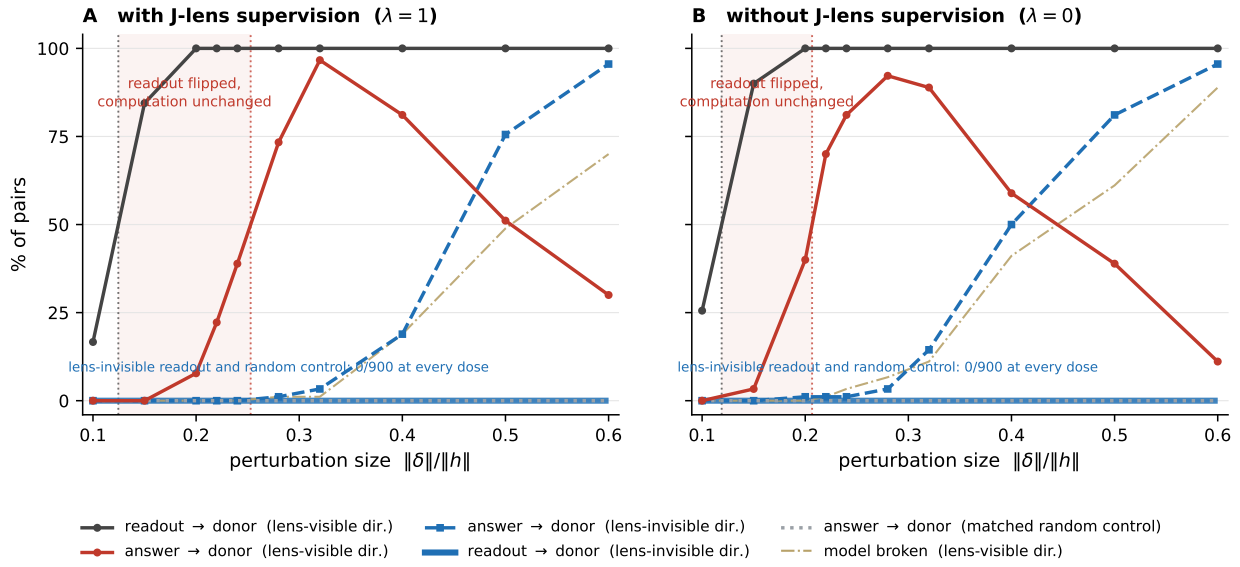


Figure 2: The threshold gap is not an artefact of J-lens supervision. Relation lmk, three seeds, 90 pairs per cell. (A) With J-lens supervision ($\lambda = 1$) and (B) without ($\lambda = 0$). Readout thresholds are 0.125 and 0.119; answer thresholds are 2.03× and 1.74× those values respectively. In both conditions the lens-invisible direction transplants answers without the readout pointing at the donor (0/900), and the matched random control produces no answer flips (0/900). Breakage along $d_\perp$ occurred in 1 of 900 trials in the unsupervised condition. The higher top-1 stability at large magnitude under supervision (900/900 vs 875/900) is a direct consequence of supervision sharpening the readout.

sharp, and an unsupervised readout (p_{bridge} ranging from 0.02 to 0.96 across items) is more easily displaced by a large perturbation. The claim in §4.3 that the lens registers no change at all is therefore stated for the supervised condition; without supervision it holds up to $\alpha = 0.40$.