SEPARATING REPRESENTATIONAL REDUNDANCY FROM CAUSAL RESPONSIBILITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Neural network interpretations often treat decodability as evidence of functional use. If a linear classifier predicts a task variable from a component's hidden state, that component may be taken both to contain and contribute the information. This work separates those properties. A component may encode task information the model does not use, while another may control the output without linearly encoding the answer. Two architectural mechanisms organise the distinction. Transport spreads task information between components and determines where it can be recovered. Selection governs which available components affect the output. In two transport sweeps, selection has its largest effect after transport creates usable alternatives but before the original component loses measured responsibility. A follow-up statistic and decision rule for this non-monotonic pattern were fixed after the initial result and confirmed on fresh seeds. Across recurrent, graph, convolutional, transformer, and mixture-of-experts networks, greater transport capacity is accompanied by broader recoverability with little change in task accuracy. Fixed-weight route cuts isolate transport where such interventions are available. In a graph network with its body frozen, the task variable is decoded from components other than its source with 0.88 accuracy, yet source removal reduces the trained readout to chance. Retraining only the readout recovers 0.87 accuracy, showing that the information was available to a new readout but unused by the original one. Conversely, transplanting a pointer node redirects the answer even though the answer remains linearly undecodable there. Representational availability and causal responsibility are distinct and should be measured separately when attributing functions to network components.

1 INTRODUCTION

Mechanistic interpretability asks whether an internal state merely contains task information or whether the model uses it. Linear probes localise recoverable information by training a classifier on intermediate activations (Alain & Bengio, 2016). Successful decoding establishes availability, not use. Probe conclusions depend on decoder capacity and controls, and properties the trained model does not need can remain recoverable (Hewitt & Liang, 2019; Belinkov, 2022; Ravichander et al., 2021). Decodability and causal influence can therefore disagree within one model and task (Huang & Chang, 2025). This distinction affects mechanistic interpretability, AI auditing, robustness analysis, model compression and architecture design. It also matters to AI safety and security whenever internal attributions guide monitoring or intervention.

Causal interventions such as activation patching, causal mediation, interchange interventions, causal tracing and path patching measure how predictions change when internal states or paths are modified (Vig et al., 2020; Geiger et al., 2021; Meng et al., 2022; Goldowsky-Dill et al., 2023; Wang et al., 2023; Chan et al., 2022; Zhang & Nanda, 2024). They answer a complementary question. Ablation measures the dependence of the current computation, not whether equivalent information exists elsewhere. Redundant and compensatory pathways make that difference important. Alternative attention heads may compensate for removed name-copying heads, and self-repair can recruit downstream computation (Wang et al., 2023; McGrath et al., 2023; Rushing & Nanda, 2024). Neither high decodability nor a small ablation effect, taken alone, identifies where a computation resides.

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.

The question is which architectural variables control this mismatch. *Transport* moves task information between candidate components through lateral messages, graph message passing, receptive-field growth, self-attention or multi-expert routing. It determines where information becomes recoverable. Architecture sweeps vary transport capacity together with computation, while route cuts test transport at fixed weights. *Selection* determines which available components affect the output and whether transported information remains a decodable by-product or enters the realised computation. The central claim is that varying transport and selection moves the gap between availability and responsibility systematically.

A frozen graph representation makes the distinction concrete (Section 5.1). The non-source nodes decode the queried variable with 0.88 accuracy against 0.25 chance, yet source removal reduces the trained-readout accuracy from 0.969 to 0.240. With the body frozen, retraining only the readout on the remaining nodes recovers 0.865. The same representation is informative outside the source and sufficient for a new readout, although the original readout does not successfully use it after source removal.

Across recurrent, graph, convolutional, transformer and mixture-of-experts networks, greater transport makes the queried variable recoverable from more components with little change in task accuracy. Selection has its largest effect after transport creates usable alternatives but before the original component loses responsibility. The separation also appears in citation networks and three frozen pretrained language models. Reversing the search order by ranking intervention effects before measuring encoded content finds a pointer that controls retrieval while answer decoding remains at chance. A search ordered by answer decodability would miss it.

The work makes three contributions. First, it separates representational availability, current causal responsibility and recoverable sufficiency in the same frozen representation. Second, it identifies transport as an architectural control of availability and downstream selection as a control of responsibility whose measured effect is strongest at intermediate transport capacity. Third, it shows that decodability-based attribution is incomplete in both directions. Information may be decodable where the original model does not use it, while a component may control behaviour without encoding the answer. Appendix J records every claim, including the predictions that failed.

2 Related work

One useful division in mechanistic interpretability separates methods that ask what a representation contains from those that ask which components affect a computation (Sharkey et al., 2025). Since features may be superposed across neurons (Elhage et al., 2022), the first group has moved from reading individual units to learning overcomplete dictionaries with sparse autoencoders (Huben et al., 2024). Recent studies question whether these dictionaries recover task-relevant or canonical features (Kantamneni et al., 2025; Korznikov et al., 2026). Whether features are read from individual neurons or learned directions, recovering a feature still does not establish that the model uses it. The second group asks which components carry a computation. Circuits have been established by hand for induction and for indirect object identification (Olsson et al., 2022; Wang et al., 2023), then found automatically by pruning the edges whose removal changes behaviour (Conmy et al., 2023), and aligned with high-level variables in rotated subspaces by distributed alignment search (Geiger et al., 2024). Search of that kind already orders components by causal effect, so the reversal in Section 5.4 is not a new procedure. The new result concerns the conclusion drawn from that procedure. The component identified in Section 5.4 controls retrieval through an address while answer decodability remains at chance. A search ordered by answer decodability would miss it.

Probes identify where a property is recoverable, but probe accuracy alone does not establish that the model uses it. Control tasks show that accuracy can reflect information learned by the probe itself (Hewitt & Liang, 2019), while task-irrelevant properties may remain highly decodable from a model's representations (Ravichander et al., 2021; Belinkov, 2022). Intervention studies demonstrate both possible outcomes. Giulianelli et al. (2018) use gradients from a diagnostic classifier to alter a long short-term memory network's (LSTM's) hidden states and improve subject-verb agreement, providing evidence that the recovered signal can participate in the computation. By contrast, Elazar et al. (2021) remove linearly decodable property information through iterative nullspace projection and find that conventional probe accuracy does not predict its importance to model behaviour. An unresolved disagreement of that kind is what a hidden variable governing the size of the gap

would produce, and naming such a variable is the aim here. Interventional methods go beyond correlational evidence by testing how outputs change when internal states or computational paths are altered. These include causal mediation analysis (Vig et al., 2020), interchange interventions within the causal abstraction framework (Geiger et al., 2021), causal tracing (Meng et al., 2022), and activation or path patching (Wang et al., 2023). These methods also have failure modes. Apparent neuron-level concepts can depend on the evaluation corpus and can arise along random directions (Bolukbasi et al., 2021). Naive subspace activation patching can produce the intended behavioural change by activating a dormant pathway through a component that is normally disconnected from the output (Makelov et al., 2024). Probes and interventions answer complementary questions, so the quantities defined below are reported separately instead of being combined into a single score.

Closest to the present motivation are two recent dissociations. Huang & Chang (2025) show a layerwise mismatch between decodability and causality in counting vision transformers and treat the two as complementary measurements. Sharma et al. (2026) train transformers on bracket sequences whose hierarchical variables are known exactly. They find that depth and distance remain decodable, yet ablating the corresponding probe-aligned subspaces barely affects behaviour. Instead, behaviour depends on the attention edge that routes each closing bracket to its opener. The dissociation is documented by both studies within a fixed architecture, and its existence is assessed. Here, the architectural variables determining its magnitude and direction are instead identified through direct manipulation. The routing mechanism found to be decisive in the bracket study is treated below as an instance of selection. It is isolated within a single architecture in that study, whereas it is varied across five here.

Related literature approaches this separation through self-repair techniques also. Ablating an apparently important component in a pretrained transformer can produce a smaller effect than its direct contribution would suggest because downstream components compensate (McGrath et al., 2023; Rushing & Nanda, 2024; Wang et al., 2023). These studies examine backup computation elicited by interventions in fixed models, where compensation complicates attribution. The experiments here manipulate two conditions that precede such compensation. Transport determines whether alternative information is already available, while selection determines whether the readout uses it. This distinction separates alternatives already present in the representation from compensatory computation induced by the intervention. The frozen-body experiments establish the former. Alternative information remains present in an unchanged representation and becomes useful when only the readout is retrained.

Redundancy has been studied both as a source of robustness and as an obstacle to attribution (Morcos et al., 2018). Related work examines when network components become modular (Filan et al., 2021). The present framework suggests that transport and selection can shape both properties. A formal statement of the attribution problem is provided by Méloux et al. (2025) using networks small enough for exhaustive enumeration. Explanations are non-identifiable because several circuits can reproduce the same behaviour and several subspaces can align with the same algorithm. Redundant availability is one way this ambiguity can arise, and the experiments below identify transport as one mechanism that produces it. Oversmoothing (Li et al., 2018) is different because a factor can spread before node representations become homogenised. The citation-network experiment tests the responsibility associated with that spreading. Load balancing in mixture-of-experts networks is widely used (Shazeer et al., 2017; Fedus et al., 2022), the controlled expert-routing appendix tests one attribution cost that accuracy alone does not reveal.

The word causal is used here for interventions on the internal computation of a trained network, and not for the identification of causal structure in a data-generating process, which is a different programme with different assumptions (Appendix D). The closest framework is causal abstraction, which treats interventions on internal representations as tests of whether a high-level description faithfully abstracts a low-level computation, and which places patching, mediation, scrubbing, tracing and concept erasure in one language (Geiger et al., 2021; 2025). The question here sits inside that programme and is narrower than which algorithm a network implements. It is what decides whether task information available at a component is used there, and causal responsibility accordingly means relative control under a stated internal intervention. The next section makes that question measurable.

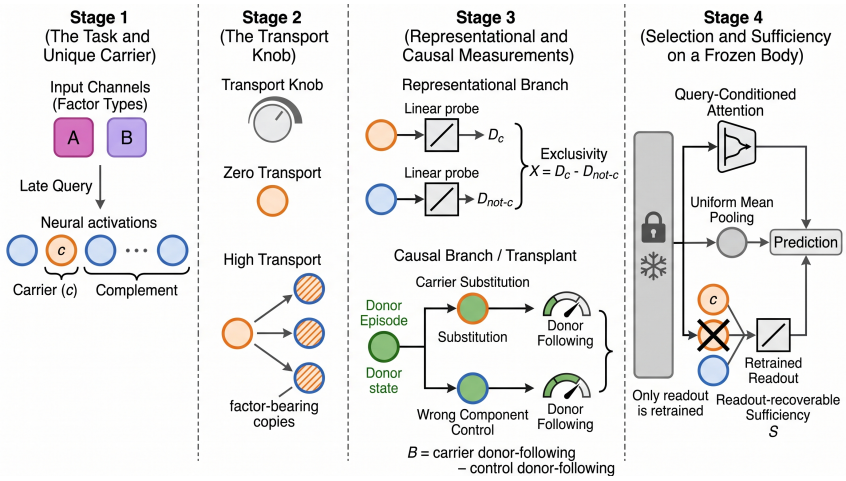

Figure 1: Randomised factor order followed by a late query requires binding by content, and the component holding the queried value is the carrier. Transport is varied while the task remains solvable. Probes measure D and X, while matched carrier and wrong-component transplants measure B on the same trained network. Selection is varied over a frozen body, and S retrains only the readout after carrier removal.

3 Measuring Availability and Responsibility

A component can hold the answer without determining the output, and a single network can contain components that play each role. The roles are distinguished using four measurements around the same trained body, applied in the order shown in Figure 1, while varying the two architectural controls introduced in Section 1.

Task, factor, carrier, and complement. Consider a binding episode that presents $A{=}2$ and $B{=}3$, then asks for the value of A. The task is to return the value bound to the queried type. The *factor* f is this value, 2 here, and has four possible values, giving a chance level of 0.25. The *carrier* is the component designated for the queried factor. It is marked by the task when components have fixed roles and nominated on separate data when components are exchangeable. The *complement*, written $\neg c$, contains all other components. Factor information is *decodable* from a state when the stated probe predicts f above chance. This does not imply that the state literally stores the answer. It may contain an embedding or distributed code. A component may instead carry only the answer's address. Intervening on such a component can redirect the computation even when f is undecodable there (Section 5.4).

Computation as a causal model. A trained network is a deterministic computation whose internal states are manipulable variables, which makes an intervention on one of them well defined. Let $h_1, \ldots, h_n$ be the component states at the point of measurement and c the carrier index. At the downstream selection stage, the architectures studied can be written schematically as

$$\hat{y} \;=\; g\big(\textstyle\sum_{i=1}^{n} \alpha_i h_i\big)\,, \tag{1}$$

where g is the readout map and α_i is the weight assigned by the downstream access mechanism. This abstracts the selection stage, not the entire network. Under query-conditioned selection, α itself depends on the states, and influence can follow paths that pooling alone does not name. Transport describes which h_i make f recoverable. Selection describes which h_i the output depends on through α. The two need not coincide. A copy with $\alpha_i = 0$ is available but not read, so it is not a functional backup for the current readout. Selection covers gated readouts, query-conditioned attention, mixture-of-experts routers, and attention between positions in frozen pretrained transformers.

The four quantities. Decodability is the accuracy of a linear probe $\hat{f}_i$ fitted to predict the factor from one component's state,

$$D_i \;=\; \Pr\Big[\hat{f}_i(h_i) = f\Big]\,. \tag{2}$$

The probe uses closed-form ridge regression, making D_i independent of gradient-based probe optimisation. Availability below means availability to that stated decoder class. What a richer decoder could recover is a different quantity, and what some readout could run on is S. *Exclusivity*, X, is the margin by which the carrier out-decodes the rest of the network,

$$X \;=\; D_c - D_{\neg c}, \tag{3}$$

and it falls when transport places copies elsewhere. *Transplant responsibility*, denoted B and called *break* below, is a difference of two follow rates under matched transplants. Writing h'_c for a donor's carrier state and f' for that donor's answer, and w for a wrong component of the same kind,

$$B \;=\; \Pr[\hat{y} = f' \mid \mathrm{do}(h_c \leftarrow h'_c)] \;-\; \Pr[\hat{y} = f' \mid \mathrm{do}(h_w \leftarrow h'_c)]\,, \tag{4}$$

counted only on episodes where $f' \neq f$, since a transplant that does not change the answer cannot show whether it was used. The conditions differ only in where identical content is injected, so the contrast isolates location. *Recoverable sufficiency*, S, asks what the rest of the network could support if it were read differently,

$$S \;=\; \Pr[\hat{y}_{g'} = f \mid \mathrm{do}(h_c \leftarrow \varnothing)]\,, \tag{5}$$

where the body is frozen and only the readout is retrained, from g to g'. It examines whether the complement can support a new output map, not whether the original model used it. The complement state entering $D_{\neg c}$ is the mean over non-carrier components in the graph, convolutional and transformer settings and the concatenation of them in the slot setting, which gives the slot complement more capacity and makes exclusivity harder to achieve there, so the two families are not comparable on the magnitude of X.

Interpreting the quantities. Availability and responsibility are distinct, so each quantity supports a different conclusion.

$$\text{decodable} \not\Rightarrow \text{sufficient} \not\Rightarrow \text{used} \not\Rightarrow \text{uniquely used}.$$

Each implication has a counterexample below, and the reverse dissociation also occurs. A decoding-based search can find readable but unused components while missing poorly decodable components that control the output, such as addresses. The four quantities separate probe-readable from selector-usable redundancy. D and X measure information recoverable by a decoder, S measures information an alternative readout could use, and B measures what the trained network currently uses. A copy can be readable and usable yet unused. As probe quantities, D and X inherit known limitations (Belinkov, 2022; Hewitt & Liang, 2019) and support no causal claims. B and S instead intervene on the forward pass using real donor states rather than noise (Zhang & Nanda, 2024). B measures relative control under matched substitution, not necessity. Necessity is measured separately by removal and reported alongside B where required (Appendix D).

Three regimes. At matched task accuracy, three regimes emerge. In the localised regime, carrier removal is not overcome and a retrained readout finds no alternative. In the redundant but carrier-dependent regime, alternatives exist but the carrier remains dominant. In the redundant and substitutable regime, carrier responsibility dissolves. The regimes are ordered by S instead of D, because substitution requires a usable, not merely readable, alternative. The unobserved fourth corner is not ruled out by these measures (Figure 6). Since selection is measured relative to the carrier, its contrast is largest in the second regime.

4 Experimental setup

Architectures and tasks. The binding episode of Section 3 is instantiated in five settings with different transport mechanisms. In the recurrent setting, a *slot* is one of eight parallel hidden-state vectors. Slots are updated recurrently and may exchange lateral messages. Each constructed task remains solvable at the lowest and highest transport settings, allowing transport to vary without becoming necessary for solving the task. Four additional settings use existing models or data. two citation networks, three pretrained language models, a frozen vision transformer (ViT) used only for measurement, and natural images with published annotations. Table 1 lists the knobs and controls for each.

Interventions and controls. Availability is measured with the probe of Equation 2 and responsibility by transplantation. A donor state is inserted into the carrier and into a wrong component of the same type. The contrast in Equation 4 isolates insertion location while controlling for the magnitude and plausibility of the content. Donors are selected from the evaluation distribution, by a fixed cyclic permutation within each batch, restricted to episodes with the same query but a different answer. Donor and recipient states thus share a marginal distribution. Sufficiency is measured by ablating the carrier and retraining only the readout while the body remains frozen. Setting-specific rules and a robustness check across five control definitions appear in Appendix E.

Nomination for exchangeable components. No carrier is designated in advance in the slot and expert settings. Decoding nominates it, so an error also misplaces the intervention. Nomination uses a disjoint discovery split, and stability is assessed by resampling (Appendix E). Section 5.4 excludes the probe from nomination. Their wrong-component control is fixed halfway around the arbitrary eight-component index, $(c + 4) \bmod 8$. This is called the *antipodal* slot or expert. "Antipodal" describes only this index choice. It does not imply geometric opposition in the learned representation.

Measurement exceptions. On citation networks, transplantation targets node inputs, whereas the classifier reads final representations by construction. At the pointer in Section 5.4, redirection replaces B because the pointer carries no answer. It measures where the transplant moves the read over recipient values against the same wrong-component control.

Reporting. Within-seed paired contrasts are reported as means with intervals and seed sign counts (Appendix I). Preregistered sign and magnitude criteria govern effect claims. Cells below the task viability floor are excluded from causal interpretation (Appendix E).

Setting	Transport knob	Selection knob	Carrier	Wrong component
Recurrent slots (8)	message width w	not varied	nominated slot	fixed antipodal slot
Graph network	depth L	attention, mean, source dropout	task-marked node	random non-carrier node
Convolutional	receptive field	attention, mean, source dropout	task-marked block	random disjoint block
Transformer	attention depth L	attention, mean, source dropout	task-marked node	random non-carrier node
Expert routing (8)	routing k	sharp vs uniform gate	nominated expert	fixed antipodal expert
Indirection graphs	read before vs after routing	not varied	pointer, value as control	random wrong node
Cora, CiteSeer	propagation depth L	self, neighbours, dropout	target node features	random neighbour
Three language models	attention into carrier cut	answer position's access cut	repeated subject mention	non-name position
WikiText agreement	the same cuts	the same cuts	subject token	attractor noun
ResNet-18, frozen ViT	stage or layer depth	attention, mean, carrier dropout	quadrant, region or box	random disjoint patches

Table 1: The upper block uses constructed binding tasks and the lower uses existing settings. Slots and experts have eight exchangeable components, with a nominated carrier and an antipodal control. Every wrong component receives identical injected content. Full rules appear in Appendix E.

5 Experiments

5.1 Transport and selection in one setting

In a message-passing graph network, transport is the layer count $L \in \{0, 4\}$ and selection is the readout. Either can vary while the other remains fixed. The task-marked source node is the carrier in this setting.

Raising L from zero to four drops X from $+0.595$ to $+0.106$ and raises $D_{\neg c}$ from 0.38 to 0.88 in all fifteen paired seeds, while accuracy remains unchanged at 0.97 (Table 2). The factor becomes readable in more places without a detected change in task accuracy.

The readout determines whether copies acquire measured responsibility (Table 3, Figure 4). Default query-conditioned attention leaves the source with high relative responsibility after transport reduces X. Training attention with source-node dropout leaves little ($B = +0.033$), an interaction of -0.413 in all fifteen paired seeds. Uniform mean pooling is architecturally flat but not trained to use complements, and gives only -0.059. Flatness alone does not recruit copies.

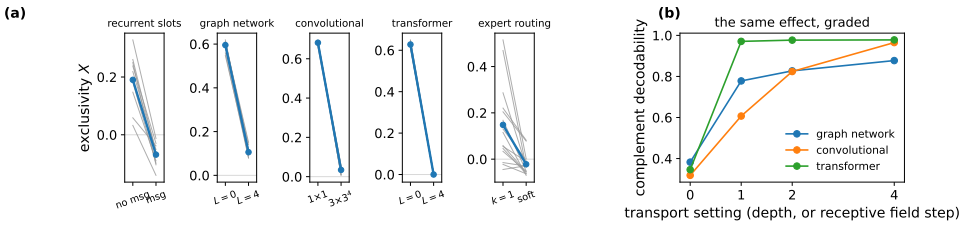

Figure 2: Transport controls probe-measured availability in five architecture families. (a) Exclusivity at each family's lowest and highest transport setting. Faint lines show seeds and the heavy line joins means. Accuracy is flat within each family. (b) The same effect when depth controls transport.

Joint encoder and readout training confounds their effects. To separate them, one body per depth is trained with the default readout, then frozen while only the readout is retrained with or without source dropout. At depth four, this reduces B by -0.130 in all fifteen seeds at matched accuracy. At depth zero it has no effect because no copies exist. Readout training can move responsibility on an unchanged body only where transport has created alternatives, with about a quarter of the end-to-end effect (-0.130 against -0.503).

One frozen representation at depth four gives the sharpest dissociation (Table 4). The complement decodes the factor, source ablation drops the original readout to chance, and retraining only the readout restores accuracy. The representation is decodable, unused and sufficient after adaptation. At depth zero, retraining recovers nothing, distinguishing a recoverable complement from an absent one.

5.2 How far transport carries

The mechanism extends beyond message passing. Across all five media in Table 1, X falls without a detectable accuracy difference (Figure 2, Table 2). These capacity sweeps establish generality but not causality because deeper stacks also add computation. The inference-time cuts below isolate transport without retraining.

The recurrent slot setting isolates the mechanism most directly by varying write discipline and lateral messages independently. Write discipline is constrained when one slot may write each factor and permissive when every slot may. Messages carry the effect. B is near $+0.80$ without messages and falls near $+0.08$ with them, while accuracy remains between 0.988 and 0.998. Averaged over write disciplines, the no-message minus message contrast is $+0.628$ on break and $+0.150$ on exclusivity, while the write contrast is small. Zeroing the message tensor at evaluation reduces complement decodability from 0.995 to 0.842 under permissive writing and from 1.000 to 0.806 under constrained writing, in all fifteen seeds for each. The copies are maintained by messages at inference and are not simply remnants of training. Together with the attention cuts in Section 5.3, this provides direct evidence that removing a route removes availability at fixed weights.

Bandwidth produces a stepwise effect. One unit of message width costs $+0.245$ of exclusivity and wider channels add nothing. A tested capacity law for this cliff failed. The result provides no evidence for a bandwidth law and is more consistent with a threshold for the presence of a lateral channel. It does not establish that explanation. The four remaining media reproduce the effect but saturate at different rates. These rates are not comparable across families because the tasks, dimensions, components, decoders and connectivity differ. With three factor types instead of two, causal carriers still form while exclusivity weakens because the complement contains carriers of the other factors. Causal localisation can strengthen as representational exclusivity weakens (Appendix A).

5.3 Outside the constructed setting

The separation is next tested where the task was not designed around an identifiable answer holder. The same measurements are applied across depths in real graphs and transformers frozen before measurement.

Citation networks and homophily. Across viable depths, B falls from $+0.097$ to $+0.015$ on Cora and from $+0.075$ to $+0.006$ on CiteSeer (Sen et al., 2008) (Figure 3). Propagation reduces the measured responsibility of a node's own features. At depth two and matched accuracy, changing only the readout determines whether the node's features matter. Break is $+0.065$ on Cora under the standard head and near zero under neighbourhood or dropout readouts.

X is near zero at every viable depth because each neighbourhood already decodes its node's label as well as the node itself. The probe measure is silent where the responsibility measure is not. This discrepancy reflects homophily and should not be generalised to graphs. Lowering Cora's edge homophily η while preserving every degree, feature and label raises X at depth two from $+0.000$ at $\eta = 0.81$ to $+0.080$ at $\eta = 0.4$, with a corresponding rise in B (Figure 3c). Before propagation, the rewirings are indistinguishable. Probing is uninformative when the data are already redundant.

Frozen language models. Layerwise measurements in a fixed model are observational. Two inference-time cuts restore manipulation. One blocks attention into the carrier from below and the other blocks the answer position's access from above. Depth is compared across the three models.

Availability emerges with depth. X is near zero at the embeddings and falls by the middle layers as the complement becomes decodable. Blocking attention into the carrier below a deep layer restores X by about $+0.32$ to $+0.47$ in every model, demonstrating transport in networks not trained for this study. Cutting the answer position's access to the carrier costs substantial accuracy early and almost none late, while B likewise falls from high to zero. S separates from both. With the carrier ablated from layer ℓ upwards, a linear readout on the frozen final layer recovers the factor late but fails early (Table 5). Carrier decodability can remain perfect while dependence on the carrier vanishes.

The same depthwise pattern appears on unmodified WikiText agreement items, where the carrier is the subject token and the control is the intervening attractor noun (Linzen et al., 2016; Merity et al., 2017). Subject B is high early and near zero late, and both the transport cut and sufficiency split reproduce. The selection interaction has the predicted sign but falls below the magnitude threshold. On corpus text, selection replicates in direction, but its magnitude does not meet the criterion (Appendix H).

5.4 Responsibility without decodability

The preceding experiments nominate components by representation before testing them causally, so they can find only components that hold the answer. Reversing the order removes this restriction. Every node is intervened on and ranked by causal effect before decodability is measured at the ranked sites.

Each graph contains a value node holding the answer, a pointer naming that node, and distractors of the same form. Address and answer are drawn independently. Their estimated mutual information is 0.0056 bits, while answer entropy is 1.999 bits. The model reaches 0.969 accuracy. Measurements are taken at the encoder output before routing, while the pointer still holds only an address.

Transplanting a donor pointer should redirect the prediction to the value named by the donor's address, which is neither the donor's nor the recipient's answer. Redirection counts this event. At the pointer, B is $+0.082$, below the standing magnitude threshold, while redirection is $+0.454$. Answer decodability is at chance (0.243 against 0.250), although the address decodes at 1.000. The pointer controls retrieval while the answer remains linearly undecodable at its state. The value node provides a positive control. Decoding would miss the pointer. In four of five seeds, a distractor has higher answer decodability, although both remain at chance.

After routing, the same battery no longer separates the roles because the answer has been copied into the pointer. Responsibility without decodability is visible only before transport. Transport spreads the factor and erases the distinction exposed by reversed nomination. The dissociation closes in both directions.

5.5 Where the effects stop

Selection at intermediate transport. The comparison pairs a default query-conditioned readout with a source-dropout readout trained to use the complement. At the largest transport setting, the

planned selection contrast fails in both the convolutional and transformer families. It is -0.053 and $+0.010$, with nine and five of fifteen seeds agreeing in sign, respectively. These results are uninformative because the default readout retains only $+0.070$ and $+0.004$, leaving selection almost nothing to remove. This diagnosis was based on the initial endpoint results and motivated a non-monotonic prediction. The statistic and decision rule were then fixed before fresh test seeds were drawn. Selection should move responsibility only when transport has produced alternatives while the carrier retains responsibility. On fifteen fresh seeds, the prespecified statistic W, defined as the mean selection effect at the two interior transport settings minus its mean at the two endpoints, is $+0.371$ in the convolutional family and $+0.267$ in the transformer. Both exceed the preregistered $+0.10$ threshold with the predicted sign in all fifteen paired seeds. Every setting lies on its predicted side, and S is at chance where no alternative exists to recruit (Figure 5). Changing only the selector on a fixed representation preserves this effect in the transformer at $+0.155$ in fourteen of fifteen seeds. One frozen representation read by two readouts yields different responsibility. This test is unavailable in the convolutional family because one cell falls below the viability floor. The endpoint effects are attenuated. The expert-routing evidence remains inconclusive (Appendices B and C).

Local and global labels. A CIFAR-10 label (Krizhevsky, 2009) is a property of the whole image, and B at the best-decoding ResNet-18 region stays small at every stage. On a composed task with the queried object confined to one quadrant, the same body and battery give $+0.370$ at the final stage against $+0.023$. The whole-image value is seed-sensitive in magnitude where the contrast is not, so the boundary rather than an absolute zero is the claim. On unmodified Microsoft Common Objects in Context (COCO) images (Lin et al., 2014), where an annotator's box defines the carrier, break is real but an order of magnitude smaller ($+0.036$ against $+0.341$), and two of four predictions fail. A frozen vision transformer trained for neither the task nor its query retains its carrier across depth. The battery tracks whether the queried variable is local and whether the body was trained for it. It does not produce an effect indiscriminately (Appendix F).

6 Discussion

A representation does not determine what a network depends on. In graph and transformer settings, one frozen representation read by two readouts yields different responsibility at constant accuracy. Capacity sweeps and fixed-weight route cuts show that transport governs availability across five media and in frozen pretrained transformers. Selection moves responsibility only while usable alternatives coexist and the carrier retains responsibility. This asymmetry partly reflects the measure. B scores control of the original carrier, while another intervention might redistribute responsibility among alternatives.

Report D, X, B and S together, or identify what was not measured. When D is high and B is low, ask whether transport created copies and whether the selector uses them (Table 4, Appendix D). The networks are small, the converse dissociation appears in one setting designed to contain it, and the pretrained evidence compares model depths without manipulating them. Responsibility is reproducible for a fixed network and intervention, but is not intrinsic to a representation. It depends on transport, the readout and the intervention.

7 Conclusion and future work

Transport governs availability, while selection governs control. Across five architecture families and frozen pretrained models, separate measurements of D, X, B and S expose readable but unused copies and responsible components that do not encode the answer.

In mechanistic interpretability, this separates representation discovery from functional attribution. AI auditing, safety evaluation and security analysis should pair probes with interventions. Robustness and architecture design should test usable redundancy rather than equating copies with backups. Compression and pruning can use B to locate dependencies and S to test retained states under a new readout. Distillation may need to preserve routing and selection, not only decoded features. This remains untested.

Future work should test these uses at scale, with learned routers, distribution shift and deliberate evasion.

REPRODUCIBILITY STATEMENT

Confirmatory experiments were specified before execution. Their protocols fixed viability floors, sign rules, magnitude thresholds and the interpretation of both outcomes before training the constructed settings or reading pretrained activations. Full dated protocols, amendments and unchanged source documents are supplied with the code at `https://anonymous.4open.science/r/research_representation_causality-3115`. Follow-up experiments motivated by earlier results were specified before execution, while analyses defined after inspecting the same data are labelled exploratory. Table 6 records this status alongside confirmed, failed and unavailable outcomes. Appendix G documents conditions that failed to train and are excluded from interpretation. One script reruns every experiment in dependency order and verifies every table cell, effect size and seed count against its source run.

AI USE STATEMENT

Generative AI tools were used for online source discovery, feedback on manuscript structure and exposition, prose editing for clarity and concision, assistance with software and anonymous artifact preparation, support for reproduction checks, and generation of Figure 1 from an author-specified scientific description. No generative AI output was treated as experimental evidence. All AI-assisted text, code, figures, and artifact changes were checked against the source code, stored results, experimental records, and cited sources. The authors take responsibility for the study design, analysis, claims, and final content.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

Tolga Bolukbasi, Adam Pearce, Ann Yuan, Andy Coenen, Emily Reif, Fernanda Viégas, and Martin Wattenberg. An interpretability illusion for BERT. *arXiv preprint arXiv:2104.07143*, 2021.

Lawrence Chan, Adrià Garriga-Alonso, Nicholas Goldowsky-Dill, Ryan Greenblatt, Jenny Nitishinskaya, Ansh Radhakrishnan, Buck Shlegeris, and Nate Thomas. Causal scrubbing: A method for rigorously testing interpretability hypotheses, 2022. AI Alignment Forum.

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.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition. *Transformer Circuits Thread*, 2022. `https://transformer-circuits.pub/2022/toy_model/index.html`.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

Daniel Filan, Stephen Casper, Shlomi Hod, Cody Wild, Andrew Critch, and Stuart Russell. Clusterability in neural networks. *arXiv preprint arXiv:2103.03386*, 2021.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pp. 9574–9586, 2021.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah 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, 2024.

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.

C. Lee Giles, Kurt D. Bollacker, and Steve Lawrence. CiteSeer: An automatic citation indexing system. In *Proceedings of the Third ACM Conference on Digital Libraries*, pp. 89–98, 1998.

Mario Giulianelli, Jack Harding, Florian Mohnert, Dieuwke Hupkes, and Willem Zuidema. Under the hood: Using diagnostic classifiers to investigate and improve how language models track agreement information. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 240–248, 2018.

Nicholas Goldowsky-Dill, Chris MacLeod, Lucas Sato, and Aryaman Arora. Localizing model behavior with path patching. *arXiv preprint arXiv:2304.05969*, 2023.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pp. 2733–2743, 2019.

Lianghuan Huang and Yingshan Chang. Causality $\neq$ Decodability, and Vice Versa: Lessons from Interpreting Counting ViTs. *arXiv preprint arXiv:2510.09794*, 2025.

Robert Huben, Hoagy Cunningham, Logan Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations*, pp. 7827–7845, 2024.

Subhash Kantamneni, Joshua Engels, Senthooran Rajamanoharan, Max Tegmark, and Neel Nanda. Are sparse autoencoders useful? a case study in sparse probing. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 29018–29049, 2025.

Anton Korznikov, Andrey Galichin, Alexey Dontsov, Oleg Y. Rogov, Ivan Oseledets, and Elena Tutubalina. Sanity checks for sparse autoencoders: Do SAEs beat random baselines? *arXiv preprint arXiv:2602.14111*, 2026.

Nikolaos Kougioulis, Nikolaos Gkorgkolis, MingXue Wang, Bora Caglayan, Dario Simionato, Andrea Tonon, and Ioannis Tsamardinos. Large causal models for temporal causal discovery. *arXiv preprint arXiv:2602.18662*, 2026.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Qimai Li, Zhichao Han, and Xiao-Ming Wu. Deeper insights into graph convolutional networks for semi-supervised learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, pp. 740–755, 2014.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535, 2016.

Sridhar Mahadevan. Large causal models from large language models. *arXiv preprint arXiv:2512.07796*, 2025.

Aleksandar Makelov, Georg Lange, Atticus Geiger, and Neel Nanda. Is this the subspace you are looking for? an interpretability illusion for subspace activation patching. In *International Conference on Learning Representations*, 2024.

Andrew Kachites McCallum, Kamal Nigam, Jason Rennie, and Kristie Seymore. Automating the construction of internet portals with machine learning. *Information Retrieval*, 3(2):127–163, 2000.

Thomas McGrath, Matthew Rahtz, János Kramár, Vladimir Mikulik, and Shane Legg. The hydra effect: Emergent self-repair in language model computations. *arXiv preprint arXiv:2307.15771*, 2023.

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.

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, pp. 17359–17372, 2022.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations*, 2017.

Ari S Morcos, David GT Barrett, Neil C Rabinowitz, and Matthew Botvinick. On the importance of single directions for generalization. In *International Conference on Learning Representations*, 2018.

Niklas Muennighoff, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Jacob Morrison, Sewon Min, Weijia Shi, Pete Walsh, Oyvind Tafjord, Nathan Lambert, Yuling Gu, Shane Arora, Akshita Bhagia, Dustin Schwenk, David Wadden, Alexander Wettig, Binyuan Hui, Tim Dettmers, Douwe Kiela, Ali Farhadi, Noah A. Smith, Pang Wei Koh, Amanpreet Singh, and Hannaneh Hajishirzi. Olmoe: Open mixture-of-experts language models. In *International Conference on Learning Representations*, 2025.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *Transformer Circuits Thread*, 2022. `https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html`.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2nd edition, 2009.

Oleg Platonov, Denis Kuznedelev, Michael Diskin, Artem Babenko, and Liudmila Prokhorenkova. A critical look at the evaluation of GNNs under heterophily: Are we really making progress? In *International Conference on Learning Representations*, 2023.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. Probing the probing paradigm: Does probing accuracy entail task relevance? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 3363–3377, 2021.

Cody Rushing and Neel Nanda. Explorations of self-repair in language models. In *International Conference on Machine Learning*, 2024.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor, Brian Gallagher, and Tina Eliassi-Rad. Collective classification in network data. *AI Magazine*, 29(3):93–106, 2008.

Lee Sharkey, Bilal Chughtai, Joshua Batson, Jack Lindsey, Jeff Wu, Lucius Bushnaq, Nicholas Goldowsky-Dill, Stefan Heimersheim, Alejandro Ortega, Joseph Bloom, Stella Biderman, Adria Garriga-Alonso, Arthur Conmy, Neel Nanda, Jessica Rumbelow, Martin Wattenberg, Nandi Schoots, Joseph Miller, Eric J. Michaud, Stephen Casper, Max Tegmark, William Saunders, David Bau, Eric Todd, Atticus Geiger, Mor Geva, Jesse Hoogland, Daniel Murfet, and Tom McGrath. Open problems in mechanistic interpretability. *arXiv preprint arXiv:2501.16496*, 2025.

Aryan Sharma, Cutter Dawes, and Shivam Raval. Dissociating decodability and causal use in bracket-sequence transformers. *arXiv preprint arXiv:2604.22128*, 2026.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pp. 12388–12401, 2020.

Kevin 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.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations*, 2024.

A PER LEVEL TABLES

Table 2: Transport reduces exclusivity in every medium tested, at accuracy that does not distinguish the settings. Fifteen seeds per cell except the recurrent slots (ten). The final column counts paired seeds in the predicted direction.

Family	Transport medium	Accuracy	X low	X high	Compl. decode	Seeds
Recurrent slots	lateral messages	0.99	$+0.190$	-0.069	$0.84 \to 0.99$	$10/10$
Graph network	message passing depth	0.97	$+0.595$	$+0.106$	$0.38 \to 0.88$	$15/15$
Convolutional	receptive field	1.00	$+0.682$	$+0.034$	$0.32 \to 0.97$	$15/15$
Transformer	attention depth	0.97	$+0.628$	-0.000	$0.35 \to 0.98$	$15/15$
Expert routing	routing multiplicity	0.97	$+0.148$	-0.022	n/a	$14/15$

Table 3: The selection manipulation in the graph substrate, quoted in Section 5.1. On the left, selection training intervention, whole model trained. On the right, the same manipulation with the body frozen and only the readout retrained, against a control that retrains it normally. ΔB is the within seed break difference. Fifteen seeds per cell.

Readout	B at $L{=}0$	B at $L{=}4$	ΔB	Frozen body	$L{=}0$	$L{=}4$
attention	$+0.549$	$+0.459$	-0.090	readout retrained	$+0.564$	$+0.445$
mean pooling	$+0.555$	$+0.406$	-0.149	readout, source dropout	$+0.564$	$+0.316$
attention, source dropout	$+0.536$	$+0.033$	-0.503	difference	-0.000	-0.130

Recurrent slots, message bandwidth. At widths $0, 1, 2, 4, 8$ break is $+0.881, +0.244, +0.098, +0.143, +0.066$ with accuracy between 0.976 and 0.993, and exclusivity is $+0.231, -0.015, -0.024, -0.026, -0.023$. Almost the whole of both effects is the first unit of width. Exclusivity drops by $+0.245$ there (95% confidence interval (CI) $+0.176$ to $+0.315$, $14/15$ seeds) and is flat after, and break drops by two thirds and then wanders between $+0.07$ and $+0.14$ without further trend, the endpoint difference being $+0.815$ over the whole ladder and $+0.637$ over its first step. The intermediate widths are described as a plateau rather than a

Table 4: Readout recoverable sufficiency in the graph substrate, fifteen seeds, quoted in Section 5.1. The body is identical in each row. Only the readout differs across the last two columns. The three columns are three different questions about one representation.

Transport	Full accuracy	Source ablated, trained readout	Source ablated, readout retrained
$L = 0$	0.976	0.252	0.248
$L = 4$	0.969	0.240	0.865

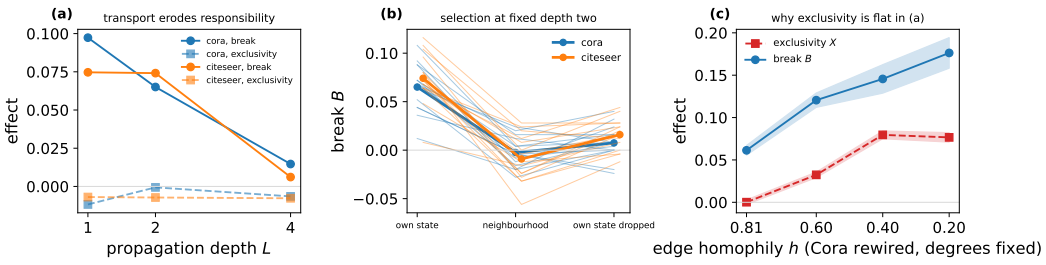

Figure 3: Both mechanisms on natural data, fifteen seeds per cell, for Section 5.3. (a) Transport erodes the causal responsibility of a node's own features, while exclusivity sits flat at zero throughout because a node's neighbourhood already decodes its label as well as the node itself. (b) At fixed transport and matched accuracy, changing only what the readout may consult removes that responsibility entirely. Each point is a seed. (c) The flat trace in (a) is a fact about homophily rather than about graphs. Rewiring Cora's edges to lower edge homophily, with every degree, feature and label unchanged, lifts exclusivity off zero and raises break with it. Bands are one standard error over fifteen seeds.

decline. The five seed sample read as a slow monotone decay after the cliff, and the fifteen seed sample does not order $w = 2$ below $w = 4$. At evaluation, zeroing the message tensor reduces complement decodability from 0.995 to 0.842 under permissive writing and from 1.000 to 0.806 under constrained writing, in all fifteen seeds for each.

The failed capacity-law follow-up (Experiment 18; E18). The cliff above is one tested as a capacity threshold by adding Gaussian noise to unit-power messages and measuring the critical width with three seeds per cell. At width one, break rises from $+0.201$ at $\sigma = 0$ to $+0.527$ at $\sigma = 4$, while exclusivity changes by only $+0.002$ and critical width does not move. A larger payload also fails to require greater width, with the registered direction observed in $0/3$ seeds. The proposed capacity law therefore fails. Only the separation of responsibility from exclusivity under noise is retained as descriptive evidence.

Recurrent slots, three factors. Constrained accuracy is 0.968, break $+0.852$, exclusivity $+0.075$. Permissive accuracy is 0.956, break $+0.020$, exclusivity -0.177. The primary statistic is the paired per seed gap, $+0.832$ with five of five seeds positive. The count of factor seed pairs above threshold ($14/15$ against $0/15$) is descriptive only, since factors within a seed are not independent.

Graph network, depth sweep. At depths $0, 1, 2, 4$, accuracy is $0.976, 0.973, 0.974, 0.969$, break $+0.552, +0.512, +0.484, +0.458$, exclusivity $+0.595, +0.203, +0.156, +0.106$, and complement decodability $0.38, 0.78, 0.83, 0.88$. Zeroing messages at inference in the trained depth four model returns complement decodability to 0.424 and exclusivity to $+0.560$ in all fifteen seeds at accuracy 0.969.

Convolutional substrate. Under the default attention readout, at kernel 1 and 3×3 depths $1, 2, 4$, exclusivity is $+0.682, +0.382, +0.170, +0.034$, complement decodability $0.32, 0.61, 0.82, 0.97$, break $+0.658, +0.547, +0.378, +0.070$, and accuracy between 0.999 and 1.000. Under attention trained with source dropout, break at the same settings is $+0.666, +0.130, +0.042, +0.025$, and

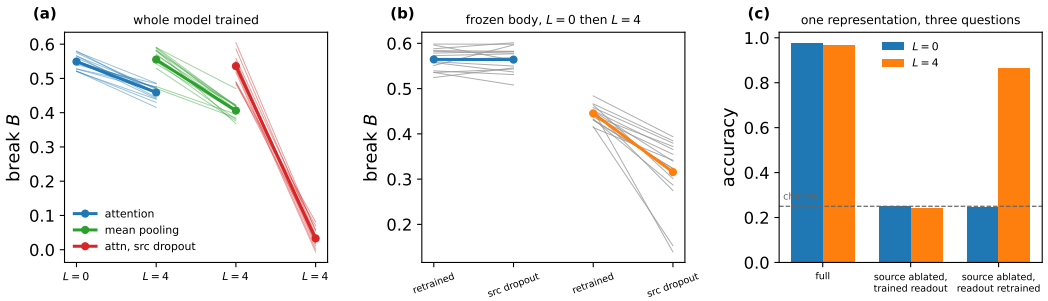

Figure 4: Selection, with one line per seed throughout. (a) Whole model trained. At identical transport, a readout trained to rely on complements leaves the source with little measured responsibility, while the default attention readout retains it and architectural flatness alone does little. (b) The same manipulation with the encoder and message layers frozen, so only the readout differs. Readout training alone still redistributes responsibility where transport has created alternatives (-0.130) and does nothing where it has not (-0.000), though the isolated effect is far smaller than in (a). (c) One frozen representation answering three different questions, the factor is decodable from the complement, unusable by the readout the network was trained with, and recoverable once only the readout is adapted.

under mean pooling $+0.500, +0.240, +0.081, -0.010$ at accuracies between 0.997 and 0.999. The source-dropout cell at 3×3 depth four is bimodal. Two of fifteen seeds reach 0.62 and 0.64, while the other thirteen exceed 0.996. Its mean of 0.950 passes the registered floor. Requiring both the mean and four fifths of seeds to clear the floor changes no verdict in the paper, so the cell is retained but its optimisation is reported as unreliable.

Two preregistered cells held out by their own decision rule. A residual bypass raises break by $+0.250$ at 3×3 depth four (95% CI $+0.172$ to $+0.328$, $15/15$ seeds). Its registered inclusion test required this advantage to shrink with kernel size, but it instead rises from $+0.165$ to $+0.547$. The largest kernel also fails its viability floor. The protocol therefore excludes both cells from the paper's claims, leaving the transformer counterexample unexplained rather than assigning it an unsupported mechanism (Appendix J).

Transformer substrate. Under the default attention readout, at $0, 1, 2, 4$ layers, exclusivity is $+0.628, +0.008, +0.001, -0.000$, complement decodability $0.35, 0.97, 0.98, 0.98$, break $+0.588, +0.373, +0.212, +0.004$, and accuracy between 0.970 and 0.975. Under mean pooling, break is $+0.534, +0.000, -0.000, +0.000$.

Transport by selection factorial, viability. All six cells of the graph factorial of Section 5.1 sit between 0.964 and 0.976 accuracy, and the frozen body arms are matched at 0.972 against 0.968.

Citation networks. The benchmark versions follow the LINQS preprocessing described by Sen et al. (2008), derived from the Cora collection (McCallum et al., 2000) and CiteSeer (Giles et al., 1998). Accuracy by propagation depth is 0.854 at $L = 1$ falling to 0.810 at $L = 4$ on Cora and 0.717 to 0.703 on CiteSeer, then collapsing to 0.300 and 0.237 at $L = 8$, which is why the depth eight cells are excluded as trainability failures. Exclusivity is between -0.012 and -0.001 at every viable depth on both, the neighbourhood decoding a node's label as well as the node itself does (0.89 against 0.89 on Cora, 0.78 against 0.77 on CiteSeer).

Homophily, observed and manipulated. The heterophilous half of Section 5.3 uses the same battery with only the data loader changed. The `roman-empire` and `amazon-ratings` benchmarks are taken from Platonov et al. (2023). On `roman-empire`, only $L = 1$ clears the viability floor, so the depth prediction is unavailable. Propagation nevertheless reduces break by $+0.600$ from input features to $L = 1$ in all fifteen seeds. `amazon-ratings` fails its floor at every depth and is not interpreted. In the Cora rewiring, edge homophily matches each target within 0.001 while degree,

features and labels remain fixed. Lowering homophily from 0.81 to 0.2 raises exclusivity by $+0.076$ in all fifteen seeds, above the registered $+0.05$ threshold. The gap opens because neighbourhood decodability falls, confirming homophily as the source of the original probe null.

Natural images, decodability trajectories. In the composed image task the carrier decodes at 0.496 at the first residual stage rising to 0.923 at the last, the complement at 0.239 rising to 0.862. In the frozen vision transformer the carrier decodes at 0.54 at layer 2 rising to 0.72 at layer 11, the complement at 0.29 rising to 0.47.

Indirection generator. Beyond the 0.0056 bits of mutual information quoted in Section 5.4, a classifier ignoring the pointer entirely reaches 0.442, so the address is required rather than merely helpful.

B THE SELECTION WINDOW LADDER

The selection-window follow-up (Experiment 24; E24) exists because the family sweep's selection contrast failed at the endpoints of the convolutional and transformer ladders, and the explanation the manuscript offered for that failure was a hypothesis it had not tested. Selection can only redistribute responsibility where transport has produced alternatives to recruit and the carrier still has responsibility to lose, so the effect should be non-monotonic in transport rather than uniform. That is a shape, and `docs/e24_selection_window_preregistration.md` fixes it as one scalar per family before the seeds testing it were drawn,

$$W \;=\; \tfrac{1}{2}\left[\Delta B(T_1) + \Delta B(T_2)\right] \;-\; \tfrac{1}{2}\left[\Delta B(T_0) + \Delta B(T_3)\right],$$

with $\Delta B(T) = B(\text{anchored}) - B(\text{recruited})$ at rung T, a threshold of $+0.10$ and a rule of twelve of fifteen paired seeds. The protocol states in its own second sentence that the archived seeds 0 to 14 were already measured and their shape already known, so what it protects is the statistic, the decision rule and the choice of which cells enter it, not the hypothesis. Seeds are 100 to 114, disjoint from every seed used elsewhere in this work, and nothing already reported was remeasured.

The two ladders. Convolutional, at kernel 1 and 3×3 depths $1, 2, 4$: anchored break $+0.645, +0.488, +0.416, +0.083$, recruited break $+0.694, +0.118, +0.030, +0.022$, so ΔB is $-0.048, +0.369, +0.386, +0.061$ at $3, 15, 15, 15$ of fifteen seeds, at accuracies from 0.996 to 1.000. Transformer, at $0, 1, 2, 4$ layers: anchored break $+0.590, +0.387, +0.186, +0.017$, recruited $+0.566, -0.001, -0.002, -0.001$, so ΔB is $+0.023, +0.387, +0.189, +0.018$ at $11, 15, 15, 12$ of fifteen, at accuracies from 0.970 to 0.976. Every one of the sixteen cells clears the viability floor.

The four predictions. W is $+0.371$ (95% CI $+0.322$ to $+0.421$, $d_z = +4.15$) in the convolutional family and $+0.267$ ($+0.226$ to $+0.309$, $d_z = +3.56$) in the transformer, both at fifteen of fifteen seeds against the registered threshold of $+0.10$. Both are confirmed. The rung by rung prediction is confirmed at all eight rungs: each endpoint inside $|\Delta B| \leq 0.10$ and each interior rung above it at twelve of fifteen or better. The sufficiency ladder is confirmed in both families, $S(T_3; G) - S(T_0; G)$ being $+0.746$ and $+0.725$ at fifteen of fifteen, on ladders of $0.250, 0.951, 0.975, 0.997$ and $0.249, 0.968, 0.974, 0.974$. The first entry of each is the one that matters: at zero transport a retrained readout recovers the factor at chance, so the account's claim that there is nothing to recruit there is measured rather than assumed. Here S is written with its readout class G attached: query-conditioned attention pooling and a head retrained for thirty epochs on a frozen body, because sufficiency is relative to what is allowed to read the complement.

The frozen body arm, and where it is unavailable. The remaining prediction asks whether the window survives holding the representation fixed: one body per rung, trained once with the anchored readout and frozen, with only the readout reinitialised and retrained under the two access conditions. In the transformer it is confirmed, $W_{\text{fr}} = +0.155$ ($+0.111$ to $+0.199$, fourteen of fifteen seeds), on a ladder of $+0.014, +0.207, +0.132, +0.014$. It is smaller than the end-to-end W of $+0.267$, which is the relation the graph substrate already showed: selection alone moves responsibility, and joint adaptation amplifies what it can move.

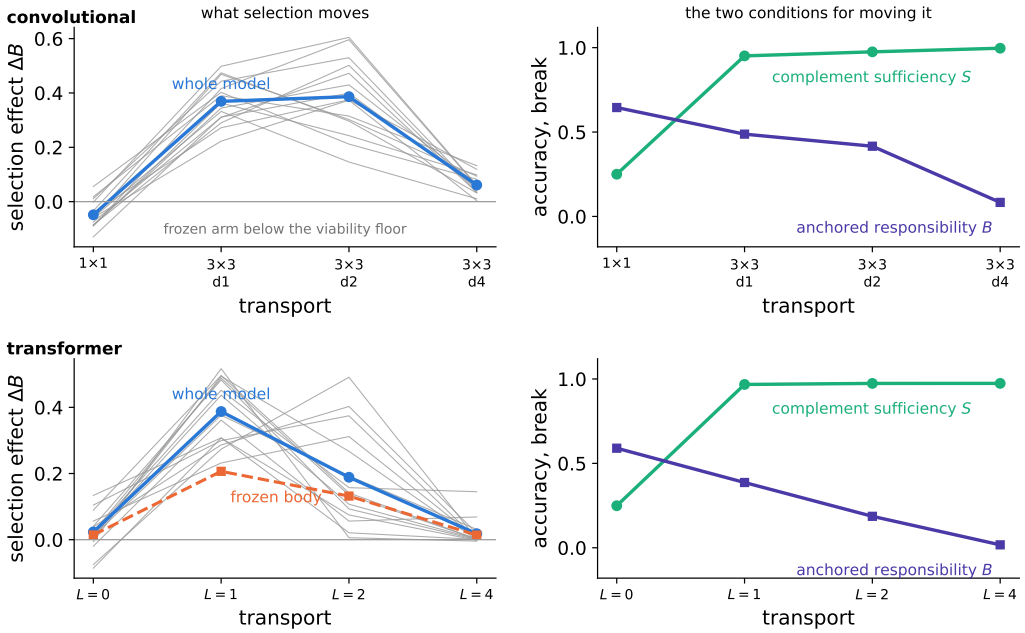

Figure 5: The selection window on fresh seeds. Left, the effect of changing the readout at each transport setting, faint lines seeds and the heavy line their mean, with the frozen body version dashed where its cells are viable. Right, the two conditions the account requires: complement sufficiency S, which is what there is to recruit, and anchored responsibility B, which is what there is to move. They cross, and the selection effect is largest between the crossings. The convolutional frozen arm is not drawn because one of its cells falls below the viability floor and the protocol excludes it.

In the convolutional family it is unavailable rather than null. Retraining the readout under carrier dropout on the frozen 3×3 depth one body fails in three of fifteen seeds, which reach chance, and the cell mean of 0.845 is under the floor. The protocol fixed the consequence in advance: an unviable cell is excluded and W_{fr} is reported as unavailable for the family rather than computed from the rungs that survived. For the record those rungs give $+0.012$ at six of fifteen, which is not claimed here. The failure is specific to freezing, since the same cell trained end to end reaches 0.998. Learning to route around the carrier is harder when the representation cannot move to help, which is the asymmetry the window describes rather than a counterexample to it.

C Expert routing, secondary cells

The full factorial crosses routing multiplicity $k \in \{1, 2, 4, \mathrm{soft}\}$ with a sharp query-conditioned gate and a uniform mean combine, at accuracies between 0.963 and 0.967. Break under the sharp combine is $+0.553, +0.363, +0.274, +0.374$ and under the flat combine $+0.447, +0.275, +0.119, +0.318$. Break is a substantially noisier instrument here than elsewhere. It does not reliably register the transport effect that exclusivity registers cleanly, eight of fifteen seeds against fourteen of fifteen. The interaction test was extended from five to fifteen seeds under a decision rule fixed in advance, giving a minimum detectable effect of 0.190.

Adding a switch style load balancing auxiliary loss at $k = 1$, an objective widely used in sparse mixture-of-experts training, lowers break from $+0.553$ to $+0.406$ (thirteen of fifteen seeds, minimum detectable effect 0.083) and exclusivity by 0.156 (fourteen of fifteen), at accuracy unchanged at 0.960. The scope of this observation should be stated carefully. It is established in a controlled synthetic binding task with eight experts, and it identifies a plausible interpretability cost of a routine engineering choice that should be tested in larger deployed models, rather than a demonstrated property of them.

A dose response. A single on-off comparison cannot say whether the balancing objective causes the loss of localisation, so its coefficient is swept over $\lambda \in \{0, 0.001, 0.01, 0.03, 0.1\}$ at fifteen seeds. Utilisation entropy rises from 1.341 to 2.077 nats, with per-seed Spearman $+0.987$ in $15/15$ seeds. Exclusivity falls from $+0.148$ to -0.030, meeting its registered rule at $13/15$, while break falls from $+0.553$ to $+0.395$ but misses its rule at $11/15$. Accuracy changes by only $+0.002$. The effect saturates with entropy by $\lambda = 0.01$.

Routing per token rather than per episode. The factorial above routes episodes, unlike sparse language models. A token-routed transformer gives the same transport result. Exclusivity falls from $+0.705$ at $k = 1$ to $+0.001$ at $k = 2$, a paired difference of -0.704 (95% CI -0.709 to -0.698, $5/5$ seeds), while accuracy remains between 0.966 and 0.973. The balancing effect does not transfer. Although the loss raises utilisation entropy by $+0.25$ to $+0.44$ nats, exclusivity changes by at most 0.002 and break by at most 0.029. Because this layer sums expert outputs, individual expert identity is not preserved downstream. The per-episode localisation result is therefore specific to architectures in which one expert supplies the layer output.

The same question of a deployed sparse mixture. The observation above is also tested on OLMoE-1B-7B (Muennighoff et al., 2025), with sixty-four experts, top-eight routing and 0.998 task accuracy. Because summed expert outputs hide expert identity downstream, the intervention reroutes the carrier position to its next eight experts and applies the same swap at a control position. Carrier-minus-control accuracy is $+0.002$, -0.003 and $+0.000$ at layers 1, 8 and 14. The intervention changes the answer margin in 89%, 97% and 64% of rows, so it lands despite the behavioural null. Expert identity at one position is therefore not necessary in this model on this task. Summed routed content remains outside the instrument's scope.

The substrate failed its viability floor on first construction, at approximately 0.61 in every cell. Before any interpretable metric was computed, a disjoint smoke seed exposed three implementation faults in query conditioning, straight-through routing and the flat combine. Correcting them changed no prediction. Table 7 records the repair.

D READING THE TWO MECHANISMS

This appendix expands the worked case of Section 3 for readers who want the two mechanisms spelled out before meeting them in five substrates at once. Nothing here is a result. It fixes what the words are doing.

Why this is not causal discovery. Structural causal modelling and causal representation learning infer relations among observed or latent variables (Pearl, 2009; Schölkopf et al., 2021), while large causal models extend that programme across datasets or textual claims (Kougioulis et al., 2026; Mahadevan, 2025). Here the factor and computational graph are known, and internal states can be overwritten directly. The target is therefore control within a known computation, not causal structure in the world.

Three meanings of factor bearing. A component may encode the factor directly or through a transformed and possibly distributed code. Both cases are decodable and are the targets of Sections 5.2 through 5.3. A component may instead carry a causally upstream address or routing instruction. The answer is then undecodable even though an intervention redirects the computation. Section 5.4 detects this third case by ranking components causally before decoding them.

Why B is not necessity. Necessity is what removal measures, $N_{\mathcal{C}} = \mathrm{Perf}(M) - \mathrm{Perf}(M \mid do(\mathcal{C} = \varnothing))$, and it is a different quantity from the transplant responsibility B of Section 3. A component may be necessary without producing donor following, and donor following can be inflated by a substitution that lands off the manifold the network was trained on. B measures relative causal control under a matched substitution, in which the donor state is a real state of the same component on a different episode and the wrong component control receives the same treatment. Necessity is reported alongside it wherever the paper leans hardest on a causal claim, at Table 4, at the attention cuts of Section 5.3, and at the pointer of Section 5.4, where the two quantities are furthest apart.

What availability establishes. Availability means only that the stated decoder can recover f from a component's state. It does not show that the trained readout uses the component, that the component is causally important or necessary, or that its state can replace the carrier. Those conclusions require the separate measurements in Section 3.

Where selection lives. Selection is the role played by whatever sets the component weights α_i. Gated readouts set them directly, query-conditioned attention sets them per episode, and expert routers assign them across experts. Pooling fills the same role in the convolutional and vision substrates. A component can therefore remain unused despite carrying useful information whenever its weight is near zero. The term *readout mechanism* is used because a classifier head applied after fixed pooling selects feature directions rather than components. The relevant selector is the operation that still has separate access to those components.

Why transport alone does not redistribute responsibility. If transport copies a value from h_1 to h_2 while the readout remains $\hat{y} = g(h_1)$, both states decode the value but only h_1 controls the output. Recruiting the copy requires selection to give h_2 nonzero weight. This is why the frozen-body experiment in Section 5.1 is decisive. Retraining only the readout restores accuracy after carrier ablation, showing that the information already existed and only its route to the output changed.

Why the tasks are constructed. Transport reads as a cause of availability only when the task remains solvable across transport settings. Otherwise reduced transport is confounded with task failure. Constructed tasks also fix the carrier through the data-generating process rather than through a probe. The citation networks, natural images, corpus text and frozen pretrained models in Table 1 then test whether the conclusions survive without either guarantee.

E MEASUREMENT DETAILS

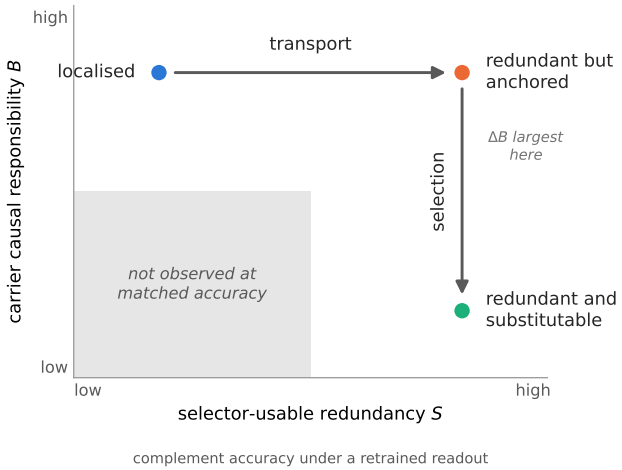

Figure 6: The regimes the two mechanisms reach. The horizontal axis is selector-usable redundancy, S of Equation 5, and not the probe-readable redundancy D and X measure; the two come apart, and it is the first that substitution needs. Transport moves a network rightwards by creating alternatives a readout could run on. Selection moves it downwards by recruiting them, and the contrast that measures it is largest near the top of that arrow, while the carrier still has responsibility to lose (Appendix B). Both axes are what an instrument registered rather than what exists, so the lower left is labelled as unobserved and not as impossible. S near zero says the retrained readout of Equation 5 could not recover the task from the measured complement, which a nonlinear or differently decomposed alternative could still evade, and B near zero is relative control under a matched transplant rather than necessity. The pointer of Section 5.4 is the standing warning that a component can be decisive while the instrument aimed at answers reads nothing.

Wrong component rule. Break is a difference against a wrong component transplant, and the rule differs by substrate. The recurrent-slot and routed-expert substrates each have $n = 8$ exchangeable components with no intrinsic ordering. Their control is fixed halfway around the arbitrary index, $(c + n/2) \bmod n = (c + 4) \bmod 8$; this is the *antipodal* component defined in Section 4, not a claim about representational geometry. The control is a uniformly random non-carrier node in the graph and transformer substrates, a uniformly random block that overlaps neither factor in the convolutional substrate, a uniformly random neighbour of the target node on Cora, and a non-name position in the language models. In the vision substrates, it is drawn uniformly from patches outside both the recipient's and donor's carrier regions, because a control inside the donor's object would place content where it belongs. In every case the wrong component receives *identical injected content* to the carrier condition, so the contrast isolates location rather than the magnitude or distribution of what is written. Because break is defined by this control, the choice of wrong component is consequential, so it was checked directly. On the trained graph checkpoints, break was recomputed under five rules while everything else was held fixed. A uniformly random non-carrier node, an average over five independent random draws, the non-carrier node whose state norm is closest to the carrier's, the non-carrier node whose state is most cosine-similar to the carrier's (the hardest available decoy), and the other factor's source node (a semantically loaded control). The result is invariant. At $L = 0$ the five rules give $+0.6114$, $+0.6118$, $+0.6117$, $+0.6104$ and $+0.6117$. At $L = 4$ they give $+0.5093$, $+0.5093$, $+0.5091$, $+0.5085$ and $+0.5093$. The largest deviation from the reported rule in any individual seed is 0.004, two orders of magnitude below the effects the paper claims. The reported effects therefore do not depend on comparing against uninformative controls, which is what one would expect given that every condition injects identical donor content and differs only in location.

The claiming convention. An effect is claimed only when at least four fifths of paired seeds agree in sign and the mean magnitude exceeds 0.10 on the break scale. The convention is a secondary check on the effect sizes rather than what makes a result real, since the paired means and intervals of Appendix I carry the evidence.

The magnitude bar on natural data. The 0.10 magnitude threshold of the claiming convention above is applied to the constructed tasks only. On natural data break spans roughly a tenth of the range it does on constructed tasks, so a bar set for one scale would reject every effect on the other before any of them had been tested. Paired effects and their intervals are reported there instead, and the sign rule still applies.

Donor matching. Donors are drawn from the same evaluation distribution and restricted to episodes with the same query as the recipient and a different answer, so that following the donor is diagnostic. Donors are formed by a fixed cyclic permutation within batch rather than by resampling, which keeps the marginal distribution of donor states identical to that of recipient states.

Complement decoder. $D_{\neg c}$ is not defined identically across substrates, and the difference matters. In the graph, convolutional and transformer substrates the complement is the *mean* representation over non-carrier components. In the recurrent slot substrate it is the *concatenation* of all non-carrier slots, which gives the complement considerably more capacity and therefore makes exclusivity harder to achieve. Slot exclusivity is thus measured under a stricter test than the others, and the two families should not be compared on exclusivity magnitude.

Carrier identification. In the graph, convolutional and transformer substrates the carrier is the component the task marks as holding the queried factor, so it is fixed by the data-generating process and not estimated. In the recurrent slot and expert routing substrates components are exchangeable and the carrier is the one with highest decoding accuracy on a discovery split, ties broken by lowest index, with every reported quantity computed on a disjoint evaluation split. On the citation networks the carrier is the target node itself and no nomination is involved.

Transplant fidelity in the composed image task. The failure of two of the composed-image experiment's four preregistered predictions (Experiment 20; E20) traces to the content rule rather than to the substrate. The write places the donor's channel mean over the queried quadrant's positions, and that resembles what the layer would itself have produced only where the map is coarse. At

64×64 the recipient follows neither donor nor control (0.129 against 0.027). At 8×8 it follows them at 0.774 against 0.404. Break therefore rises with depth in that task because the instrument becomes faithful with depth, and the two cannot be separated within the cells run. The representational rung is unaffected, since decoding does not write anything.

Where nomination applies it was checked directly, by resampling the discovery split two hundred times per cell with replacement, re-nominating on each resample, and recording the share of resamples that agree with the modal nomination. A single stable carrier gives 1.000 and a nomination driven by discovery noise gives about 0.125, there being eight components. Over sixty nominated cells agreement averages 0.819 and twenty nine are unanimous, far above the noise null ($+0.694$, 95% CI $+0.631$ to $+0.756$, $t_{59} = 22.2$, all sixty above it). The check does not, however, come out uniformly, and the pattern in how it fails is the paper's own thesis applied to its own methodology. Agreement is 1.000 in the constrained slot substrate and 0.494 in the permissive one, a paired difference of $+0.506$ (95% CI $+0.308$ to $+0.705$, five of five seeds). Across the two by two, the message main effect on nomination stability is $+0.389$ (95% CI $+0.206$ to $+0.572$, $t_4 = 5.91$, five of five). Transport does not merely make the factor decodable elsewhere. It makes the question *which component holds the factor* progressively ill posed, because the ranking that answers it is then a ranking among near equals. Two consequences are worth stating. Nomination stability is itself a symptom of the phenomenon under study, and could be used as a cheap probe free diagnostic for it. And the quantities the paper reports at high transport are conditioned on a nomination that is only about half reproducible, which is a reason to prefer the transplant based quantities there. Break at a component nominated by noise is small because the component is not causal, which is the correct answer arrived at without needing the nomination to be stable.

F THE VISION BOUNDARY IN FULL

Both paragraphs below are summarised in Section 5.5.

Unmodified images. COCO images with human-annotated boxes push the same question off the construction. The comparison is made on a frozen ViT body rather than the ResNet of Section 5.5, and at that body's last layer break at the annotated box is real but an order of magnitude smaller than on the composed task ($+0.036$ against $+0.341$). Two of the four predictions for that experiment failed. Selection still replicates. On the same frozen ViT body, query-conditioned attention yields higher break than mean pooling or carrier dropout. The size of the effect is sensitive to how locality was built. The existence of a selection effect is not.

A body not trained for the factor. A frozen ViT trained on neither the composed query task nor the COCO query keeps its carrier across depth. Exclusivity stays flat and break falls only mildly, while sufficiency still rises with depth. Partial transport without collapse is the expected pattern when the body was not trained to manufacture redundancy for the measured factor.

G CELL TYPE SUBSTITUTIONS THAT FAILED TO TRAIN

An attempt was made to vary the recurrent cell of the slot substrate in order to test whether the transport result depended on leaky dynamics. The attempt failed, and the failure is instructive enough to record. With a tanh cell neither condition learned the binding task. With a gated recurrent cell the permissive condition trained to 0.995 and showed exactly the expected signature of unconstrained transport, with break at zero and complement decodability at 1.00, but the constrained condition reached only 0.21, below the 0.25 chance floor. Because a substrate that cannot perform the task cannot localise it, the constrained gated cell is uninformative and no conclusion about cell types can be drawn from this pair.

This failure determined the design of the graph experiments. Rather than substituting components and hoping both conditions train, the graph task was constructed to be solvable at every transport setting, so that the transport variable could be swept without confounding it with trainability. The discipline is recommended generally. A causal localisation experiment needs a viability floor and a task whose solvability does not depend on the variable under study.

H PRETRAINED MODEL DETAILS

Table 5: The ladder in three frozen pretrained language models and one frozen vision transformer, at relative depths, five seeds per cell, quoted throughout Section 5.3. X cut is exclusivity when attention into the carrier is blocked below that layer, the inference-time transport manipulation. sel. cost is the accuracy lost when the answer position may not attend to the carrier from that layer up, against the wrong position control. S is what a readout trained on the frozen final layer recovers with the carrier ablated. Every language model follows the same course, with a localised early phase, a redundant but anchored middle, and a late phase in which the carrier is fully decodable but has negligible measured responsibility. The vision model, which was trained on neither this task nor its query, does not. Its carrier stays exclusive and stays responsible. For the ViT, the inference-time cuts are omitted. The selection cut has no answer position to target, while the transport-cut diagnostic is sensitive to the software environment and is therefore not interpreted.

Model	Layer	D_{car}	D_{comp}	X	B	X cut	sel. cost	S
GPT-2 small	1/11	1.00	0.49	$+0.51$	$+0.75$	$+0.47$	$+0.29$	0.51
GPT-2 small	4/11	1.00	0.91	$+0.09$	$+0.58$	$+0.13$	$+0.36$	0.56
GPT-2 small	7/11	1.00	0.99	$+0.01$	$+0.51$	$+0.25$	$+0.35$	0.97
GPT-2 small	10/11	1.00	0.98	$+0.02$	$+0.00$	$+0.48$	-0.00	0.98
GPT-2 medium	2/23	0.80	0.50	$+0.30$	$+0.71$	$+0.36$	$+0.19$	0.50
GPT-2 medium	8/23	1.00	1.00	$+0.00$	$+0.65$	$+0.12$	$+0.29$	0.60
GPT-2 medium	15/23	1.00	0.99	$+0.01$	$+0.05$	$+0.48$	$+0.01$	0.99
GPT-2 medium	21/23	1.00	0.94	$+0.06$	$+0.00$	$+0.49$	$+0.00$	0.99
Pythia-160m	1/11	0.99	0.51	$+0.49$	$+0.70$	$+0.48$	$+0.40$	0.50
Pythia-160m	4/11	1.00	0.62	$+0.38$	$+0.63$	$+0.46$	$+0.43$	0.50
Pythia-160m	7/11	1.00	0.96	$+0.04$	$+0.42$	$+0.42$	$+0.18$	0.95
Pythia-160m	10/11	1.00	0.76	$+0.24$	$+0.00$	$+0.48$	$+0.00$	0.95
ViT-B/16	2/11	0.54	0.29	$+0.24$	$+0.41$	–	–	0.10
ViT-B/16	6/11	0.70	0.44	$+0.27$	$+0.39$	–	–	0.22
ViT-B/16	11/11	0.72	0.47	$+0.26$	$+0.34$	–	–	0.47

In a fresh-repository reproduction with empty results under PyTorch 2.13.0 and Transformers 5.14.1, the carrier-decoding change under the ViT transport cut was $+0.084$ rather than the archived -0.600. This sign reversal affects only the uninterpreted diagnostic. The reported ViT depth, selection and sufficiency results retain their directions.

The ladder's zero point sits below the first row of Table 5. Read at the token embeddings, before any attention has run, exclusivity is $+0.034$ in GPT-2 small, $+0.025$ in GPT-2 medium and $+0.032$ in Pythia-160m. The factor is not yet anywhere in particular, so everything the table shows is built by the layers above.

The two transition statistics quoted in Section 5.3 are the first layer at which complement decodability exceeds 0.90, and the first layer beyond which break stays below 0.10. Both thresholds were chosen after the trajectories in Table 5 were seen, so the between model comparison is preregistered while its statistic is post hoc.

The generated table in `pretrained.tex` records every registered contrast as a paired seed difference with its interval, standardised effect size and sign count. It remains part of the supplementary source rather than being repeated in the PDF. Some entries coincide when carrier decodability is saturated because exclusivity is then the negative of complement decodability.

The interventions are implemented in `redundancy/pretrained.py` as context managers over the frozen model, an additive attention mask for the two route cuts, and a residual write or zeroing at a position range for the transplant and the ablation. Prompts are grouped by tokenised length rather than padded, so that a position index means the same thing in every row of a batch. A twin whose tokenisation differs in length from its item cannot serve as that item's donor and is dropped, which is why the reported prompt count per seed is below the two hundred and fifty six generated. The corpus-agreement experiment (Experiment 23; E23) uses items that vary in length

far more than templates do, so those are right padded instead and every position the experiment names is carried per row. That is exact rather than approximately harmless because these bodies are causal. A real token attends only to positions at or before its own, and every pad sits after every real token, so no real token's state can depend on one. It is checked rather than argued, in `tests/test_reproducibility.py`, against the same rows run unpadded.

Decodability in this appendix and in Section 5.3 is held out, k fold, with the ridge strength chosen inside each fold. In sample ridge reads 1.00 off a seven hundred and sixty eight dimensional residual from two hundred and fifty six prompts whatever the residual contains, which would have made exclusivity a statement about dimensionality rather than about the model.

I Statistical summary

All claimed contrasts are paired within seed. Reported intervals are Student intervals on paired means, with bootstrap intervals agreeing throughout. Sign counts follow the preregistered direction, and minimum detectable effects state what each null excludes. The generated table in `statistics.tex` remains in the supplementary source and contains every mean, interval, standardised effect, sign count and minimum detectable effect. Headline values are reported with their corresponding claims in the main text and preceding appendices.

J Experimental provenance and deviations

Fifteen dated protocols cover the experiments in this paper. Each fixes its predictions, viability floors, sign rules, magnitude thresholds and the meaning of both outcomes, and each closes with a clause forbidding any edit to those once its first run has begun. The documents travel whole and unedited with the code and the run manifests, at `https://anonymous.4open.science/r/research_representation_causality-3115`. This appendix records what they license, not what they say.

Table 6 gives the status of every claim. Four statuses are used. *Confirmatory* means the hypothesis and the analysis were fixed before the relevant data existed. *Prospective follow-up* means the experiment was motivated by an earlier result but was specified before its own runs. *Exploratory* means the analysis was defined after inspecting the same data it is applied to. *Unavailable* means the intervention or a viable comparison could not be constructed.

Table 7 lists every departure from an original plan. Two substrates were repaired after failing a viability floor, in both cases before any interpretable metric was computed and on a smoke seed disjoint from the evaluation seeds, with no prediction edited. One further protocol made its own inclusion conditional on an interaction that then went the other way, so the material it covers is absent from this paper's claims while the document is still shipped.

One further note concerns magnitudes rather than plans. Re-executing the whole-image cells of Section 5.5 from empty results in a second software environment moved the final-stage break from $+0.023$ to $+0.106$, with seventeen rather than twenty of twenty stage-by-seed cells below 0.10. The cell is seed- and environment-sensitive in magnitude while the contrast with the composed task is not, which is why the text reports it as a contrast rather than an absolute zero.

Table 6: Claim provenance. *Where* gives the section that reports the claim, or the appendix letter where the claim is reported only there. *Manip.* separates a manipulated variable from one only observed. *Status* records when the analysis was fixed relative to the data it is applied to.

Claim	Where	Manip.	Status	Outcome
Transport controls availability, five media	5.2	✓	confirmatory	confirmed
Bandwidth cliff is a step, not a gradient	5.2	✓	confirmatory	confirmed
The cliff as a capacity law	5.2	✓	confirmatory	**failed**
Selection × transport, graph	5.1	✓	confirmatory	confirmed
Readout alone, with the body frozen	5.1	✓	prosp. follow-up	confirmed
Decodable, unused, readout recoverable	5.1	✓	confirmatory	confirmed
Both mechanisms, citation networks	5.3	✓	confirmatory	confirmed
Homophily is why the probe is silent	5.3	✓	confirmatory	confirmed
The same, on released heterophilous graphs	5.3	obs.	confirmatory	1 of 2 viable
Transport and selection cuts, three LMs	5.3	✓	confirmatory	confirmed
The same on corpus agreement items	H	✓	confirmatory	1 of 4 failed
Phases at matched relative depth	5.3	obs.	exploratory	confirmed
Responsibility without decodability	5.4	✓	confirmatory	confirmed
Selection, convolutional and transformer	5.5	✓	confirmatory	**failed**, floor
The same at fixed transport	5.5	✓	exploratory	confirmed
The selection window, fresh seeds	B	✓	confirmatory	confirmed, both
the window with the body frozen	B	✓	confirmatory	1 of 2 unavailable
the sufficiency ladder behind it	B	✓	confirmatory	confirmed, both
Selection, expert routing	C	✓	confirmatory	undetermined
Load balancing lowers localisation	C	✓	confirmatory	confirmed, bounded
Whole image label, no carrier	5.5	✓	confirmatory	confirmed null
Localised label, the carrier returns	5.5	✓	confirmatory	2 of 4 failed
Annotated regions in COCO images	F	✓	confirmatory	2 of 4 failed
Frozen vision transformer keeps its carrier	F	✓	confirmatory	1 of 4 unavailable

Table 7: Deviations from the original plans, and what each cost.

Experiment	Original plan	Change	Timing	Effect on claim
Seed extension	5 seeds	15 seeds on claim-bearing cells	After preliminary results, before the added runs	Changed three verdicts
Frozen body selection	Not in the initial factorial	Added isolation experiment	Before its runs	Confirmed
Conv. and transformer selection	Highest transport setting	Added fixed transport comparisons	After the endpoint failed at floor	Exploratory confirmation
The same, replicated	Nothing registered on the shape	Window statistic W registered on fresh seeds	Before seeds 100–114 were drawn	Confirmed, both families
Selection window, frozen conv. body	Registered as a secondary endpoint	None; the cell fell below the floor	After the runs, on the rule fixed before them	W_{fr} unavailable
Pointer substrate	First substrate failed its floor	Generator, architecture and one metric repaired	Before the amended run	Confirmed
Expert routing substrate	Failed its floor on first run	Three implementation faults fixed	Before any interpretable metric	Undetermined
COCO regions	Assumed sample count and chance level	Sampling and floor corrected	Before training and evaluation	Two predictions failed
Relative depth statistic	Not specified	Statistic defined on the observed trajectories	After the data	Exploratory