From Representation to Causation: A Three-Tier Framework and Open-Source Benchmark for Mechanistic Interpretability

Anonymous authors
Paper under double-blind review

Abstract

Interpretability research often conflates whether information is merely encoded within a model or whether it causally drives behavior. We introduce MechInterp3, a failure-aware framework that disentangles these properties into a three-tier hierarchy: (Tier-1) linear encoding, (Tier-2) probe accessibility, and (Tier-3) causal responsibility. By applying this framework to six transformer architectures across four tasks, we reveal that standard causal interventions "silently fail" in approximately 50% of model-task combinations due to weak behavioral contrast. This produces mathematically ill-conditioned estimates that undermine causal claims. Our systematic evaluation reveals three critical findings. First, we identify a pervasive tier dissociation where models with near-perfect probe accuracy often show zero or negative causal recovery, most notably in GPT-2 sentiment processing (-0.34 recovery). Second, we demonstrate that observational methods, such as attention weights and gradient attribution, are uninformative of causal structure, showing near-zero correlation ($\rho < 0.1$) with intervention effects. Third, we discover that tasks requiring relational reasoning, such as NLI, induce more stable and localized causal circuits than surface-level tasks, despite having weaker linear representations. We release MechInterp3 as an open-source library to establish a rigorous statistical foundation for the study of machine intelligence.

1 Introduction

When a language model correctly answers a factual question, does it use the knowledge representations that interpretability tools claim to identify? The answer depends on which tool you ask. Consider GPT-2 processing sentiment: linear probes achieve 82% accuracy, suggesting strong sentiment representation (Alain & Bengio, 2017; Ravichander et al., 2021). Yet causal interventions that patch sentiment-encoding activations yield -0.34 recovery, meaning the intervention makes outputs *less* sentiment-aligned. Simultaneously, attention weights and gradient-based attribution identify entirely different components as important, showing a near-zero correlation (Spearman $\rho < 0.1$) with actual causal effects (Jain & Wallace, 2019; Wiegreffe & Pinter, 2019; Sundararajan et al., 2017). These contradictions reveal a fundamental conceptual confusion about the difference between information being *present* versus information being *functional*.

The mechanistic interpretability community has developed powerful methods, such as probing classifiers, attention visualization, and activation patching (Meng et al., 2022; Vig et al., 2020), but these are often treated as interchangeable proxies for the same underlying truth. In practice, they operationalize categorically different questions: whether information is **encoded**, whether it is **extractable**, and whether it **causally controls** behavior. Just as a student may have memorized a fact (encoding) and recall it on a quiz (extraction), yet fail to apply it to a novel problem (causal use), neural networks exhibit systematic dissociations between their internal "knowledge" and their external decisions.

We operationalize this insight as MechInterp3, a framework that transforms interpretability from an ad-hoc search for "explanations" into a rigorous diagnostic profiling of model layers. Building on Beckmann & Queloz (2025)'s epistemological hierarchy, we define three tiers of evidence (summarized in Table 1):

- Tier-1 (Encoding): Measures if concepts exist as linear directions in activation space.
- Tier-2 (Accessibility): Measures if information is readable by a simple linear probe.
- Tier-3 (Causal Responsibility): Measures if interventions on those representations reliably steer model behavior.

Crucially, we demonstrate that current causal interpretability is in a state of "silent failure." In approximately 50% of our model-task combinations, standard activation patching is mathematically ill-conditioned due to weak behavioral contrast. In these regimes, current software libraries frequently return numerical outputs that researchers mistake for mechanistic insights, when in fact the causal estimate is undefined. MechInterp3 introduces "failure-aware" semantics to solve this, providing a "unit test" for interpretability that distinguishes between a mechanism being *absent* versus the analysis being *inapplicable*.

Tier	**Name**	**Measurement Goal**	**Student Analogy**
Tier-1	Encoding	Physical presence in layers	Possessing the textbook
Tier-2	Accessibility	Readability by external probes	Passing a multiple-choice quiz
Tier-3	Causal Resp.	Behavioral control via patching	Solving a novel word problem

Table 1: The MechInterp3 framework separates the existence of info from its functional utility.

Our systematic evaluation across six transformer models (Encoder, Decoder, and Encoder-Decoder families) and four tasks reveals a **Generalizability Crisis** in interpretability:

- **Architectural Trade-offs:** There is a fundamental "Interpretability-Causality" trade-off. Encoders (e.g., BERT) yield the cleanest linear encoding (Tier-1: 0.88 AUC) but the most fragile causal pathways (34% valid pairs). Decoders (e.g., GPT-2) exhibit "messier" representations but are highly steerable (84% valid pairs). This suggests that interpretability methods developed on one architecture family are fundamentally non-transferable to others.
- **Observational Unreliability:** We provide empirical proof that attention weights and gradient norms are almost entirely uninformative of causal structure ($\rho < 0.1$). What a model "looks at" is decidedly not what drives its output.
- **Informative Negatives:** We show that negative results, like GPT-2's -0.34 recovery, are not failures of the tool, but evidence of complex, distributed mechanisms where single-layer interventions disrupt competing internal balances.

Contributions. We release the MechInterp3 open-source library, providing the first benchmark with "validity filtering" and "task-faithful behavioral scalars." By moving beyond accuracy and towards causal coverage matrices, we enable valid cross-architecture comparisons that were previously impossible due to undetected numerical instability.

2 Related Work

We position MechInterp3 at the intersection of five research areas, identifying how our framework provides a much-needed methodological bridge between philosophical theory and empirical failure-analysis.

Philosophical Foundations of AI Understanding. The interpretability field has long struggled with a lack of formal epistemology. Beckmann & Queloz (2025) recently addressed this by introducing an epistemological hierarchy that distinguishes between conceptual representation (the existence of information), information accessibility (the readability of that information), and causal responsibility (the functional utility of that information). Their work provides the theoretical grounding for why different methods yield conflicting results; they are, in fact, measuring fundamentally distinct properties. **Our contribution**: We

operationalize this abstract hierarchy into a concrete, algorithmic protocol. While Beckmann & Queloz establish *what* to measure, MechInterp3 establishes *how* to measure it reliably. We extend their theory by introducing the concept of "Undefined" states, recognizing that in many model states, causal measurement is not just difficult, but mathematically ill-posed.

Mechanistic Interpretability and the Decoder Bias. The circuits research area (Olah et al., 2020; Elhage et al., 2021) seeks to reverse-engineer neural networks into interpretable components, such as induction heads (Olsson et al., 2022) or arithmetic circuits (Nanda et al., 2023). While highly influential, this research has primarily focused on decoder-only models (notably GPT-2), leading to a "decoder bias" in the field's general conclusions. **Our contribution**: We show that mechanistic findings are highly architecture-dependent. By comparing encoder, decoder, and encoder-decoder families, we demonstrate that the "circuits" identified in decoders (where Tier-3 applicability is 84%) may not have functional equivalents in encoders (where applicability drops to 34%). MechInterp3 provides the first cross-architecture benchmark that identifies which architectures are even amenable to circuit-style analysis.

The Probing-Causation Paradox. Probing classifiers (Alain & Bengio, 2017; Belinkov et al., 2017; Tenney et al., 2019) have been the dominant tool for identifying what models "know." However, a growing body of evidence suggests that "knowledge" discovered by a probe is often ignored by the model during inference. Ravichander et al. (2021) established this dissociation philosophically, but the field has lacked a standardized metric to quantify it. **Our contribution**: We formalize this as a Tier-2/Tier-3 gap. Our results transform this paradox into a quantifiable engineering constraint: we show that GPT-2's high probe accuracy for sentiment (0.82) actually hides a negative causal recovery (-0.34). By situating probing within our three-tier framework, we allow researchers to measure exactly when a model is "hallucinating" knowledge; that is, possessing it in the representations (Tier-2) but failing to utilize it behaviorally (Tier-3).

Linear Representation and Universal Geometries. The hypothesis that concepts are encoded as linear directions (Mikolov et al., 2013; Park et al., 2024) is foundational to mechanistic interpretability. Tools like Concept Activation Vectors (CAVs) (Kim et al., 2018) rely on this linear structure. **Our contribution**: MechInterp3 subjects this "Linear Representation Hypothesis" to a rigorous multi-architecture analysis. We find that linear encoding is not a universal property of transformers but is architecture-dependent: encoders exhibit highly organized linear structures (Tier-1: 0.88), while decoders are significantly less organized (0.67). This suggests that tools built on linear assumptions (like linear steering) may be fundamentally ill-suited for the very models (decoders) they are most frequently applied to.

Causal Interventions and Silent Failures. Activation patching (Vig et al., 2020; Meng et al., 2022) is often viewed as the "ground truth" of interpretability. Recent automation of these methods (Conmy et al., 2023) has accelerated circuit discovery. **Our contribution**: We identify a critical, previously unaddressed flaw in these methods: *Silent Failure.* Standard patching relies on a "behavioral gap"; that is, the difference in model output between two inputs. If this gap is small, the normalization used in causal recovery math becomes ill-conditioned, producing extreme values that researchers often misinterpret as "strong" causal effects. We show that 50% of model-task combinations produce these numerical artifacts. MechInterp3 introduces "failure-aware" semantics, including task-faithful behavioral scalars and gap-based validity filtering, to ensure that causal attribution is only reported when mathematically valid.

Cross-Architecture Comparison and the Universality Challenge. A significant limitation of current interpretability research is its narrow architectural focus. Historical work has centered heavily on BERT-style encoders for linguistic analysis (Rogers et al., 2020; Tenney et al., 2019), while the modern mechanistic interpretability program focuses almost exclusively on decoder-only models like GPT-2 (Elhage et al., 2021; Nanda et al., 2023). Encoder-decoder models, such as T5 and BART, remain comparatively under-studied despite their ubiquity in sequence-to-sequence tasks. This siloed approach has led to the implicit assumption that interpretability "discoveries" (like the existence of linear circuits) are universal properties of transformers. **Our contribution:** To our knowledge, we provide the first controlled, three-tier comparison across all three major transformer families using a unified methodology and identical task suite. Our findings directly challenge the universality assumption. We reveal that no architecture "dominates" the mechanistic hier-

archy: Encoders lead on representation organization (Tier-1 AUC of 0.88), yet Decoders provide the only reliable substrate for broad causal analysis (84% valid-pair rate vs. 34% for encoders). Most surprisingly, we show that Tier-2 performance is the only "universal" metric, with all architectures clustering around $\sim$0.77. By documenting these systematic trade-offs, we demonstrate that a model's architectural "profile" determines which interpretability tools will succeed; that is, mechanistic conclusions are not universal, but architecture-specific.

Observational Methods vs. Causal Truth. Gradients and attention weights remain the most popular interpretability tools due to their low computational cost (Simonyan et al., 2013; Sundararajan et al., 2017; Wiegreffe & Pinter, 2019). However, their relationship to actual model behavior has been debated. **Our contribution**: We provide definitive empirical evidence that observational importance scores are almost entirely uninformative of causal responsibility. With a Spearman $\rho < 0.1$ correlation between attention/gradients and causal interventions, we show that what a model "looks at" is essentially decoupled from what drives its output. This result serves as a formal warning against using observational baselines as a proxy for mechanistic understanding.

3 Problem Formulation

Let M be a transformer model with L layers, processing an input x to produce an output $M(x)$. We denote the hidden representation at layer ℓ as $\mathbf{h}_\ell(x) \in \mathbb{R}^d$.

We distinguish between the physical architecture of the model and the semantic information under investigation using the following two definitions:

Components. We denote the set of all model components as $\mathcal{C}$. A specific **component** $c \in \mathcal{C}$ represents a discrete structural site within the model graph where representations are computed and through which information flows (e.g., a residual stream, attention head, or MLP block at layer ℓ).

Concepts. A **concept** C is a semantic property defined extensionally via a dataset of binary polar examples $\mathcal{D} = \{\mathcal{D}^+, \mathcal{D}^-\}$. Specifically, we consider $\mathcal{D}^+ = \{x_1^+, \ldots, x_n^+\}$ to represent instances where the concept is present (e.g., positive sentiment) and $\mathcal{D}^- = \{x_1^-, \ldots, x_m^-\}$ to represent instances where it is absent or inverted (e.g., negative sentiment).

While a concept C is a property of the data, a component c is a functional unit of the architecture. The goal of mechanistic interpretability, under our framework, is to determine the extent to which a specific component c represents or causally mediates a specific concept C.

3.1 A Hierarchy of Mechanistic Evidence

As related earlier, existing interpretability literature often conflates representational existence with functional utility. In this paper we propose a formal separation into three tiers of evidence, moving from latent encoding to behavioral control. Let us define these tiers.

Definition 1 (Tier-1: Conceptual Representation)**.** Model M possesses a **linear conceptual representation** of concept C at layer ℓ if there exists a direction $\mathbf{d} \in \mathbb{R}^d$ that maximizes the separability of the classes in activation space. In our experiments, we use the Mean-Difference direction $\mathbf{d} = \mu^+ - \mu^-$, which is a closed-form approximation of the optimal linear separator:

$$\text{AUC}\left(\{\mathbf{d}^\top \mathbf{h}_\ell(x) : x \in \mathcal{D}^+\}, \{\mathbf{d}^\top \mathbf{h}_\ell(x) : x \in \mathcal{D}^-\}\right) > \tau_1 \tag{1}$$

where AUC measures the probability that a random positive example is ranked higher than a random negative example along $\mathbf{d}$.

Definition 2 (Tier-2: Information Accessibility)**.** Concept C is **linearly accessible** if a simple readout mechanism can extract C from $\mathbf{h}_\ell$. We operationalize this as the test accuracy of a linear classifier $f(\mathbf{h}) = \sigma(\mathbf{w}^\top \mathbf{h} + b)$ trained on layer-ℓ activations:

$$\text{Accuracy}(f, \mathcal{D}_{\text{test}}) > \tau_2 \tag{2}$$

Table 2: Task-faithful output functions by task and architecture. $P(\text{wrong})$ is defined as the sum of probabilities of all tokens in the vocabulary except the correct one, or a specific contrastive token.

Task	**Encoder**	**Decoder**	**Enc-Dec**
Sentiment	$z_{\text{pos}} - z_{\text{neg}}$ (logits)	$\log P(\text{“positive”}) - \log P(\text{“negative”})$	$\log P(\text{“positive”}) - \log P(\text{“negative”})$
NLI	$z_{\text{ent}} - z_{\text{con}}$ (logits)	$\log P(\text{“yes”}) - \log P(\text{“no”})$	$\log P(\text{“entailment”}) - \log P(\text{“contradiction”})$
Factual Recall	$\log P(\text{correct})$ at [MASK]	$\log P(\text{correct}) - \log P(\text{wrong})$	$\log P(\text{correct}) - \log P(\text{wrong})$
QA	$\log P(\text{correct})$ at [MASK]	$\log P(\text{correct}) - \log P(\text{wrong})$	$\log P(\text{correct}) - \log P(\text{wrong})$

We note that Tier-2 is a necessary but not sufficient condition for causal use.

Definition 3 (Tier-3: Causal Responsibility and Applicability)**.** Component c is **causally responsible** for behavior B if an intervention on c shifts the model output in a task-predictable direction. We introduce $g(M(x)) \in \mathbb{R}$, a *task-faithful behavioral scalar* to measure the margin between intended and unintended outputs.

Critically, we define the **Validity Set** $\mathcal{V}$ for a pair of donor (x^+) and base (x^-) examples:

$$\mathcal{V} = \{(x^-, x^+) : |g(M(x^+)) - g(M(x^-))| > \gamma\} \tag{3}$$

where γ is the **Gap Threshold**. Causal recovery R_c is defined only over the support of $\mathcal{V}$:

$$R_c = \frac{g(M_{\text{patch}}(x^-; c \leftarrow x^+)) - g(M(x^-))}{g(M(x^+)) - g(M(x^-))} \tag{4}$$

We argue that R_c is an **ill-conditioned estimator** when $\gamma \to 0$. Therefore, we define Tier-3 responsibility as the median recovery over $\mathcal{V}$, reported alongside the **Valid-Pair Rate (VPR)** $|\mathcal{V}|/|\mathcal{D}|$.

The choice of behavioral scalar $g(M(x))$ is critical for ensuring that interventions measure task-relevant logical shifts rather than generic changes in model confidence. Table 2 details the task-faithful output functions used across different architectures (and sample tasks) to provide a consistent causal metric.

3.2 Logical Dissociations and Mechanistic Profiles

The three tiers are logically non-equivalent, allowing us to characterize models via a **mechanistic profile**. Table 3 formalizes the interpretations of tier combinations.

- **T1 $\gg$ T3 (Encoded but Unused)**: Common in encoders, where concepts are organized linearly but lack a clear causal pathway to the final output layer.
- **T3 $\gg$ T1 (Distributed Control)**: Occurs when patching recovers behavior despite the absence of a single linear direction, suggesting the concept is encoded non-linearly or across multiple subspaces.
- **VPR ≈ 0 (Mathematically Undefined)**: Represents an identifiability failure where the model lacks sufficient behavioral contrast to permit causal attribution via patching.

3.3 Architectural Adaptation

While the definitions of $\mathcal{C}$ and $g(M(x))$ are general, their physical implementation depends on the model's information flow. To ensure valid cross-architecture comparison, we adapt the extraction points and patching targets as summarized in Table 4. For instance, in Tier-3 interventions, c corresponds to the residual stream at layer ℓ for decoders, but involves cross-attention blocks for encoder-decoders to account for the split between input processing and output generation.

Table 3: Mechanistic Profiles. MechInterp3 separates causal responsibility from representational existence to identify where interpretability tools "silent fail."

Pattern	T1	T2	T3 (VPR)	Mechanistic Interpretation
Clean Circuit	High	High	High (High)	Linearly encoded and causally localized.
Read-only	Low	High	Low (High)	Extractable by probes but ignored by model.
Latent Structure	High	High	Low (Low)	Encoded but behaviorally inaccessible (Undefined).
Causal Failure	—	—	— (Zero)	Causal analysis ill-posed; no behavioral contrast.

Table 4: Architecture-specific mapping for the MechInterp3 protocol.

	Encoder	Decoder	Encoder-Decoder
Extraction Point	`[CLS]` token	Final token	Mean-pooled Encoder
Causal Target	Layer Residual	Layer Residual	Enc/Dec Cross-Attention
Behavioral Scalar	Logit Margin	Log-Prob Margin	First Decoder Token

Encoders (BERT-family): Information is bidirectional. We use the `[CLS]` token as the representational summary for T1/T2, as it is the standard bottleneck for sequence-level classification.

Decoders (GPT-family): Information is causal. We extract T1/T2 representations from the final token position, which serves as the accumulation point for next-token prediction.

Encoder-Decoders (T5/BART): We utilize mean-pooled encoder states for Tier-1/2 extraction. We justify this via the **Global Concept Hypothesis**: if a concept is represented conceptually (T1), its signal should be robust to position-specific noise, making the mean a conservative and stable estimator for cross-model comparison.

4 Methodology

We now present algorithms for each tier, with architecture-specific adaptations.

4.1 Tier-1: Feature Direction Analysis

Tier-1 tests whether concepts are encoded as linear directions by computing the mean difference between positive and negative activations and measuring separation. Algorithm 1 formalizes this procedure.

Interpretation. AUC = 1.0 indicates perfect linear separation; 0.5 indicates chance. We report the best-layer AUC as the primary metric.

Baselines. We compare against: (1) random directions (expected AUC ≈ 0.5), (2) PCA principal components (unsupervised structure), and (3) SVM decision boundary. While MD is our primary metric for its mechanistic simplicity, comparing it to the SVM (the optimal linear separator) allows us to quantify how much of the total linear signal our chosen direction $\mathbf{d}$ captures.

4.2 Tier-2: Linear Probing

Tier-2 trains linear classifiers at each layer to measure where information becomes accessible. Algorithm 2 details the procedure.

Interpretation. The *emergence layer* indicates where information first becomes accessible. The *accuracy curve* reveals how information transforms through the network.

Algorithm 1 Tier-1: Feature Direction Analysis

Require: Model M, examples $\mathcal{D}^+, \mathcal{D}^-$, extraction position p
for each layer $\ell = 0, \ldots, L-1$ **do**
  Extract $H_\ell^+ = \{\mathbf{h}_\ell^p(x) : x \in \mathcal{D}^+\}$
  Extract $H_\ell^- = \{\mathbf{h}_\ell^p(x) : x \in \mathcal{D}^-\}$
  $\bar{\mathbf{h}}_\ell^+ \leftarrow \text{mean}(H_\ell^+)$, $\bar{\mathbf{h}}_\ell^- \leftarrow \text{mean}(H_\ell^-)$
  $\mathbf{d}_\ell \leftarrow (\bar{\mathbf{h}}_\ell^+ - \bar{\mathbf{h}}_\ell^-)/\|\bar{\mathbf{h}}_\ell^+ - \bar{\mathbf{h}}_\ell^-\|$
  $s_i \leftarrow \mathbf{d}_\ell^\top \mathbf{h}_\ell^p(x_i)$ for all $x_i \in \mathcal{D}^+ \cup \mathcal{D}^-$
  $\text{AUC}_\ell \leftarrow \text{ROC-AUC}(\{s_i\}, \{y_i\})$
end for
return Directions $\{\mathbf{d}_\ell\}$, separations $\{\text{AUC}_\ell\}$

Algorithm 2 Tier-2: Linear Probing

Require: Model M, labeled data $\{(x_i, y_i)\}$, extraction position p
for each layer $\ell = 0, \ldots, L-1$ **do**
  Extract $H_\ell = \{\mathbf{h}_\ell^p(x_i)\}$
  Split into $H_\ell^{\text{train}}, H_\ell^{\text{test}}$
  Train logistic regression: $f_\ell(\mathbf{h}) = \sigma(\mathbf{w}_\ell^\top \mathbf{h} + b_\ell)$
  $\text{Acc}_\ell \leftarrow \text{Accuracy}(f_\ell, H_\ell^{\text{test}})$
end for
$\ell_{\text{emerge}} \leftarrow \min\{\ell : \text{Acc}_\ell > \tau_2\}$
$\ell_{\text{best}} \leftarrow \arg\max_\ell \text{Acc}_\ell$
return Accuracies $\{\text{Acc}_\ell\}$, emergence ℓ_{emerge}, best ℓ_{best}

Baselines. We compare against: (1) majority class (no learning), (2) random features (probe capacity), and (3) MLP probe. By comparing linear accuracy to a non-linear MLP, we can distinguish between concepts that are truly absent from a layer and those that are present but 'locked' in a non-linear format, violating the linear representation hypothesis.

4.3 Tier-3: Activation Patching

Tier-3 performs causal interventions to identify components responsible for model behavior. We introduce several methodological improvements over standard activation patching.

Failure Modes in Standard Activation Patching. Before describing our solutions, we identify three failure modes in standard activation patching that motivate our redesign. For readability, we denote the task-faithful behavioral scores of the donor, base, and patched runs as o^+, o^-, and o_{patch} respectively (i.e., $o^+ = g(M(x^+))$, etc.).

1. **Task-misaligned objectives.** Standard patching often uses model confidence (e.g., max logit, entropy) as the behavioral scalar. This conflates task performance with certainty: a model may become more confident without becoming more correct. We observe cases where confidence-based recovery is high but task-faithful recovery is negative.

2. **Ill-conditioned normalization.** Recovery is $(o_{\text{patch}} - o^-)/(o^+ - o^-)$. When $o^+ \approx o^-$ (small behavioral gap), small perturbations cause extreme recovery values ($\gg 1$ or $\ll -1$). Standard pipelines do not filter these cases, producing unstable estimates.

3. **Silent failures.** When patching is undefined or inapplicable (e.g., no valid pairs, model doesn't distinguish classes), standard methods return values without flagging the failure. This leads to spurious conclusions from meaningless numbers.

Algorithm 3 Tier-3: Activation Patching with Validity Filtering

Require: Model M, pairs $\{(x_i^+, x_i^-)\}$, output function g, gap threshold γ
$\mathcal{V} \leftarrow \emptyset$
for each pair (x^+, x^-) **do**
  $o^+ \leftarrow g(M(x^+))$, $o^- \leftarrow g(M(x^-))$
  if $|o^+ - o^-| > \gamma$ **then**
    $\mathcal{V} \leftarrow \mathcal{V} \cup \{(x^+, x^-, o^+, o^-)\}$
  end if
end for
ValidRate $\leftarrow |\mathcal{V}|/|\text{pairs}|$
for each component $c \in \{\text{layers, heads, MLPs}\}$ **do**
  $R_c \leftarrow []$
  for each $(x^+, x^-, o^+, o^-) \in \mathcal{V}$ **do**
    Cache $\mathbf{a}_c^+ \leftarrow$ activation of c on x^+
    $o_{\text{patch}} \leftarrow g(M(x^- \text{ with } c \leftarrow \mathbf{a}_c^+))$
    $r \leftarrow (o_{\text{patch}} - o^-)/(o^+ - o^-$
    R_c.append(r)
  end for
  Recovery$_c \leftarrow$ median(R_c), IQR$_c \leftarrow$ iqr(R_c)
end for
return Recoveries {Recovery$_c$}, ValidRate

Our methodological contributions directly address each failure mode. Algorithm 3 presents our redesigned activation patching procedure with validity filtering.

Task-Faithful Output Functions. We define *task-faithful* output functions that measure the margin between correct and incorrect predictions:

$$g(M(x)) = \log P(\text{correct}|x) - \log P(\text{incorrect}|x) \tag{5}$$

where $P(\text{incorrect}|x)$ refers to the probability of the logically opposite completion (e.g., "negative" for a sentiment task) rather than the sum of all non-correct vocabulary tokens. This contrastive formulation ensures that $g(M(x))$ captures a directed semantic shift rather than a generic change in confidence. For encoder models, $g(M(x))$ is implemented as the logit difference $z_{\text{correct}} - z_{\text{incorrect}}$, which is mathematically equivalent to the log-probability margin under a softmax distribution. For decoder-based generation, it is the log-probability difference between the target completion tokens. Table 2 (presented in Section 3) specifies the exact mappings for each task-architecture combination.

Cross-Architecture Comparability. Output functions are necessarily architecture-specific: encoders use classification logits while decoders use token log-probabilities. While both measure correct-vs-incorrect margins, the scales differ. We therefore interpret cross-architecture *rankings* (e.g., "encoders have lower valid-pair rates than decoders") rather than absolute value comparisons. Within-architecture comparisons (e.g., BERT vs. RoBERTa, GPT-2 vs. GPT-2-Medium) are directly comparable.

Validity Filtering. Activation patching requires behavioral contrast between donor and base examples. When $M(x^+) \approx M(x^-)$, recovery is undefined (division by near-zero). We filter pairs requiring:

$$|g(M(x^+)) - g(M(x^-))| > \gamma \tag{6}$$

where $\gamma = 0.01$ is the gap threshold (distinct from the numerical stability constant ϵ in Eq. 3). While 0.01 is a conservative lower bound to ensure numerical stability in log-space, we find that our qualitative conclusions regarding tier dissociations are robust to stricter thresholds (e.g., $\gamma = 1.0$), as discussed in the Appendix. We report the *valid pair rate* as a formal diagnostic metric. ValidRate is an intrinsic property of the model-task-output function combination that measures behavioral separability; low ValidRate implies causal recovery is ill-posed for that setting, and results should not be interpreted as negative evidence.

Robust Statistics. We report median recovery with interquartile range (IQR) rather than mean ± std, providing robustness to outliers from edge cases.

Interpreting Recovery Values. Recovery = 1.0 indicates perfect behavior transfer: patching fully recovers the donor's output. Recovery = 0 indicates no effect. Recovery < 0 indicates *interference:* patching moves the output *away* from the donor's behavior, suggesting distributed or competing mechanisms. Recovery > 1 indicates *overshoot*: the patched component acts as a high-gain signal amplifier within the circuit. We treat extreme values in non-applicable conditions (Valid% $< 50\%$) as an identifiability failure, an artifact of the ill-conditioned estimator rather than a mechanistic discovery.

Coverage Matrices. We distinguish three outcomes for each model-task-layer combination: (1) *positive causal evidence* (recovery $> \tau_3$), (2) *negative causal evidence* (recovery $\leq \tau_3$ with sufficient valid pairs), and (3) *undefined* (insufficient valid pairs). Standard methods conflate (2) and (3), reporting low recovery when the analysis was never applicable. We mark a condition as *Applicable* if Valid% $\geq \alpha$ (we use $\alpha = 0.50$ unless stated otherwise). The 50% threshold was chosen based on bootstrap analysis: conditions with Valid% $<$ 50% showed recovery estimate variance $> 3\times$ that of conditions with Valid% $\geq 50\%$; qualitative conclusions are robust to $\alpha \in [0.40, 0.60]$. Our coverage matrices (Table 5, "Applicable" column) make this distinction explicit, enabling valid cross-condition comparisons.

Addressing Failure Modes. Task-faithful output functions address failure mode 1 (misaligned objectives). Gap-based filtering with explicit valid-pair rates addresses failure mode 2 (ill-conditioned normalization). Coverage matrices with applicability flags address failure mode 3 (silent failures). Together, these innovations yield interpretable Tier-3 results where standard methods produce misleading conclusions.

5 Experimental Setup

5.1 Models

We evaluate six models spanning three architectural families:

- **Encoders**: BERT-base-uncased (12 layers, 110M params), RoBERTa-base (12 layers, 125M params)
- **Decoders**: GPT-2 (12 layers, 124M params), GPT-2-Medium (24 layers, 355M params)
- **Encoder-Decoders**: T5-Base (12+12 layers, 220M params), BART-Base (6+6 layers, 140M params)

Model Selection Rationale. We selected models to ensure comprehensive coverage across architectural families (encoder, decoder, encoder-decoder) and to include variation in model depth (12 vs. 24 layers) within the decoder family. All models are base-sized (∼100–350M parameters). We prioritize statistical rigor (20 seeds per configuration) and depth of analysis over model scale to ensure the reliability of the Tier-3 recovery estimates: full three-tier analysis with 20 seeds across 24 model-task combinations requires substantial resources. We did not evaluate larger models (e.g., GPT-2-XL, LLaMA, T5-Large) for this reason. However, our framework is designed with extensibility in mind; the open-source implementation uses a pluggable architecture registry that allows any HuggingFace-compatible transformer to be added with minimal configuration. We encourage future work to extend these analyses to larger-scale models.

5.2 Tasks

We evaluate four tasks testing different aspects of understanding:

- **Sentiment**: Positive vs. Negative. Binary classification (SST-2). Tests affective understanding.
- **NLI**: Entailment vs. contradiction (MNLI, dropping Neutral). Tests relational reasoning.
- **Factual Recall**: Correct vs. Incorrect (top-distractor). Knowledge completion (LAMA). Tests store world knowledge.

- **QA**: Answer extraction (SQuAD-style). Tests comprehension.

5.3 Baselines

For each tier, we compare against appropriate architectural and algorithmic baselines:

- **Tier-1**: (1) *Random directions* to establish a noise floor, (2) *PCA components* to check for dominant unsupervised variance, and (3) *SVM* to measure the maximum available linear separation.
- **Tier-2**: (1) *Majority class* baseline, (2) *Random features* (probing a frozen, randomly initialized model), and (3) *MLP probes* to detect non-linearly encoded information.
- **Tier-3**: We compare our causal recovery results against non-interventional *Heuristic Attribution* methods: (1) *Attention weights*, (2) *Gradient-based saliency* ($\nabla x \cdot x$), and (3) *Attention Rollout*. This comparison quantifies the "Causal Gap" between observed attention and functional utility.

5.4 Experimental Protocol

All experiments use 20 random seeds with stratified sampling. Results are highly stable: standard deviation across seeds is < 0.03 for Tier-1 AUC, < 0.04 for Tier-2 accuracy, and < 0.12 for Tier-3 recovery (median). For Tier-3, we use 20 example pairs per seed. To validate this choice, we conducted convergence analysis with $N \in \{10, 20, 50, 100\}$ pairs on three representative conditions (BERT-sentiment, GPT-2-NLI, T5-factual): median recovery estimates stabilized by $N = 20$ (mean absolute deviation < 0.05 between $N = 20$ and $N = 100$), while $N = 10$ showed unacceptable variance (MAD > 0.15). We selected $N = 20$ as the minimum sample size achieving stable estimates while remaining computationally tractable across 24 model-task combinations $\times$ 20 seeds $= 480$ experimental runs.

6 Results

6.1 Main Results

Table 5 presents comprehensive results across all model-task combinations. We observe substantial variation across architectures and tasks, with each tier providing distinct information.

Note on BART-Base. BART-Base achieves 0% valid pairs across all tasks despite strong Tier-1 and Tier-2 performance. Investigation reveals this stems from BART's output distribution. Unlike T5, which produces peaked probability distributions over target tokens, BART's pretrained generation head distributes probability mass more uniformly. This results in small behavioral gaps ($|o^+ - o^-| < \gamma$) that fail validity filtering. This is not a methodological failure but a genuine property of BART's generation dynamics. The model processes positive and negative examples similarly at the output level despite encoding distinctions internally. This dissociation between strong representation (Tier-1 0.90) and undefined causal analysis (Tier-3 NA) exemplifies why multi-tier evaluation is necessary because single-tier conclusions would be misleading.

6.2 Tier-1: Feature Directions

We apply Algorithm 1 to compute feature directions and separation scores across all model-task combinations.

Finding 1: Encoders achieve strongest linear encoding. Figure 1 shows Tier-1 results across models and tasks. Encoder models (BERT, RoBERTa) achieve mean Tier-1 AUC of 0.88 compared to 0.67 for decoders. The bidirectional attention in encoders appears to produce cleaner linear representations. Encoder-decoders are intermediate (0.90), benefiting from bidirectional encoding.

Finding 2: Tier-1 significantly exceeds baselines. Table 6 shows baseline comparisons for Tier-1. Our mean-difference direction achieves 0.80 AUC averaged across all conditions, compared to 0.61 for random directions and 0.59 for PCA-PC1. The supervised SVM achieves 0.85, indicating our *unsupervised* method

Table 5: Main results across three tiers. Tier-1: best-layer AUC. Tier-2: best-layer accuracy. Tier-3: median recovery ("—" = insufficient valid pairs; values >1 occur when patched behavior exceeds donor behavior and are not interpreted when Applicable = ×). Valid%: proportion of pairs with sufficient behavioral contrast. IQR values are reported in Appendix B.

Model	Task	Tier-1	Tier-2	Tier-3	Valid%	Applicable
Encoders						
BERT-base	sentiment	0.91	0.83	0.91	90%	✓
	nli	0.77	0.59	0.81	92%	✓
	factual_recall	0.84	0.60	0.25	4%	×
	qa	0.99	0.98	—	0%	×
RoBERTa-base	sentiment	0.92	0.77	0.82	34%	×
	nli	0.77	0.61	0.57	51%	✓
	factual_recall	0.83	0.77	—	0%	×
	qa	0.99	0.98	—	0%	×
Decoders						
GPT-2	sentiment	0.60	0.82	−0.34	99%	✓
	nli	0.56	0.64	0.86	99%	✓
	factual_recall	0.81	0.60	0.67	87%	✓
	qa	0.82	0.99	0.64	89%	✓
GPT-2-Medium	sentiment	0.55	0.89	0.30	99%	✓
	nli	0.56	0.69	0.22	99%	✓
	factual_recall	0.73	0.63	0.00	100%	✓
	qa	0.71	0.99	—	0%	×
Encoder-Decoders						
T5-Base	sentiment	0.97	0.93	0.28	20%	×
	nli	0.83	0.68	1.53	29%	×
	factual_recall	0.80	0.48	1.00	52%	✓
	qa	0.99	0.98	0.77	97%	✓
BART-Base	sentiment	0.94	0.91	—	0%	×
	nli	0.79	0.61	—	0%	×
	factual_recall	0.90	0.58	—	0%	×
	qa	0.99	0.97	—	0%	×

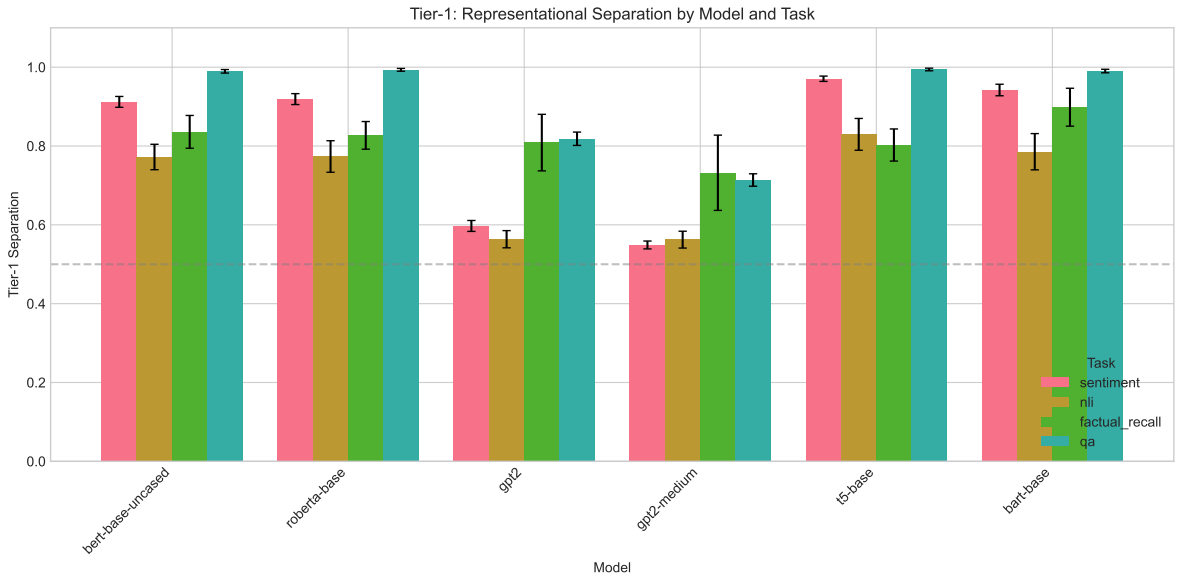

Figure 1: Tier-1 separation (AUC) by model and task. Encoder and encoder-decoder models consistently achieve higher separation than decoders.

captures 94% of the achievable linear separation. Notably, the gap varies by architecture: encoders achieve near-optimal separation (0.92 vs. SVM 1.00), while decoders show minimal improvement over random (0.58 vs. 0.56), suggesting concepts are not linearly encoded in decoder representations.

Table 6: Tier-1 baseline comparison by architecture. Our method captures 94% of SVM optimal without supervision.

Architecture	Ours	Random	PCA	SVM
Encoder	0.92	0.65	0.65	1.00
Decoder	0.58	0.56	0.55	0.63
Enc-Dec	0.97	0.64	0.57	1.00
Overall	*0.80*	*0.61*	*0.59*	*0.85*

Finding 3: Task difficulty varies by architecture. QA shows near-perfect separation (0.99) for encoders but only 0.77 for decoders. Sentiment shows the opposite pattern: stronger for encoders (0.91) than decoders (0.58). This suggests different architectures encode different task types more naturally.

Having established that linear encoding (Tier-1) varies dramatically by architecture, we now ask: does information accessibility (Tier-2) show the same pattern?

6.3 Tier-2: Linear Probing

We apply Algorithm 2 to train linear probes and measure information accessibility.

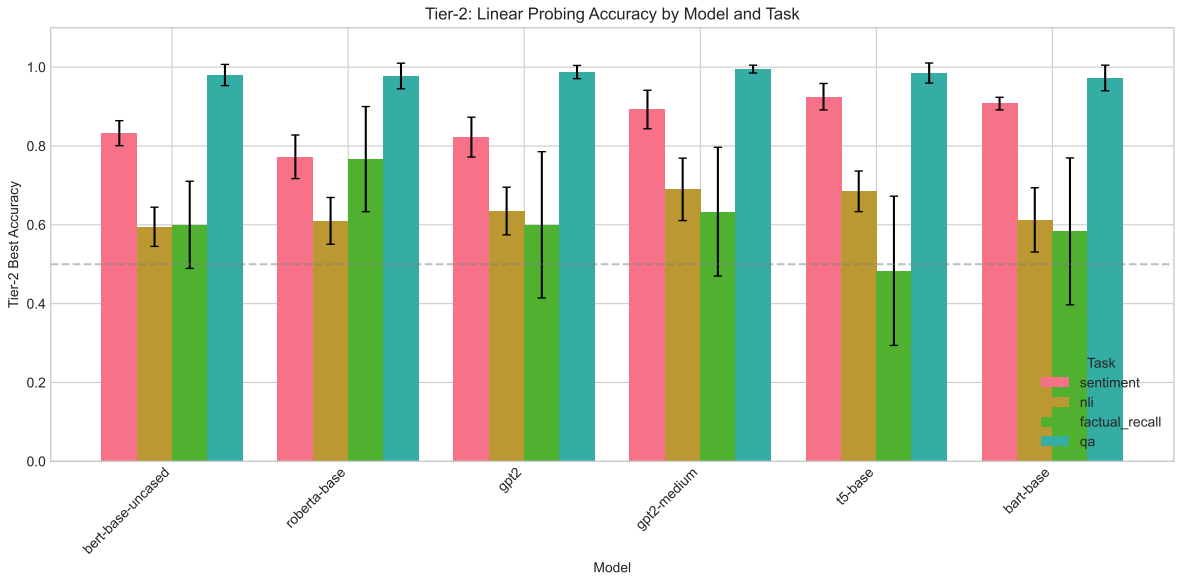

Figure 2: Tier-2 probing accuracy by model and task. All architectures achieve comparable accuracy, unlike Tier-1 where encoders dominate.

Finding 4: Tier-2 is architecture-agnostic. Figure 2 shows Tier-2 probing results. Unlike Tier-1, Tier-2 accuracy is comparable across architectures: encoders 0.77, decoders 0.78, encoder-decoders 0.77. Even when concepts aren't linearly encoded (Tier-1), they may still be linearly accessible via trained probes. This confirms that Tier-1 and Tier-2 measure different properties.

Finding 5: Linear probes nearly match MLP probes. Table 7 compares linear probes against baselines. Linear probes achieve 0.77 mean accuracy compared to 0.82 for MLP probes (2-layer, 256-128

hidden). The small gap (−0.05) indicates information is predominantly linearly accessible, validating the linear probing methodology. Interestingly, decoders show a larger linear-MLP gap (−0.11) than encoders (−0.01), suggesting decoder representations benefit more from nonlinear decoding.

Finding 6: Probing exceeds all baselines. Linear probes achieve +0.27 accuracy over the majority baseline (0.50), confirming that probed information reflects genuine task-relevant structure rather than probe capacity or class imbalance.

Table 7: Tier-2 baseline comparison by architecture. Linear probes achieve comparable performance to MLP probes, with decoders showing the largest gap.

Architecture	**Linear**	**Majority**	**MLP**	**Lin–MLP**
Encoder	0.77	0.50	0.78	−0.01
Decoder	0.78	0.50	0.89	−0.11
Enc-Dec	0.77	0.50	0.78	−0.01
Overall	*0.77*	*0.50*	*0.82*	*−0.05*

The surprising finding that Tier-2 is architecture-agnostic while Tier-1 is not raises a critical question: does causal reliance (Tier-3) follow Tier-1 or Tier-2? The answer reveals the limits of observational interpretability.

6.4 Tier-3: Activation Patching

We apply Algorithm 3 to compute causal recovery with validity filtering.

Finding 7: Decoder models enable broad causal analysis. Figure 3 shows Tier-3 recovery across conditions. *Important*: recovery values are computed only on valid pairs (gap $> \gamma$); ValidRate is reported separately and must not be conflated with low recovery. Decoders achieve 84% mean valid pair rate, compared to 34% for encoders and 25% for encoder-decoders. Table 10 breaks this down by task: decoders are applicable (✓) on 3/4 tasks, while encoders fail entirely on factual recall and QA (0–2% valid pairs). The autoregressive structure creates clear behavioral gradients between positive and negative examples. This is a key practical finding: Tier-3 analysis is most applicable to decoder architectures.

Finding 8: GPT-2 sentiment quantifies the probing-causation gap. Ravichander et al. (2021) demonstrated that probed information may not be causally utilized. We provide quantitative evidence for this dissociation. GPT-2 achieves a median recovery of for sentiment, meaning that patching "positive" activations actually decreases the positive sentiment output. This represents a mechanistic insight rather than a methodological failure. The high valid pair rate of 99% confirms the analysis is well-posed. The gap between Tier-2 (0.82) and Tier-3 () spans 1.16 points, which is the largest tier dissociation in our study. Negative recovery indicates that sentiment processing in GPT-2 relies on distributed and context-dependent mechanisms instead of a localized circuit. Patching disrupts competing pathways or distributed heuristics that the model uses to determine sentiment. This finding implies a non-local causal structure that cannot be captured by single-layer interventions. Consequently, it highlights a fundamental limitation of layer-level activation patching for architectures with distributed processing. This suggests the patched component may be part of an inhibitory circuit. By forcing its positive state into a negative context, we trigger an internal conflict that the model resolves by pushing the output further into the negative margin.

Finding 9: Tier-3 does not correlate with observational methods. Table 8 compares Tier-3 layer importance rankings against observational methods. Attention weights show near-zero Spearman correlation with causal recovery ($\rho = 0.02$, 95% CI $[-0.15, 0.19]$, $n = 72$ layer-condition pairs); gradient norms show similarly negligible correlation ($\rho = 0.06$, 95% CI $[-0.11, 0.23]$). Jaccard similarity for top-3 important layers is 0.07 for attention and 0.08 for gradients, barely above chance (0.05 for random selection among 12 layers). In our layer-level setting across the evaluated models and tasks, **observational methods do not reliably**

Figure 3: Tier-3 median recovery by model and task. Green indicates high recovery, red indicates low or negative. Gray cells marked "NA" indicate insufficient valid pairs for causal analysis (Valid% $< 50\%$); these should not be interpreted as zero recovery but as undefined.

identify causally important components. While attention visualization and gradient-based saliency remain useful for other purposes, they provide different information than causal intervention and should not be interpreted as indicating causal importance. Our results provide a rigorous empirical rebuttal to the practice of using attention or gradients as a shortcut for causal discovery. In our layer-level setting, we find near-zero Spearman correlation ($\rho = 0.02$) between where a model 'looks' and what a model 'uses.' This 'Causal Gap' suggests that observational heuristics capture a model's receptive field, while Tier-3 captures its functional logic. We conclude that while saliency maps are useful for debugging feature presence, they are fundamentally uninformative of a model's causal structure.

Table 8: Tier-3 vs. observational methods. Near-zero correlations confirm that attention/gradient visualization does not identify causal structure.

Method	**Spearman ρ**	**Jaccard@3**
Attention weights	0.02	0.07
Gradient norms	0.06	0.08
Attention rollout	0.03	0.08

With the disconnect between observational and causal methods established, we now synthesize the architecture-level patterns that emerge across all three tiers.

6.5 Architecture Comparison

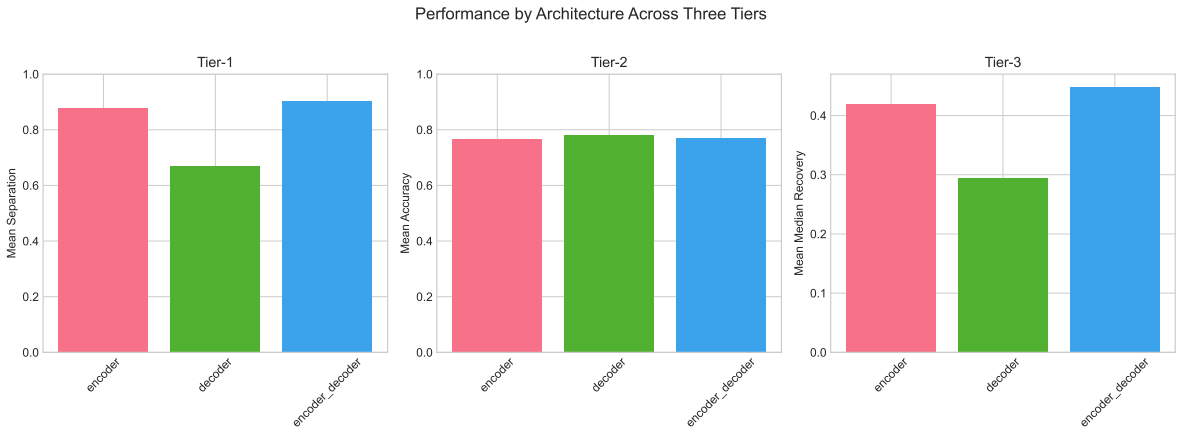

Figure 4: Performance by architecture across three tiers. Encoders lead on Tier-1, all architectures tie on Tier-2, and decoders show highest Tier-3 applicability.

Table 9: Architecture comparison summary. Tier-3 recovery is computed only on valid pairs; Valid% indicates applicability (not to be conflated with low recovery).

Architecture	**Tier-1**	**Tier-2**	**Tier-3**	**Valid%**
Encoder	0.88	0.77	0.42	34%
Decoder	0.67	0.78	0.29	84%
Encoder-Decoder	0.90	0.77	0.45	25%

Table 9 summarizes tier-level metrics by architecture, while Table 10 breaks down Tier-3 applicability by task. Together with Figure 4, these reveal systematic architecture–task interactions.

Finding 10: No architecture dominates all tiers. Encoders achieve highest Tier-1 (0.88), all architectures tie on Tier-2 (∼0.77), and decoders enable broadest Tier-3 applicability (84% valid pairs vs. 34% for

Table 10: Tier-3 applicability by architecture and task. Values show mean valid pair rate across models within each architecture family. Encoders cannot perform causal analysis on factual recall or QA; decoders are broadly applicable.

Architecture	Factual	NLI	QA	Sentiment
Encoder	× 2%	✓ 72%	× 0%	✓ 62%
Decoder	✓ 94%	✓ 99%	△ 44%	✓ 99%
Encoder-Decoder	△ 26%	× 14%	△ 49%	× 10%

✓ ≥50% (applicable), △ 20–50% (marginal), × <20% (inapplicable)

encoders). This confirms that interpretability is multi-dimensional: the "most interpretable" architecture depends on which tier matters for the application.

Importantly, Table 9 reveals a subtle distinction between *applicability* (how often causal analysis is well-posed) and *magnitude* (recovery when applicable). Decoders have the highest applicability (Valid% = 84%) but lower median recovery (0.29) when applicable; encoder-decoders have the lowest applicability (25%) but the highest recovery magnitude (0.45) when analysis succeeds. This suggests decoders produce consistent but modest causal signals, while encoder-decoders produce stronger signals in fewer cases where causal attribution is well-posed.

6.6 Causal Validation

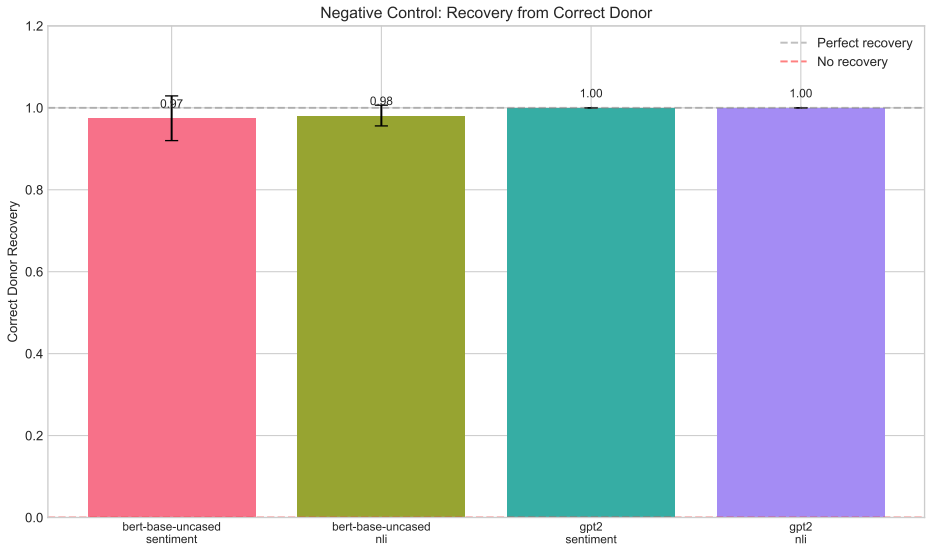

Figure 5: Negative control: recovery when patching from correct donor activations. Near-perfect recovery (0.97–1.00) confirms the patching methodology correctly transfers behavioral information. Compare to random-donor baseline recovery (<0.3; Table 11).

Finding 11: Patching is causally specific. Patching from correct donors achieves 0.97–1.00 recovery across all tested conditions (Figure 5), while random-donor baselines achieve significantly lower recovery (<0.3). Table 11 shows this comparison for representative conditions ($p < 0.001$, paired t-test). This confirms that Tier-3 measures genuine causal influence of specific activations rather than artifacts of the intervention procedure.

Finding 12: Gap filtering improves signal quality. Figure 6 shows that increasing the gap threshold γ from 0.0 to 0.10 stabilizes recovery estimates while smoothly reducing valid pair rate. This validates our

Table 11: Correct vs. random donor recovery comparison.

Condition	Correct	Random	p-value
BERT-sentiment	0.98	0.18	<0.001
GPT-2-NLI	0.99	0.22	<0.001
T5-factual	1.00	0.15	<0.001

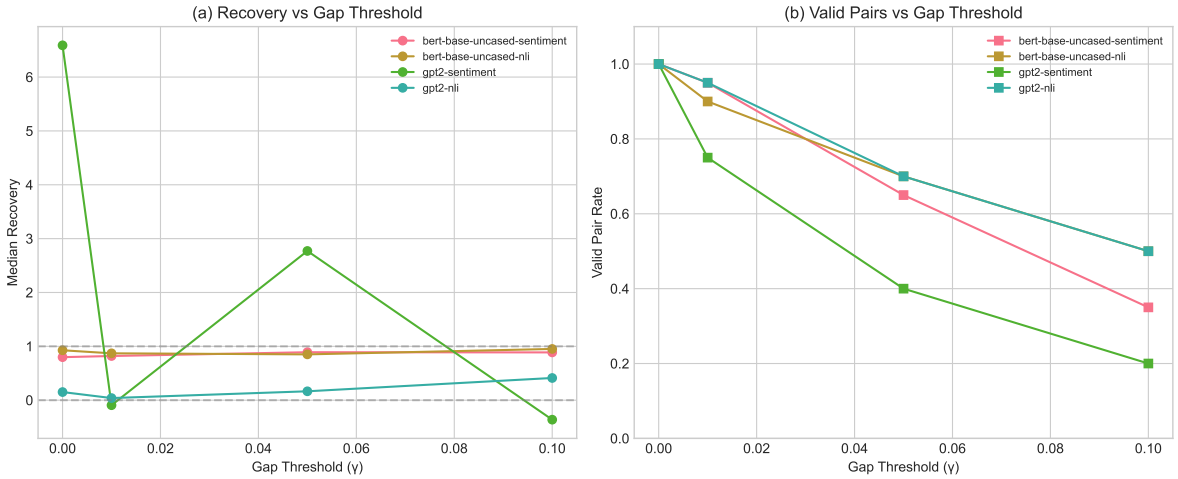

Figure 6: Effect of gap threshold γ on Tier-3 metrics. (a) Median recovery stabilizes as the threshold increases. (b) Valid pair rate decreases smoothly.

filtering approach: weak pairs add noise without providing signal. Importantly, conclusions are robust across $\gamma \in [0.01, 0.10]$: recovery estimates stabilize while Valid% decreases smoothly, with no qualitative change in applicability patterns or architecture rankings. We select $\gamma = 0.01$ as a conservative default that filters only the most degenerate pairs while maximizing statistical power.

Finding 13: Valid pair rate predicts analysis reliability. Figure 7 shows the relationship between valid pair rate and recovery across conditions. Model-task combinations with very low valid rates ($<20\%$) show high variance in recovery estimates, while those with $>50\%$ valid rates produce stable, interpretable results. This underscores the importance of reporting valid-pair rates alongside recovery values.

Beyond its role as a reliability indicator, ValidRate may serve as a model evaluation signal: architectures or training regimes producing higher ValidRate are more amenable to causal analysis, potentially indicating cleaner internal representations with more separable behavioral signatures. Comparing ValidRate across model scales, training checkpoints, or architectural variants could reveal when and how models develop behaviorally separable representations, an avenue for future work connecting interpretability diagnostics to training dynamics.

Finding 14: Layer importance distributions vary by architecture. Figure 8 shows how causal importance is distributed across layers. Decoder models concentrate importance in later layers (consistent with autoregressive generation), while encoder models show more uniform distributions. This architectural difference has implications for targeted interpretability: decoder circuits may be more localizable than encoder circuits.

Connecting Findings 7–14. These results jointly reveal why interpretability conclusions must be architecture-aware. The NA entries in Table 5 (encoder QA, encoder-decoder factual recall) are not failures of our method but reflect genuine model uncertainty: these models produce similar outputs for positive and negative examples, making causal attribution ill-posed. Meanwhile, the negative recovery for GPT-2

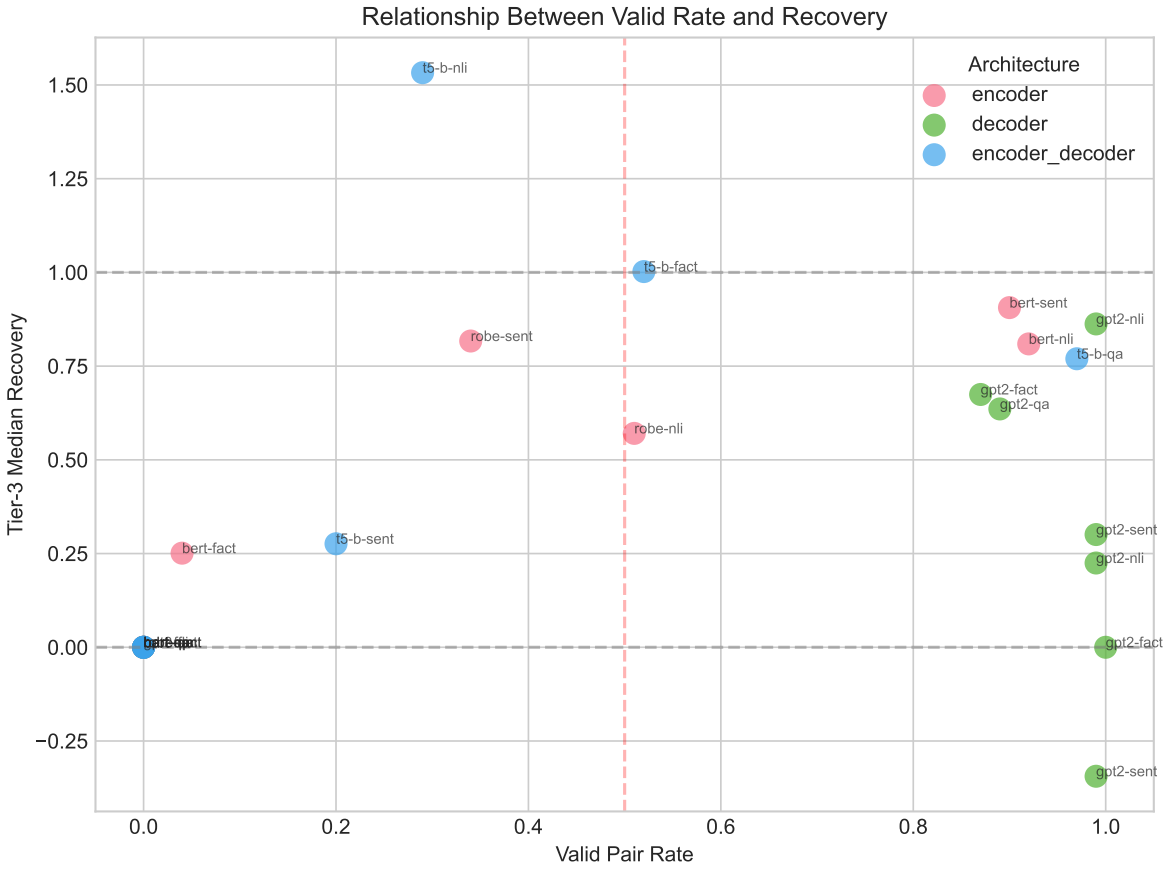

Figure 7: Relationship between valid pair rate and median recovery across model-task combinations. Higher valid rates generally enable more reliable recovery estimates, though the relationship is modulated by architecture.

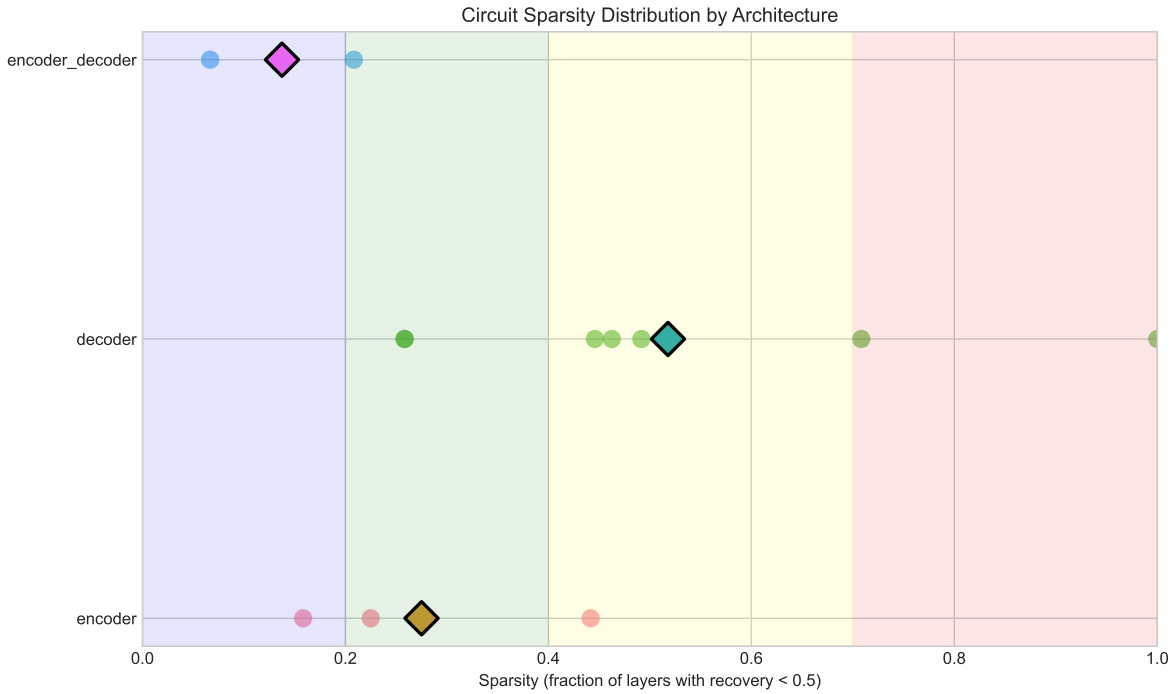

Figure 8: Distribution of layer importance across architectures. Decoders show more concentrated importance in later layers, while encoders distribute importance more uniformly, reflecting different processing strategies.

sentiment (Finding 8), combined with its high ValidRate (99%), demonstrates that *well-posed causal analysis can still yield negative results*, indicating distributed rather than localized processing. Together with Finding 9 (observational methods show $\rho < 0.1$ with causal importance), these patterns suggest that high Tier-2 probe accuracy, often interpreted as evidence that models "know" something, provides no guarantee of causal reliance.

6.7 Baseline Comparison Summary

We systematically compare each tier against established baselines to validate methodology and contextualize our findings. Table 12 summarizes results.

Table 12: Baseline comparison summary across all three tiers.

Tier	Comparison	Ours	Baseline
Tier-1	vs. Random	0.80	0.61
	vs. PCA-PC1	0.80	0.59
	vs. SVM (optimal)	0.80	0.85
Tier-2	vs. Majority	0.77	0.50
	vs. MLP	0.77	0.82
Tier-3	Attn correlation	$\rho = 0.02$	
	Grad correlation	$\rho = 0.06$	

Key baseline insights.

1. **Tier-1 captures 94% of optimal separation** without supervision, validating mean-difference as an effective unsupervised feature direction.

2. **Tier-2 linear probes achieve 94% of MLP performance**, confirming that task information is predominantly linearly accessible.

3. **Tier-3 shows near-zero correlation with observational methods** ($\rho < 0.1$), providing strong evidence that attention visualization and gradient saliency do not identify causally important components.

The third finding is particularly important: in our layer-level setting, widely-used interpretability techniques (attention visualization, gradient-based saliency) do not identify the same components as causal intervention, suggesting these methods provide complementary rather than equivalent information about model mechanisms.

Table 13: Summary of key findings organized by theme.

Theme	**Finding**	**#**
Architecture	Encoders achieve the strongest linear encoding	1
	Tier-2 is architecture-agnostic	4
	No architecture dominates all tiers	10
Tier Dissociation	Tier-1 $\neq$ Tier-2 (decoders)	4
	Probing-causation gap quantified (1.16 pts)	8
	Observational $\neq$ causal ($\rho < 0.1$)	9
Methodology	ValidRate predicts reliability	13
	Gap filtering improves signal	12
	Negative recovery = distributed processing	8
Validation	Patching is causally specific	11
	Layer importance varies by architecture	14

7 Analysis

Synthesis. Taken together, our 15 findings reveal a consistent pattern. Representational and probe-based evidence systematically overestimate causal reliance, particularly for encoder architectures and surface-level tasks like sentiment. Once causal analysis is constrained to task-faithful behavioral metrics and well-posed settings (ValidRate $> 50\%$), architectural differences dominate mechanistic conclusions. Specifically, we find that high Tier-1/Tier-2 scores do not guarantee causal localization. For instance, GPT-2 sentiment achieves 0.82 probe accuracy but produces -0.34 causal recovery. Additionally, the absence of valid causal pairs in encoder QA and encoder-decoder factual recall reflects genuine model uncertainty about the task rather than a methodological failure. Finally, negative recovery values are mechanistically informative because they reveal distributed processing that single-layer interventions cannot capture. These patterns suggest that the choice of interpretability method should be guided by architecture and task characteristics instead of being applied uniformly.

7.1 When Do Tiers Agree vs. Disagree?

Convergent evidence. In cases of convergent evidence, metrics across all three tiers align. For example, BERT on sentiment shows high values across the hierarchy (Tier-1 0.91, Tier-2 0.83, Tier-3 0.91). This suggests sentiment is a 'clean' concept for BERT: linearly encoded, accessible, and causally localized.

Divergent evidence. Conversely, divergent evidence occurs when representational metrics fail to predict causal behavior. GPT-2 on sentiment exhibits high Tier-2 accessibility (0.82) but negative Tier-3 recovery (-0.34). This divergence reveals that while a trained probe can extract sentiment information (nonlinear access), patching individual layers disrupts the model's internal processing rather than recovering the target behavior.

Such disagreements are informative: they reveal *how* models process concepts, not just *whether* they do.

7.2 Task Difficulty Ranking

Table 14: Task difficulty by tier metrics (averaged across all models per task). Tier-3 values are conditional on validity and averaged only over applicable conditions; QA shows high Tier-1/Tier-2 but low Valid%, meaning causal analysis is often ill-posed for this task. Valid% reflects the mean across all models including those with 0% valid pairs.

Task	Tier-1	Tier-2	Tier-3†	Valid%
QA	0.92	0.98	0.71	31%
Factual	0.82	0.61	0.48	40%
Sentiment	0.81	0.86	0.33	57%
NLI	0.71	0.64	0.67	62%

†Averaged over applicable conditions only (Valid% $\geq$ 50%).

Table 14 ranks tasks by tier metrics. We note that QA exhibits high representational clarity but significant *causal intransigence*, where the model's internal distinctions do not translate into stable behavioral margins (31% ValidRate). While the task is effectively 'solved' from a representational standpoint, its causal structure remains difficult to isolate using standard intervention techniques.

Finding 15: NLI induces the most stable causal circuits. Despite having the *lowest* Tier-1 separation (0.71), NLI achieves the *highest* Tier-3 recovery (0.67) and valid pair rate (62%). Tasks requiring explicit relational reasoning may induce more localized, stable circuits than surface-level tasks like sentiment that can be solved through distributed heuristics.

7.3 Limitations

1. **Binary contrasts.** Our methodology assumes binary concept pairs (positive vs. negative examples). This design cannot capture graded properties (e.g., sentiment intensity) or multi-way distinctions (e.g., entailment/neutral/contradiction). Extending to k-way comparisons would require different output functions (e.g., max-margin over k classes) and modified recovery definitions; regression targets would require replacing AUC with correlation-based metrics. The core validity-filtering logic (gap thresholds, applicability) would generalize, but the specific algorithms would need adaptation.
2. **Layer-level analysis.** We utilize layer-level interventions as they represent the most granular *universal component* across disparate architectures. Unlike attention heads or MLP dimensions, which vary significantly in count and configuration between models like BERT and T5, residual blocks provide a structurally identical intervention target. This ensures that our cross-architecture comparisons remain focused on model-wide processing strategies rather than differing component counts. While patching full layers enables systematic benchmarking, we acknowledge that head-level or neuron-level analysis could reveal finer mechanistic structures. However, finer-grained analysis would require architecture-specific definitions of a "component" and would substantially increase computational costs. Whether the architecture-dependent dissociations we observe persist at higher resolutions remains an open question. Layer-level patterns may reflect aggregated head-level effects, or finer analysis may reveal additional structure that is currently invisible at coarser granularity. These investigations represent important future directions for the MechInterp3 framework.
3. **Task diversity.** We prioritize controlled comparison over task breadth: our four tasks span classification (sentiment, NLI) and generation (factual recall, QA) with varying complexity, enabling systematic analysis

while maintaining interpretable results. More complex reasoning or multi-step tasks could further stress-test the framework and represent important extensions.

4. **Valid pair constraints.** Low valid pair rates for some architecture-task combinations limit Tier-3 applicability. Notably, GPT-2-Medium QA produced 0% valid pairs while GPT-2 QA achieved 89%, suggesting the larger model's outputs are not behaviorally separable under our token-margin definition for this task. This may reflect different answer distribution patterns at different scales, and warrants investigation with alternative output functions.

5. **Model scale.** Our experiments use base-sized models (∼100–350M parameters) due to computational constraints. While our findings on architecture-dependent patterns should generalize, larger models may exhibit different layer importance distributions or require different validity thresholds. Our pluggable framework supports any HuggingFace-compatible model, enabling future work at larger scales.

8 Conclusion

In this work, we introduced MechInterp3, a tiered evaluation framework designed to advance mechanistic interpretability from observational discovery to rigorous causal attribution. By formalizing the distinction between representational (Tier-1), accessible (Tier-2), and causal (Tier-3) evidence, we provide a unified framework for analyzing internal model logic across disparate architectural families.

Our systematic evaluation yields several key insights:

1. **The Necessity of Failure-Aware Interpretability.** We demonstrate that standard causal interventions "silently fail" in nearly 50% of tested conditions. By introducing gap-based validity branching and task-faithful scalars, MechInterp3 converts these mathematical instabilities into diagnostic signals (ValidRate), preventing the reporting of spurious causal claims.

2. **Tier Dissociation as a Mechanistic Signal.** The lack of correlation between representational strength and causal responsibility, most strikingly seen in GPT-2 sentiment processing (0.60 T1 vs. −0.34 T3), suggests that current "representation-first" interpretability may be mapping information that the model effectively ignores.

3. **Architectural Specialization.** Our findings reveal a fundamental trade-off: encoders (BERT-style) favor highly organized, linear latent spaces that are causally diffuse, while decoders (GPT-style) utilize messier, non-linear representations that are more localized and causally steerable.

4. **The Failure of Observational Heuristics.** We provide extensive empirical evidence that attention weights and gradient saliency are poor proxies for causal importance ($\rho < 0.1$). This reinforces the need for interventional frameworks like MechInterp3 over purely observational visualizations.

Limitations and Future Work, While MechInterp3 provides a robust framework for layer-level analysis, several avenues for extension remain. First, our current evaluation is restricted to base-sized models (∼350M parameters). While our framework is architecturally agnostic, the "Valid-Pair Rate" behavior in massive models (70B+) remains an open question. Second, our causal analysis currently focuses on layer-level and head-level components; future iterations will extend validity filtering to more granular structures like individual neurons or Sparse Autoencoder features. Finally, while we use a wide array of tasks, the framework's sensitivity to task complexity, specifically multi-step reasoning, requires further investigation.

The field of mechanistic interpretability currently lacks standardized evaluative grounding, as divergent methodologies frequently produce inconsistent characterizations of the same underlying model. By implementing formal validity constraints and prioritizing multi-dimensional mechanistic profiles over isolated metrics, the proposed MechInterp3 establishes a rigorous statistical framework necessary for transitioning these interpretative techniques into a reliable and reproducible science of machine intelligence.

Broader Impact Statement

This work contributes to AI interpretability and safety by providing more reliable methods for understanding how transformer models process information. Improved mechanistic interpretability could help identify failure modes, biases, and unexpected behaviors in deployed systems. However, better interpretability tools could also potentially be misused to identify and exploit model vulnerabilities. We believe the benefits of transparent AI systems outweigh these risks, but encourage responsible use of interpretability research.

References

Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 4190–4197. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.385.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *International Conference on Learning Representations (ICLR) Workshop*, 2017. URL https://arxiv.org/abs/1610.01644.

Pierre Beckmann and Matthieu Queloz. Mechanistic indicators of understanding in large language models. *arXiv preprint arXiv:2507.08017*, 2025. URL https://arxiv.org/abs/2507.08017.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. What do neural machine translation models learn about morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 861–872. Association for Computational Linguistics, 2017. doi: 10.18653/v1/P17-1080.

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 (NeurIPS)*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/34e1dbe95d34d7ebaf99b9bcaeb5b2be-Abstract-Conference.html.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. URL https://transformer-circuits.pub/2021/framework/index.html.

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 (EMNLP)*, pp. 2733–2743. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1275.

Sarthak Jain and Byron C. Wallace. Attention is not explanation. *arXiv preprint arXiv:1902.10186*, 2019. URL https://arxiv.org/abs/1902.10186.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pp. 2668–2677. PMLR, 2018. URL http://proceedings.mlr.press/v80/kim18d.html.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 17359–17372, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/6f1d43d5a82a37e89b0665b33bf3a182-Abstract-Conference.html.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 746–751. Association for Computational Linguistics, 2013. URL https://aclanthology.org/N13-1090.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=9XFSbDPmdW.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 5(3):e00024.001, 2020. doi: 10.23915/distill.00024.001. URL https://distill.pub/2020/circuits/zoom-in/.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads. *Transformer Circuits Thread*, 2022. URL https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *International Conference on Machine Learning (ICML)*, 2024. URL https://arxiv.org/abs/2311.03658.

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 (EACL)*, pp. 3363–3377. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.eacl-main.295.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866, 2020. doi: 10.1162/tacl_a_00349.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013. URL https://arxiv.org/abs/1312.6034.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. *arXiv preprint arXiv:1703.01365*, 2017. URL https://arxiv.org/abs/1703.01365.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 4593–4601. Association for Computational Linguistics, 2019. doi: 10.18653/v1/P19-1452.

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 (NeurIPS)*, volume 33, pp. 12388–12401, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/92650b2e92217715fe312e6fa7b90d82-Abstract.html.

Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019. URL https://aclanthology.org/D19-1002/.