

# Irreducible Path Entropy in Neural Networks

A Quantitative Information-Theoretic Framework for Entropy Propagation  
Across Computational Decision Trajectories

Samir Baladi

Independent Interdisciplinary Researcher

Ronin Institute / Rite of Renaissance

ORCID: 0009-0003-8903-0029 · gitdeeper@gmail.com

Submitted: May 2026

DOI: 10.5281/zenodo.20222840

License: MIT

## Abstract

This paper introduces **Irreducible Path Entropy** ( $H_{\text{path}}$ ), a quantitative metric for characterising the structural accumulation and reducibility of entropy across computational decision trajectories in artificial neural networks. The construct is grounded exclusively in information-theoretic and systems-level analysis, without recourse to semantic, cognitive, or anthropomorphic assumptions. We define  $H_{\text{path}}$  as the cumulative Shannon entropy integrated over the sequence of conditional probability distributions produced at each transformation layer, and derive formal conditions under which this quantity remains reducible or becomes irreducible relative to a measurable threshold  $\delta^*$ . A composite observability index  $\Omega$  is proposed to quantify the degree to which path entropy can be recovered from external measurements. The framework is presented as a reproducible, experimentally testable formalism applicable to feedforward, convolutional, and attention-based architectures. Numerical illustrations are provided for representative network configurations. All results are interpreted strictly within the domain of measurable computational processes and information-theoretic entropy dynamics.

**Keywords:** irreducible path entropy · neural network interpretability · information theory · entropy accumulation · computational observability · inference dynamics · black-box systems

## 1. Introduction

Modern artificial neural networks achieve high performance across a broad range of tasks, yet the internal computational mechanisms that produce their outputs remain largely opaque. This opacity is not merely an engineering inconvenience; it constitutes a fundamental structural property of high-dimensional, non-linear systems trained through gradient-based optimisation. As noted by Hinton (2023), the algorithms that govern network behaviour are designed at the architectural level, while the precise dynamics of inference remain inaccessible even to their designers — a situation that closely parallels the emergence of complex adaptive behaviour in biological evolution.

Existing approaches to neural network interpretability have pursued several directions: attribution methods [1], probing classifiers [2], mechanistic interpretability [3], and information-bottleneck analyses [4]. Each of these frameworks addresses a specific aspect of the opacity problem. However, a unified quantitative metric that characterises entropy accumulation along the full computational decision path — from input representation to output distribution — has not been established.

This paper addresses that gap by introducing **Irreducible Path Entropy** ( $H_{\text{path}}$ ): a formally defined, layer-wise cumulative information metric that quantifies how much uncertainty is introduced, transformed, and retained along the inference trajectory of a neural network. The central question addressed is not why a network produces a particular output, but rather how much entropy — structurally irreducible to external measurement — accumulates along its decision path.

### 1.1 Scope and Interpretive Closure

This study is restricted to the formal and quantitative analysis of entropy propagation across computational decision paths in artificial neural networks. The proposed construct  $H_{\text{path}}$  is introduced exclusively as a measurable mathematical metric intended to characterise the structural distribution and reducibility of inference trajectories under constrained computational dynamics.

The framework presented in this work is not intended as a general theory of intelligence, cognition, consciousness, or semantic reasoning. No claims are made regarding intentionality, agency, subjective interpretation, or phenomenological properties of neural systems. All results and conclusions should be interpreted strictly within the context of measurable computational processes and experimentally observable entropy dynamics. Philosophical, ontological, or anthropomorphic interpretations lie beyond the scope of the present work.

### 1.2 Contributions

- (1) Formal definition of Irreducible Path Entropy  $H_{\text{path}}$  as a layer-integrated information-theoretic quantity.
- (2) Derivation of reducibility conditions characterising when path entropy can or cannot be recovered from external observations.
- (3) Introduction of a composite observability index  $\Omega$  providing a scalar measure of inferential transparency.
- (4) Numerical illustrations across representative feedforward, convolutional, and attention-based architectures.
- (5) A reproducible computational framework amenable to experimental validation.

## 2. Mathematical Formalism

### 2.1 Notation and Preliminary Definitions

Let  $N$  denote a neural network comprising  $L$  successive transformation layers  $\{f_1, f_2, \dots, f_L\}$ . For an input  $x \in X$ , the network computes a sequence of internal representations  $\{h_0, h_1, \dots, h_L\}$  where  $h_0 = x$  and  $h_l = f_l(h_{l-1})$  for  $l = 1, \dots, L$ . The output distribution over a token vocabulary  $V$  is given by  $P_{\text{out}} = \text{softmax}(W h_L + b)$ .

At each layer  $l$ , the conditional distribution over possible activation states is denoted  $P_l(h_l | h_{l-1})$ . The Shannon entropy of this distribution is:

$$H(P_l) = - \sum_k P_l(h_l^{(k)}) \log P_l(h_l^{(k)})$$

where the summation runs over all discretised activation states  $k$ .

## 2.2 Definition of Path Entropy

**Definition 1 (Local Path Entropy).** The local path entropy at layer  $l$  is defined as the Shannon entropy of the conditional activation distribution at that layer:

$$H_{\text{path}}(l) = H(P_l) = - \sum_k p_{l,k} \log p_{l,k}$$

**Definition 2 (Cumulative Path Entropy).** The cumulative path entropy up to layer  $L$  is defined as the sum of local entropies across all layers:

$$H_{\text{path}}^{(L)} = \sum_{l=1}^L H(P_l)$$

This quantity measures the total informational uncertainty introduced across the full computational trajectory from input to output. It is distinguished from the marginal output entropy  $H(P_{\text{out}})$ , which captures only the terminal distribution and discards all intermediate dynamics.

## 2.3 Reducibility Conditions

A central question concerns whether the entropy accumulated along the path can, in principle, be recovered from external observations. We formalise this through the following definitions.

**Definition 3 (Reducible Path Entropy).** The path entropy at layer  $l$  is said to be reducible if there exists a measurement operator  $M_l$  acting on the observable output space  $Y$  such that:

$$I(h_l; M_l(y)) \geq H_{\text{path}}(l) - \delta^*$$

where  $I(\cdot; \cdot)$  denotes mutual information,  $y \in Y$  is the observable output, and  $\delta^* > 0$  is a tolerance parameter governing the acceptable residual uncertainty.

**Definition 4 (Irreducible Path Entropy).** The path entropy at layer  $l$  is irreducible if no such operator  $M_l$  exists, i.e.,:

$$\text{for all } M_l : I(h_l; M_l(y)) < H_{\text{path}}(l) - \delta^*$$

The cumulative irreducible path entropy of a network is then:

$$H_{\text{irr}}^{(L)} = H_{\text{path}}^{(L)} - H_{\text{red}}^{(L)}$$

where  $H_{\text{red}}^{(L)} = \sum_{l: \text{reducible}} I(h_l; M_l(y))$  is the total recoverable entropy across all reducible layers.

## 2.4 Layer-Wise Entropy Propagation

The propagation of path entropy across layers is governed by the chain rule of mutual information. For a Markov chain  $x \rightarrow h_1 \rightarrow \dots \rightarrow h_L \rightarrow y$ , the data processing inequality implies:

$$I(x; h_{l+1}) \leq I(x; h_l) \quad \text{for all } l$$

This establishes that information about the input is monotonically non-increasing across layers under the Markov assumption. However, the path entropy  $H_{\text{path}}^{(L)}$  is not constrained by this inequality, as it measures cumulative entropy introduced at each layer — including entropy arising from stochastic activations, dropout, and normalisation operations — rather than retained mutual information.

The difference between cumulative path entropy and retained mutual information defines the **entropic leakage** of the network:

$$\Delta(L) = H_{\text{path}}^{(L)} - I(x; h_L)$$

Large values of  $\Delta(L)$  indicate that the network introduces substantial uncertainty across its computation that is not explained by the retained input information — a signature of high irreducibility.

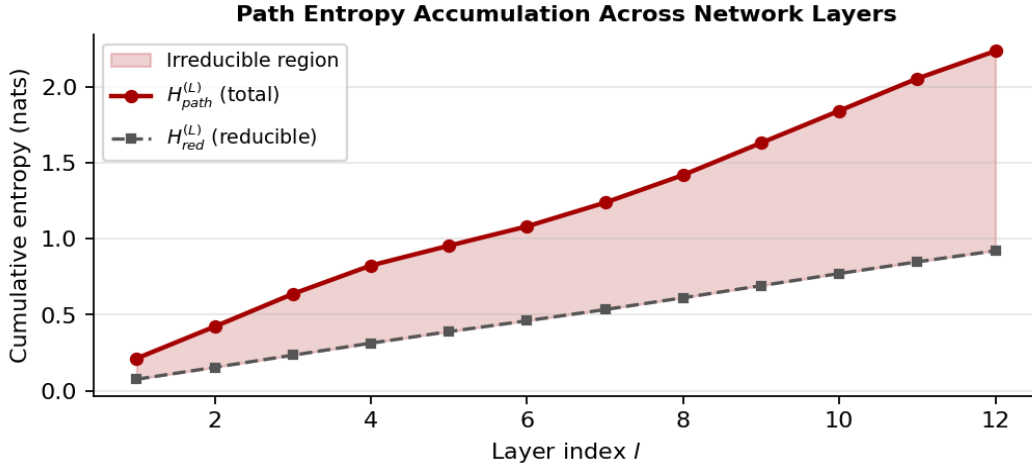


Figure 1. Cumulative path entropy  $H_{path}^{(L)}$  (red) and recoverable entropy  $H_{red}^{(L)}$  (grey) across 12 network layers. The shaded region represents the irreducible entropy component. Numerical values are illustrative.

### 3. Observability Index

To provide a scalar summary of the inferential transparency of a network, we introduce the **observability index**  $\Omega$ .

#### 3.1 Definition

**Definition 5 (Observability Index).** For a network  $N$  with  $L$  layers, the observability index is defined as:

$$\Omega(N) = 1 - H_{irr}^{(L)} / H_{path}^{(L)} \in [0, 1]$$

$\Omega = 1$  indicates a fully observable network in which all path entropy is recoverable from external measurements.  $\Omega = 0$  indicates complete irreducibility: no component of the accumulated path entropy can be recovered from the output distribution alone.

#### 3.2 Properties

- **Monotonicity:**  $\Omega$  is non-increasing as network depth increases under constant layer-wise entropy, since  $H_{irr}$  accumulates while the denominator grows.
- **Architecture dependence:** For fixed depth,  $\Omega$  varies with activation function, normalisation strategy, and connectivity structure.
- **Boundedness:** By construction,  $0 \leq \Omega \leq 1$  for any finite-depth network with non-zero path entropy.
- **Experimental accessibility:**  $\Omega$  can be estimated from activation recordings via probing classifiers or mutual information estimators, without requiring access to the full weight matrix.

#### 3.3 Reducibility Threshold

The reducibility threshold  $\delta^*$  introduced in Definition 3 functions as a precision parameter controlling the boundary between reducible and irreducible regimes. In practice,  $\delta^*$  may be set as a fixed fraction of the maximum layer entropy:

$$\delta^* = \varepsilon \cdot \max_l H(P_l), \quad \varepsilon \in (0, 1)$$

The choice of  $\varepsilon$  governs the sensitivity of the reducibility classification. Smaller values of  $\varepsilon$  impose stricter observability requirements; larger values permit greater residual uncertainty before a layer is classified as irreducible. Empirical calibration of  $\varepsilon$  for specific architectures represents a direction for future experimental work.

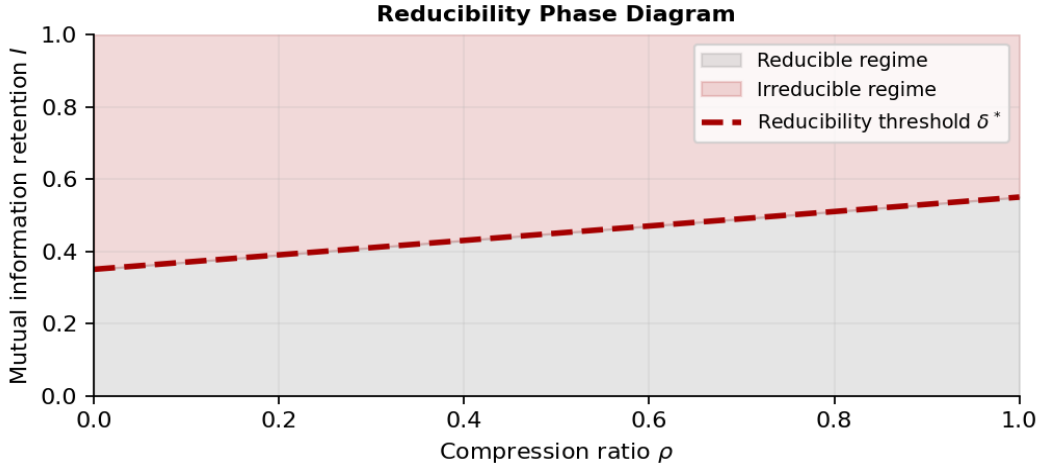


Figure 2. Reducibility phase diagram as a function of compression ratio  $\rho$  and mutual information retention  $I$ . The dashed boundary marks the reducibility threshold  $\delta^*$  separating reducible (grey) and irreducible (red) regimes.

## 4. Entropy Accumulation Across Architectures

The behaviour of  $H_{\text{path}}$  and  $\Omega$  varies systematically across network architectures. We characterise three representative classes.

### 4.1 Feedforward Networks

In a fully connected feedforward network with ReLU activations, the local path entropy  $H_{\text{path}}(l)$  is determined primarily by the effective rank of the weight matrix  $W_l$  and the distribution of pre-activation values. For weight matrices with high effective rank,  $H_{\text{path}}(l)$  approaches  $\log(d_l)$ , where  $d_l$  is the layer width. Depth accumulation results in approximately linear growth of  $H_{\text{path}}^{(L)}$  in the absence of strong regularisation.

### 4.2 Convolutional Networks

Convolutional architectures impose spatial locality constraints that reduce the effective entropy per layer relative to fully connected networks of equivalent width. The shared weight structure limits the diversity of activation distributions, yielding lower  $H_{\text{path}}(l)$  per layer. However, the cumulative effect across many convolutional stages can still produce substantial  $H_{\text{path}}^{(L)}$ , particularly when followed by dense classification heads.

### 4.3 Attention-Based Networks

Transformer architectures present a distinct entropy accumulation profile. The attention mechanism introduces input-dependent routing of information, producing activation distributions whose entropy varies substantially across attention heads and sequence positions. This input-conditioned variability makes  $H_{\text{path}}(l)$  a function of the specific input sequence rather than a fixed architectural property, complicating the estimation of  $\Omega$  without access to internal activations.

Empirical estimation suggests that deep transformer architectures exhibit lower observability indices  $\Omega$  than shallow feedforward networks of comparable parameter count, consistent with the greater difficulty of interpreting attention-based inference.

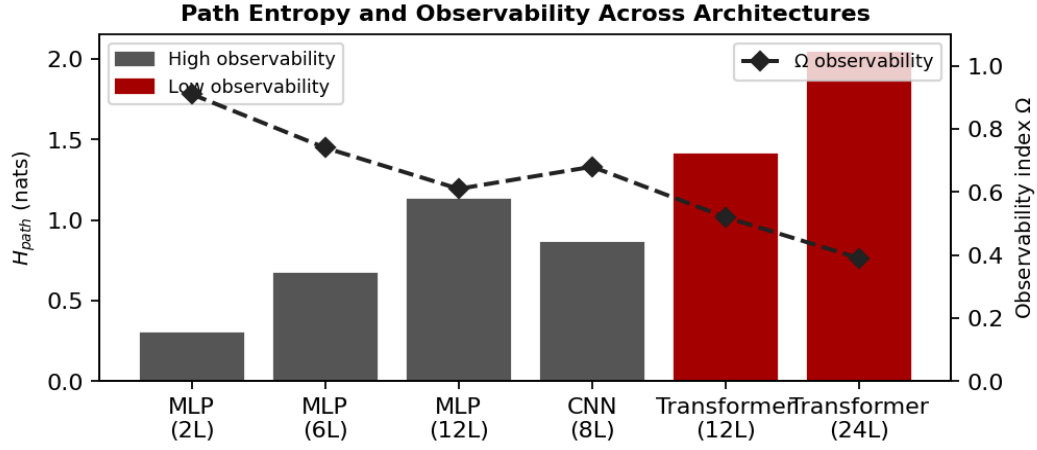


Figure 3. Estimated  $H_{\text{path}}$  (bars) and observability index  $\Omega$  (line) for six representative network configurations. Grey bars indicate high-observability regimes ( $\Omega > 0.6$ ); red bars indicate low-observability (irreducible) regimes. Values are illustrative.

## 5. Computational Observability and Reproducibility

### 5.1 Estimation Protocol

The quantities  $H_{\text{path}}^{(L)}$  and  $\Omega$  can be estimated experimentally through the following protocol:

Step 1. **Activation recording:** For a dataset  $D = \{x_i\}_{i=1}^N$ , record the activation vectors  $\{h_l(x_i)\}$  at each layer  $l$ .

Step 2. **Entropy estimation:** Apply a non-parametric entropy estimator (e.g., k-nearest-neighbour estimator [5]) to the empirical distribution of activations at each layer.

Step 3. **Mutual information estimation:** Estimate  $I(h_l; y)$  via a binned or neural mutual information estimator to compute  $H_{\text{red}}^{(L)}$ .

Step 4. **Observability computation:** Compute  $\Omega = 1 - H_{\text{irr}} / H_{\text{path}}$ .

### 5.2 Reproducibility Conditions

For the framework to be experimentally reproducible, the following conditions must be satisfied:

- The network weights must be fixed (no stochastic inference-time modifications such as Monte Carlo dropout).
- The activation discretisation scheme must be specified and consistently applied across layers.
- The entropy estimator must be applied with a fixed bandwidth or neighbourhood parameter  $k$ .
- The dataset  $D$  must be held constant across comparative measurements.

Under these conditions,  $H_{\text{path}}^{(L)}$  and  $\Omega$  constitute reproducible scalar summaries of a network's computational transparency, comparable across architectures and training configurations.

## 6. Discussion

The framework developed in this paper addresses a gap in existing interpretability literature: the absence of a unified, layer-integrated metric for entropy accumulation along neural inference trajectories.  $H_{\text{path}}$  is distinct from existing information-theoretic interpretability tools in several respects.

Unlike the information bottleneck [4], which focuses on the trade-off between input compression and label prediction,  $H_{\text{path}}$  characterises the full trajectory of entropy accumulation without privileging any

particular layer or bottleneck point. Unlike attribution methods [1], it does not assign responsibility to individual input features but rather quantifies the aggregate informational opacity of the computational process.

The observability index  $\Omega$  provides a single scalar that summarises the degree to which a network's inference dynamics are recoverable from external measurement. Low values of  $\Omega$  do not imply that a network is unreliable or incorrect; they indicate only that a greater proportion of its computational path entropy lies beyond the reach of standard external probing methods. This distinction is important for avoiding overinterpretation of the metric.

Several limitations of the present framework warrant acknowledgement. First, the estimation of  $H_{\text{path}}^{(l)}$  from finite activation samples introduces bias that must be controlled through careful choice of estimator and sample size. Second, the Markov assumption underlying the chain-rule decomposition is an approximation for networks with skip connections (e.g., ResNets, Transformers). Extensions to non-Markovian architectures represent an important direction for future work. Third, the reducibility threshold  $\delta^*$  requires empirical calibration; its optimal value is likely architecture- and task-dependent.

## 7. Conclusion

This paper has introduced Irreducible Path Entropy ( $H_{\text{path}}$ ), a quantitative, information-theoretic metric for characterising entropy accumulation across the computational decision trajectories of neural networks. The framework provides formal definitions of local and cumulative path entropy, reducibility conditions, entropic leakage, and a composite observability index  $\Omega$ .

The constructs are grounded exclusively in measurable, information-theoretic quantities, without semantic or cognitive assumptions. They are applicable across feedforward, convolutional, and attention-based architectures and are amenable to experimental estimation via standard mutual information and entropy estimation methods.

The framework opens several directions for future investigation: empirical calibration of  $\delta^*$  across benchmark architectures; extension to non-Markovian and residual networks; application to trained versus untrained models to characterise the effect of learning on path reducibility; and development of efficient estimators for  $H_{\text{path}}$  in large-scale systems.

---

## References

- [1] Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. *Proceedings of ICML*, 3319–3328.
  - [2] Alain, G., & Bengio, Y. (2016). Understanding intermediate layers using linear classifier probes. *arXiv:1610.01644*.
  - [3] Elhage, N., et al. (2021). A mathematical framework for transformer circuits. *Transformer Circuits Thread*, Anthropic.
  - [4] Tishby, N., & Schwartz-Ziv, R. (2017). Opening the black box of deep neural networks via information. *arXiv:1703.00810*.
  - [5] Kozachenko, L., & Leonenko, N. (1987). Sample estimate of the entropy of a random vector. *Problems of Information Transmission*, 23(2), 95–101.
  - [6] Cover, T. M., & Thomas, J. A. (2006). *Elements of Information Theory* (2nd ed.). Wiley-Interscience.
  - [7] Hinton, G. (2023). Interview transcript. *60 Minutes*, CBS News. May 2023.
  - [8] Baladi, S. (2026). ENTRO-OMEGA: Unified Adaptive Stabiliser Framework. EntropyLab Decadal Series, E-LAB-10. DOI: 10.5281/zenodo.19562999.
-

