MECHANISTIC LOCALIZATION IS DISTRIBUTION-RELATIVE: RECEIVER SUFFICIENCY, CONTEXT INVARIANCE, AND THE SUPPORT SHADOW

Anonymous authors
Paper under double-blind review

ABSTRACT

Mechanistic localization is distribution-relative. A component set does not determine the optimal rule obtainable from the internal representation it exposes: even with the network, receiver, phenomenon, and literal support fixed, changing only probability weights can reverse the Bayes-optimal rule. We therefore include the realized probability state in the localization problem and distinguish a structural location K, a receiver-level presentation (K, h)—the selected receiver set together with a decoder—and the induced action $[hQ_K]_\mu$. The receiver map determines what state is exposed; an admissible decoder class determines how that state may be used. We take all measurable receiver-level decoders as the maximal receiver-sufficiency class and recover architecture-specific continuation classes by restriction. This separation yields two different laws. Passive receiver refinement can only improve predictive adequacy, but it does not order context invariance. Under log loss, the value of revealing context is $I(Y_\Phi; C \mid Q_K(X))$, and adding receiver j changes this value by $I(Y_\Phi; Q_j \mid Q_K, C) - I(Y_\Phi; Q_j \mid Q_K)$; either sign occurs. Exact localization is strictly coarser. In finite 0–1 settings, zero risk forgets probability weights and reduces to support-level factorization, functional dependency, and decision-reduct structure. Fixing the network does not fix this exact localization family: one Boolean OR network with fixed coordinate receivers realizes every nonempty upward-closed exact localization family by varying support alone. A circuit label is therefore not a complete mechanistic localization claim without the realized distribution, access, admissible downstream use, phenomenon, and adequacy criterion under which it is asserted.

1 INTRODUCTION

A circuit-localization claim names a subset of internal components and asks whether the exposed state through those components suffices for a phenomenon. Decision-theoretic representation theory asks whether an exposed representation is Bayes-sufficient for a specified distribution and loss (Sevetlidis, 2026). That fixes the probability law while asking what the representation makes available. Mechanistic localization then asks what survives when the network, receiver, phenomenon, loss, and literal support stay fixed while the realized probability law changes.

Even the smallest finite example shows that the receiver-level rule need not survive. Let

$$X = \{0,1\}^2, \qquad \Phi(x_1, x_2) = x_1, \qquad Q_2(x_1, x_2) = x_2.$$

For $0 < \eta < 1/2$, let μ_η put mass $(1-\eta)/2$ on each diagonal point $00, 11$ and mass $\eta/2$ on each off-diagonal point. Let ν_η reverse those weights. The represented state space, receiver, phenomenon, and literal support are unchanged:

$$\operatorname{supp} \mu_\eta = \operatorname{supp} \nu_\eta = X.$$

Under 0–1 loss, however, the optimal rule through the same receiver reverses,

$$h_{\mu_\eta}(z) = z, \qquad h_{\nu_\eta}(z) = 1 - z,$$

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

and has error η in either state. Their equal mixture is uniform; there Q_2 carries no predictive information about Φ and the minimum error is $1/2$. Passing from the probability law to its support erases exactly the information that selects the receiver-level rule. Thus neither a component set nor a support-relative circuit determines the receiver-level realization selected by the probability distribution. We therefore treat that distribution as part of the realized computational state.

Recent results often grouped under "circuit variability" vary different objects. Circuit structure changes under input variation, resampling, and prompt rephrasing, while transfer and interchange tests can make some structural differences functionally weak (Bayat Makou et al., 2026; Wu et al., 2026). A behavior can admit several sparse or otherwise admissible circuit descriptions (Chen et al., 2026; Méloux et al., 2025); formal verification can separately certify robustness or minimality for a declared candidate (Hadad et al., 2026). Mediator choice changes the exposed internal state, and interventions can move that state away from the natural distribution without changing learned parameters (Mueller et al., 2026; Grant et al., 2026). These are not one kind of instability: they change different arguments of the mechanistic claim.

Existing formalisms already expose some of these arguments. Causal abstraction fixes maps between computational systems; verification-based discovery certifies a declared circuit on an explicit domain; MIB fixes benchmark tracks and evaluation targets for circuit and causal-variable localization (Geiger et al., 2025; Hadad et al., 2026; Mueller et al., 2025). Here we hold the ambient network, receiver access, phenomenon, admissible receiver-to-action class, and decision criterion fixed and vary the realized probability state. This isolates a precise question: which parts of a localization claim survive a change in the law governing the realized states?

A deterministic network transports a source law μ_0 to a realized probability state at every represented cut. At one such cut, a localization specification fixes the realized state μ, receiver access Q_K, a phenomenon Φ, admissible downstream use $\mathscr{H}_K$, and an adequacy criterion ℓ. Operationally, $Q_K(X)$ is the internal representation made available to the analysis, while a decoder $h \in \mathscr{H}_K$ specifies how that representation is used to predict or realize the target phenomenon. The main theory takes all measurable receiver-level decoders as the maximal receiver-sufficiency class and obtains architecture-specific localization by restricting $\mathscr{H}$.

For $h \in \mathscr{H}_K$, the resulting hierarchy is

$$\boxed{K \longrightarrow (K, h) \longrightarrow [hQ_K]_\mu.} \tag{1}$$

In ordinary terms, these are the selected internal variables, those variables together with a decoder, and the predictor that this decoder induces on the realized distribution. Location is not presentation, and presentation is not realized action. The first object records where the analysis reads; the second records one admissible rule applied to that state; the third records the induced action on the realized population up to almost-sure equality. The opening example holds K fixed while changing the optimal presentation and realized action. A component set is therefore not a complete localization specification.

This distinction yields two different laws. Inside one fixed passive access family,

$$K \subseteq K' \quad \Longrightarrow \quad \delta_\mu(K') \leq \delta_\mu(K) :$$

more exposed receiver state cannot worsen optimal prediction. Context-invariant realization obeys no corresponding monotonicity law. If C is a context random variable jointly distributed with X, then under log loss

$$\Gamma_K^{\log} = I(Y_\Phi; C \mid Q_K(X)),$$

and adding receiver j changes this quantity by

$$\boxed{\Gamma_{K\cup\{j\}}^{\log} - \Gamma_K^{\log} = I(Y_\Phi; Q_j \mid Q_K, C) - I(Y_\Phi; Q_j \mid Q_K).} \tag{2}$$

Either sign occurs. Receiver refinement therefore orders predictive adequacy but does not order context invariance: more internal information can make a common rule easier or harder to sustain across distributions.

Exact localization is a strictly coarser problem. Almost-sure exact realization depends only on measure class; in finite 0–1 problems, zero risk depends only on support and becomes support-level

factorization, functional dependency, and decision-reduct structure. This coarse-graining destroys probability-weight information by construction. It also destroys any hope that architecture alone selects a canonical exact circuit at the receiver-sufficiency level. We prove that one fixed Boolean OR network, with fixed coordinate receivers and a fixed Boolean target, realizes every nonempty upward-closed exact localization family by varying support alone. The paper therefore establishes both sides of the distribution-relative picture: quantitative receiver adequacy has a monotone refinement law, while context invariance and exact circuit identity remain distribution-relative. The appendices carry the architecture-specialization result, longer finite geometry, the categorical formulation, exact descent constructions, and a map from the localization specification to representative empirical discovery procedures.

2 REALIZED LOCALIZATION AND PREDICTIVE REFINEMENT

Probability state is not metadata attached to a neural computation. It determines which receiver-level rule is optimal and therefore belongs in the realized localization problem. We include it in the process before defining receiver sufficiency.

Realized process. Let $X_0, \dots, X_L$ be standard Borel represented state spaces of a deterministic feed-forward network. Write

$$F^{\theta}_{q\rightsquigarrow r} : X_q \to X_r, \qquad A^{\theta}_q := F^{\theta}_{0\rightsquigarrow q},$$

for its continuations and prefixes. A source law $\mu_0 \in \mathsf{Prob}(X_0)$ induces a realized probability state

$$\mu_q = (A^{\theta}_q)_*\mu_0, \qquad \mu_r = (F^{\theta}_{q\rightsquigarrow r})_*\mu_q \quad (q \le r). \tag{3}$$

Thus one input distribution induces a probability state on the represented space at every cut of the same fixed network. We call (X_q, μ_q) the realized state at cut q, and the sequence of these transported states the realized neural process. Appendix A packages this transport categorically as the Grothendieck lift of the probability functor and records its formal properties.

Equation (3) describes source-law variation through one fixed network. An internal intervention can instead supply a new boundary law at a represented cut; after that state is supplied, the same downstream construction applies to the rebased suffix. This paper studies the resulting realized process. Modeling the intervention operation that connects the natural and altered processes is a different comparison problem.

Declared access and presentations. Localization is relative to an access interface. Fix a represented cut—an internal layer or interface at which the analysis reads the computation—and suppress its index. A mechanistic protocol declares a finite receiver family of measurable maps

$$Q_j : X \to Z_j, \qquad j \in J,$$

into standard Borel receiver spaces. A receiver is any declared internal variable or readout used by the localization analysis; neurons, attention heads, sparse features, edges, and causal mediators instantiate different choices of Q_j (Mueller et al., 2025; 2026). Receiver access need not be coordinate-local: learned projections or distributed subspaces are also admissible receiver maps (Geiger et al., 2024; Wu et al., 2023). The protocol also declares a target phenomenon $\Phi : X \to Y$ into a standard Borel target space Y, representing the behavior or quantity to be localized. Writing the represented state as a random variable $X \sim \mu$, set $Y_\Phi := \Phi(X)$. For $K \subseteq J$, set

$$Z_K = \prod_{j\in K} Z_j, \qquad Q_K = \langle Q_j\rangle_{j\in K}.$$

If $K \subseteq K'$, coordinate projection $p_{K'K} : Z_{K'} \to Z_K$ satisfies

$$Q_K = p_{K'K}Q_{K'}. \tag{4}$$

These choices define different receiver interfaces rather than interchangeable descriptions of one fixed access map.

Fix a standard Borel action space A and measurable loss $\ell : Y \times \mathsf{A} \to [0,\infty]$. Define the maximal receiver-level decoder class

$$\mathsf{Dec}_{\mathsf{A}}(K) := \{h : Z_K \to \mathsf{A} \mid h \text{ measurable}\}.$$

A concrete downstream architecture or continuation protocol can instead declare a restricted decoder hypothesis class $\mathscr{H}_K \subseteq \mathsf{Dec}_{\mathsf{A}}(K)$. These are distinct sources of limitation: Q_K is the declared receiver access, $\mathscr{H}_K$ is the admissible downstream use, and ℓ is the adequacy criterion. The main theory takes $\mathscr{H}_K = \mathsf{Dec}_{\mathsf{A}}(K)$ to isolate receiver sufficiency itself. Architecture-specific localization is obtained by restricting $\mathscr{H}$.

A receiver-level presentation is a receiver set together with one admissible decoder, (K, h) with $h \in \mathscr{H}_K$. Two presentations are state-relative equivalent when

$$(K, h) \equiv_\mu (K', h') \quad \Longleftrightarrow \quad hQ_K = h'Q_{K'} \quad \mu\text{-a.s.}$$

Thus presentations are identified when they induce the same predictor on the realized population, even if they use different internal variables. This is the formal version of equation (1). Different locations can induce the same realized action: on $X = \{0, 1\}$, the accesses $Q_1(x) = x$ and $Q_2(x) = 1 - x$ paired with $h_1(z) = z$ and $h_2(z) = 1 - z$ satisfy $h_1 Q_1 = h_2 Q_2$ pointwise. The opening example gives the converse: one fixed location supports different optimal rules under different states. Location, presentation, and realized action are therefore distinct levels of mechanistic description. Appendix A records the presentation category, architecture-specific decoder subfamilies, and measure-class properties.

Receiver Bayes risk. For an admissible rule h, define its receiver risk by

$$R_\mu(K, h) := \mathbb{E}_\mu\big[\ell\big(\Phi(X), h(Q_K(X))\big)\big].$$

For a declared decoder family $\mathscr{H}$, define

$$\delta_{\mu,\mathscr{H}}(K) := \inf_{h \in \mathscr{H}_K} R_\mu(K, h). \tag{5}$$

The maximal receiver-sufficiency risk used throughout the main results is

$$\delta_\mu(K) := \delta_{\mu,\mathsf{Dec}_{\mathsf{A}}}(K) = \inf_{h: Z_K \to \mathsf{A}} R_\mu(K, h). \tag{6}$$

For fixed μ, this is exactly the Bayes risk of the representation exposed by K; Bayes-sufficient representation theory studies the corresponding fixed-law decision problem (Sevetlidis, 2026). Here K ranges over a declared internal receiver family and the paper compares the resulting risk profile across realized probability states. For a tolerance $\varepsilon \geq 0$, write

$$\mathcal{R}_{\mu,\varepsilon} := \{K \subseteq J : \delta_\mu(K) \leq \varepsilon\}.$$

The full risk profile and the thresholded circuit family are different objects.

Here refinement is passive: K' exposes additional receiver variables from the same realized computation rather than modifying the computation itself. Equation (4) exhibits the smaller receiver state as a deterministic garbling of the larger one, so receiver inclusion is a Blackwell comparison of experiments (Blackwell, 1951; Fritz et al., 2023).

Proposition 2.1 (Receiver inclusion is Blackwell refinement)**.** *If $K \subseteq K'$, then*

$$\delta_\mu(K') \leq \delta_\mu(K).$$

Proof. For any rule $h : Z_K \to \mathsf{A}$, let $h' = h \circ p_{K'K}$. Equation (4) gives $h'Q_{K'} = hQ_K$ pointwise, so both rules have the same risk. Taking infima gives the claim. □

The proof uses only closure under forgetting added receiver coordinates. Hence every architecture-specific family satisfying

$$h \in \mathscr{H}_K \quad \Longrightarrow \quad h \circ p_{K'K} \in \mathscr{H}_{K'} \qquad (K \subseteq K')$$

inherits the same monotonicity law for $\delta_{\mu,\mathscr{H}}$. In ML terms, the decoder class available at K' must be able to ignore the newly exposed coordinates. Appendix A states this as a subfunctor result. Passive receiver access therefore has a monotonicity law whenever the admissible downstream class respects receiver refinement. If K instead specifies an ablation, patching, retention, or rerun protocol, changing K can change the state passed to the continuation; structural inclusion alone then does not imply Blackwell refinement.

Distribution relativity does not make receiver adequacy unstable. Let

$$d_{\mathrm{TV}}(\mu,\nu) := \sup_B |\mu(B) - \nu(B)|.$$

If $0 \le \ell \le 1$, then

$$\sup_{K \subseteq J} |\delta_\mu(K) - \delta_\nu(K)| \le d_{\mathrm{TV}}(\mu,\nu). \tag{7}$$

Indeed, for each fixed rule the pointwise loss lies in $[0,1]$, so its expectations under μ and ν differ by at most $d_{\mathrm{TV}}(\mu,\nu)$; infimizing and exchanging μ,ν gives (7). Thus the entire receiver-risk profile is 1-Lipschitz in total variation, even though optimizer identities and thresholded circuit families can change discontinuously when a decision boundary is crossed. Appendix B gives the corresponding threshold-stability and finite polyhedral selection geometry.

3 Context-invariant realization

Predictive sufficiency and context-invariant realization are different properties. Receiver refinement monotonically improves the first. It need not improve the second. Continue with the maximal receiver decoder class from section 2; fix X, the receiver family, Φ, the action space A, and the loss ℓ, and vary only the probability law.

Let C be an experimental-context random variable jointly distributed with X—for example, an input population, domain, prompt family, or other distributional condition—with $\Pr(C = c) = \lambda_c > 0$ and $X \mid C = c \sim \mu_c$. Write

$$\bar{\mu} = \sum_c \lambda_c \mu_c.$$

When C is observed, each context may use its own optimal rule through Q_K. When C is hidden, one rule must act on the pooled law. We measure the cost of requiring one receiver-level decoder to serve all contexts by

$$\Gamma_K := \delta_{\bar{\mu}}(K) - \sum_c \lambda_c \delta_{\mu_c}(K) \ge 0. \tag{8}$$

It is the excess Bayes risk incurred when the context label is hidden.

Proposition 3.1 (Context invariance is a common-optimum condition)**.** *Assume the context and pooled minima are attained. Then*

$$\Gamma_K = 0 \quad \Longleftrightarrow \quad \bigcap_c \mathrm{Opt}_{\mu_c}(K) \neq \varnothing,$$

where $\mathrm{Opt}_{\mu_c}(K)$ *is the set of risk-minimizing rules through* Q_K.

Proof. A rule optimal in every context attains equality in equation (8). Conversely, if equality holds and h^* minimizes pooled risk, then

$$\sum_c \lambda_c R_{\mu_c}(K, h^*) = \sum_c \lambda_c \delta_{\mu_c}(K).$$

Every term on the left is at least the corresponding optimum, and every λ_c is positive. Hence h^* is optimal in every context. □

Nothing in this argument depends on maximality of the decoder class. For any fixed architecture-specific family $\mathscr{H}_K$, replacing δ_μ by $\delta_{\mu,\mathscr{H}}$ gives the same common-optimum criterion with optimizers taken inside $\mathscr{H}_K$. The criterion therefore isolates the internal invariance question at either level: once the receiver map and admissible downstream class are fixed, does one decoder remain optimal across realized states? This differs from invariant prediction and invariant risk minimization, which impose related common-predictor conditions in different problem settings (Arjovsky et al., 2019; Ruan et al., 2022).

Log loss makes the maximal receiver-sufficiency penalty exact. Take $\mathsf{A} = \mathsf{Prob}(Y)$ and use

$$\ell_{\log}(y, r) = -\log r(y).$$

Because the maximal class contains the Bayes action, the optimal rule is the conditional law of Y_Φ, so the pooled and context-aware risks are conditional entropies. Since log loss is unbounded, the bounded-loss TV statement from section 2 is not invoked here.

Theorem 3.2 (Context value and receiver increment under log loss)**.** *Assume Y_Φ and C have finite ranges, $\mathsf{A} = \mathsf{Prob}(Y)$, and $\ell = \ell_{\log}$. Then*

$$\Gamma_K^{\log} = I(Y_\Phi; C \mid Q_K(X)). \tag{9}$$

For $j \notin K$, writing $Z = Q_K(X)$ and $W = Q_j(X)$,

$$\boxed{\Gamma_{K\cup\{j\}}^{\log} - \Gamma_K^{\log} = I(Y_\Phi; W \mid Z, C) - I(Y_\Phi; W \mid Z).} \tag{10}$$

The increment can be positive or negative.

Proof. For log loss,

$$\delta_{\bar{\mu}}^{\log}(K) = H(Y_\Phi \mid Z), \qquad \sum_c \lambda_c \delta_{\mu_c}^{\log}(K) = H(Y_\Phi \mid Z, C),$$

which gives equation (9). Also,

$$\Gamma_{K\cup\{j\}}^{\log} - \Gamma_K^{\log} = I(Y_\Phi; C \mid Z, W) - I(Y_\Phi; C \mid Z).$$

Expanding $I(Y_\Phi; C, W \mid Z)$ in the two chain-rule orders gives equation (10). □

Equation (9) gives a direct interpretation: $\Gamma_K^{\log}$ is exactly the target-relevant information carried by context that remains after the receiver representation is known. Equation (10) then rules out any general monotonicity of context invariance under receiver refinement. More internal information can make a common rule easier to sustain or harder to sustain across distributions.

Both failures occur in elementary binary systems. For an increase, let C and Y_Φ be independent fair bits, let $N \sim \mathrm{Bernoulli}(p)$ be independent with $0 < p < 1/2$, and set

$$Q_1 = Y_\Phi \oplus N, \qquad Q_2 = Y_\Phi \oplus C.$$

With $K = \{1\}$, C is independent of (Y_Φ, Q_1), so $\Gamma_{\{1\}}^{\log} = 0$. Marginally, Q_2 is also independent of (Y_Φ, Q_1); pooled predictive risk does not improve. After conditioning on Q_2, however, $C = Y_\Phi \oplus Q_2$, and

$$\Gamma_{\{1,2\}}^{\log} = I(Y_\Phi; C \mid Q_1, Q_2) = H_2(p) > 0.$$

The added receiver has made the optimal decoder context-dependent.

For a decrease, let $C = Y_\Phi$ be nondegenerate, begin from a constant receiver, and add $Q_j = Y_\Phi$. Then

$$\Gamma_{\varnothing}^{\log} = I(Y_\Phi; C) = H(Y_\Phi) > 0, \qquad \Gamma_{\{j\}}^{\log} = I(Y_\Phi; C \mid Y_\Phi) = 0.$$

Here the added receiver removes the common-rule penalty completely.

The opening two-bit construction gives the same separation under 0–1 loss. The context-specific rules are z and $1 - z$, their equal mixture has error $1/2$, and

$$\Gamma_{\{2\}} = \frac{1}{2} - \eta > 0$$

even though the two context laws have identical support. Exact support localization therefore cannot see the distinction. The next section identifies precisely what remains after the probability weights are discarded.

4 The exact support shadow and fixed-network non-canonicity

The distributional theory distinguishes full-support laws that favor opposite receiver-level rules. Exact localization cannot. We now take the maximal receiver-sufficiency problem to zero loss, isolating what the access itself preserves before any further restriction on downstream implementation. Exact localization is obtained after the realized probability state is coarse-grained to its null-set structure:

$$\boxed{\text{probability state} \longrightarrow \text{measure class} \underset{\text{finite discrete}}{\cong} \text{support} \longrightarrow \text{functional dependency / reduct.}} \tag{11}$$

The first arrow forgets probability weights while retaining null sets; in finite discrete spaces those null sets are equivalent to support.

A receiver family K exactly realizes Φ under μ if some measurable $h : Z_K \to Y$ satisfies $\Phi = hQ_K$ μ-almost surely. Write

$$\mathcal{R}_\Phi^{\mathrm{a.s.}}(\mu) := \{K \subseteq J : \exists h \text{ with } \Phi = hQ_K \ \mu\text{-a.s.}\}.$$

Proposition 4.1 (Exact localization depends only on measure class)**.** *If $\mu \sim \nu$ have the same null measurable sets, then*

$$\mathcal{R}_\Phi^{\mathrm{a.s.}}(\mu) = \mathcal{R}_\Phi^{\mathrm{a.s.}}(\nu).$$

Proof. For fixed K and h, the disagreement set $\{x : \Phi(x) \neq h(Q_K(x))\}$ is null under μ exactly when it is null under ν. □

Now assume X, Y, and the realized receiver images are finite, take $\mathsf{A} = Y$, use 0–1 loss, and let $S = \operatorname{supp} \mu$. Define

$$\mathcal{R}_\Phi(S) := \{K \subseteq J : \exists h : Q_K(S) \to Y \text{ with } \Phi|_S = hQ_K|_S\},$$

and for $f : S \to T$ write $\mathrm{Eq}(f) = \{(x, x') \in S^2 : f(x) = f(x')\}$. We call this exact support-level object the support shadow because it records which target distinctions survive on the realized support while forgetting how probability mass is distributed across that support.

Proposition 4.2 (Finite exact localization is the support shadow)**.** *For every $K \subseteq J$,*

$$\delta_\mu(K) = 0 \iff \Phi|_S \textit{ factors through } Q_K|_S \iff \mathrm{Eq}(Q_K|_S) \subseteq \mathrm{Eq}(\Phi|_S). \tag{12}$$

For each pair with $\Phi(x) \neq \Phi(x')$, let

$$E(x, x') := \{j \in J : Q_j(x) \neq Q_j(x')\}, \qquad \mathcal{H}_\Phi(S) := \{E(x, x') : x, x' \in S,\ \Phi(x) \neq \Phi(x')\}.$$

Each $E(x, x')$ is exactly the set of receiver coordinates capable of distinguishing one target-relevant pair. Then

$$\mathcal{R}_\Phi(S) = \mathrm{Hit}(\mathcal{H}_\Phi(S)), \tag{13}$$

where Hit *denotes the family of hitting sets.*

Proof. Zero 0–1 risk on finite support is equality at every point of positive mass. A decoder through Q_K exists exactly when Q_K never identifies two support points that Φ distinguishes. Equivalently, every phenomenon-distinguished pair is separated by at least one receiver in K. □

Equation (12) is the exact access test: if it fails, no downstream architecture receiving only $Q_K(X)$ can realize Φ on S. A restricted decoder class can impose an additional implementation constraint even when it holds. Accordingly, the architecture-specific exact family

$$\mathcal{R}_\Phi^{\mathscr{H}}(S) := \{K \subseteq J : \exists h \in \mathscr{H}_K \text{ with } \Phi|_S = hQ_K|_S\}$$

satisfies $\mathcal{R}_\Phi^{\mathscr{H}}(S) \subseteq \mathcal{R}_\Phi(S)$.

The finite zero-loss coarse-graining lands exactly on functional-dependency and decision-reduct structure. Equation (12) is the database functional dependency $K \to d$ when the receivers are condition attributes and Φ is the decision attribute (Beeri et al., 1984). Its inclusion-minimal realizing sets are the corresponding decision reducts / minimal transversals of the discernibility family (Skowron & Rauszer, 1992; Demetrovics et al., 2014). This identifies classical reduct structure as the exact support shadow of the distribution-relative localization problem.

Support restriction is antitone. If $S \subseteq T$, then

$$\mathcal{R}_\Phi(T) \subseteq \mathcal{R}_\Phi(S), \tag{14}$$

because a decoder valid on T remains valid on S. The inclusion-minimal family

$$\mathcal{M}_\Phi(S) := \mathrm{Min}_\subseteq \mathcal{R}_\Phi(S)$$

is generally an antichain rather than a singleton. Minimality therefore provides no canonical exact circuit. Appendix C records the blocker formulation, support functoriality, and the finite empirical g_3 correspondence.

Fixing the network does not fix the exact localization family. The question is how much variation remains once the architecture, target, and receiver interface are all frozen. Within the upward-closure constraint already forced by passive receiver inclusion, one fixed elementary network realizes every admissible family by support variation alone. Let $J \neq \varnothing$ and fix

$$\mathsf{N}_J : \{0,1\}^J \xrightarrow{\mathrm{id}} \{0,1\}^J \xrightarrow{\mathrm{OR}} \{0,1\}, \tag{15}$$

with coordinate receivers $Q_j(v) = v_j$ and phenomenon $\Phi(v) = \mathrm{OR}(v)$. Classical candidate-key and reduct constructions already show broad realizability of Sperner systems (Demetrovics, 1979; Demetrovics et al., 2014). Here the architecture, binary coordinate interface, and target function are all frozen; only the admitted support changes. For a family $\mathcal{F} \subseteq 2^J$, let $\mathrm{Hit}(\mathcal{F})$ be the sets intersecting every member of $\mathcal{F}$. For $\mathcal{A} \subseteq 2^J$, write

$$\uparrow \mathcal{A} := \{K \subseteq J : \exists A \in \mathcal{A} \text{ with } A \subseteq K\}$$

for its upward closure. If $F \subseteq 2^J$ is nonempty and upward closed, define

$$\sigma_{\mathrm{OR}}(F) := \{0^J\} \cup \{\mathbf{1}_E : E \in \mathrm{Hit}(F)\}. \tag{16}$$

Theorem 4.3 (Fixed-network support universality)**.** *For every nonempty upward-closed $F \subseteq 2^J$,*

$$\boxed{\mathcal{R}_\Phi(\sigma_{\mathrm{OR}}(F)) = F.} \tag{17}$$

Consequently every nonempty antichain $\mathcal{A} \subseteq 2^J$ occurs as $\mathcal{M}_\Phi(S)$ for some support S of this one fixed network.

Proof. Every $E \in \mathrm{Hit}(F)$ is nonempty because $F \neq \varnothing$. The phenomenon-distinguished pairs are therefore the anchor 0^J against witnesses $\mathbf{1}_E$, and their receiver-separation sets are exactly E. Hence the discernibility hypergraph of $\sigma_{\mathrm{OR}}(F)$ is $\mathrm{Hit}(F)$. For an upward family, $\mathrm{Hit}(\mathrm{Hit}(F)) = F$: if $K \notin F$, upward closure implies every member of F intersects $J \setminus K$, so $J \setminus K \in \mathrm{Hit}(F)$ and K fails to hit it. Equation (13) gives (17). For an antichain $\mathcal{A}$, take $F = \uparrow \mathcal{A}$. □

This exhausts the upward-closure constraint imposed by passive receiver inclusion: at the exact receiver-sufficiency level, every nonempty upward-closed localization family occurs in the same fixed network under support variation alone, even though the architecture, target, and receiver granularity remain fixed. Architecture therefore does not determine a canonical exact localization family. Restricting the admissible decoder class defines a finer architecture-specific problem; it does not change the universality result for the information already exposed at the receiver interface. Appendix D gives stronger order-profile consequences. The theorem is a structural boundary on exact localization; empirical multiplicity in a particular language model can have additional causes (Méloux et al., 2025; Chen et al., 2026).

5 Implications for mechanistic localization

A component set is not a complete mechanism specification at the receiver-sufficiency level. A localization claim fixes the realized probability state μ, receiver or mediator map Q, phenomenon Φ, admissible decoder class $\mathscr{H}$, and decision criterion ℓ; a selection threshold ε and concrete rule h also belong to the claim when used. Empirically, these are the input or prompt distribution, mediator or representation being read, downstream architecture class, target behavior, adequacy criterion, selection threshold, and concrete receiver-level rule. Different experimental choices therefore act on different arguments of the problem:

Table 1 lists arguments of the localization claim itself, not the algorithm used to search for it. Fixing $\Lambda = (\mu, Q, \mathscr{H}, \Phi, \ell, \varepsilon)$ fixes a population target such as $\mathcal{R}_{\mu,\varepsilon}$; a discovery procedure applied to finite data and model access instead returns an approximation such as $\widehat{K}_A$. Reported circuits can therefore differ because the localization specification differs or because sampling, optimization, pruning, or attribution procedures approximate the same target differently. Appendix F develops this distinction.

Different reported component subsets do not by themselves establish different realized mechanisms: studies may differ in μ, Q, $\mathscr{H}$, Φ, ℓ, ε, or the induced process. Conversely, the same subset does not

What changes	Argument	What can change
input population or prompt distribution	μ	optimal rule, receiver risk, context value, selected circuit
mediator or representation granularity	Q	the localization problem itself
downstream architecture or continuation class	decoder class $\mathscr{H}$	realizable receiver-level rules, optimal risk, selected localization
target behavior	Φ	which distinctions receivers must preserve
loss or adequacy criterion	ℓ	optimal action and receiver risk
selection threshold	ε	the discrete reported circuit family
passive receiver set	K	predictive risk, monotonically under refinement
intervention protocol	realized process/state	the induced experiment; passive Blackwell order need not apply

Table 1: Different experimental changes act on different arguments of a mechanistic-localization claim.

establish the same receiver-level realization, because the optimal rule and induced action can vary with μ. The hierarchy $K \to (K, h) \to [hQ_K]_\mu$ makes both failures explicit.

The Blackwell order applies when

$$Q_K = p_{K'K} Q_{K'}$$

on the same realized state. Ablation, patching, retention, or rerun protocols can instead change the state entering downstream computation, so structural inclusion alone does not define a Blackwell refinement; comparison then requires a garbling or simulation relation between the induced experiments. Appendix F makes this requirement explicit for intervention-defined discovery procedures.

The maximal decoder class is the downstream-implementation-independent outer layer of receiver sufficiency. It separates two obstructions. An *access obstruction*, $\delta_\mu(K) > 0$, means no measurable downstream rule receiving only $Q_K(X)$ can attain zero risk: the required distinction is already lost at the receiver interface. A *decoder-class obstruction*, $\delta_\mu(K) = 0$ but $\delta_{\mu,\mathscr{H}}(K) > 0$, means the distinction reaches the receiver but the selected architecture cannot realize the required map. These are different mechanistic failures and should not be collapsed into one notion of "circuit insufficiency."

Causal necessity and intervention-defined mechanisms add another axis. Restricting $\mathscr{H}$ changes which functions may act on a fixed representation; patching or ablation can change the representation distribution itself. Causal abstraction and formal patching supply comparison structure for that problem (Geiger et al., 2025; Hadad et al., 2026); cross-network work fixes still another comparison problem (Sun & Toneva, 2026). Receiver granularity, downstream use class, and induced process are therefore separate declared parts of the analysis.

At zero loss, context invariance becomes a literal local-to-global compatibility problem. If contexts S_i each admit an exact decoder through one fixed Q_K, a decoder on $\bigcup_i S_i$ exists exactly when the local decoders agree on every receiver-image overlap

$$Q_K(S_i) \cap Q_K(S_j).$$

These can exceed the images of $S_i \cap S_j$ because noninjective access can identify previously distinct states. Appendix E gives the descent formulation. Thus Γ_K and the exact global-section obstruction are two regimes of the same question: whether context-specific receiver rules extend to one rule.

Architecture-level receiver formalisms fix the represented process, exposed distinctions, and admissible downstream use before the probability state and adequacy criterion are chosen (Rosario Freytes, 2026). Here those interfaces are fixed and the realized state varies. Mechanistic localization is therefore a statement about a realized computation under declared access and admissible use, not a property of a bare component subset.

AI USE STATEMENT

In this work, generative AI tools were used to help develop the theoretical and conceptual framework; formulate and refine mathematical claims; explore, draft, and check proof arguments; design validation methodology; and interpret formal and finite validation results. They were also used for literature

search and synthesis, generating and editing code for finite regression tests, and drafting and editing the manuscript. They were not used to generate or clean empirical datasets, conduct qualitative data analysis, assist with translation, or run empirical experiments; those tasks are not applicable to this theoretical study. All AI-assisted mathematical claims were checked against explicit arguments or finite regressions where applicable, literature claims were checked against cited sources, and the authors reviewed the final text and take responsibility for the ideas, claims, proofs, citations, code, and presentation.

Reproducibility statement

The paper is theoretical. Assumptions and proofs of the two headline results appear in the main text. Auxiliary proofs, finite structural results, exact-shadow constructions, and supporting derivations are supplied in the appendices so that the formal claims can be audited independently.

References

Samson Abramsky and Adam Brandenburger. The sheaf-theoretic structure of non-locality and contextuality. *New Journal of Physics*, 13(11):113036, 2011. doi: 10.1088/1367-2630/13/11/113036. URL https://arxiv.org/abs/1102.0264.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019. URL https://arxiv.org/abs/1907.02893.

Alireza Bayat Makou, Jingcheng Niu, Subhabrata Dutta, and Iryna Gurevych. Many circuits, one mechanism: Input variation and evaluation granularity in circuit discovery. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=Ur42Xx1QPv. Featured Certification.

Catriel Beeri, Martin Dowd, Ronald Fagin, and Richard Statman. On the structure of Armstrong relations for functional dependencies. *Journal of the ACM*, 31(1):30–46, 1984. doi: 10.1145/2422.322414.

David Blackwell. Comparison of experiments. In Jerzy Neyman (ed.), *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, pp. 93–102. University of California Press, 1951. doi: 10.1525/9780520411586-009.

Xi Chen, Mingyu Jin, Jingcheng Niu, Yutong Yin, Jinman Zhao, Bangwei Guo, Dimitris N. Metaxas, Zhaoran Wang, Yutao Yue, and Gerald Penn. All circuits lead to Rome: Rethinking functional anisotropy in circuit and sheaf discovery for LLMs. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*. PMLR, 2026. URL https://arxiv.org/abs/2605.12671.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-0719. URL https://papers.neurips.cc/paper_files/paper/2023/hash/34e1dbe95d34d7ebaf99b9bcaeb5b2be-Abstract-Conference.html.

János Demetrovics. On the equivalence of candidate keys with Sperner systems. *Acta Cybernetica*, 4(3):247–252, 1979. URL https://cyber.bibl.u-szeged.hu/index.php/actcybern/article/view/3183.

János Demetrovics, Vu Duc Thi, and Nguyen Long Giang. On finding all reducts of consistent decision tables. *Cybernetics and Information Technologies*, 14(4):3–10, 2014. doi: 10.1515/cait-2014-0001.

Tobias Fritz. A synthetic approach to Markov kernels, conditional independence and theorems on sufficient statistics. *Advances in Mathematics*, 370:107239, 2020. doi: 10.1016/j.aim.2020.107239. URL https://doi.org/10.1016/j.aim.2020.107239.

Tobias Fritz, Tomáš Gonda, Paolo Perrone, and Eigil Fjeldgren Rischel. Representable Markov categories and comparison of statistical experiments in categorical probability. *Theoretical Computer Science*, 961:113896, 2023. doi: 10.1016/j.tcs.2023.113896. URL https://doi.org/10.1016/j.tcs.2023.113896.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah D. Goodman. Finding alignments between interpretable causal variables and distributed neural representations. In *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pp. 160–187. PMLR, 2024. URL https://proceedings.mlr.press/v236/geiger24a.html.

Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, and Thomas Icard. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, 26(83):1–64, 2025. URL https://www.jmlr.org/papers/v26/23-0058.html.

Satchel Grant, Simon Jerome Han, Alexa Tartaglini, and Christopher Potts. Addressing divergent representations from causal interventions on neural networks. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/133e588e1429f9f1e25b215da145580e-Abstract-Conference.html.

Itamar Hadad, Guy Katz, and Shahaf Bassan. Formal mechanistic interpretability: Automated circuit discovery with provable guarantees. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/b5afe13494c825089b1e3944fdaba212-Abstract-Conference.html.

Jyrki Kivinen and Heikki Mannila. Approximate inference of functional dependencies from relations. *Theoretical Computer Science*, 149(1):129–149, 1995. doi: 10.1016/0304-3975(95)00028-U.

David Lindner, János Kramár, Sebastian Farquhar, Matthew Rahtz, Thomas McGrath, and Vladimir Mikulik. Tracr: Compiled transformers as a laboratory for interpretability. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2301.05062.

Maxime Méloux, Silviu Maniu, François Portet, and Maxime Peyrard. Everything, everywhere, all at once: Is mechanistic interpretability identifiable? In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/51157713f1f1f572a05647e1aef1b1ed-Abstract-Conference.html.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://arxiv.org/abs/2202.05262.

Aaron Mueller, Atticus Geiger, Sarah Wiegreffe, Dana Arad, Iván Arcuschin, Adam Belfki, Yik Siu Chan, Jaden Fried Fiotto-Kaufman, Tal Haklay, Michael Hanna, Jing Huang, Rohan Gupta, Yaniv Nikankin, Hadas Orgad, Nikhil Prakash, Anja Reusch, Aruna Sankaranarayanan, Shun Shao, Alessandro Stolfo, Martin Tutek, Amir Zur, David Bau, and Yonatan Belinkov. MIB: A mechanistic interpretability benchmark. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 45069–45108. PMLR, 2025. URL https://proceedings.mlr.press/v267/mueller25a.html.

Aaron Mueller, Jannik Brinkmann, Millicent Li, Samuel Marks, Koyena Pal, Nikhil Prakash, Can Rager, Aruna Sankaranarayanan, Arnab Sen Sharma, Jiuding Sun, Eric Todd, David Bau, and Yonatan Belinkov. The quest for the right mediator: Surveying mechanistic interpretability for NLP through the lens of causal mediation analysis. *Computational Linguistics*, 52(1):331–378, 2026. doi: 10.1162/coli.a.572. URL https://aclanthology.org/2026.cl-1.10/.

Luis F. Rosario Freytes. Architecture before the formula: Individuating neural architecture beyond the composite map. *arXiv preprint arXiv:2601.11618*, 2026. doi: 10.48550/arXiv.2601.11618. URL https://arxiv.org/abs/2601.11618.

Yangjun Ruan, Yann Dubois, and Chris J. Maddison. Optimal representations for covariate shift. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=Rf58LPCwJj0.

Vasileios Sevetlidis. Bayes-sufficient representations in supervised learning. *arXiv preprint arXiv:2606.04045*, 2026. URL https://arxiv.org/abs/2606.04045.

Andrzej Skowron and Cecylia Rauszer. The discernibility matrices and functions in information systems. In *Intelligent Decision Support: Handbook of Applications and Advances of the Rough Sets Theory*, pp. 331–362. Springer Netherlands, 1992. doi: 10.1007/978-94-015-7975-9_21.

Alan Sun and Mariya Toneva. Tracking equivalent mechanistic interpretations across neural networks. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/412e7a0143cb33d3a7a9b6abe8d90470-Abstract-Conference.html.

Aaquib Syed, Can Rager, and Arthur Conmy. Attribution patching outperforms automated circuit discovery. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 407–416. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.blackboxnlp-1.25. URL https://aclanthology.org/2024.blackboxnlp-1.25/.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: A circuit for indirect object identification in GPT-2 small. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2211.00593.

Frank Zhengqing Wu, Francesco Tonin, and Volkan Cevher. Demystifying variance in circuit discovery of LLMs. *arXiv preprint arXiv:2606.16920*, 2026. doi: 10.48550/arXiv.2606.16920. URL https://arxiv.org/abs/2606.16920.

Zhengxuan Wu, Atticus Geiger, Thomas Icard, Christopher Potts, and Noah D. Goodman. Interpretability at scale: Identifying causal mechanisms in alpaca. In *Advances in Neural Information Processing Systems*, volume 36, pp. 78205–78226, 2023. doi: 10.52202/075280-3418. URL https://papers.neurips.cc/paper_files/paper/2023/hash/f6a8b109d4d4fd64c75e94aaf85d9697-Abstract-Conference.html.

Yiyu Yao and Yan Zhao. Attribute reduction in decision-theoretic rough set models. *Information Sciences*, 178(17):3356–3373, 2008. doi: 10.1016/j.ins.2008.05.010. URL https://doi.org/10.1016/j.ins.2008.05.010.

Yan Zhao, S. K. Michael Wong, and Yiyu Yao. A note on attribute reduction in the decision-theoretic rough set model. In *Rough Sets and Current Trends in Computing*, volume 5306 of *Lecture Notes in Computer Science*, pp. 61–70. Springer, 2008. doi: 10.1007/978-3-540-88425-5_7.

Wojciech Ziarko. Variable precision rough set model. *Journal of Computer and System Sciences*, 46 (1):39–59, 1993. doi: 10.1016/0022-0000(93)90048-2.

A Formal details for realized processes and receiver presentations

The main text defines realized state transport operationally through deterministic maps and pushforwards. This appendix records the categorical formulation of that transport together with formal properties of receiver-level mechanism presentations. It also treats architecture-specific decoder classes as restrictions of the maximal receiver-level decoder functor: the main text's decoder hypothesis classes appear here as subfunctors. The construction below uses only deterministic measurable maps and pushforwards; for the broader categorical-probability setting, see Fritz (2020).

A.1 State transport and the realized neural process

Let **SBor** denote the category of standard Borel spaces and measurable deterministic maps. For a measurable map $f : X \to Y$, pushforward defines

$$f_* : \mathsf{Prob}(X) \to \mathsf{Prob}(Y), \qquad (f_*\mu)(B) = \mu(f^{-1}(B)).$$

Thus

$$\mathsf{Prob} : \mathbf{SBor} \longrightarrow \mathbf{Set} \tag{18}$$

is a covariant functor.

Fix represented cuts

$$[L] = \{0 < 1 < \cdots < L\}$$

regarded as a category, and let

$$X^\theta : [L] \to \mathbf{SBor}$$

be a deterministic feed-forward neural process. Write

$$F^\theta_{q \rightsquigarrow r} := X^\theta(q \to r), \qquad A^\theta_q := F^\theta_{0 \rightsquigarrow q}.$$

For a source state $\mu_0 \in \mathsf{Prob}(X_0)$, set

$$\mu_q := (A^\theta_q)_* \mu_0.$$

The category of elements $\int \mathsf{Prob}$ has objects (X, μ) and morphisms

$$f : (X, \mu) \to (Y, \nu)$$

with $f : X \to Y$ measurable and $f_*\mu = \nu$.

Proposition A.1 (The realized neural process)**.** *For every deterministic neural process X^θ and source state μ_0, the pushed-forward states define a unique lift*

$$\widetilde{X}^{\theta,\mu_0} : [L] \longrightarrow \int \mathsf{Prob}$$

whose projection to **SBor** *is X^θ and whose value at cut q is (X_q, μ_q).*

Proof. For $q \le r$,

$$(F^\theta_{q \rightsquigarrow r})_* \mu_q = (F^\theta_{q \rightsquigarrow r})_* (A^\theta_q)_* \mu_0 = (A^\theta_r)_* \mu_0 = \mu_r.$$

Hence every network arrow is a morphism in $\int \mathsf{Prob}$. The source state determines every downstream state through the unique arrow $0 \to q$, so the lift is unique. □

Proposition A.2 (Equivalent states remain equivalent downstream)**.** *Let μ and ν be probability measures on X with the same null measurable sets. For every measurable deterministic map $f : X \to Y$,*

$$f_*\mu \sim f_*\nu.$$

Consequently, if two source states lie in the same measure class, their induced states lie in the same measure class at every represented cut of a deterministic neural process.

Proof. For every measurable $B \subseteq Y$,

$$(f_*\mu)(B) = 0 \iff \mu(f^{-1}(B)) = 0 \iff \nu(f^{-1}(B)) = 0 \iff (f_*\nu)(B) = 0.$$

Apply the result to each prefix map A^θ_q. □

A.2 Receiver refinement as a presentation category

Fix a finite declared receiver family J at one represented cut, with measurable maps

$$Q_j : X \to Z_j.$$

Write

$$\mathbb{B}_J := (2^J, \subseteq)$$

for the Boolean receiver poset regarded as a category. For $K \subseteq J$, let

$$Z_K := \prod_{j \in K} Z_j, \qquad Q_K := \langle Q_j \rangle_{j \in K},$$

and for $K \subseteq K'$ let

$$p_{K'K} : Z_{K'} \to Z_K$$

be coordinate projection. Fix a standard Borel action space A and define

$$\mathsf{Dec}_{\mathsf{A}}(K) := \{h : Z_K \to \mathsf{A} \mid h \text{ measurable}\}.$$

Receiver refinement pulls a rule on K back to one on K',

$$\mathsf{Dec}_{\mathsf{A}}(K) \longrightarrow \mathsf{Dec}_{\mathsf{A}}(K'), \qquad h \longmapsto h \circ p_{K'K}. \tag{19}$$

These maps are functorial, so

$$\mathsf{Dec}_{\mathsf{A}} : \mathbb{B}_J \to \mathbf{Set}$$

is a functor. Its category of elements $\int \mathsf{Dec}_{\mathsf{A}}$ has objects (K, h) with $K \subseteq J$ and $h : Z_K \to \mathsf{A}$.

A refinement arrow preserves the induced action pointwise. If $K \subseteq K'$ and $h' = h \circ p_{K'K}$, then

$$h' Q_{K'} = h Q_K. \tag{20}$$

A.3 Architecture-specific decoder subfunctors

The maximal decoder functor separates receiver access from any particular downstream implementation. Categorically, refinement-compatible decoder hypothesis classes are exactly subfunctors of the maximal decoder functor,

$$\mathscr{H} \hookrightarrow \mathsf{Dec}_{\mathsf{A}}.$$

Concretely, this consists of subsets

$$\mathscr{H}_K \subseteq \mathsf{Dec}_{\mathsf{A}}(K)$$

for each $K \subseteq J$ such that

$$h \in \mathscr{H}_K \quad \Longrightarrow \quad h \circ p_{K'K} \in \mathscr{H}_{K'} \qquad (K \subseteq K'). \tag{21}$$

This is exactly the closure needed for passive receiver refinement to preserve the admissible downstream class.

For a probability state μ and measurable loss $\ell : Y \times \mathsf{A} \to [0, \infty]$, define

$$\delta_{\mu,\mathscr{H}}(K) := \inf_{h \in \mathscr{H}_K} \mathbb{E}_\mu\big[\ell\big(\Phi(X), h(Q_K(X))\big)\big].$$

Proposition A.3 (Refinement monotonicity for decoder subfunctors)**.** *If $\mathscr{H} \hookrightarrow \mathsf{Dec}_{\mathsf{A}}$ is a refinement-compatible decoder subfunctor and $K \subseteq K'$, then*

$$\delta_{\mu,\mathscr{H}}(K') \le \delta_{\mu,\mathscr{H}}(K).$$

Proof. For $h \in \mathscr{H}_K$, closure (21) gives $h' := h \circ p_{K'K} \in \mathscr{H}_{K'}$. Equation (20) gives $h' Q_{K'} = h Q_K$ pointwise, so the two rules have the same risk. Taking the infimum over $h \in \mathscr{H}_K$ gives the claim. □

The maximal case is $\mathscr{H} = \mathsf{Dec}_{\mathsf{A}}$. Smaller subfunctors specialize the receiver-sufficiency problem to restricted downstream computation classes without changing the access maps themselves. A family that is not closed under (21) need not inherit the refinement monotonicity law.

In the finite exact setting, define

$$\mathcal{R}_\Phi^{\mathscr{H}}(S) := \{K \subseteq J : \exists h \in \mathscr{H}_K \text{ with } \Phi|_S = h Q_K|_S\}.$$

Then

$$\mathcal{R}_\Phi^{\mathscr{H}}(S) \subseteq \mathcal{R}_\Phi(S). \tag{22}$$

Hence failure of the maximal factorization condition is an access obstruction shared by every downstream architecture using only $Q_K(X)$, while membership in $\mathcal{R}_\Phi(S)$ need not imply membership in $\mathcal{R}_\Phi^{\mathscr{H}}(S)$ for a restricted architecture class.

A.4 State-relative presentation equivalence

For a probability state μ on X, define

$$(K,h) \equiv_\mu (K',h') \quad \Longleftrightarrow \quad hQ_K = h'Q_{K'} \qquad \mu\text{-almost surely.}$$

Proposition A.4 (State-relative presentation equivalence)**.** *For every probability state μ, the relation $\equiv_\mu$ is an equivalence relation on receiver-level presentations with common action space A. Moreover:*

(i) every refinement arrow in $\int \mathsf{Dec}_{\mathsf{A}}$ relates equivalent presentations pointwise and therefore under every state;

(ii) if $\mu \sim \nu$ have the same null sets, then $\equiv_\mu$ and $\equiv_\nu$ coincide;

(iii) if $\bar{\mu} = \sum_{c=1}^m \lambda_c \mu_c$ with every $\lambda_c > 0$, then

$$(K,h) \equiv_{\bar{\mu}} (K',h') \quad \Longleftrightarrow \quad (K,h) \equiv_{\mu_c} (K',h') \text{ for every } c.$$

Proof. Reflexivity, symmetry, and transitivity are inherited from almost-sure equality of measurable maps. Part (i) follows from (20). For part (ii), the disagreement set

$$B := \{x : h(Q_K(x)) \neq h'(Q_{K'}(x))\}$$

is measurable and is μ-null if and only if it is ν-null. For part (iii),

$$\bar{\mu}(B) = \sum_{c=1}^m \lambda_c \mu_c(B).$$

With all coefficients strictly positive, this vanishes exactly when every $\mu_c(B)$ vanishes. □

These properties justify the main text's hierarchy

$$K \longrightarrow (K,h) \longrightarrow [hQ_K]_\mu$$

and make architecture-specific specialization explicit without requiring the categorical details to interrupt the main argument.

B Finite distribution geometry

This appendix records finite-space consequences of the distribution-relative receiver Bayes risk that are useful for structural selection analysis but not needed for the main theorem spine.

Assume throughout that X, Y, and all receiver images are finite and specialize to

$$\mathsf{A} = Y, \qquad \ell(y,\hat{y}) = \mathbf{1}[y \neq \hat{y}].$$

B.1 Threshold stability away from a boundary

For a tolerance $\varepsilon \in [0,1]$, define

$$\gamma_{\mu,\varepsilon} := \min_{K \subseteq J} |\delta_\mu(K) - \varepsilon|.$$

Corollary B.1 (Localization-family stability radius)**.** *If $\gamma_{\mu,\varepsilon} > 0$ and*

$$d_{\mathrm{TV}}(\mu,\nu) < \gamma_{\mu,\varepsilon},$$

then

$$\mathcal{R}_{\mu,\varepsilon} = \mathcal{R}_{\nu,\varepsilon}.$$

Hence their inclusion-minimal localization antichains are also equal.

Proof. The main-text TV-Lipschitz theorem implies

$$|\delta_\mu(K) - \delta_\nu(K)| < \gamma_{\mu,\varepsilon}$$

for every K. Every defect value therefore remains on the same side of the threshold ε. □

B.2 Finite receiver risk is piecewise linear

For fixed K, write

$$p^{\mu}_{z,y} := \mu\{x : Q_K(x) = z,\ \Phi(x) = y\}.$$

Proposition B.2 (Finite receiver risk is ordinary Bayes classification error)**.** *For every receiver family K,*

$$\delta_\mu(K) = 1 - \sum_{z \in Z_K} \max_{y \in Y} p^{\mu}_{z,y}.$$

In particular, $\mu \mapsto \delta_\mu(K)$ is concave and piecewise linear on the probability simplex $\Delta(X)$.

Proof. For a rule $h : Z_K \to Y$,

$$R_\mu(K, h) = 1 - \sum_{z \in Z_K} p^{\mu}_{z,h(z)}.$$

Each receiver value can be optimized independently, so the largest correct mass at z is $\max_y p^{\mu}_{z,y}$. □

B.3 Polyhedral selection geometry

Proposition B.3 (Finite localization cells)**.** *Fix a finite receiver family J and tolerance $\varepsilon \in [0, 1]$. There is a finite polyhedral subdivision of $\Delta(X)$ such that on the relative interior of every cell:*

(i) the optimizer set $\mathrm{Opt}_\mu(K)$ is constant for every $K \subseteq J$;

(ii) the thresholded localization family $\mathcal{R}_{\mu,\varepsilon}$ is constant;

(iii) the inclusion-minimal localization antichain

$$\mathcal{M}_{\mu,\varepsilon} := \mathrm{Min}_{\subseteq}\, \mathcal{R}_{\mu,\varepsilon}$$

is constant.

Proof. For fixed K and h, the map

$$\mu \longmapsto R_\mu(K, h)$$

is affine. Partition the finite decoder set into classes whose members induce identical affine risk functionals. Distinct classes can exchange optimizer status only on nontrivial tie hyperplanes

$$R_\mu(K, h) = R_\mu(K, h'),$$

and threshold membership can change only on nontrivial hyperplanes

$$R_\mu(K, h) = \varepsilon.$$

The finite arrangement of these hyperplanes together with the faces of $\Delta(X)$ yields a finite polyhedral subdivision. On each relative interior the ordering and equality pattern of all distinct risk functionals, and their signs relative to ε, are fixed. The optimizing equivalence classes are therefore fixed; every decoder inside one such class enters or leaves the optimizer set together, so the full set $\mathrm{Opt}_\mu(K)$ is fixed as well. The thresholded family and its inclusion-minimal members are then fixed on the same cell. □

We use *localization cell* for a cell of such a subdivision and reserve *input regime* or *context* for the scientific population being analyzed. The distinction avoids overloading the word "regime."

The main text uses only the consequence needed for interpretation: quantitative receiver risks vary continuously under TV shift, while optimizer identities and thresholded selections can change discretely at decision boundaries.

C Further structure of the exact support shadow

This appendix records exact-support structure that is useful for auditing and extending the support shadow but is not required for the main narrative.

C.1 SUPPORT TRANSPORT

For finite discrete spaces define

$$\operatorname{supp}\mu := \{x : \mu(x) > 0\}.$$

Proposition C.1 (Finite support shadow)**.** *If $f : (X, \mu) \to (Y, \nu)$ is a morphism in the finite discrete subcategory of $\int \mathsf{Prob}$, then*

$$\operatorname{supp}\nu = f(\operatorname{supp}\mu).$$

Consequently support defines a functor from finite state-decorated spaces to finite sets by restricting each deterministic map to the support of its source state.

Proof. For $y \in Y$,

$$\nu(y) = (f_*\mu)(y) = \sum_{x : f(x) = y} \mu(x),$$

which is positive exactly when some $x \in \operatorname{supp}\mu$ satisfies $f(x) = y$. □

C.2 MEASURE-CLASS INVARIANCE OF EXACT REALIZATION

Let

$$\mathcal{R}_\Phi^{\text{a.s.}}(\mu) := \{K \subseteq J : \exists h : Z_K \to Y \text{ measurable with } \Phi = hQ_K \ \mu\text{-a.s.}\}.$$

Proposition C.2 (Exact localization factors through measure class)**.** *If $\mu \sim \nu$, then*

$$\mathcal{R}_\Phi^{\text{a.s.}}(\mu) = \mathcal{R}_\Phi^{\text{a.s.}}(\nu).$$

Proof. For fixed K and h, let

$$B_{K,h} := \{x : \Phi(x) \neq h(Q_K(x))\}.$$

The decoder realizes Φ exactly under μ iff $B_{K,h}$ is μ-null, which under $\mu \sim \nu$ is equivalent to being ν-null. Existence of a realizing decoder is therefore the same under both laws. □

Together with Proposition A.2, this gives a process-level coarse-graining: equivalent source states remain equivalent at every deterministic cut, and exact almost-sure localization cannot distinguish their different positive weights.

C.3 SUPPORT RESTRICTION AND TRUTH-VALUED LOCALIZATION

For finite $S \subseteq X$, write

$$\mathcal{R}_\Phi(S) := \{K \subseteq J : \exists h : Q_K(S) \to Y \text{ with } \Phi|_S = hQ_K|_S\}.$$

Proposition C.3 (Support restriction)**.** *If $S \subseteq T$, then*

$$\mathcal{R}_\Phi(T) \subseteq \mathcal{R}_\Phi(S).$$

Proof. Any decoder that works on T still works after restriction to S. □

Write

$$\mathbf{Supp}(X) := (\mathcal{P}(X), \subseteq), \qquad \mathbf{2} = \{0 < 1\},$$

and let $\mathbf{Up}(\mathbb{B}_J)$ denote the upward-closed subsets of the Boolean receiver poset $\mathbb{B}_J$, ordered by inclusion. Exact support localization can be represented as

$$\mathsf{Loc}_{\Phi,Q} : \mathbf{Supp}(X)^{\text{op}} \times \mathbb{B}_J \longrightarrow \mathbf{2}, \tag{23}$$

with $\mathsf{Loc}_{\Phi,Q}(S, K) = 1$ exactly when $K \in \mathcal{R}_\Phi(S)$. Currying gives

$$\mathscr{R}_{\Phi,Q} : \mathbf{Supp}(X)^{\text{op}} \longrightarrow \mathbf{Up}(\mathbb{B}_J), \qquad S \longmapsto \mathcal{R}_\Phi(S). \tag{24}$$

This truth-valued object remembers exact factorization on support and forgets all probability weights.

C.4 Discernibility clutters and blockers

For a map $f : S \to T$, write

$$\mathrm{Eq}(f) := \{(x, x') \in S^2 : f(x) = f(x')\}.$$

For a phenomenon-distinguished pair $x, x' \in S$, define

$$E(x, x') := \{j \in J : Q_j(x) \neq Q_j(x')\},$$

and the distinction hypergraph

$$\mathcal{H}_\Phi(S) := \{E(x, x') : x, x' \in S,\ \Phi(x) \neq \Phi(x')\}.$$

For a hypergraph $\mathcal{H} \subseteq 2^J$, let

$$\mathrm{Hit}(\mathcal{H}) := \{K \subseteq J : K \cap E \neq \varnothing \text{ for every } E \in \mathcal{H}\}.$$

The main text records the equivalence

$$K \in \mathcal{R}_\Phi(S) \iff \mathrm{Eq}(Q_K|_S) \subseteq \mathrm{Eq}(\Phi|_S) \iff K \text{ hits every } E \in \mathcal{H}_\Phi(S).$$

Deleting hyperedges that strictly contain another gives the clutter

$$\mathcal{C}_\Phi(S) := \mathrm{Min}_\subseteq \mathcal{H}_\Phi(S).$$

If

$$b(\mathcal{C}) := \mathrm{Min}_\subseteq \mathrm{Hit}(\mathcal{C})$$

is the blocker, then the inclusion-minimal exact localizations satisfy

$$\mathcal{M}_\Phi(S) = b(\mathcal{C}_\Phi(S)). \tag{25}$$

This is classical transversal / decision-reduct structure. It makes explicit why an inclusion-minimal circuit need not be the unique minimal circuit.

C.5 Uniform empirical defect and the classical g_3 error

Suppose $S = \{x_1, \ldots, x_n\}$ carries the uniform empirical law and use 0–1 loss. For each receiver value z, let

$$n_{z,y} := \big|\{x \in S : Q_K(x) = z,\ \Phi(x) = y\}\big|.$$

The finite Bayes formula gives

$$\delta_{\mathrm{unif}(S)}(K) = 1 - \frac{1}{n} \sum_z \max_y n_{z,y}. \tag{26}$$

To make the functional dependency $K \to d$ exact by deleting rows, one may retain at most one decision value inside each Q_K-equivalence class. The largest exact subtable therefore keeps $\sum_z \max_y n_{z,y}$ rows. Hence (26) is exactly the classical g_3 deletion error for an approximate functional dependency (Kivinen & Mannila, 1995).

Variable-precision and decision-theoretic rough-set models likewise replace exact reduct criteria by probabilistic and cost/risk-sensitive ones (Ziarko, 1993; Zhao et al., 2008; Yao & Zhao, 2008). The paper's distributional content lies upstream of this finite empirical shadow: the probability state is transported through one neural process and receiver refinement is compared with context invariance before these coarse-grainings are taken.

D Further fixed-network universality consequences

This appendix records consequences of the fixed-OR support-universality theorem that are stronger than needed for the main narrative.

For a support S of the fixed OR network, define its *exact localization order* by

$$\lambda_\Phi(S) := \min\{|K| : K \in \mathcal{R}_\Phi(S)\}. \tag{27}$$

The family $\mathcal{R}_\Phi(S)$ is nonempty because the full coordinate set J always determines the OR output on S.

Corollary D.1 (Arbitrary monotone localization-order profile)**.** *Let P be a finite poset and let*

$$d : P \to \{0, \dots, |J|\}$$

satisfy

$$p \le q \Longrightarrow d(p) \le d(q).$$

Then the fixed network

$$\mathsf{N}_J : \{0,1\}^J \xrightarrow{\mathrm{id}} \{0,1\}^J \xrightarrow{\mathrm{OR}} \{0,1\}$$

admits supports S_p such that

$$p \le q \Longrightarrow S_p \subseteq S_q, \qquad \lambda_\Phi(S_p) = d(p). \tag{28}$$

Proof. Set

$$F_p := \{K \subseteq J : |K| \ge d(p)\}$$

and

$$S_p := \sigma_{\mathrm{OR}}(F_p).$$

Each F_p is nonempty and upward closed. If $p \le q$, then $d(p) \le d(q)$ and hence

$$F_q \subseteq F_p.$$

Every hitting set of F_p therefore hits F_q, so

$$\mathrm{Hit}(F_p) \subseteq \mathrm{Hit}(F_q).$$

By the definition of σ_{OR}, this gives

$$S_p \subseteq S_q.$$

Theorem 4.3 gives $\mathcal{R}_\Phi(S_p) = F_p$, whose smallest members have cardinality exactly $d(p)$. Thus $\lambda_\Phi(S_p) = d(p)$. □

The result strengthens Theorem 4.3 from support-wise universality to arbitrary finite monotone profiles of minimum exact receiver-set size along nested support diagrams. Thus even the evolution of minimum exact localization size across nested supports is not determined by architecture alone.

E ACCESS-IMAGE DESCENT FOR EXACT MECHANISMS

This appendix develops the local-to-global form of exact context compatibility induced by receiver access.

Suppose every context individually admits an exact receiver-level decoder. It is still possible that no single decoder works across their union. The obstruction does not have to appear at the source: a noninjective internal access can identify states from different contexts that were distinct before the access. Local exact descriptions must therefore be compared on the overlaps created by the receiver itself.

E.1 RECEIVER ACCESS CAN CREATE NEW CROSS-CONTEXT COLLISIONS

For a map $f : X \to Y$ in $\mathbf{Set}$, direct image and inverse image define

$$\Sigma_f(A) := f(A), \qquad f^*(B) := f^{-1}(B),$$

with

$$\Sigma_f \dashv f^*. \tag{29}$$

Direct image preserves unions but only satisfies

$$f(A \cap B) \subseteq f(A) \cap f(B). \tag{30}$$

The inclusion can be strict exactly because f may identify a point of $A \setminus B$ with a point of $B \setminus A$. Thus two disjoint or weakly overlapping source regimes can acquire a larger overlap after a network prefix or receiver access.

For mechanistic compatibility, the relevant overlap is therefore

$$Q_K(S_i) \cap Q_K(S_j),$$

not merely $S_i \cap S_j$.

E.2 Exact context compatibility is descent on the access-image cover

Let

$$S = \bigcup_{i \in I} S_i$$

be a finite cover and let

$$L : S \twoheadrightarrow U = L(S)$$

be an access. Put

$$U_i := L(S_i).$$

Assume $L|_{S_i}$ exactly realizes $\Phi|_{S_i}$ for every i. Each context then has a unique local decoder

$$h_i : U_i \to Y$$

with

$$\Phi|_{S_i} = h_i L|_{S_i}.$$

The local decoder domains form a cover

$$U = \bigcup_i U_i.$$

Theorem E.1 (Exact context compatibility is an access-image global-section condition)**.** *Under the preceding assumptions, L exactly realizes Φ on S if and only if*

$$h_i|_{U_i \cap U_j} = h_j|_{U_i \cap U_j} \qquad \textit{for all } i, j. \tag{31}$$

When a global decoder exists, it is unique on U.

Proof. If $h : U \to Y$ is a global decoder, uniqueness of factorization through the surjections $L|_{S_i} : S_i \twoheadrightarrow U_i$ gives $h_i = h|_{U_i}$, so compatibility is necessary.

Conversely, if the local decoders agree on every overlap, define

$$h(u) := h_i(u)$$

for any i with $u \in U_i$. The overlap condition makes this well defined, and for $x \in S_i$,

$$h(L(x)) = h_i(L(x)) = \Phi(x).$$

Surjectivity gives uniqueness on U. □

The theorem is ordinary function gluing. The mechanistic identification is the domain on which it must be applied: the cover is generated by access images, not by source contexts before the access.

E.3 Sheaf formulation

On the poset of subsets of U, define

$$\mathscr{D}_Y(V) := \{h : V \to Y\}$$

with restriction along inclusions. For the ordinary union coverage, $\mathscr{D}_Y$ is the usual sheaf of Y-valued functions. The local decoders are sections

$$h_i \in \mathscr{D}_Y(U_i),$$

and Theorem E.1 says that exact global realization is equivalent to their extension to one section of $\mathscr{D}_Y(U)$.

This has the same local-to-global form used in sheaf-theoretic contextuality (Abramsky & Brandenburger, 2011), but the objects are different. Here the cover is generated by access images of input regimes through one deterministic receiver interface, and the local data are decoder functions constrained by one phenomenon. No measurement scenario, no-signalling condition, or hidden-variable semantics is assumed. The shared mathematics is the global-section obstruction; the mechanistic content is the access-induced cover.

E.4 Descent as the zero-loss endpoint of context value

Assume finite spaces with

$$\mathsf{A} = Y, \qquad \ell(y, \hat{y}) = \mathbf{1}[y \neq \hat{y}].$$

Let μ_i have full support S_i and suppose

$$\delta_{\mu_i}(K) = 0$$

for every context. Let

$$\bar{\mu} = \sum_i \lambda_i \mu_i, \qquad \lambda_i > 0, \quad \sum_i \lambda_i = 1.$$

Then $\operatorname{supp} \bar{\mu} = \bigcup_i S_i$.

Corollary E.2 (Zero context value becomes exact descent)**.** *For a fixed receiver family K, the following are equivalent:*

(i) $\delta_{\bar{\mu}}(K) = 0$*;*

(ii) the contexts admit one common zero-risk decoder through Q_K;

(iii) the local exact decoders agree on all access-image intersections $Q_K(S_i) \cap Q_K(S_j)$;

(iv) the local decoder sections extend to a global section of $\mathscr{D}_Y$ on $Q_K(\bigcup_i S_i)$;

(v) Q_K exactly realizes Φ on $\bigcup_i S_i$.

Proof. Finite zero defect is equivalent to exact realization on support. Since every local optimum is zero and every mixture coefficient is positive, a zero-risk mixture decoder is exactly a rule with zero risk in every context, giving (i)⇔(ii). Theorem E.1 gives the equivalence of (ii), (iii), and (v), while the sheaf condition gives (iii)⇔(iv). □

For positive loss, Γ_K measures how much Bayes risk is saved by revealing context. At zero loss, that quantitative question collapses to the exact extension problem above. The positive-loss context-value problem therefore has an exact local-to-global endpoint at zero loss.

F Localization specifications and empirical discovery procedures

This appendix maps representative mechanistic-interpretability workflows onto the main text's distinction between a localization specification and the procedure used to search it.

F.1 Population localization versus a discovered circuit

Fix a localization specification

$$\Lambda = (\mu, Q, \mathscr{H}, \Phi, \ell, \varepsilon).$$

At the maximal receiver-sufficiency level, this determines the population family

$$\mathcal{R}_{\mu,\varepsilon} = \{K \subseteq J : \delta_\mu(K) \leq \varepsilon\},$$

and an architecture-specific specification analogously determines the family defined by $\delta_{\mu,\mathscr{H}}$. A practical discovery procedure instead receives finite data, model access, and its own optimization or approximation rule. Denote such a procedure by A and a reported component set by $\widehat{K}_A$.

Reported localizations can vary because the specification itself changes, because finite sampling changes an empirical estimate, or because search, pruning, attribution, or optimization returns a different approximation to the same population target. The last source is algorithmic; it is not another argument of Λ.

F.2 Representative methods in the receiver formalism

Table 2 records how several established methods instantiate different parts of the formalism. They retain distinct intervention semantics, objectives, and search spaces.

Method or domain	Relevant operation	Location in the present formalism
ACDC (Conmy et al., 2023)	searches computational-graph edges using activation-patching-based pruning	a discovery procedure over structural candidates; its dataset and metric instantiate parts of the empirical localization problem, while pruning and thresholding belong to the approximation/search layer
Edge Attribution Patching (Syed et al., 2024)	uses a linear approximation to activation patching to score edges before pruning	an attribution/search approximation layered on an intervention-defined localization problem
DAS / Boundless DAS (Geiger et al., 2024; Wu et al., 2023)	learns alignments between causal variables and distributed linear subspaces	a non-coordinate receiver presentation Q together with an optimization procedure for finding that presentation
Tracr (Lindner et al., 2023)	compiles known programs into transformer weights with known internal structure	a controlled model family on which localization procedures can be compared against known computational structure
IOI (Wang et al., 2023)	studies a natural language behavior in GPT-2 small using causal interventions and projections	a concrete choice of task/population, phenomenon, receiver access, and intervention protocol

Table 2: Representative empirical methods and domains mapped to the distinction between a localization specification and the procedure used to estimate or search it.

Distributed localization therefore lies inside the receiver formalism rather than outside it. A coordinate projection is one admissible access map; a learned projection or rotated subspace is another. Changing the latter changes Q and hence changes the localization problem being posed. The optimization used to find that projection is a separate layer.

F.3 Passive refinement under intervention protocols

Proposition 2.1 uses a specific comparison structure. For passive access and $K \subseteq K'$, the deterministic projection

$$p_{K'K} : Z_{K'} \to Z_K, \qquad Q_K = p_{K'K} Q_{K'},$$

exhibits the smaller receiver state as a garbling of the larger one. The monotonicity conclusion follows from this map, not from structural inclusion alone.

An ablation, patching, retention, or rerun protocol may instead induce a different experiment E_K for each structural choice K. Then $K \subseteq K'$ does not by itself supply a map from $E_{K'}$ to E_K that reproduces the smaller experiment. A Blackwell-style monotonicity statement requires an actual garbling or simulation relation between those induced experiments. Greedy or pruning-based search under intervention semantics therefore cannot inherit Proposition 2.1 merely from inclusion of the retained component sets; the required comparison relation has to be established for the intervention protocol being used.

This separates two questions that are often operationally coupled: whether one receiver state contains enough information for the target, and whether a particular intervention/search procedure preserves an ordering between candidate experiments. The first is the receiver-sufficiency problem studied in the main theory. The second belongs to the semantics of the discovery protocol.

F.4 Controlled and semantic test domains

Tracr, IOI, and CounterFact/causal tracing provide natural controlled or semantic domains for empirical tests of these distinctions (Lindner et al., 2023; Wang et al., 2023; Meng et al., 2022). Such tests instantiate a localization specification together with a discovery procedure; the structural results above characterize the localization object before that empirical layer is chosen.