A Theory of Conditional Collapse under Low-Rank Weight-Space Ablations: I. The Single-Block Theory and Synthetic Validation

Abdallah Khemais
ISITCOM, University of Sousse

July 2026

Abstract

Activation patching and weight-space ablation are both used to argue that a component of a network is causally responsible for a behavior, yet they act on different objects: one forward pass, versus the parameters behind every forward pass. We ask when they agree.

We study an idealized model in which a conditional computation is carried additively through a residual stream, $F(x) = F_0(x) + \sum_i \alpha_i(x)v_i$, and read out by a linear functional, and prove three exact results. First, deleting a subset of the carriers v_i collapses a matched input pair onto one and the same unconditional output *if and only if* the removal is symmetric on the pair and leaves no contrast outside it; the resulting error is deterministic, with polarity the sign of the pair's mean margin. Since those two conditions are exact equalities, we also give the exact error identity when they hold only approximately, so the criterion degrades gracefully rather than describing a null set. Second, patching a carrier moves the readout by that carrier's donor–receiver *contrast*, whereas ablating it moves the readout by its *absolute level* at the receiver. Neither bounds the other, and we construct matched pairs on which every single-carrier patch flips the decision while no single-carrier ablation does: the redundancy regime in which patching overstates importance and ablation understates it. Third, for an attention head composed with its own layer's normalization and MLP, we derive an exact first-order formula, with a provably second-order remainder, for the interaction term the idealized model sets to zero, and show it vanishes identically whenever the MLP alone is ablated but not, in general, when a head is; we connect it to the idealized selector's own error against the true network.

Small transformers trained on a synthetic conditional task illustrate all three predictions. Over thirty-nine ablation configurations the measured interaction is strongly rank-correlated with how well the idealized model predicts the edited network's behavior (Spearman -0.83); a clean separation we first read off fourteen of those configurations does *not* survive on the remaining twenty-five, and we report the weaker monotone claim that does. A second task and architecture (an inverse, value-to-key recall the first task never requires, on a smaller network) reproduces the same monotone relationship, the same patch-equals-weight-edit exactness, and a further instance of the polarity reversal.

The single-block interaction result derived here extends past one residual block, and the synthetic validation reported here is tested against a real pretrained model, in a companion analysis that takes the present theory further along both axes.

1 Introduction

Mechanistic interpretability certifies that a component of a network is causally responsible for a behavior by intervening on it, and two families of intervention dominate. *Activation patching* replaces a component's activation, inside a single forward pass, by the value it takes on a second input, and asks whether the model's decision follows [1, 4, 5]. *Weight-space ablation* instead edits the parameters (zeroing a component, or projecting a low-rank direction out of a weight matrix [10]) and asks the same question of the edited network. Both are read as evidence for the same informal claim, that a given component carries a given behavior, and both are frequently interpreted through a common lens of causal abstraction [2].

The two interventions are nevertheless different operators on different objects: patching alters one realized computation, ablation alters the parameters that generate every computation. Nothing in current practice specifies when they should agree, and there is direct evidence that they need not. The *Hydra effect* [7] documents that ablating an attention layer in a language model causes downstream layers to change their behavior and compensate; an ablation-based importance score is therefore measured on a network that has already partly repaired the damage, and can fall far below the importance the same component is assigned by an intervention that does not give the rest of the network that opportunity. Redundant, superposed codes [6] make the same mismatch structurally likely rather than accidental.

This paper makes the mismatch exact. We work in an idealized model, the *abstract conditional model* of Section 4, in which a conditional computation is carried additively through a residual stream,

$$F(x) \;=\; F_0(x) + \sum_{i=1}^{k} \alpha_i(x)\, v_i,$$

with F_0 the unconditional part, v_i fixed directions, α_i scalar coefficients (we call the pair (v_i, α_i) a *carrier*), and read out by a linear functional, $s(x) = \psi(F(x)) + b$, whose sign is the decision. Fix a *matched pair* (x_A, x_B), two inputs that should receive the same answer through opposite branches of the conditional, with margins $g_A = s(x_A) > 0 > g_B = s(x_B)$, and write $\beta_i := \psi(v_i)$. Our results are then the following, in the order in which they are proved.

Conditional collapse (Section 5). Deleting the carriers indexed by a subset S maps *both* members of the pair to one and the same value $\bar{s} = \frac{1}{2}(g_A + g_B)$ if and only if two conditions hold: the removed mass is symmetric on the pair ($\bar{q}_S = 0$), and no contrast survives outside S ($\Psi_S = 0$). When they hold, exactly one input is misclassified, its error is deterministic rather than noisy, the surviving branch is $\operatorname{sign} \bar{s}$, and any two subsets satisfying the criterion produce the identical collapse.

Patching–ablation dissociation (Section 6). Patching carrier i from donor to receiver moves the readout by exactly $\beta_i \delta_i$ with $\delta_i := \alpha_i(x_A) - \alpha_i(x_B)$ (the carrier's *contrast*), whereas ablating carrier i moves it by exactly $-\beta_i \alpha_i(x_B)$ (the carrier's *absolute level* at the receiver). Neither quantity bounds the other. We construct matched pairs with $n \geq 2$ carriers on which every single-carrier patch flips the decision while no single-carrier ablation does, which is precisely the regime that produces patch recoveries exceeding 1 (overshoot) alongside ablations that appear to show the component is dispensable: the shape of the self-repair observations above, obtained here with no repair mechanism at all.

Nonlinear interaction (Section 7). The idealized model treats carriers as independent. For the one composition where that is architecturally false (an attention head and its own layer's MLP, separated by an RMSNorm that reads the head's output), we compute the error exactly:

$$\Delta(x) \;=\; (I - Q)\Big[-Dg\big(r_1(x)\big)\eta(x) - R(x)\Big], \qquad \|R(x)\| \leq \Lambda \|\eta(x)\|^2,$$

where $\eta(x)$ is the head's ablated write-in, Q the MLP's ablation projector, g the normalization–MLP composition, and the displayed remainder bound holds under the bounded-curvature hypothesis stated in Theorem 3, a hypothesis whose constant Λ can in fact be exhibited in closed form, from the trained weights alone, as a companion analysis to this one shows, so that it need not be assumed unverified. The remainder is provably second order, and Δ vanishes identically whenever the MLP alone is ablated ($\eta(x) \equiv 0$), but not in general otherwise: ablating the head alone perturbs its own block's MLP output regardless of whether the MLP is separately touched, and we show precisely what this costs the first two theorems' predictions when applied to the true network rather than the idealized one.

Section 8 illustrates all three predictions on small transformers trained on a synthetic conditional task. That section is deliberately minimal: its purpose is to show that the predictions are realized in a trained network rather than vacuous under the stated assumptions, not to conduct a validation campaign. It also reports a negative result we consider load-bearing. An apparent clean separation, on which the interaction magnitude split fourteen ablation configurations into those the idealized model describes and those it does not with an empty band between, does not survive when the same frozen criterion is applied to twenty-five further configurations from the same networks; what survives is a strong monotone association rather than a threshold, and we state the claim at that strength. A second finding is more structural than surprising, and we report it for the same reason: the robust collapse criterion of Section 5.1 is a *sufficient* tolerance condition, and on these networks it is met by only one matched pair in ten, and by no single configuration on more than 71.7% of its pairs, because the condition compares two quantities of unequal, architecturally fixed scale, not because the underlying collapse fails to occur, which it does on four times as many pairs as the criterion certifies. To check that neither the theory's predictions nor this section's own negative results are an artifact of the marker task specifically, Section 8.1 repeats the checks above on a second, freshly trained instance with a different conditional mechanism (inverse recall rather than a fixed relabeling) and a smaller architecture; the predictions replicate, including a further instance of the polarity reversal, though the carrier count needed widening past this section's own threshold to give the interaction/fidelity check enough points to be informative, itself reported rather than smoothed over. The single-block result of Theorem 3 itself extends past one block, to an exact identity for the interaction between two arbitrarily distant layers, and is tested, together with the collapse and dissociation predictions above, on an emergent circuit found, not designed, in a real pretrained model; we develop both extensions in a companion analysis rather than here, so as to keep the present paper to the single-block theory and its synthetic-task validation. Section 9 separates what is established from what the model leaves open.

2 Related Work

Causal interventions as an interpretability method. Patching an activation from one run into another, and reading the change in the output, originates in causal mediation analysis applied to language models [1] and was scaled into causal tracing for factual recall [4] and into full circuit reconstructions [5]. Refinements restrict the intervention to individual paths through the computation graph [8], automate the search over candidate edges [9], and examine how sensitive the resulting conclusions are to the choice of corrupted input and evaluation metric [12]. Causal abstraction supplies the semantics under which such an experiment counts as evidence that a network implements a given high-level algorithm [2], including when the high-level variable is realized in a distributed, rotated subspace rather than in a single neuron [13]. The present paper takes this methodology as given: we do not propose a new intervention or a new search procedure, and the abstract conditional model of Section 4 is an abstraction in exactly the sense of [2], specialized to a binary readout and an additively carried conditional.

Weight-space edits and the additive residual stream. A separate line of work intervenes on parameters instead of activations, most directly by projecting a single low-rank direction out of the weight matrices that write into the residual stream [10]. The decomposition $F = F_0 + \sum_i \alpha_i v_i$ we start from is the standard reading of the residual stream as a sum of component write-ins [3], with the v_i playing the role of the output directions of individual heads or MLPs. What the present paper adds is not this decomposition but a criterion, internal to it, for when *removing* a subset of the write-ins forces the readout to a single value on two inputs that previously separated.

When an intervention misleads. Several results establish that a causal intervention can support a conclusion the network does not warrant. Subspace activation patching can flip a model's behavior through a direction that the unperturbed model does not use, an interpretability illusion that survives the usual sufficiency and necessity checks [11]. Ablating a component can be compensated by downstream components, so that measured importance reflects the network's self-repair as much as the component's role [7], and redundant or superposed codes [6] give a representational reason to expect exactly this. These are, respectively, an argument that patching can overstate and an argument that ablation can understate. Our contribution is to place both in one model and make the gap between them an identity rather than a caution: Theorem 2 names the two quantities the two interventions actually measure, $\beta_i\delta_i$ and $\beta_i\alpha_i(x_B)$, and Corollary 4 exhibits, by explicit construction, matched pairs on which every single-carrier patch is sufficient while no single-carrier ablation is necessary.

What is new here. Relative to the above, this paper contributes (i) an *if and only if* criterion for a weight-space ablation to collapse a conditional onto a single unconditional branch, together with the polarity of the resulting error (Theorem 1); (ii) a constructive separation between patching sufficiency and ablation necessity within a single model, rather than as two separately observed empirical phenomena (Corollary 4); and (iii) an exact, second-order-bounded formula for the interaction that the additive picture omits whenever an ablated head's own layer's MLP is not itself ablated away entirely (Theorem 3), which identifies the single architectural mechanism through which the first two results cease to be exact, together with a closed-form bound on the curvature constant controlling its remainder, established in a companion analysis, so that the second-order claim is checkable on a given trained network rather than conditional on an unverified hypothesis.

3 Problem Formulation

We study conditional computations implemented by residual neural networks. Our objective is not to analyze a particular transformer architecture but to characterize, in an abstract setting, the effect of removing a low-dimensional weight-space subspace supporting a learned conditional computation.

Let a residual block act as $x_{l+1} = x_l + F(x_l; W)$, where $W \in \mathbb{R}^m$ collects the trainable parameters of the components writing into the stream. By a *low-rank weight-space ablation* we mean an edit that projects a low-dimensional subspace

$$\mathcal{S} \;=\; \mathrm{span}\{u_1, \ldots, u_p\}, \qquad p \ll m,$$

out of those parameters, with $P_{\mathcal{S}}$ the orthogonal projector onto $\mathcal{S}$: for a weight matrix this reads $W \mapsto (I - P_{\mathcal{S}})W$ or $W \mapsto W(I - P_{\mathcal{S}})$ according to which side of the map the removed subspace acts on. The concrete instances used below are the head and MLP edits written out in Remark 1 and Section 7, in which $\mathcal{S}$ is spanned by directions estimated from data. This is the operator whose effect the paper analyzes, and it is not the operator activation patching applies: patching alters the activations realized on one forward pass, leaving W intact, whereas the edit above changes the map generating every forward pass.

What we analyze. Working directly with W' is unwieldy and specific to the architecture. We therefore analyze the induced change in the *residual stream*: Section 4 posits that the block's output decomposes additively over the components that write into it, and Sections 5–6 study the deletion of a subset of those write-ins. The two views are connected concretely, not in general but for the specific edits we use, in Remark 1: ablating an attention head by projecting a direction out of its output projection subtracts exactly one term of the form $v\,\alpha(x)$ from the residual stream, which is precisely the deletion operation Section 5 formalizes. Everything after the present section is stated in the output space; the weight-space operator above is what realizes it.

Two features of a *conditional* computation matter for what follows. It is carried additively, so removing it is a subtraction rather than a retraining; and it is what makes the block's output differ between two inputs that the network must nevertheless answer identically. We do not assume it is gated off outside some region of input space: the coefficients $\alpha_i(x)$ of Section 4 are unrestricted scalars, and in the construction of Corollary 4 every carrier is active on both members of the pair. The question the paper answers is when deleting such a computation forces the two inputs to the same answer, and how that differs from what patching the same component would have suggested.

4 Abstract Conditional Model

This section introduces the abstract mathematical model underlying the subsequent theoretical analysis. The objective is to isolate the essential geometric structure responsible for conditional collapse independently of any specific neural architecture.

4.1 Residual Decomposition

Assume that the residual mapping admits the decomposition

$$F(x) \;=\; F_0(x) + \sum_{i=1}^{k} \alpha_i(x)\, v_i,$$

where F_0 denotes the unconditional computation, $v_i \in \mathbb{R}^d$ are fixed feature directions, and $\alpha_i(x)$ are scalar selector functions. The vectors $\mathcal{V} = \{v_1, \ldots, v_k\}$ represent the directions supporting the conditional computation. We call the pair (v_i, α_i) the i-th *carrier*, a term we also use for its concrete realizations (an attention head or an MLP writing into the stream) in Sections 7 and 8.

4.2 Low-Rank Support

We assume $\dim \operatorname{span}(\mathcal{V}) = r$ with $r \ll d$, so the conditional computation occupies only a low-dimensional subspace of the representation space. We denote this support by $\mathcal{C} = \operatorname{span}(\mathcal{V})$.

The two integers k and r need not coincide: the carriers are k given directions, not a basis, and $r = \dim \mathcal{C} \le k$ with strict inequality whenever they are linearly dependent. Every subset-indexed statement below (Sections 5 and 6) is therefore stated relative to a *fixed, arbitrary indexing* $i = 1, \ldots, k$ of the given carriers, which is the indexing an experimenter fixes when choosing which components to intervene on. We neither assume the v_i independent nor claim that the results are invariant under a change of spanning set for $\mathcal{C}$.

4.3 Deleting a Subset of Carriers

The intervention analyzed in the rest of the paper deletes a sub-collection of carriers. For an index subset $S \subseteq \{1, \ldots, k\}$, write

$$F_S(x) \;:=\; F_0(x) + \sum_{i \notin S} \alpha_i(x)\, v_i, \tag{1}$$

so that the carriers in S contribute nothing while F_0 and every other carrier are left exactly as they were. This is the object Section 5 analyzes.

Deleting terms is not the same as projecting the output onto $\mathcal{C}_S^{\perp}$, where $\mathcal{C}_S := \mathrm{span}\{v_i : i \in S\}$: the projection would also strip from the *surviving* carriers whatever component they have inside $\mathcal{C}_S$, and the two operations agree for every choice of coefficients only when the surviving directions are orthogonal to $\mathcal{C}_S$. Taking $S = \{1, \dots, k\}$ makes the point in the other direction: applying $I - P_{\mathcal{C}}$ to the whole sum annihilates it identically, since the sum lies in $\mathcal{C}$ by construction, so that extreme case returns F_0 and says nothing about proper subsets. Term deletion, not subspace projection, is what a weight-space edit of a component performs.

Remark 1 (From a weight edit to a deleted term)**.** *The deletion* (1) *is exactly what a low-rank edit of one component's own weights does to the residual stream. In the notation of Section 7, an attention head with pre-projection activation $a(x)$ and output projection W writes $Wa(x)$ into the stream; projecting a unit direction u out of that projection, $W \mapsto W(I - uu^\top)$, changes the write-in by exactly*

$$Wa(x) \;-\; W(I - uu^\top)a(x) \;=\; (Wu)\,\langle u, a(x)\rangle,$$

that is, it deletes one term $v\,\alpha(x)$ with $v := Wu$ and $\alpha(x) := \langle u, a(x)\rangle$; a projector of rank ℓ deletes ℓ such terms, one per direction of its range. This is an identity, not an approximation. What is not *an identity, and is the subject of Section 7, is the further assumption that deleting one component's term leaves every other component's term unchanged.*

4.4 Selector Representation

Writing the selector coefficients as $\alpha(x) = (\alpha_1(x), \dots, \alpha_k(x)) \in \mathbb{R}^k$ and $V = [v_1, \dots, v_k]$, the residual computation is compactly $F(x) = F_0(x) + V\alpha(x)$. This matrix form makes explicit the separation between the geometry of the supporting directions (V) and the selector responsible for activating them (α), the central abstraction used throughout the remainder of the paper.

4.5 Standing Assumptions

The following assumptions frame all subsequent theoretical results.

Assumption 1 (Residual Additivity)**.** *The conditional computation enters the network through an additive residual connection, so that $F(x) = F_0(x) + \sum_{i=1}^{k} \alpha_i(x)v_i$ with the v_i independent of x.*

Assumption 2 (Low-Rank Support)**.** *The conditional component has finite-dimensional support: a subspace $\mathcal{C}$ with $\dim(\mathcal{C}) \ll d$.*

Assumption 3 (Stable Base Computation)**.** *The unconditional component F_0 remains invariant under the considered deletion: F_0 in* (1) *is the same function of x before and after the intervention.*

Assumptions 1 and 3 are load-bearing: together they are what licenses writing the ablated computation as (1), and they are invoked at that point in every proof of Section 5. Assumption 2 is not. No theorem below uses $\dim(\mathcal{C}) \ll d$ quantitatively; the assumption records *why* one expects a conditional to be carried by a few directions rather than by an arbitrary linear functional of the stream: it is a modeling commitment that makes the objects of Section 4.3 the ones an experimenter can actually enumerate and intervene on, and we state it as such rather than pretend the results depend on it. Together the three assumptions avoid architecture-specific details and define the framework within which the following theorems are established.

5 Conditional Collapse

The model of Section 4 describes how a conditional computation is embedded into a residual stream. We now ask how the network's discrete decision responds to a low-rank ablation of that computation. Throughout this section we restrict attention to *matched pairs*: two inputs intended to receive the same decision through opposite branches of the conditional. All results below are conditional on a single additional assumption relating the residual stream to the network's output.

Assumption 4 (Linear readout)**.** *There exist a linear functional $\psi \in (\mathbb{R}^d)^*$ and a bias $b \in \mathbb{R}$ such that the network's binary decision on $F(x)$ is determined by the sign of*

$$s(x) \;:=\; \psi\big(F(x)\big) + b,$$

branch A being selected when $s(x) > 0$ and branch B when $s(x) < 0$.

The boundary case $s(x) = 0$ is left undefined, and the same convention applies to the ablated selector s_S and to the mean margin $\bar{s}$ introduced below: we assume throughout that none of these vanishes on the pairs under consideration. We do not claim this is generic in any sense we have verified; it is a convention that keeps every statement below a statement about strict inequalities, and a pair violating it is simply outside the scope of the results.

Definition 1 (Matched pair and margins)**.** *A* matched pair *is a pair of inputs (x_A, x_B) for which the network is correct exactly when*

$$g_A \;:=\; s(x_A) \;>\; 0 \;>\; s(x_B) \;=:\; g_B.$$

We call g_A, g_B the margins *of the pair and $\Phi := g_A - g_B > 0$ its* total contrast*.*

Note that g_A, g_B and Φ, and likewise $\bar{q}_S$, Φ_S, Ψ_S and $\bar{s}$ below, are attached to the fixed pair (x_A, x_B) and carry no free input argument; only quantities genuinely evaluated at a variable input, such as $\alpha_i(x)$, $s(x)$ and $s_S(x)$, are written with one.

By Assumptions 1 and 3, ablating a subset $S \subseteq \{1, \ldots, k\}$ of carriers replaces F by F_S of (1): the deleted carriers contribute nothing and everything else, F_0 included, is unchanged. Write $\beta_i := \psi(v_i)$ for the fixed scalar through which carrier i acts on the readout. The *ablated selector* is

$$s_S(x) \;=\; \psi\big(F_S(x)\big) + b \;=\; s(x) - q_S(x), \qquad q_S(x) \;:=\; \sum_{i \in S} \beta_i \, \alpha_i(x),$$

where $q_S(x)$ is the *removed mass.* On a matched pair, split the removed mass into symmetric and antisymmetric parts,

$$\bar{q}_S \;:=\; \tfrac{1}{2}\big(q_S(x_A) + q_S(x_B)\big), \qquad \Phi_S \;:=\; q_S(x_A) - q_S(x_B),$$

so that $q_S(x_A) = \bar{q}_S + \Phi_S/2$ and $q_S(x_B) = \bar{q}_S - \Phi_S/2$.

Lemma 1 (Exact flip criterion)**.** *For every matched pair and every ablated subset S: branch A flips to B if and only if $\bar{q}_S + \Phi_S/2 > g_A$, and branch B flips to A if and only if $\bar{q}_S - \Phi_S/2 < g_B$.*

Proof. $s_S(x_A) = g_A - q_S(x_A) = g_A - \bar{q}_S - \Phi_S/2$; branch A flips exactly when this is negative. The claim for B follows symmetrically from $s_S(x_B) = g_B - \bar{q}_S + \Phi_S/2$. □

Because each carrier is removed exactly rather than through a possibly imperfect projection, ablating S never leaves a partial residue of the carriers it contains. What determines collapse is therefore not how completely S removes its own carriers (always exact here) but whether S accounts for the *entire* contrast between x_A and x_B. Define the *uncaptured contrast*

$$\Psi_S \;:=\; \psi\big(F_0(x_A) - F_0(x_B)\big) + \sum_{i \notin S} \beta_i \big(\alpha_i(x_A) - \alpha_i(x_B)\big),$$

the portion of Φ carried by everything S does not touch: the unconditional part F_0 together with every carrier left in place. A direct computation gives $\Phi_S = \Phi - \Psi_S$: what S removes is exactly what remains once the uncaptured contrast is subtracted from the total.

Theorem 1 (Conditional collapse)**.** *For a given matched pair* (x_A, x_B) *and ablated subset* S*, define*

$$\bar{s} \;:=\; \tfrac{1}{2}\big(g_A + g_B\big).$$

Then $s_S(x_A) = s_S(x_B) = \bar{s}$ if and only if

(i) $\bar{q}_S = 0$ (the removal is symmetric on the pair)*, and*

(ii) $\Psi_S = 0$ (no format contrast survives outside S).

When (i)–(ii) hold, both inputs of the pair are mapped to the same branch after ablation: A if $\bar{s} > 0$*, B if* $\bar{s} < 0$*. We call this an* unconditional collapse *with* polarity sign $\bar{s}$.

Proof. ($\Leftarrow$) Condition (ii) gives $\Phi_S = \Phi$. Substituting into $q_S(x_A) = \bar{q}_S + \Phi_S/2$ and $q_S(x_B) = \bar{q}_S - \Phi_S/2$ together with condition (i) gives $q_S(x_A) = \Phi/2$ and $q_S(x_B) = -\Phi/2$. Hence $s_S(x_A) = g_A - \Phi/2 = \frac{1}{2}(g_A + g_B) = \bar{s}$, and symmetrically $s_S(x_B) = g_B + \Phi/2 = \bar{s}$.

($\Rightarrow$) From $s_S(x_A) = g_A - q_S(x_A)$ and $s_S(x_B) = g_B - q_S(x_B)$, subtracting gives $s_S(x_A) - s_S(x_B) = \Phi - \big(q_S(x_A) - q_S(x_B)\big) = \Phi - \Phi_S = \Psi_S$; if $s_S(x_A) = s_S(x_B)$ this forces $\Psi_S = 0$, which is (ii). Averaging the same two identities gives $\frac{1}{2}\big(s_S(x_A) + s_S(x_B)\big) = \bar{s} - \bar{q}_S$; if this common value equals $\bar{s}$, then $\bar{q}_S = 0$, which is (i). □

Corollary 1 (Purity)**.** *Under the hypotheses of Theorem 1, if* $\bar{s} \neq 0$ *then exactly one input of the pair is misclassified after ablation, and its error is deterministic: the collapsed decision equals the branch that was already correct for the* other *input of the pair. No third outcome is possible.*

Corollary 2 (Idealized invariance)**.** *If two subsets* S, S' *each satisfy the hypotheses of Theorem 1 for the same matched pair, then* $s_S(x_A) = s_{S'}(x_A)$ *and* $s_S(x_B) = s_{S'}(x_B)$*: they induce the same collapse, with the same polarity.*

Remark 2. *Corollary 2 is a statement about the idealized model alone: under Assumption 4, the outcome of any qualifying ablation is pinned down by the pair's own margins* g_A, g_B*, not by which qualifying subset realizes it. An empirical instance in which two ablations that each look complete produce* different *polarities on the same trained network is therefore evidence that at least one of them fails hypothesis (i) or (ii). Two mechanisms can produce that failure and they are worth keeping apart: the true computation may depart from Assumption 4's additive, linearly-read picture, which Section 7 identifies and bounds for one structurally motivated nonlinearity; or the hypotheses may fail* inside *the linear model, with* $\bar{q}_S$ *or* Ψ_S *simply not small. Corollary 3 below makes the second alternative quantitative, and Section 8 finds it is the one that actually occurs on the instance where we observe a reversal.*

5.1 A robust form

Conditions (i) and (ii) of Theorem 1 are exact equalities between real numbers computed from a trained network, so they hold on a set of measure zero and no experiment ever satisfies them. Stated only in that form the theorem is a statement about an event that never occurs. It is worth recording that nothing in it depends on the equalities being exact: the same computation gives an error term, and the error term is an identity rather than an estimate.

Proposition 1 (Exact error decomposition and robust collapse)**.** *For every matched pair and every subset* S*, with no hypothesis whatsoever,*

$$s_S(x_A) - \bar{s} \;=\; -\bar{q}_S + \tfrac{1}{2}\Psi_S, \qquad s_S(x_B) - \bar{s} \;=\; -\bar{q}_S - \tfrac{1}{2}\Psi_S. \tag{2}$$

Consequently, if $|\bar{q}_S| \le \varepsilon_1$ *and* $|\Psi_S| \le \varepsilon_2$*, then*

$$\big|s_S(x_A) - \bar{s}\big| \;\le\; \varepsilon_1 + \tfrac{1}{2}\varepsilon_2, \qquad \big|s_S(x_B) - \bar{s}\big| \;\le\; \varepsilon_1 + \tfrac{1}{2}\varepsilon_2, \qquad \big|s_S(x_A) - s_S(x_B)\big| \;\le\; \varepsilon_2,$$

and if in addition $|\bar{s}| > \varepsilon_1 + \frac{1}{2}\varepsilon_2$*, then both inputs of the pair are mapped to the branch* $\operatorname{sign}\bar{s}$*, exactly one of them is misclassified, and its error is deterministic. That is, the conclusions of Theorem 1 and Corollary 1 survive verbatim with the exact collapse* $s_S(x_A) = s_S(x_B) = \bar{s}$ *weakened to agreement within* ε_2*. Theorem 1 is the case* $\varepsilon_1 = \varepsilon_2 = 0$*.*

Proof. The proof of Theorem 1 already establishes $s_S(x_A) - s_S(x_B) = \Psi_S$ and $\frac{1}{2}\big(s_S(x_A) + s_S(x_B)\big) = \bar{s} - \bar{q}_S$; solving these two linear equations for the individual terms gives (2). (Directly: $s_S(x_A) - \bar{s} = g_A - q_S(x_A) - \big(g_A - \frac{1}{2}\Phi\big) = \frac{1}{2}(\Phi - \Phi_S) - \bar{q}_S$, using $q_S(x_A) = \bar{q}_S + \frac{1}{2}\Phi_S$ and $\Psi_S = \Phi - \Phi_S$; the computation for x_B is symmetric.) The three bounds are the triangle inequality applied to (2) and to their difference. For the last claim, write $z := s_S(x_A)$; from $|z - \bar{s}| \le \varepsilon_1 + \frac{1}{2}\varepsilon_2 < |\bar{s}|$ we get $z > \bar{s} - |\bar{s}| = 0$ when $\bar{s} > 0$ and $z < \bar{s} + |\bar{s}| = 0$ when $\bar{s} < 0$, so $\operatorname{sign} z = \operatorname{sign}\bar{s}$, and likewise for $s_S(x_B)$. Since the pair is matched, the branch $\operatorname{sign}\bar{s}$ is the correct answer for exactly one of the two inputs, which gives purity and determinacy as in Corollary 1. □

Corollary 3 (Robust invariance)**.** *If* S *and* S' *both satisfy* $|\bar{q}| \le \varepsilon_1$ *and* $|\Psi| \le \varepsilon_2$ *on the same matched pair, then* $\big|s_S(x) - s_{S'}(x)\big| \le 2\varepsilon_1 + \varepsilon_2$ *for* $x \in \{x_A, x_B\}$*, and if* $|\bar{s}| > \varepsilon_1 + \frac{1}{2}\varepsilon_2$ *they induce the* same *polarity. Contrapositively, if two subsets collapse the same pair onto* opposite *branches, they cannot both satisfy the hypothesis at that tolerance: at least one has* $\varepsilon_1 + \frac{1}{2}\varepsilon_2 \ge |\bar{s}|$*.*

Proof. Subtracting the first identity of (2) for S and for S' gives $s_S(x_A) - s_{S'}(x_A) = (\bar{q}_{S'} - \bar{q}_S) + \frac{1}{2}(\Psi_S - \Psi_{S'})$, whose modulus is at most $2\varepsilon_1 + \varepsilon_2$; the computation at x_B differs only in the sign of the second group. The polarity claim and its contrapositive follow from the last part of Proposition 1, which assigns both subsets the polarity $\operatorname{sign}\bar{s}$, a quantity attached to the pair, not to the subset. □

Corollary 3 is the quantitative form of the observation in the remark above, and it is the one that can be checked: a polarity reversal between two ablations of the same network is not merely qualitative evidence that some hypothesis fails, but a lower bound, $\varepsilon_1 + \frac{1}{2}\varepsilon_2 \ge |\bar{s}|$, on how badly it fails for at least one of them. Section 8 measures $\bar{q}_S$, Ψ_S and $\bar{s}$ on the trained instances and reports how often the hypothesis of Proposition 1 is actually available.

6 Patching–Ablation Dissociation

Activation patching and weight-space ablation are often reported side by side as two ways of probing the same causal object: both are described as "removing" or "restoring" a carrier's contribution, and both are read as evidence about whether that carrier is causally load-bearing. We show that, even in the idealized model of Sections 4–5, the two interventions are governed by different quantities and can disagree in a specific, predictable way: a carrier can be *sufficient* to flip a decision under patching while no single-carrier ablation is *necessary* to flip the same decision. This section makes that distinction exact.

Definition 2 (Patching a carrier)**.** *For a matched pair* (x_A, x_B) *(*x_A *the* donor*,* x_B *the* receiver*) and a single carrier* i*,* patching i *replaces* $\alpha_i(x_B)$ *by* $\alpha_i(x_A)$ *inside the receiver's computation, leaving every other carrier and* F_0 *evaluated at* x_B*. The patched selector is*

$$s_i^{\to B}(x_B) \;:=\; s(x_B) + \beta_i\delta_i, \qquad \delta_i \;:=\; \alpha_i(x_A) - \alpha_i(x_B).$$

Theorem 2 (Patching–ablation dissociation)**.** *For every carrier* i *and matched pair* (x_A, x_B)*:*

(a) Patching i *into the receiver changes the selector by exactly* $\beta_i\delta_i$*,* independently of F_0, of every other carrier, and of how the total contrast Φ is distributed among carriers*; the receiver flips if and only if* $\beta_i\delta_i > -g_B$*.*

(b) *Ablating i alone changes the selector by exactly $-\beta_i\alpha_i(x_B)$, a quantity that depends on the receiver's own coefficient $\alpha_i(x_B)$, not on the contrast δ_i; the receiver flips if and only if $-\beta_i\alpha_i(x_B) > -g_B$.*

In particular, patching sufficiency and ablation necessity are governed by two generally unrelated quantities, $\beta_i\delta_i$ and $\beta_i\alpha_i(x_B)$, and neither bounds the other without further assumptions.

Proof. Part (a) is Definition 2 together with the sign condition for branch B in the sense of Assumption 4. Part (b) is Lemma 1 specialized to $S = \{i\}$, for which $\bar{q}_{\{i\}} = \frac{1}{2}(\beta_i\alpha_i(x_A) + \beta_i\alpha_i(x_B))$ and $\Phi_{\{i\}} = \beta_i\delta_i$, giving $\bar{q}_{\{i\}} - \Phi_{\{i\}}/2 = \beta_i\alpha_i(x_B)$ after cancellation. □

Corollary 4 (Sufficiency without necessity)**.** *There exist matched pairs and families of $n \geq 2$ carriers for which every single-carrier patch flips the receiver while no single-carrier ablation does.*

Proof by explicit construction. Fix $n \geq 2$, set $\beta_i = 1$, $\alpha_i(x_B) = -\frac{1}{2}$ and $\alpha_i(x_A) = 1$ for $i = 1, \ldots, n$ (so $\delta_i = \frac{3}{2}$), and choose F_0 with $\psi(F_0(x_A)) = \psi(F_0(x_B))$ (e.g. $F_0 \equiv 0$) and b so that $g_B = -1$. Every single patch satisfies $\beta_i\delta_i = \frac{3}{2} > 1 = -g_B$ and therefore flips the receiver by part (a). Every single ablation satisfies $-\beta_i\alpha_i(x_B) = \frac{1}{2} < 1 = -g_B$ and therefore does not, by part (b). Because $\psi(F_0(x_A)) = \psi(F_0(x_B))$, the margins are exactly $g_A = -1 + \frac{3}{2}n$ and $g_B = -1$, so consistency ($g_A > 0$) holds for every $n \geq 2$, and the construction exhibits a family of $n \geq 2$ carriers satisfying both requirements simultaneously. □

Remark 3 (Discussion)**.** *The mechanism behind Corollary 4 is redundancy, not an artifact of the construction: if n carriers each carry enough* contrast *δ_i to flip the receiver on their own, patching any one of them injects that full contrast and flips it: activation-level sufficiency requires nothing about the other $n - 1$ carriers. Ablating one of the n carriers, in contrast, removes only its own* absolute *level $\alpha_i(x_B)$ at the receiver; if the receiver's correct decision is instead supported by the combination of what remains (the untouched carriers and F_0), removing one contributor changes nothing about whether the others still suffice.*

This is the regime that produces a patch-recovery ratio $\kappa > 1$, i.e. a single-site patch that overshoots *the donor's output rather than merely reaching it: the patched contrast exceeds what is strictly required to cross the margin, precisely because the same margin is, in the unablated network, being protected redundantly from the ablation side. Section 8 reports both halves of this signature on the same trained network (a single-site patch with $\kappa > 1$ at a site whose weight-level ablation nonetheless leaves the majority of receiver-format inputs answered correctly), which is an instance of Theorem 2, not a separate phenomenon. It is also, we note, what an observer with access to only one of the two interventions would report as an anomaly: overshooting recovery from the patching side, or an apparently dispensable component from the ablation side.*

7 Nonlinear Interaction Theorem

Sections 5 and 6 treat each carrier's contribution $\alpha_i(x)v_i$ as exact and independent of every other carrier: ablating a set S simply deletes the corresponding terms from the sum, leaving the rest untouched. This section identifies precisely when that independence itself is exact, and exhibits the architectural mechanism through which two carriers of the same residual block fail to satisfy it.

7.1 A concrete two-carrier composition

Vectors are columns throughout this section, and every operator acts on the left. Consider two carriers written into the same residual stream by a single block. The first is an attention head with pre-projection activation $a(x) \in \mathbb{R}^{d_\mathrm{h}}$ and output projection $W \in \mathbb{R}^{d \times d_\mathrm{h}}$, contributing

$Wa(x)$ to the post-attention residual $r_1(x) := x + \mathrm{ao}(x)$. The second is the block's own MLP M, applied *after* a normalization layer N acting on r_1. Writing $g := M \circ N$ for the composition of the normalization and the MLP (this g is unrelated to the pair margins g_A, g_B of Section 5; context disambiguates throughout), the second carrier's realized value is $g(r_1(x))$, a function *of the first carrier's output*, not an independent term.

Ablating the head with an orthogonal projector $P = UU^\top$ on $\mathbb{R}^{d_\mathrm{h}}$ replaces W by $W(I-P)$, equivalently subtracting $\eta(x) := WPa(x)$ from r_1 (Remark 1); ablating the MLP itself with an orthogonal projector Q on $\mathbb{R}^d$ replaces $M(z)$ by $(I-Q)M(z)$. A *frozen-activation* estimate of the resulting change in the MLP's contribution (treating the two carriers as independent, in the sense of Section 5) computes $g(r_1(x))$ once, applies $(I-Q)$ if the MLP itself is ablated, and otherwise leaves the head's ablation with no effect on it; the *true* computation instead recomputes the MLP from whatever residual the head's ablation actually produces, then applies $(I-Q)$. Their difference,

$$\Delta(x) \;:=\; (I-Q)\Big[g\big(r_1(x) - \eta(x)\big) - g\big(r_1(x)\big)\Big],$$

is exactly the interaction that the independent-carrier model of Sections 4–6 sets to zero by assumption. We now derive $\Delta(x)$ exactly and bound it.

7.2 The interaction term

Let $N(r) := \rho(r)\,\gamma \odot r$ with $\rho(r) := (\frac{1}{d}\|r\|^2 + \varepsilon)^{-1/2}$ denote root-mean-square normalization with scale γ and $\varepsilon > 0$.

Lemma 2 (Normalization Jacobian). *N is C^∞ on all of $\mathbb{R}^d$, with*

$$DN(r) \;=\; \rho(r)\,\mathrm{diag}(\gamma)\Big(I_d - \tfrac{\rho(r)^2}{d} rr^\top\Big), \qquad \|DN(r)\|_{\mathrm{op}} \;\le\; \|\gamma\|_\infty\,\rho(r).$$

Proof. With $T(r) := \frac{1}{d}\|r\|^2 + \varepsilon \ge \varepsilon > 0$, $\rho = T^{-1/2}$ is smooth everywhere and $\partial\rho/\partial r_j = -\frac{1}{d}\rho^3 r_j$. Since $N_i(r) = \gamma_i \rho(r) r_i$, $\partial N_i/\partial r_j = \gamma_i\rho\big(\delta_{ij} - \frac{\rho^2}{d} r_i r_j\big)$, which is the displayed matrix. The matrix $I_d - \frac{\rho^2}{d} rr^\top$ is symmetric with eigenvalue $\frac{d\varepsilon}{\|r\|^2 + d\varepsilon} \in (0,1]$ along r and eigenvalue 1 on $r^\perp$, hence operator norm 1; combined with $\|\,\mathrm{diag}(\gamma)\|_{\mathrm{op}} = \|\gamma\|_\infty$ this gives the bound. □

Theorem 3 (First-order interaction formula)**.** *For every x,*

$$\Delta(x) \;=\; (I-Q)\Big[-Dg\big(r_1(x)\big)\,\eta(x) \;-\; R(x)\Big],$$

$$R(x) \;:=\; \int_0^1 \Big[Dg\big(r_1(x) - t\eta(x)\big) - Dg\big(r_1(x)\big)\Big]\eta(x)\,dt,$$

where $Dg = DM(N(r))\,DN(r)$ by the chain rule. The identity is exact for every x, every $\eta(x)$, and every Q. Moreover g is C^∞ (Lemma 2 for N; M is a composition of linear maps and smooth elementwise nonlinearities), so $\|R(x)\| \le L(x)\|\eta(x)\|$ with $L(x) := \sup_{t\in[0,1]} \|Dg(r_1(x) - t\eta(x)) - Dg(r_1(x))\|_{\mathrm{op}}$ finite; if in addition $\|D^2 g\|_{\mathrm{op}} \le \Lambda$ on the segment joining $r_1(x) - \eta(x)$ and $r_1(x)$, then $\|R(x)\| \le \Lambda\|\eta(x)\|^2$: the interaction is second order in the size of the perturbation.

Proof. Let $\phi(t) := g(r_1(x) - t\eta(x))$, so $\phi'(t) = -Dg(r_1(x) - t\eta(x))\eta(x)$ and, by the fundamental theorem of calculus, $g(r_1(x) - \eta(x)) - g(r_1(x)) = \phi(1) - \phi(0) = \int_0^1 \phi'(t)\,dt$. Adding and subtracting $Dg(r_1(x))\eta(x)$ inside the integral and left-multiplying by $(I-Q)$ gives the displayed identity. The bound on $R(x)$ follows from $\|R(x)\| \le \sup_t \|Dg(r_1(x) - t\eta(x)) - Dg(r_1(x))\|_{\mathrm{op}}\|\eta(x)\|$; if $D^2 g$ is bounded by Λ on the segment, the mean value inequality gives $\|Dg(r_1(x) - t\eta(x)) - Dg(r_1(x))\|_{\mathrm{op}} \le \Lambda t\|\eta(x)\| \le \Lambda\|\eta(x)\|$ for all $t \in [0,1]$, hence $L(x) \le \Lambda\|\eta(x)\|$. □

Corollary 5 (Single-carrier exactness). *$\Delta(x) \equiv 0$ identically, for every x, if $\eta(x) \equiv 0$ (in particular whenever only the MLP, and not the head, is ablated), since both terms inside the bracket of Theorem 3 then vanish. The same holds trivially if $Q = I$, since the prefactor $(I - Q)$ annihilates the bracket regardless of η (a case of no practical interest under Assumption 2, recorded only for completeness). Ablating the head* alone *($Q = 0$, generically $\eta(x) \neq 0$) satisfies neither condition: Theorem 3 still applies and bounds $\Delta(x) = g(r_1(x) - \eta(x)) - g(r_1(x))$, but nothing forces this quantity to vanish, and it generically does not: ablating a head changes its own layer's MLP output whether or not the MLP is separately ablated. Proposition 3 below makes precise what this costs the idealized model's own prediction in that case.*

Proof. If $\eta(x) = 0$: the bracket in Theorem 3 is $0 - 0 = 0$ independently of Q. If $Q = I$: the prefactor $(I - Q)$ is the zero operator, independently of the bracket. Neither is implied by $Q = 0$ alone. $\square$

Corollary 5 concerns $\Delta(x)$, the quantity the idealized model omits. A different single-carrier exactness statement, easily confused with it, concerns instead the *measurement protocol* used to estimate Δ empirically, and holds for a head just as much as for an MLP.

Proposition 2 (Single-carrier patch–edit equivalence). *Fix an input x and a single component whose write-in to the residual stream is $Wc(x)$, with $c(x)$ its own activation and W its output matrix. Ablate it by an orthogonal projector, applied either on the activation side (P, so that the intended write-in is $W(I - P)c(x)$) or on the stream side (Q, giving $(I - Q)Wc(x)$). Then the* weight-edit *route (edit W and run a fresh forward pass) and the* frozen-activation *route (leave all weights intact, replace this component's realized output by its projected value, and recompute everything downstream) assign identical values to every node of the network.*

Proof. $c(x)$ is computed strictly upstream of W, so editing W does not change it. On the activation side, applying the projector to the activation gives write-in $W\big((I - P)c(x)\big)$ and applying it to the matrix gives $\big(W(I - P)\big)c(x)$; these are the same vector. On the stream side, $(I - Q)\big(Wc(x)\big) = \big((I - Q)W\big)c(x)$ likewise. The two routes therefore agree on this component's write-in, every other weight is untouched, and hence every downstream node is computed from identical inputs by identical maps. $\square$

Remark 4. *Proposition 2 holds for* any *projector, in particular for one estimated from data: both routes use the same projector, so an error in choosing it is common to the two and cancels in their difference. A measured gap between the two routes on a single-carrier ablation is therefore neither evidence about $\Delta(x)$ nor evidence about how well the ablated direction was estimated, a point Section 8 returns to with measurements.*

Proposition 3 (Propagation to the readout). *Suppose that the block of Section 7 feeds the readout directly (Remark 6 discusses what is and is not available when it does not), i.e. $F(x) = r_1(x) + g(r_1(x)) + E(x)$ where $E(x)$ collects $F_0(x)$ and every carrier not drawn from this block. Let $S \subseteq \{1, \ldots, k\}$ consist only of carriers drawn from this block's own head and/or its own MLP (so that no carrier in S lies upstream of $r_1(x)$ and $E(x)$ is unaffected by ablating S): $\eta(x)$ is the head's contribution if the head's carrier is in S and 0 otherwise, and Q is the MLP's ablation projector if the MLP's carrier is in S and 0 otherwise (Remark 1 identifies both as genuine elements of $\{v_1, \ldots, v_k\}$). Then the idealized selector $s_S(x)$ of Section 5 and the selector $s_S^{\mathrm{true}}(x)$ of the genuinely weight-edited network agree up to exactly the interaction term of Theorem 3:*

$$s_S^{\mathrm{true}}(x) \;=\; s_S(x) \;+\; \psi\big(\Delta(x)\big), \qquad \big|\psi(\Delta(x))\big| \;\le\; \|\psi\|_{\mathrm{op}}\Big(\|Dg(r_1(x))\|_{\mathrm{op}}\|\eta(x)\| + L(x)\|\eta(x)\|\Big),$$

with $L(x)$ as in Theorem 3, and $\big|\psi(\Delta(x))\big| \le \|\psi\|_{\mathrm{op}}\big(\|Dg(r_1(x))\|_{\mathrm{op}}\|\eta(x)\| + \Lambda\|\eta(x)\|^2\big)$ under that theorem's additional $\|D^2 g\|_{\mathrm{op}} \le \Lambda$ hypothesis. In particular $s_S^{\mathrm{true}}(x) = s_S(x)$ exactly whenever Corollary 5's sufficient conditions hold, and Lemma 1 and Theorem 1, stated for s_S, transfer *to the true network without error in that case; otherwise they transfer with an error bounded as above, in particular whenever only the head, and not the MLP, is ablated.*

Proof. By Assumptions 1 and 3, $E(x)$ is the same function of x in the clean network and in $F_S(x)$; by the hypothesis that S draws only from this block, it is also unaffected by the true weight edit. The idealized model (1) deletes the head's carrier exactly and the MLP's carrier by subtracting its removed mass computed on the *clean* residual $r_1(x)$: every surviving carrier's coefficient, the MLP's included, is by (1) the same function of x as in the unablated network:

$$F_S(x) \;=\; F(x) - \eta(x) - Qg(r_1(x)) \;=\; \big(r_1(x) - \eta(x)\big) + (I - Q)g\big(r_1(x)\big) + E(x).$$

The true, weight-edited network instead recomputes the MLP on the genuinely ablated residual before applying its own ablation:

$$F_S^{\text{true}}(x) \;=\; \big(r_1(x) - \eta(x)\big) + (I - Q)g\big(r_1(x) - \eta(x)\big) + E(x).$$

Subtracting, $F_S^{\text{true}}(x) - F_S(x) = (I - Q)\big[g(r_1(x) - \eta(x)) - g(r_1(x))\big] = \Delta(x)$ by definition. Applying ψ and adding b gives $s_S^{\text{true}}(x) = s_S(x) + \psi(\Delta(x))$; the stated bounds follow from Theorem 3's bounds on $\|\Delta(x)\|$, $|\psi(\Delta(x))| \le \|\psi\|_{\text{op}}\|\Delta(x)\|$, and $\|I - Q\|_{\text{op}} \le 1$ since Q is an orthogonal projector. □

Remark 5 (Interpretation and limits)**.** *Corollary 5 and Proposition 3 together close the loop opened by Corollary 2 in Section 5: two ablated subsets can satisfy the idealized collapse hypotheses exactly, and hence agree with the true network, whenever $\psi(\Delta(x)) = 0$ for each, which $\eta(x) \equiv 0$ (an MLP-alone ablation) guarantees; otherwise the idealized selector and the true one differ by exactly $\psi(\Delta(x))$, which is bounded and generically nonzero, though not necessarily so at every x: $\Delta(x)$ may happen to lie in $\ker\psi$ there. Ablating a head at all, whether or not its own layer's MLP is separately ablated, therefore generically (though not necessarily) perturbs the idealized prediction; only an MLP-alone ablation is seen exactly for every x. Two limitations bound the scope of this result and are not closed here (a third, the absence of a closed form for the curvature constant, is closed in a companion analysis): Proposition 3 tracks $\Delta(x)$'s effect on the selector exactly only when this block feeds the readout directly, and propagating it through further nonlinear layers when it does not (the general case Remark 6 discusses only at the level of $\|\Delta(x)\|$, not of $\psi(\Delta(x))$) remains open; and the two-carrier composition considered here (one attention head and its own layer's MLP) does not by itself cover interactions between carriers in different layers. A companion analysis shows exactly how far the same technique reaches into that case: the same-block discrepancy at each touched layer is isolated exactly, in closed form only at the layer that feeds the readout directly, and what is left over (both the propagated same-block effect at every other touched layer and the genuinely cross-layer part) is isolated, not hidden, as a single separately measurable remainder.*

Remark 6 (Propagation across layers)**.** *The analysis above stops at the output of one block; in a deep network, $\Delta(x)$ is written into the residual stream and carried forward by every subsequent layer. Two effects govern its fate, both already implicit in the tools developed here. Every subsequent normalization layer acts on it through the same Jacobian bound as Lemma 2: since $\rho(r)$ shrinks as $\|r\|$ grows, normalization attenuates a fixed-size perturbation more strongly on residual streams of larger norm, one layer later. Every subsequent linear map, in turn, can amplify it by up to its own operator norm. A perturbation surviving L further layers is therefore bounded, heuristically, by a product of L alternating attenuation and amplification factors, none of which is guaranteed to stay below one. Whether this product typically grows, shrinks, or is absorbed by the residual stream's own accumulation on trained weights is an empirical question the present model does not answer. A companion analysis turns this into a sharper question than "how does a perturbation propagate": it gives an exact decomposition that isolates the entire cross-layer effect, for an arbitrary ablated subset spanning any number of layers, as one remainder term* *separate from the same-block discrepancy at each touched layer, closed form only at the layer feeding the readout directly and pinned to zero at any layer whose own carriers are MLP-only, without bounding either open piece there either, which we leave, along with the present question, for future work.*

8 Illustrative Experiments

The theorems of Sections 5–7 are statements about an idealized model. This section checks that their predictions (collapse under joint ablation, patch/ablation dissociation, and an interaction term that vanishes when the MLP alone is ablated but not, in general, otherwise) are realized in an actual trained network, rather than being vacuous under the stated assumptions. We report the minimum needed to make this check, not a general empirical study of the underlying task.

Task and vocabulary. We reuse the setting and the measurements of the empirical study these theorems were written to explain, restated here in the minimum detail needed to read the table and the two figures. Small transformers are trained on a synthetic associative-recall task: a context of key–value pairs is followed by a two-token query, a marker m_A or m_B and a displayed token, and the answer is the value $v(k)$ of the true key k. The marker selects the surface format. Under m_A the displayed token *is* the true key; under m_B it is $\sigma(k)$ for a fixed derangement σ, so the network must apply σ^{-1} before looking the value up. Both formats demand the same answer, so a competent network implements exactly a conditional of the kind formalized in Section 4, and the two formats are the two branches A and B of Definition 1.

Each branch has one *systematic* wrong answer, which is what makes a collapse observable rather than merely a drop in accuracy. On format A the wrong reading is the *inverted* one, $v(\sigma^{-1}(k))$: the network applies σ^{-1} where it should not. On format B it is the *literal* one, $v(\sigma(k))$: the network reads the displayed token at face value. All rates below are measured on filtered evaluation distributions on which the wrong reading is a well-defined token distinct from the correct one, so "inverted-A rate 0.93" means the ablated network returned that specific token on 93% of filtered format-A inputs.

Carriers and ablation configurations. Across five independently trained instances, a greedy activation-patching search identifies a small set of components (*carriers*, in the sense of Section 4) whose activation swap between formats flips the answer, a redundant code of the kind reported elsewhere in superposed and self-repairing networks [6, 7]. Each carrier is a single attention head or a single MLP, i.e. the concrete realization of a direction v_i. We then ablate carriers directly in weight space, projecting the low-rank subspace of their donor–receiver activation deltas out of the corresponding weight matrices, exactly as in Remark 1: an instance of the deletion (1) with the directions estimated from data. Three subsets S are used per instance: **D1**, the single strongest carrier alone; **DJ**, all first-layer carriers jointly (two of them, in every instance here); and **DJA**, all carriers, including those in deeper layers. The three are nested, $S_{\mathrm{D1}} \subset S_{\mathrm{DJ}} \subseteq S_{\mathrm{DJA}}$. Seed 11 has only two carriers in total and therefore has no DJA configuration, which is why the five instances yield 14 configurations rather than 15.

Table 1 illustrates Theorem 1 directly: on four of five instances, ablation collapses the network onto one unconditional reader but not, at the tested subsets, the other. Seed 44 is the exception that matters. There D1 deletes a single carrier, the rank-one head $h3$, and collapses the network onto the inverted reader (inverted-A rate 0.93, format B retained at 0.963); DJ deletes $h3$ *together with* the same block's MLP and collapses it onto the literal reader (literal-B rate 0.977, format A retained at 0.997). The two subsets are nested, $S_{\mathrm{D1}} \subset S_{\mathrm{DJ}}$: enlarging a subset that already produces a clean collapse does not deepen that collapse but reverses its polarity. By Corollary 2 this cannot happen inside the idealized model, so at least one of the two subsets must violate hypothesis (i) or (ii). Both subsets ablate the head $h3$, so by Corollary 5 neither is guaranteed

Instance	Total rank $\sum k$	Best inverted-A rate	Best literal-B rate
inst. 2	5	.86	.01
seed 11	6	.41	.47
seed 22	4	.49	.27
seed 33	5	.87	.20
seed 44	4	.93	.977

Table 1: Redundancy (≥ 2 carriers per instance) and low rank ($\sum k \leq 6$) hold on 5/5 instances. The two rightmost columns report, for each instance, the highest rate at which ablation collapses the network onto each of the two unconditional readers of Theorem 1 (the inverted reader on format A, the literal reader on format B) over the configurations D1, DJ and DJA. Each is therefore a maximum over the 2–3 configurations available for that instance, not a single measurement, and should be read as "some tested subset achieves this" rather than as a typical value. Rates are measured on the filtered distributions; the un-ablated baselines checked on these instances sit at 0 there (seed 44's is the leftmost group of Figure 1), so each entry is a collapse rate over a zero baseline, not a difference against a nonzero one. Seed 44 is the only instance in which *both* collapses are realized on the same trained weights, each with the untouched format retained at rate ≥ 0.96, by two *nested* subsets S.

the exact agreement Proposition 3 reserves for an MLP-alone ablation; DJ adds the same block's MLP on top, giving Theorem 3's interaction term every freedom to differ in size and sign between the two configurations, which is sufficient to flip which branch the idealized hypotheses would otherwise identify.

Patching versus ablation on the same site. The dissociation of Theorem 2 is visible on a further instance of the same task, on which the patch search returns a *single* sufficient site (the first-layer MLP output). Patching that site alone gives a median recovery of $\kappa = 1.043$ over 20 independent validation pairs (values 1.04–1.22 on the pairs used for the search) and an argmax flip rate of 0.85: recovery above 1, the overshoot signature of Remark 3. Ablating the weights feeding the same site leaves the receiver format correct on 57% of inputs: patch-sufficient, but far from ablation-necessary. The error it does produce is nonetheless the predicted one: after ablation, correct plus literal accounts for 0.993 of the answers, so what the ablation removes is the conditional and not the associative lookup. Sufficiency under patching and necessity under ablation are measured on the same component of the same network and disagree, exactly as $\beta_i\delta_i$ and $\beta_i\alpha_i(x_B)$ are permitted to.

Are the collapse hypotheses ever available? Theorem 1's conditions are exact equalities and hold on a null set, so the question that can actually be asked of a trained network is the one Proposition 1 poses: are the tolerances small enough, relative to $|\bar{s}|$, to certify anything? We measured $\bar{q}_S$, Ψ_S and $\bar{s}$ per matched pair on all 39 configurations (120 pairs each, on probe seeds again disjoint from every other measurement here). The identities (2) reproduce to 0 in floating point on all 4680 pair-configurations, which checks the algebra rather than the network.

The certificate is rarely available. Taking $\varepsilon_1 = |\bar{q}_S|$ and $\varepsilon_2 = |\Psi_S|$ at their measured per-pair values, the condition $|\bar{s}| > \varepsilon_1 + \frac{1}{2}\varepsilon_2$ holds on 10.0% of pair-configurations; no configuration satisfies it on more than 71.7% of its pairs, and none on 90%. The reason is structural rather than incidental. $\bar{s}$ is the *mean* of a positive and a negative margin on a matched pair, so it is small by construction (median $|\bar{s}|/\Phi = 0.083$), while the two tolerances are of the order of the margins themselves, median $|\bar{q}_S|/\Phi = 0.14$ and $|\Psi_S|/\Phi = 0.21$. In the median the sufficient condition misses by a factor of 5.3.

It is nonetheless a *sufficient* condition, and a conservative one: the conclusion it certifies (both selectors on the side of $\bar{s}$) holds on 44.5% of pair-configurations, four times more often than the

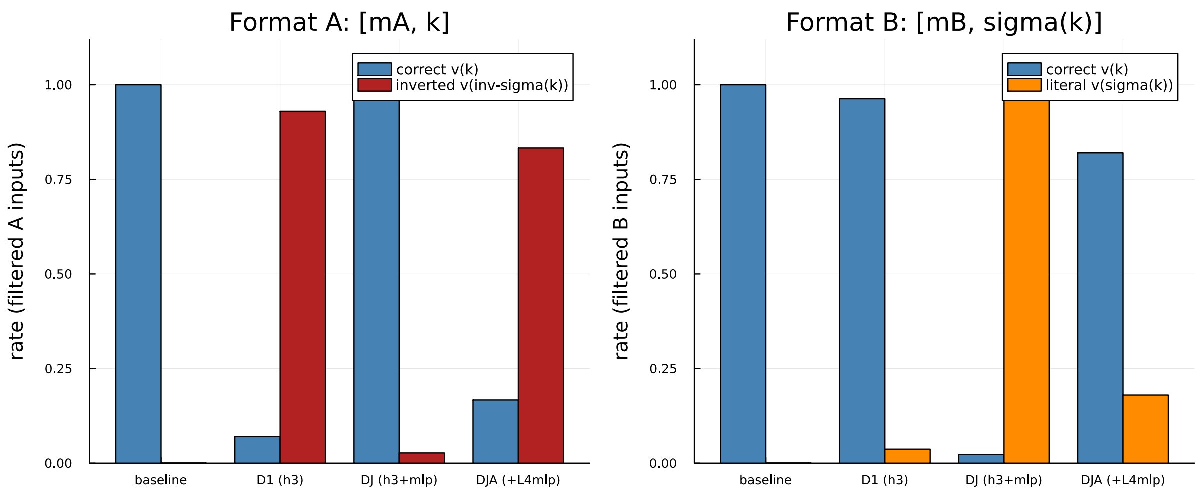

Figure 1: Seed 44: two *nested* ablated subsets ($S_{\mathrm{D1}} \subset S_{\mathrm{DJ}}$) collapse the same trained network onto *opposite* unconditional readers, each with the untouched format retained at rate ≥ 0.96. Left: filtered format-A inputs, correct answer $v(k)$ against the inverted reading $v(\sigma^{-1}(k))$. Right: filtered format-B inputs, correct answer against the literal reading $v(\sigma(k))$. The leftmost group of each panel is the un-ablated baseline. This is the phenomenon Corollary 2 shows cannot occur within the idealized model, and that Theorem 3 traces to an interaction term that Corollary 5 guarantees only an MLP-alone ablation avoids: neither D1 nor DJ qualifies for that guarantee here, since both ablate the head.

certificate fires, and no configuration was certified on at least half its pairs while collapsing on fewer than half, as Proposition 1 requires pairwise. It is also discriminating rather than uniformly pessimistic: certificate rate and realized collapse rate have Spearman $\rho = 0.66$ across the 39 configurations ($p \approx 1 \times 10^{-5}$), and mean collapse rate rises from 32% to 57% to 88% across the configurations whose certificate rate is below 5%, between 5% and 25%, and above 25%. On seed 44 it separates precisely the two configurations of Figure 1: D1, the clean single-carrier collapse, is the best-certified configuration in the entire set (71.7% of pairs certified, 96.7% collapsing), whereas DJ, the nested subset that reverses polarity, is certified on 0.8%, with $|\bar{q}_S| = 6.3$ against $|\bar{s}| = 4.2$: exactly the failure of hypothesis (i) that Corollary 3 shows any polarity reversal must exhibit.

We report this as a mixed result and prefer not to round it in either direction. The robust form is what makes Theorem 1 applicable to a real network at all, since the exact hypotheses never hold; but on these instances it certifies a minority of configurations, and its practical value here is as a diagnostic that ranks configurations by how close they are to the idealized regime, not as a guarantee covering the behavior we observe.

The interaction term. Finally we test Corollary 5 and Proposition 3 against the same 14 configurations. For each we measure two quantities. The *interaction magnitude* is the largest gap, over the probed inputs and in logit units, between the frozen-activation prediction of the ablated selector and the selector of the network whose weights were actually edited. For a *jointly* ablated head and MLP, the frozen-activation prediction holds the MLP's surviving output at its *clean-run* value (exactly the bookkeeping Assumption 3 prescribes for $F_S(x)$), while the true network recomputes it from the head-ablated residual; the gap between the two is exactly what Proposition 3 predicts, $\psi(\Delta(x))$. For a *single* ablated carrier, the standard frozen-activation protocol instead estimates the removed mass by patching that one carrier's own contribution and letting the rest of the network recompute naturally, and by Proposition 2 this reproduces the true weight edit node for node, for either carrier alone and regardless of Δ. The gap being

≈ 0 on a single-carrier configuration is therefore a fact about that measurement protocol, not, by itself, evidence that $\Delta(x) = 0$ there; Corollary 5 guarantees $\Delta(x) \equiv 0$ only for the single-carrier configurations that ablate the MLP. The *agreement rate* (C1a in Figure 2) is the fraction of inputs on which the idealized model's predicted answer matches the edited network's actual answer; a second, coarser criterion (C1b) asks whether the model's predicted inverted-A and literal-B *rates* match the observed ones to within ± 0.10. Both thresholds (0.90 and ± 0.10) were fixed in the verification script before any checkpoint was probed.

Is the single-carrier coincidence exact? The argument above says the frozen-activation protocol and the weight edit should agree exactly on a single-carrier configuration, but the directions being removed are estimated from data, and C1a on those configurations is not exactly 1. Whether the residual is the estimation error surfacing is a question about the pipeline's own numerics, so we measured it rather than argued it, on 400 held-out probe inputs per instance drawn by the procedure above but with seeds distinct from those behind any other number reported here. Over all five D1 configurations the per-input gap between the frozen-activation selector and the selector of the genuinely weight-edited network has median at most 9.6×10^{-7} and maximum 7.7×10^{-6} logits, against a median selector magnitude of 3.9–8.5 logits on the same inputs: single-precision round-off, six orders of magnitude below the signal.

No estimation error survives, and the reason is structural rather than fortunate. The patch and the weight edit apply the *same* estimated projector, one to the carrier's activation and the other to its output weight matrix; since the edit lies downstream of the activation it acts on, the two produce the same write-in exactly, and whatever error is made in estimating the subspace is made identically on both sides and cancels. That this test could have failed is worth recording: re-estimating the subspace on different data for the patch route *alone* (the configuration an estimation-error account describes) raises the same gap to a median of up to 0.69 and a maximum of 6.6 logits. An estimation error of the size these subspaces actually carry is thus easily large enough to see; the protocol simply does not expose it.

The residual between C1a and 1 therefore has a different and simpler source. Across the 2000 probe measurements there is not one input on which the ablated network returned one of the two candidate readings and the idealized model predicted the other; the entire residual is the 30 inputs (1.5%, nearly all on seed 22) whose argmax fell on a *third* token, outside the binary the prediction ranges over. On a single-carrier configuration C1a is thus exactly one minus the rate of such off-binary answers: a fact about where the ablated network's argmax lands, not a measure of disagreement with the idealized model. This reinforces rather than weakens the reading above: the D1 points of Figure 2 calibrate the measurement protocol, and it is the joint configurations that carry the evidential weight.

Does the separation hold out of sample? The band above was read off the 14 configurations that produced it, which makes it descriptive, not a validated threshold. Because the five checkpoints admit many more ablation subsets than the three per instance reported, we could test it. We froze the published band (6.78, 9.95) and the decision rule (a configuration passing both criteria must fall below the band, one failing either must fall above it), and applied them unchanged to *every* non-empty subset of each instance's carrier set: 39 configurations, of which 25 had never been used to fix anything, on a third disjoint set of probe seeds.

The rule fails. Nine of the 39 configurations violate it, six of them out-of-sample; seven land inside the supposedly empty band. Three of the original 14 change which side of the band they fall on when merely re-measured on held-out probes, so the separation is not stable even in sample. A split-half control on the same probes attributes part of this to sampling noise (3 of 39 configurations flip their pass/fail verdict between the two halves, all of them sitting near the 0.90 threshold), but not the overlap itself, which is far too large for that.

The monotone relationship, on the other hand, is robust. Interaction magnitude and C1a

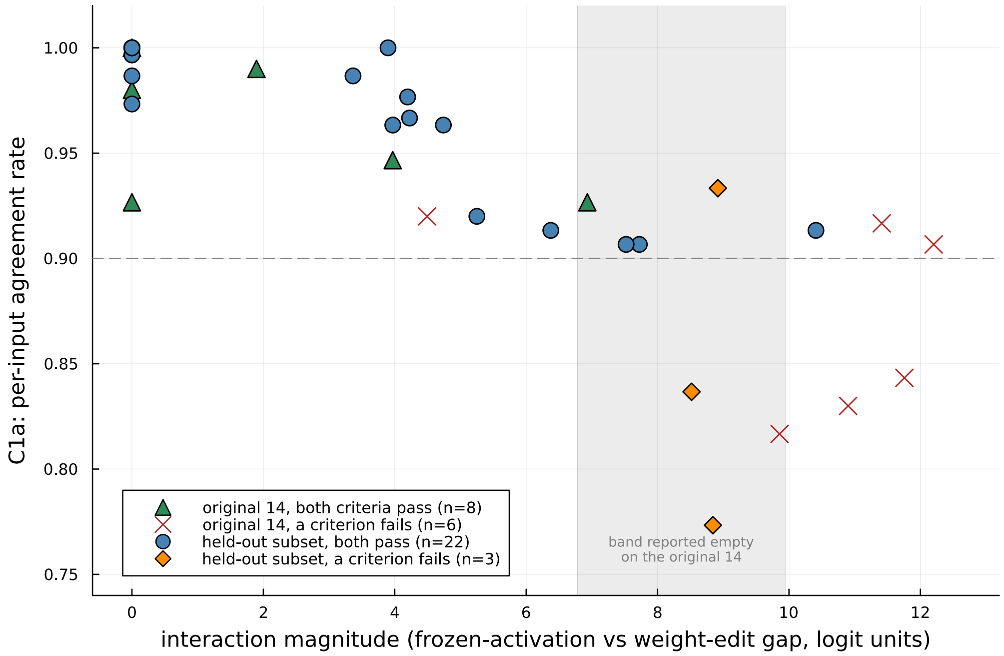

Figure 2: Agreement rate (C1a) against measured interaction magnitude over all 39 (instance, subset) configurations, on probes held out from every other number in this paper. Triangles and crosses are the 14 configurations originally reported (D1/DJ/DJA per instance); circles and diamonds are the 25 held-out subsets. The shaded region is the band that was empty on the original 14 and is used here only to show that it is *not* empty on the full set: 7 configurations fall inside it. Passing configurations reach interaction 10.41 and failing ones fall to 4.49, so the two groups overlap. What survives is monotone rather than sharp: Spearman $\rho = -0.83$ over all 39 (-0.85 on the 25 held-out ones alone), and the pass rate over the configurations with nonzero interaction falls $7/8 \to 6/8 \to 1/7$ across terciles of interaction magnitude. The five single-carrier ablations sit at interaction ≈ 0 (at most 8×10^{-6} logits, see the text) for protocol reasons rather than theoretical ones, and are not evidence either way.

have Spearman $\rho = -0.83$ over all 39 configurations (two-sided permutation test, $p < 10^{-5}$), -0.85 over the 25 held-out ones alone ($p < 10^{-5}$), and -0.81 over the 14 that are both held-out and have nonzero interaction ($p \approx 7 \times 10^{-4}$), so the association is not an artifact of the single-carrier configurations sitting at zero: if anything it is slightly tighter on the held-out subsets than on the full set, the opposite of what overfitting the original band would predict. We therefore withdraw the claim of a clean separating threshold and keep the weaker one the data support: the interaction magnitude (the same idealization gap Theorem 3 gives in closed form for a single head and its own layer's MLP, measured here for every tested subset including the ones the theorem does not cover analytically, such as carriers spanning more than one layer) predicts, monotonically and with a large effect, how far the idealized model of Sections 4–6 is from the network whose weights were actually edited. That is still the qualitative content the theory asserts (whether the idealization holds is governed by the size of a quantity derived from the architecture alone, before any data-dependent fitting), but it does not license reading a numerical cutoff off these experiments.

Two things are worth separating from that conclusion. The *measurement gap* between the frozen-activation protocol and the true edit ($\leq 8 \times 10^{-6}$ logits on all five single-carrier configurations, by the patch-equals-weight-edit identity of Proposition 2 rather than by Corollary 5 whenever the ablated carrier is a head) is not the same quantity as $\Delta(x)$ itself, which the corollary guarantees absent only when the ablated carrier is the MLP. And the thresholds themselves (C1a ≥ 0.90, C1b within ± 0.10) were fixed in the verification script before any checkpoint was probed, so the failure above is a failure of the band, not of a threshold chosen after the fact.

8.1 A second task and architecture

Everything above illustrates the theory on one task and one family of architectures. To check that the predictions are not an artifact of the marker task's specific mechanism (an externally fixed derangement applied to a displayed token, then read in a single, fixed direction), we built a second, freshly trained instance with a different conditional mechanism and a different architecture, and re-ran the two checks that do not themselves presuppose a matched-pair, two-branch structure, plus a search for a polarity-reversal analogue.

Task and architecture. The context is 4 key–value pairs from a shared vocabulary of 12 symbols (keys mutually distinct, values mutually distinct, drawn independently), mixed in random order, followed by a marker and a displayed token. Under m_{F} the displayed token is a pair's key k_t and the answer is its value v_t: ordinary associative recall. Under m_{R} the displayed token is v_t itself and the answer is k_t: a genuine *inverse* lookup, from value back to key, that m_{F} never requires and that involves no externally fixed permutation of any kind. The shared contrast axis is $s_{\mathrm{op}}(x) := \mathrm{logit}(v_t) - \mathrm{logit}(k_t)$: format F is correct when this favors v_t (positive), format R when it favors k_t (negative), structurally the same matched-pair mechanism as $s_{\mathrm{op}} = \mathrm{logit}(\mathrm{lit}) - \mathrm{logit}(\mathrm{inv})$ in Section 8, but with the two roles filled by a genuine forward/inverse recall rather than by a token relabeled under a fixed σ. The network is a 3-layer, 3-head Llama-style transformer, width 48 (head dimension 16), SwiGLU hidden width 96, smaller and shallower than the 4-layer, 4-head, width-64 networks above, on a 14-token vocabulary rather than 10. It was trained fresh, once, for this check (never on the checkpoints used elsewhere in this paper), with the same AdamW/warmup recipe as Appendix A, and reached 99.5%/98.5% accuracy on formats F/R after 2750 of a 6000-step budget: the same qualitative training dynamic (a long plateau below 35% followed by a rapid transition) as the marker task's own instances.

Carriers. The same patch-recovery sweep and $r \geq 0.25$ threshold used throughout this paper finds only two carriers on this checkpoint, both MLPs (layer 1, $r = 0.88$; layer 2, $r = 0.35$), giving 3 non-empty subsets: too few for the interaction/fidelity relationship below to mean anything statistically. The full ranked sweep shows why: after those two, recovery drops to 0.19 (layer 3's

MLP) and 0.18 (a layer-1 head), then drops again, by more than a factor of two, to 0.09 and below for every remaining site. We report both readings rather than picking one: at the paper's own threshold, this task concentrates onto two MLPs and nothing else passes; at $r \geq 0.15$ (still twice the next site down, not a threshold tuned to reach a target count) four carriers survive, spanning all three layers and including one head, giving 15 non-empty subsets. The three checks below use this second, wider carrier set, since it is the one that can actually support them; where the two readings disagree, both are reported.

Patch-equals-weight-edit exactness. On the single strongest carrier (layer 1's MLP, $r = 0.88$) and 120 fresh probes, the gap between the frozen-activation patch and the genuinely weight-edited network has median 9.5×10^{-7} and maximum 3.8×10^{-6} logits, against a median selector scale of 6.99, the same six-orders-of-magnitude, single-precision-floor gap as the five D1 configurations of Section 8 (Proposition 2 makes no reference to the task), and the falsifiability control (the same protocol with the patch route's subspace re-estimated on disjoint data) again shows what a genuine mismatch would look like: median 2.44, maximum 11.6 logits, three orders of magnitude larger.

Interaction versus fidelity. Over the 15 configurations at $r \geq 0.15$, interaction magnitude and the agreement rate C1a are anti-correlated, Spearman $\rho = -0.90$ (two-sided permutation test, $p \approx 3 \times 10^{-5}$), the same relationship, on a different task and a smaller, differently shaped network, that Section 8 reports at $\rho = -0.83$ over its own 39 configurations. (At $r \geq 0.25$ the same relationship is present on the 3 available configurations, $\rho = -1.0$, but three points do not constitute independent evidence of a monotone relationship on their own; we report it for completeness, not as a replication.) The off-binary ("other") rate is higher on this task, up to 47.5% on the largest and one other joint configuration against 1.5% on the marker task's D1 points, which is consistent with a smaller, shallower network producing more answers outside the two candidate readings once several carriers are ablated together, and is exactly the kind of number Section 8 already argues C1a's residual should track.

A polarity-reversal analogue. Scanning every nested pair of the 15 configurations against the 60 probed pairs finds eight instances, across two distinct probe pairs, in which enlarging an ablated subset that already collapses a pair cleanly reverses which branch it collapses onto: the same phenomenon Figure 1 illustrates on the marker task. The cleanest is layer-internal, exactly as on seed 44: ablating layer 1's MLP alone collapses one probed pair onto one branch, while ablating that same MLP *together with* layer 1's own head collapses it onto the other. A second instance spans more than one layer, adding layer 2's MLP to an already-collapsing layer-1-and-layer-3 pair to reverse the branch; it is not covered by this paper's own two-carrier, single-block composition (Theorem 3), and extending the analysis to carriers at different layers is left to future work. By Corollary 2, this is again evidence that at least one of each pair of subsets fails the exact collapse hypotheses, consistent with the idealized model rather than contradicting it.

What this does and does not establish. All three checks land the same way they do on the marker task, which is the point of running them: the predictions are not an artifact of one conditional mechanism or one architecture shape. It remains a single trained instance, not five, and the carrier count needed widening past this paper's own stated threshold to give the second check enough points to be informative: both are limitations we would rather state than round past. We did not attempt a second out-of-sample band test or a second robust-collapse-availability measurement here: both are properties of *how much data* a claim was checked against, which a single additional instance cannot settle either way, and re-running them here would not add evidence beyond what Section 8 already reports on that question.

9 Discussion

What is established, and how far it reaches. The flip criterion (Lemma 1), the collapse theorem (Theorem 1) and the dissociation theorem (Theorem 2) rest on Assumption 4 and on the additive decomposition of Section 4, and on nothing else. In particular none of them uses any property of transformers: they hold for any residual computation with a linear readout, and their content is arithmetic on the two numbers $\beta_i\delta_i$ and $\beta_i\alpha_i(x_B)$ that the two interventions respectively move. Proposition 1 removes the objection that the collapse criterion is stated as an exact equality and therefore never applies: the error is an identity, not an approximation, and the criterion has a tolerance form. What that form buys on real weights is measured, and reported honestly as partial, in Section 8. What they do *not* establish is that a real network satisfies the decomposition; that is the role of Theorem 3, which computes the error exactly for the composition in which the independence of carriers is architecturally false, and of Corollary 5, which shows the failure is identically absent whenever the ablated carrier is the MLP, and, by Proposition 3, exactly bounded whenever it is not, a bound whose governing constant is in fact computable in closed form from the trained weights alone, which a companion analysis to this one exhibits explicitly. The scope of the idealized model is therefore not assumed but delimited from inside it.

What the experiments establish is weaker than what a first pass suggested, and we prefer to say so in both places. The interaction magnitude tracks the idealized model's accuracy monotonically and strongly across 39 ablation configurations, but the clean numerical separation visible on the first 14 is not reproducible on the other 25. A reader should take from Section 8 that the theory identifies the right *quantity*, not that it supplies a usable decision threshold; supplying one would require establishing that the cutoff transfers across networks, which these five instances do not show.

What remains open. The theory answers *when* collapse and dissociation occur given a specific ablated subset, but not the converse question of *which* subset to choose in order to realize a target outcome: Corollary 2 shows that any two idealized-model-satisfying subsets must agree, but gives no procedure for constructing one, and does not address whether a subset can be *designed*, from measurable quantities alone, to steer the collapse polarity at will. Two further gaps are specific to Theorem 3: the analysis is not extended past the block in which the two carriers reside, though Remark 6 identifies the two mechanisms (normalization-driven attenuation and linear-map amplification) through which a quantitative account would proceed; and multi-layer carrier interactions are not covered by the two-carrier composition considered here. Each is a well-posed mathematical question rather than an open-ended empirical one, and each is a natural target for the idealized model of Section 4 to be extended rather than replaced. A fourth gap is empirical rather than mathematical: every result of Section 8 is illustrated on a small transformer trained on a synthetic conditional task, chosen because it makes the two branches of the conditional and the ground-truth answer unambiguous by construction. Mechanistic interpretability is ultimately concerned with models trained on natural language, and nothing in Sections 4–7 is specific to the synthetic task; extending the theory past a single residual block and testing it against an emergent circuit in a real pretrained model is exactly the direction a companion analysis takes, and we leave that extension, and its own genuinely mixed result, to it rather than duplicate it here. The out-of-sample failure reported in Section 8 nonetheless sharpens what any such extension would have to show. It is not enough to exhibit a correlation between interaction magnitude and idealization error; that already replicates on this paper's own synthetic instances. The open question is whether any threshold on that magnitude transfers, between configurations or between networks, and our five instances answer it negatively at the only scale we tested here, which makes the question more interesting rather than less.

10 Conclusion

Activation patching and weight-space ablation are not two measurements of one causal quantity. They are two operators on two different objects, and the useful question is not whether they agree but exactly when, and by how much they fail to when they do not. Within one idealized model this becomes a matter of arithmetic: patching moves a contrast, ablation removes a level, and the collapse theorem of Section 5 says precisely which subsets of carriers force a conditional onto a single branch and which branch survives.

The practical reading is that a negative ablation result is not, on its own, evidence that a component is unimportant, and a positive patching result is not, on its own, evidence that it is necessary, not as a caution about noisy measurement, but because the two experiments interrogate different quantities that a redundant code routinely separates. Where the idealization itself fails, the failure is not noise either: it is an interaction term of known order, with a provable zero whenever the ablated carrier is an MLP and an exact second-order bound otherwise whose constant is computable from the trained weights, and it is measurable on the network one is actually studying.

Measurable is not the same as calibrated, and we close on that distinction rather than blur it. Across thirty-nine ablation configurations on five trained networks the interaction magnitude orders the idealization's accuracy strongly and reproducibly, but the threshold that appeared to separate success from failure on the first fourteen did not survive the other twenty-five: the quantity singled out by the theory is the right one, how to read a numerical cutoff off it is not yet settled, and we would rather leave that explicit than let a tidy figure imply otherwise.

A Experimental Details and Reproducibility

Section 8 is deliberately compressed; this appendix gives what is needed to reproduce it.

Task. Key and value tokens are drawn from a vocabulary of $V = 8$ symbols, with two additional marker tokens m_A, m_B, so the model's vocabulary has 10 entries. Each sequence presents 3 key–value pairs in random order (6 tokens) followed by a marker and a displayed token, giving sequence length 8; the target is the value of the true key, read at the final position. The derangement σ is drawn once from a fixed pseudorandom seed (42) and is *identical across all instances*, so only the trained model varies between seeds.

Architecture and training. Each instance is a 4-layer decoder transformer in the Llama style: model width $d = 64$, 4 attention heads (head dimension 16), SwiGLU MLP with hidden width 128, pre-normalization by RMSNorm with $\varepsilon = 10^{-6}$, learned positional embeddings, weights in float32. There is no final normalization: the unembedding is applied directly to the residual stream, so the logit contrast is an exactly affine functional of $F(x)$ and Assumption 4 holds exactly rather than approximately for these networks. Training is 5000 optimizer steps at batch 64 with AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 10^{-8}$), peak learning rate 10^{-3}, linear warmup over 250 steps then cosine decay to $0.1\times$ peak. Only instances reaching accuracy ≥ 0.95 on *both* formats are analyzed; this gate is applied before any ablation. Four instances use initialization seed $S \in \{11, 22, 33, 44\}$ with training-data seed $1000S + 123$; the fifth ("inst. 2") was trained earlier under the same configuration with a different seed. No model was retrained for any measurement reported here: all of them read the same five checkpoints, with one exception: the second task and architecture of Section 8.1 is a genuinely separate, freshly trained instance (`notebook/bidir_recall_task_experiment.jl`, initialization seed 1, training seed 123, 2750 steps at batch 64 before the $\geq 0.97/\geq 0.97$ early-stop gate, otherwise the same optimizer recipe), never used for any other number in this paper.

Carrier selection. The candidate set is every attention head and every MLP output, $4 \times (4 + 1) = 20$ sites. On 8 matched clean pairs (seed 4242) each site is patched from donor to receiver in turn and scored by the normalized recovery of the literal-versus-inverted logit contrast; r denotes the mean over pairs. **A site is a carrier when $r \geq 0.25$.** The *carrier layer* is the shallowest layer containing a carrier. The three configurations of Section 8 are D1 (the highest-r carrier in the carrier layer, alone), DJ (all carriers in the carrier layer), and DJA (all carriers); DJA is omitted when it coincides with DJ, which is why seed 11 contributes two configurations rather than three.

Ablation subspaces. For each carrier, donor–receiver activation differences are collected at the last two sequence positions over 32 matched pairs (seed 1618), giving 64 vectors. Their SVD determines U as the leading left singular vectors, taking **the smallest k reaching 90% of the squared-singular-value energy, capped at $k \leq 4$ for a head and $k \leq 8$ for an MLP.** The weight edit is $W_O[:, \mathcal{H}] \leftarrow W_O[:, \mathcal{H}](I - UU^\top)$ for a head with column block $\mathcal{H}$, and $W_2 \leftarrow (I - UU^\top)W_2$ for an MLP. These are the concrete instances of (1) referred to in Remark 1, with the directions estimated from data.

Evaluation criteria. C1a and C1b are as defined in Section 8; the thresholds 0.90 and ± 0.10 were fixed in the verification script before any checkpoint was probed and were not revised. Rates are computed on filtered distributions on which the systematic wrong answer is a well-defined token distinct from the correct one.

Probe seeds. Every measurement added after the original run uses probe seeds disjoint from it and from each other, so that no two reported numbers share an evaluation sample:

- the original configuration run uses 90210/31337;
- the single-carrier protocol-gap measurement uses 20260726/20260727 (400 probe inputs per instance);
- the out-of-sample band test uses 20260728/20260729 (300 per configuration);
- the robust-hypothesis measurement uses 20260801/20260802 (120 matched pairs per configuration);
- the second task and architecture of Section 8.1 uses its own carrier-sweep seed 53421, delta seed 91827, and probe seeds 20260810/20260811 (60 matched pairs), all independent of every seed above and of each other, since it is a separate checkpoint with no other number in this paper to stay disjoint from.

The permutation tests use 2×10^5 resamples.

Code and data availability. All experiments run against the research repository accompanying this submission; no external data is used and no model is downloaded. Each script writes a JSON results file that the figure scripts read directly, so no number in a figure is transcribed by hand. The relevant files are:

- `notebook/marker_task_experiment.jl` — task, architecture, training.
- `notebook/marker_seed_matrix.jl` — per-seed training and carrier sweep.
- `notebook/marker_conj1_verify.jl` — the original 14 configurations, C1a/C1b and interaction magnitude.
- `notebook/verify_marker_interaction_theorem.jl` — numerical check of Theorem 3.

- `notebook/verify_d1_protocol_gap.jl` — Proposition 2, the single-carrier gap.
- `notebook/verify_config_band_oos.jl` — the 39-configuration out-of-sample test.
- `notebook/verify_robust_collapse.jl` — Proposition 1, the $\bar{q}_S/\Psi_S/\bar{s}$ measurement.
- `notebook/verify_interaction_spearman.jl` — the Spearman correlations and permutation p-values reported in Section 8, from the JSON output of the two preceding scripts, computed with tie-corrected rank correlation matching, respectively:
  - `StatsBase.corspearman` (Julia);
  - `scipy.stats.spearmanr` (Python);
  - R's `cor(method="spearman")`.

  An order-dependent, non-tie-corrected rank statistic reproduces the values reported in an earlier draft of this section and is retained in the script as a documented control.
- `notebook/bidir_recall_task_experiment.jl` — the second task's architecture, training, and checkpoint of Section 8.1.
- `notebook/verify_bidir_replication.jl` — its three checks, run at both carrier thresholds discussed there.

The five marker-task checkpoints are stored under `notebook/marker_ckpt/`, the second task's checkpoint under `notebook/bidir_ckpt/`.

References

[1] Vig, J., Gehrmann, S., Belinkov, Y., Qian, S., Nevo, D., Singer, Y., Shieber, S. (2020). Investigating Gender Bias in Language Models Using Causal Mediation Analysis. *Advances in Neural Information Processing Systems 33.*

[2] Geiger, A., Lu, H., Icard, T., Potts, C. (2021). Causal Abstractions of Neural Networks. *Advances in Neural Information Processing Systems 34.*

[3] Elhage, N., Nanda, N., Olsson, C., et al. (2021). A Mathematical Framework for Transformer Circuits. *Transformer Circuits Thread.*

[4] Meng, K., Bau, D., Andonian, A., Belinkov, Y. (2022). Locating and Editing Factual Associations in GPT. *Advances in Neural Information Processing Systems 35.*

[5] Wang, K., Variengien, A., Conmy, A., Shlegeris, B., Steinhardt, J. (2023). Interpretability in the Wild: a Circuit for Indirect Object Identification in GPT-2 small. *ICLR.*

[6] Elhage, N., et al. (2022). Toy Models of Superposition. *Transformer Circuits Thread.*

[7] McGrath, T., Rahtz, M., Kramár, J., Mikulik, V., Legg, S. (2023). The Hydra Effect: Emergent Self-Repair in Language Model Computations. *arXiv:2307.15771.*

[8] Goldowsky-Dill, N., MacLeod, C., Sato, L., Arora, A. (2023). Localizing Model Behavior with Path Patching. *arXiv:2304.05969.*

[9] Conmy, A., Mavor-Parker, A. N., Lynch, A., Heimersheim, S., Garriga-Alonso, A. (2023). Towards Automated Circuit Discovery for Mechanistic Interpretability. *Advances in Neural Information Processing Systems 36.*

[10] Arditi, A., Obeso, O., Syed, A., Paleka, D., Panickssery, N., Gurnee, W., Nanda, N. (2024). Refusal in Language Models Is Mediated by a Single Direction. *Advances in Neural Information Processing Systems 37*.

[11] Makelov, A., Lange, G., Nanda, N. (2024). Is This the Subspace You Are Looking For? An Interpretability Illusion for Subspace Activation Patching. *ICLR*.

[12] Zhang, F., Nanda, N. (2024). Towards Best Practices of Activation Patching in Language Models: Metrics and Methods. *ICLR*.

[13] Geiger, A., Wu, Z., Potts, C., Icard, T., Goodman, N. D. (2024). Finding Alignments Between Interpretable Causal Variables and Distributed Neural Representations. *Causal Learning and Reasoning (CLeaR)*.