Post-Hoc Understanding of Metaphor Processing in Decoder-Only Language Models via Conditional Scale Entropy*

Lawhori Chakrabarti[1], Jennifer Johnson-Leung[2], Bert Baumgaertner[3], Aleksandar Vakanski[4], Min Xian[5], and Boyu Zhang[1]

[1] Department of Computer Science, University of Idaho, Moscow ID 83844, USA
[2] Department of Mathematics and Statistical Science, University of Idaho, Moscow ID 83844, USA
[3] Department of Politics and Philosophy, University of Idaho, Moscow ID 83844, USA
[4] Department of Nuclear Engineering and Industrial Management, University of Idaho, Idaho Falls ID 83402, USA
[5] Department of Computer Science, University of Idaho, Idaho Falls ID 83402, USA

Abstract. Metaphor requires a language model to resolve a token whose contextual meaning diverges from its basic literal sense. Understanding how transformer models organize this reinterpretation across depth remains an open problem in mechanistic interpretability. We introduce conditional scale entropy (CSE), a wavelet-derived measure of how broadly transformer computation engages across frequency scales at each layer position. Two theorems establish that CSE is invariant to update magnitude, isolating the structural pattern of updates from their intensity. Using CSE, we find that metaphorical tokens produce significantly higher spectral breadth than literal tokens at contiguous layer positions on every decoder-only architecture tested, from 124M to 20B parameters (GPT-2 family, LLaMA-2 7B, GPT-oss 20B). The effect survives cluster-based permutation correction, recurs in the early-to-mid relative depth range across models, and converges with an independent analysis of 200 naturalistic VUA pairs. Specificity controls further show that the effect is not explained by semantic complexity or by matched propositional content. These results identify multi-scale coordination as a consistent signature of metaphorical language processing in the decoder-only architectures examined, and establish CSE as a principled tool for characterizing cross-depth structure in transformers.

Keywords: Mechanistic interpretability · Wavelet analysis · Metaphor processing · Conditional scale entropy · Transformer.

* Research reported in this publication was supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number P20GM104420.

1 Introduction

Transformer language models process metaphorical language with notable proficiency [26], yet the internal mechanisms underlying this ability remain poorly understood. When a model encounters "The lawyer is a shark," the target token must be reinterpreted from its literal referent to a figurative predicate. We observe the input and the output; the intermediate computation across the model's layers constitutes a black box. This paper asks a specific question: does metaphorical processing organize the model's internal computation differently from literal processing, and if so, how can this difference be measured?

The residual trajectory [5] of a target token is a discrete sequence of hidden states indexed by layer depth. We construct a continuous interpolant from this sequence and apply the Continuous Wavelet Transform [11] (CWT) to obtain a scalogram that resolves energy at each layer position across frequency scales. From this scalogram we derive the conditional scale entropy (CSE), a quantity that measures how broadly computation engages across scales at each depth position. Two theorems establish that CSE is invariant to update magnitude and responds only to the structural pattern of updates across layers, separating how computation is organized from how intense it is.

Experiments across five architectures (124M to 20B parameters) and two stimulus sets (25 controlled minimal pairs, 200 naturalistic VUA pairs [19]) show that metaphorical tokens produce significantly higher CSE at contiguous layer positions on every model tested, surviving cluster-based permutation correction [13]. This invariance is important empirically: among the spectral quantities examined, CSE is the only one that remains stable across architectures and stimulus regimes. Specificity controls confirm that the elevation is absent for literal sentences differing in semantic or syntactic complexity, and a matched-meaning triples experiment further shows that the effect tracks figurative mode of expression rather than propositional content alone. The framework is mathematically general; metaphor serves as the proving ground.

This paper makes three contributions: a wavelet-based framework for analysing residual-stream trajectories across transformer depth with the conditional scale entropy (CSE) as a layer-resolved measure of spectral breadth; a theoretical characterisation showing that CSE is invariant to update magnitude and ordered by majorisation of the normalised scale distribution; and empirical validation across five decoder-only architectures, naturalistic VUA data, and targeted specificity controls, where CSE proves the most stable spectral signature of metaphorical processing.

Proofs of the theoretical results, the full controlled stimulus list, and additional supporting analyses are deferred to an extended version of this paper.

2 Related Work

Metaphor identification in NLP. Computational metaphor research has predominantly addressed identification and interpretation rather than the internal

processing dynamics of neural language models. Early systems framed the task as contextual sequence labelling [7] or grounded it in selectional-preference violation [12]. Pretrained transformers substantially advanced the state of the art (for instance, MelBERT [4] pairs contextualised late interaction with the Metaphor Identification Procedure [19]), and more recent work asks whether generative LLMs can recover conceptual metaphor mappings under prompting [26]. From cognitive linguistics, the Graded Salience Hypothesis [10] and the Career of Metaphor theory [3] predict that novel metaphors demand structural alignment between source and target domains, whereas conventional metaphors are resolved by categorisation. These predictions motivate a computational measure of processing structure that operates inside the model rather than at the input–output level.

Transformer representations across depth. A complementary line of work asks how linguistic structure is organised within transformer models: BERT recapitulates the classical NLP pipeline across depth [21]; contextualised representations grow increasingly context-specific with depth [6]; metaphorical knowledge is recoverable from pretrained representations with the strongest probing signal at middle layers [2]; and probe accuracy alone is insufficient for representational claims [25]. Mechanistic accounts shift the question from what a layer encodes to how it transforms the residual stream [5, 9], and Geshkovski et al. [8] treat depth as continuous time, motivating the analysis of the hidden-state sequence as a discrete signal indexed by depth.

Spectral analysis of transformers. Frequency-domain methods have illuminated deep-network behaviour: spectral bias in training [16], information-bottleneck dynamics across layers [17], wavelet ablations of vision transformer attention maps [1], and attention entropy collapse as a deep-transformer failure mode [27]. However, wavelet decomposition has not previously been applied post hoc to hidden-state trajectories of language models. The present work fills this gap by applying the CWT to projected residual-stream trajectories and deriving the conditional scale entropy.

3 Methodology

This section presents our framework for analysing the spectral structure of transformer computation. The primary tool is the Continuous Wavelet Transform (CWT) applied to hidden-state trajectories across layers; the application to metaphor processing is described in Section 4.

3.1 Problem Setting and Notation

Consider a transformer language model with L layers and embedding dimension d. For a given input sentence, each token τ is associated with a sequence of hidden-state vectors $x_0^\tau, x_1^\tau, \ldots, x_L^\tau \in \mathbb{R}^d$, where x_0^τ is the initial embedding

and x_L^τ is the final-layer representation. The residual connection [24] yields the recurrence

$$x_{l+1}^\tau = x_l^\tau + \Delta_l(x_l^\tau), \tag{1}$$

where $\Delta_l = \mathrm{Attn}_l + \mathrm{MLP}_l$ is the combined attention and feed-forward update at layer l. Following Elhage et al. [5], we refer to the sequence $(x_0^\tau, \ldots, x_L^\tau)$ as the *residual trajectory* of token τ.

Our analysis is built on a *minimal-pair* design. Given a target lexeme w (e.g., *crushed*), we embed it in two sentences that differ only in whether w is used literally (*The weight crushed the box.*) or metaphorically (*The news crushed her hopes.*). The object of study is the pair of residual trajectories $(x_0^{\mathrm{lit}}, \ldots, x_L^{\mathrm{lit}})$ and $(x_0^{\mathrm{met}}, \ldots, x_L^{\mathrm{met}})$ for the shared target token. We ask whether the spectral structure of their respective layer-to-layer updates differs in a systematic and measurable way.

3.2 Projected Trajectory and the Contrast Direction

The residual updates $\Delta_l^\tau \in \mathbb{R}^d$ are high-dimensional vectors. Comparing them directly across conditions requires reducing them to a representation that preserves the metaphor–literal distinction while enabling frequency analysis. We achieve this by projecting onto a single direction in $\mathbb{R}^d$.

Given K minimal pairs, each contributing literal updates $\{\Delta_l^{\mathrm{lit},k}\}_{l=0}^{L-1}$ and metaphorical updates $\{\Delta_l^{\mathrm{met},k}\}_{l=0}^{L-1}$, we define the *normalized mean-difference direction*

$$v^* = \frac{\sum_{k=1}^{K}\sum_{l=0}^{L-1}\left(\Delta_l^{\mathrm{met},k} - \Delta_l^{\mathrm{lit},k}\right)}{\left\|\sum_{k=1}^{K}\sum_{l=0}^{L-1}\left(\Delta_l^{\mathrm{met},k} - \Delta_l^{\mathrm{lit},k}\right)\right\|_2}. \tag{2}$$

This is the unit vector along the aggregate update difference between metaphorical and literal processing, summed across all pairs and layers. The *projected trajectory* is $s(l) = v^{*\top} x_l^\tau$ and the *projected updates* are $\delta_l = v^{*\top}\Delta_l$, collected into the vector $\boldsymbol{\delta} \in \mathbb{R}^L$, yielding a scalar signal indexed by layer depth that can be analysed with standard signal processing tools. A leave-one-out test confirms that v^* captures a stable structural property: in every experiment, holding out a pair and re-estimating v^* from the remainder still separates the held-out pair with the correct sign. Because v^* is defined relative to a particular set of pairs, we also test sensitivity by re-estimating it from the VUA data in Section 4.4.

Two quantities follow from the projection. *Projection separation* $(\bar{\delta}_k^{\mathrm{met}} - \bar{\delta}_k^{\mathrm{lit}})$ is a first-moment check that v^* captures a meaningful contrast. The conditional scale entropy (Section 3.4) is a second-order structural quantity: two update vectors with identical mean projected differences can produce different scale entropy if their updates are organised differently across layers, so it asks whether the contrast reflects single-layer reconfiguration or multi-layer coordination.

3.3 Wavelet Decomposition of the Projected Trajectory

The projected trajectory $s(l)$ is a discrete signal of $L+1$ samples. We want to characterize its frequency content not globally but at each layer position: at layer b, which frequency scales carry energy? The Discrete Wavelet Transform decomposes a signal into octave-spaced subbands but does not localize energy along the signal's length, whereas the Continuous Wavelet Transform (CWT) produces a two-dimensional scalogram that maps energy as a function of both frequency scale a and position b [11], which is what the conditional scale entropy (Section 3.4) requires. We therefore use the CWT.

We embed the discrete signal $s(l)$ as a continuous function $\tilde{s}(t)$ via piecewise-linear interpolation on $[0, L]$, extended by constants $s(0)$ and $s(L)$ outside this interval, a standard construction that avoids artificial discontinuities [22]. We compute the CWT using the real Morlet wavelet

$$\psi(t) = e^{-t^2/2}\cos(\omega_0 t), \qquad \omega_0 = 5, \tag{3}$$

implemented by PyWavelets (`pywt.cwt` with `"morl"`). The Morlet wavelet provides good joint time–frequency localization and approximately annihilates constants ($|\hat{\psi}(0)| \approx 9.3 \times 10^{-6}$, verified numerically). The CWT at scale $a > 0$ and position $b \in \mathbb{R}$ is

$$W_\psi \tilde{s}(a, b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} \tilde{s}(t)\, \psi\left(\frac{t-b}{a}\right) dt, \tag{4}$$

and its squared modulus $|W_\psi \tilde{s}(a, b)|^2$ is the *scalogram*. We evaluate the scalogram at S integer scales $a_j = j$ for $j = 1, \ldots, S$, with $S = L + 1$. The smallest scale $a_1 = 1$ captures features that span a single layer; the largest scale $a_S = L + 1$ matches the length of the trajectory itself. Scales larger than this would extend beyond the signal and fall outside the trustworthy region defined by the cone of influence [22].

Because $\tilde{s}$ is piecewise linear, the scalogram at any fixed position b_0 is a linear function of the update vector $\boldsymbol{\delta}$. Specifically, there exists a wavelet response operator $\Phi_{b_0} \in \mathbb{R}^{S \times L}$, determined entirely by the wavelet and layer geometry, such that

$$\mathbf{z} \;=\; \Phi_{b_0}\, \boldsymbol{\delta} \in \mathbb{R}^S, \qquad |W_\psi \tilde{s}(a_j, b_0)|^2 \approx z_j^2. \tag{5}$$

A useful consequence of this decomposition: when only one layer has a nonzero update, the normalized scale distribution $z_j^2/\|\mathbf{z}\|^2$ is independent of that update's magnitude, so any observed between-condition entropy difference reflects multi-layer interaction structure, not the magnitude of any single update.

3.4 Conditional Scale Entropy

The scalogram $|W_\psi \tilde{s}(a_j, b_k)|^2$ assigns an energy value to each combination of scale a_j and position b_k. At a fixed position b_k, normalizing across scales produces a probability distribution over frequency scales:

$$p(a_j \mid b_k) = \frac{z_j^2}{\|\mathbf{z}\|^2}, \tag{6}$$

where $\mathbf{z} = \Phi_{b_k}\boldsymbol{\delta}$ as in (5). The *conditional scale entropy* is the Shannon entropy of this distribution:

$$H(\text{scale} \mid b_k) = -\sum_{j=1}^{S} p(a_j \mid b_k)\, \log\, p(a_j \mid b_k). \tag{7}$$

This quantity measures the breadth of frequency scales engaged by the model's computation at layer position b_k. High values indicate that energy is distributed across many scales simultaneously; low values indicate concentration at a single scale. The structure of this quantity is characterized by the following result.

Theorem 1 (Structure of the Conditional Scale Entropy). *Let $\boldsymbol{\delta} \in \mathbb{R}^L$ and $\mathbf{z} = \Phi_{b_0}\boldsymbol{\delta} \in \mathbb{R}^S$. Then:*

(i) $p(a_j \mid b_0) = z_j^2/\|\mathbf{z}\|^2$.
(ii) $H(\text{scale} \mid b_0) = \log S$ *if and only if* $|z_1| = \cdots = |z_S|$.
(iii) $H(\text{scale} \mid b_0) = 0$ *if and only if exactly one* $z_j \neq 0$.
(iv) $H(\text{scale} \mid b_0)$ *is invariant under* $\boldsymbol{\delta} \mapsto c\boldsymbol{\delta}$ *for any* $c \neq 0$.

Property (iv) is central: CSE is insensitive to the overall magnitude of $\boldsymbol{\delta}$ and responds only to the pattern of updates across layers. Any observed entropy difference between two conditions must therefore reflect multi-layer interaction structure, not magnitude at any individual layer. A sufficient structural condition for entropy ordering is provided by majorization.

Theorem 2 (Majorization Implies Entropy Ordering). *Let $\hat{z}_j^A = (z_j^A)^2/\|\mathbf{z}^A\|^2$ and $\hat{z}_j^B = (z_j^B)^2/\|\mathbf{z}^B\|^2$. If $\hat{\mathbf{z}}^A \prec \hat{\mathbf{z}}^B$ and $\hat{\mathbf{z}}^A$ is not a permutation of $\hat{\mathbf{z}}^B$, then $H(\text{scale} \mid b_0)^A > H(\text{scale} \mid b_0)^B$.*

Remark (one-directional implication). The converse does not hold: entropy ordering does not imply majorization. Consider $\mathbf{p} = (0.5, 0.25, 0.25)$ and $\mathbf{q} = (0.4, 0.4, 0.2)$. Direct computation gives $H(\mathbf{q}) = 1.522$ bits > 1.500 bits $= H(\mathbf{p})$. Yet the distributions are incomparable under majorization: $p_{[1]} = 0.5 > 0.4 = q_{[1]}$ violates the first partial-sum condition for $\mathbf{p} \prec \mathbf{q}$, while $q_{[1]} + q_{[2]} = 0.8 > 0.75 = p_{[1]} + p_{[2]}$ violates the condition for $\mathbf{q} \prec \mathbf{p}$. The significance of Theorem 2 is that when majorization does hold, the entropy ordering is guaranteed under every Schur-concave function — including Rényi entropies of all orders and the participation ratio — not only under Shannon entropy. The full proofs are provided in our extended preprint (arXiv:2605.21391).

3.5 The Multi-Scale Engagement Hypothesis

We now state the empirical prediction that Sections 4.3 and 4.4 test.

Table 1. Models used in the evaluation.

Model	Params	Layers	Hidden	Architecture	Precision
GPT-2 Small	124M	12	768	Dense decoder-only	FP32
GPT-2 Medium	355M	24	1024	Dense decoder-only	FP32
GPT-2 Large	774M	36	1280	Dense decoder-only	FP16
LLaMA-2 7B	7B	32	4096	Dense decoder-only	FP16
GPT-oss 20B	20B	24	2880	MoE (32 experts)	MXFP4

Multi-Scale Engagement Hypothesis. Metaphorical processing produces higher conditional scale entropy than literal processing at layer positions where the model actively reconfigures the target token's representation: $H(\text{scale} \mid b)^{\text{met}} > H(\text{scale} \mid b)^{\text{lit}}$.
By Theorem 1(iv) this inequality concerns the pattern of updates across layers, not their magnitude; Theorem 2 provides a sufficient structural condition via majorisation.

Since the reconfiguration depth is model-dependent, we identify it empirically. On the 25 controlled minimal pairs (Section 4.3) we locate the contiguous range of positions where CSE elevation is significant under cluster-based permutation correction, and call this range the *active zone* $\mathcal{A}_M$. We test this zone with two independent replications: (i) 200 naturalistic VUA pairs (Section 4.4) with v^* re-estimated from scratch on the VUA data, and (ii) four additional decoder-only architectures spanning 124M to 20B parameters, each with its own v^* and identical statistical criteria.

4 Experiments and Results

This section evaluates the Multi-Scale Engagement Hypothesis across five transformer architectures and two stimulus sets.

4.1 Experimental Design

Models. We evaluate five decoder-only transformer architectures spanning two orders of magnitude in parameter count (Table 1): GPT-2 Small/Medium/Large [15] within a single family for controlled scaling, LLaMA-2 7B [23] as a generalisation test, and GPT-oss 20B [14] as a Mixture-of-Experts model with 32 experts per layer. Hidden states are extracted from pretrained models without fine-tuning. All CWT computations use PyWavelets (`pywt.cwt` with `"morl"`) at $S = L{+}1$ integer scales.

Stimulus sets. The first stimulus set consists of 25 controlled minimal pairs in which a target lexeme appears once in a literal sentence and once in a metaphorical sentence (e.g., *The weight crushed the box / The news crushed her hopes*); each pair involves a concrete physical verb or noun used in a cross-domain mapping between a physical source and an abstract target. The shared lexeme ensures that the initial token embedding is near-identical across conditions, so differences in the residual trajectory arise from contextual processing.

The second stimulus set is drawn from the VUA All POS corpus [19], the standard benchmark for metaphor identification, providing token-level metaphor annotations on naturalistic English text spanning fiction, news, academic, and conversational genres. We access it via the `liyucheng/VUA20` dataset (160,154 annotated tokens) and extract 200 pairs in which the same word, matched by its dictionary form (lemma) so that inflected variants are treated as the same target, appears in both a metaphorically and a literally annotated sentence. We restrict to 5–35-token sentences with the target in a non-edge position, take one pair per lemma, and use a fixed random seed for reproducibility. Unlike the controlled pairs, the VUA set is dominated by conventionalised metaphors whose figurative meaning is lexically entrenched, providing a stringent test of whether CSE remains sensitive when the metaphor signal is weaker.

As a specificity control, we additionally construct 10 literal–literal pairs sharing a target word but differing in contextual complexity (e.g., *The knife cut the bread* / *The surgeon cut through the multiple layers of tissue*). These pairs contain no metaphor and test whether CSE elevation reflects general semantic processing difficulty (Section 4.5).

Evaluation quantities. The primary evaluation quantity is the layer-resolved conditional scale entropy difference

$$\Delta H_M(b) = H(\text{scale} \mid b)^{\text{met}} - H(\text{scale} \mid b)^{\text{lit}}, \tag{8}$$

computed for each pair and each layer position b. Positive values indicate broader scale engagement for metaphorical processing.

Projection separation serves as a validation check on the learned contrast direction v^*: for each pair k, a consistently positive mean projected update difference across layers confirms that v^* captures a stable metaphor–literal distinction. This quantity validates the projection but is not itself a test of the Multi-Scale Engagement Hypothesis.

Additional spectral quantities (CWT energy, H_W, $H(q)$) are summarised in Section 4.3 as supporting analyses.

Statistical testing. For the controlled experiments ($K = 25$ pairs), we use Monte Carlo sign-flip tests (10,000 permutations) at each scalogram position; for the VUA experiments ($K = 200$), one-sample t-tests on paired differences with Cohen's $d = \bar{d}/s_d$.

To correct for multiple comparisons across scalogram positions, we use cluster-based permutation testing [13]. The cluster statistic is the sum of test statistics over the largest contiguous cluster exceeding the one-sided threshold ($\alpha = 0.05$), evaluated against 5,000 sign-flipped null datasets; the cluster p-value is the fraction of null maxima meeting or exceeding the observed value.

4.2 Validation of the Contrast Direction

Three checks confirm that v^* captures a stable metaphor–literal distinction on every model tested. (i) *Projection separation* is positive for at least 21/25 con-

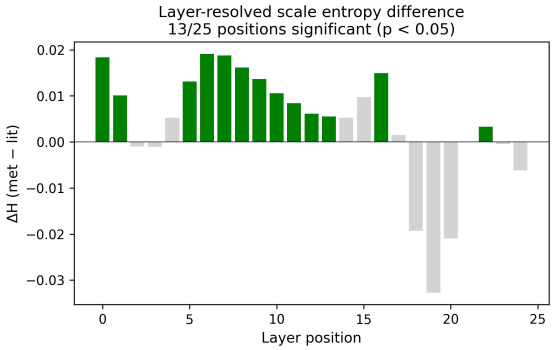

Fig. 1. Layer-resolved conditional scale entropy difference $\Delta H = H^{\text{met}} - H^{\text{lit}}$ on GPT-2 Medium (25 controlled pairs). Green bars indicate positions reaching $p < 0.05$ under Monte Carlo sign-flip testing; the cluster at positions 5–13 survives cluster-based permutation correction ($p_{\text{cluster}} = 0.007$). Absolute scale entropy spans $H \in [2.2, 3.1]$ across positions and is nearly identical between conditions, with the difference at the second decimal.

trolled pairs on every architecture (all $p < 0.001$). (ii) *Leave-one-out generalisation* on GPT-2 Medium shows that a direction learned from any 24 pairs separates the held-out pair with the correct sign for all 25 holdouts. (iii) *Stability under expansion*: v^* estimated from the original 10 pairs and from all 25 pairs has cosine similarity 0.997, indicating that the direction reflects the model rather than any particular stimulus subset.

4.3 Conditional Scale Entropy Across Architectures

Primary finding. Figure 1 shows the layer-resolved conditional scale entropy difference $\Delta H = H^{\text{met}} - H^{\text{lit}}$ on GPT-2 Medium using the full 25-pair controlled set. The metaphorical condition produces consistently higher conditional scale entropy at positions 5–13, with 13 of 25 scalogram positions reaching $p < 0.05$ under Monte Carlo sign-flip tests. The peak occurs at position 9 ($p = 0.0008$, $18/25$ pairs positive). Cluster-based permutation testing confirms the main contiguous cluster at positions 5–13 ($p_{\text{cluster}} = 0.007$), accounting for multiple comparisons across all 25 scalogram positions.

Cross-architecture consistency. Extending the analysis to four additional decoder-only architectures with the same 25-pair set and identical wavelet parameters, Table 2 and Figure 2 report the per-architecture cluster statistics. On every model tested CSE elevates significantly at a contiguous range of positions, and every cluster survives cluster-based permutation correction at $\alpha = 0.05$ ($p_{\text{cluster}} < 0.001$ on four of five models; $= 0.007$ on GPT-2 Medium). All active zones occupy the early-to-mid relative-depth range, despite absolute depth varying from 12 to 36 layers and architecture varying between dense attention

Table 2. Conditional scale entropy across architectures (25 controlled pairs, cluster-based permutation correction, 5,000 permutations). Active zones consistently occupy early-to-mid depth across all architectures.

Model	K	Entropy	Cluster (positions)	p_{cluster}	Peak p	Rel. depth
GPT-2 Small	25	Shannon	2–8	<0.001	<0.001	17–67%
GPT-2 Medium	25	Shannon	5–13	0.007	0.0008	21–54%
GPT-2 Large	25	Shannon	0–12	<0.001	<0.001	0–33%
LLaMA-2 7B	25	Shannon	5–11	<0.001	<0.001	16–34%
GPT-oss 20B	25	Shannon	0–13	<0.001	<0.001	0–54%

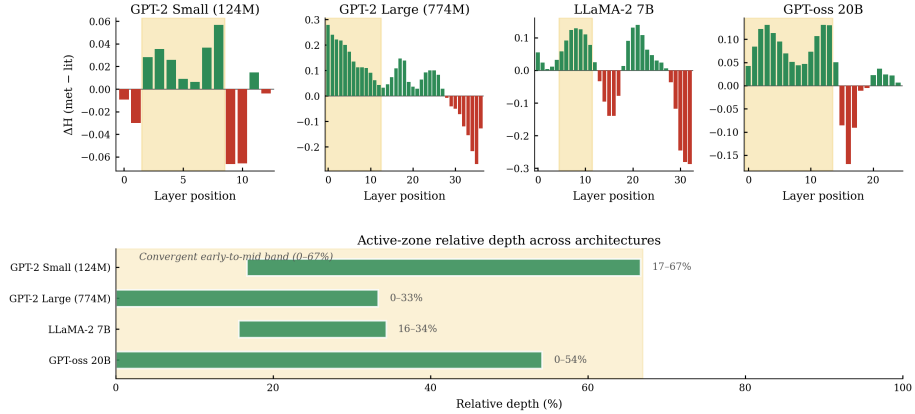

Fig. 2. Layer-resolved conditional scale entropy difference $\Delta H = H^{\text{met}} - H^{\text{lit}}$ across four decoder-only architectures (25 controlled pairs each; GPT-2 Medium shown separately in Figure 1). Green: $\Delta H > 0$; red: $\Delta H < 0$; shaded region indicates the contiguous significant cluster (cluster-based permutation test, 5,000 permutations, $\alpha = 0.05$). Bottom panel: active-zone relative depth for each architecture. Despite absolute depth varying from 12 to 36 layers, every cluster lies within the early-to-mid relative-depth band (0–67%).

(GPT-2, LLaMA) and Mixture-of-Experts routing (GPT-oss). Rényi-2 conditional entropy on GPT-2 Medium produces identical significant positions, ruling out sensitivity to distributional tail behaviour.

Robustness of CSE. Among the spectral quantities examined, only CSE generalizes reliably. Energy suppression (CWT power, met < lit) is significant on GPT-2 Medium but absent on LLaMA-2 7B and GPT-oss 20B. More strikingly, update-energy entropy $H(q)$ reverses sign on the same model when the stimulus set is expanded from 10 to 25 pairs ($p = 0.003$, met < lit $\rightarrow p = 0.002$, met > lit), a pattern mirrored across architectures; a quantity whose sign depends on stimulus composition cannot serve as a reliable signature. The attenuation of global metrics on LLaMA-2 7B parallels GPT-oss 20B, suggesting the effect is linked to model scale rather than the dense/MoE distinction. CSE is the only metric whose direction and significance are stable across pair expansion, architectures, and, as shown below, stimulus regimes.

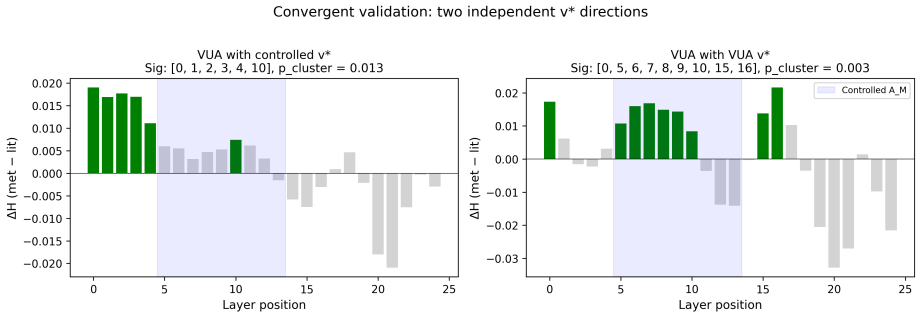

Fig. 3. Convergent validation of the VUA active zone on GPT-2 Medium (200 pairs). Left: VUA analysed with the controlled-pair v^* — the significant cluster falls at positions 0–4 ($p_{\text{cluster}} = 0.013$), shifted earlier than the controlled active zone $\mathcal{A}_M = \{5, \ldots, 13\}$ (blue shading). Right: VUA analysed with v^* independently re-estimated from the VUA data — the significant cluster shifts to positions 5–10 ($p_{\text{cluster}} = 0.003$), falling entirely within $\mathcal{A}_M$. Two independent datasets with independently estimated contrast directions thus converge on the same depth range. Green bars indicate positions reaching $p < 0.05$.

4.4 Generalization to Naturally-Occurring Metaphor

The controlled experiments use carefully constructed stimuli; we now test whether the conditional scale entropy generalizes to naturalistic metaphor drawn from an established benchmark. Stimulus construction details are described in Section 4.1; all analyses use GPT-2 Medium with identical wavelet parameters.

Projection separation. The contrast direction v^* from the 25 controlled pairs yields positive mean projected update differences for 107/200 VUA pairs ($p = 0.19$): a non-significant first-moment result, consistent with v^* not fully capturing predominantly conventionalised metaphors. As noted in Section 3.2, projection separation tests average update magnitude along v^*; a null mean shift does not preclude structural differences in the depth profile.

CSE elevation and convergent validation. Two analyses are needed because v^* from the controlled pairs may not perfectly align the depth profile for conventionalised VUA metaphors. With the controlled-pair v^*, the VUA cluster falls at positions 0–4 ($p_{\text{cluster}} = 0.013$), shifted earlier than $\mathcal{A}_M = \{5, \ldots, 13\}$ (Figure 3, left). Re-estimating v^* from the VUA data alone (cosine similarity 0.964 with the controlled-pair direction) shifts the significant cluster to positions 5–10 ($p_{\text{cluster}} = 0.003$), falling entirely within $\mathcal{A}_M$ (Figure 3, right). Two independent datasets with independently estimated contrast directions thus converge on the same depth range, confirming that the active zone reflects a stable property of how GPT-2 Medium processes figurative language rather than an artifact of any one stimulus set.

Graded effect size. The effect size for VUA metaphors ($d \approx 0.17$ at the cluster peak) is approximately half that of the controlled vivid metaphors ($d \approx 0.34$), consistent with the Graded Salience Hypothesis [10]: conventionalised metaphors, whose figurative meaning is lexically entrenched, produce a weaker but structurally identical CSE elevation compared to novel cross-domain metaphors.

Robustness of the conditional scale entropy. On GPT-2 Medium across the two stimulus sets, projection separation, CWT energy, H_W, and $H(q)$ all attenuate to non-significance on the VUA data, while only the conditional scale entropy generalizes to naturalistic data, mirroring the cross-architecture pattern of Section 4.3.

4.5 Specificity Controls

Two controls test whether CSE elevation reflects the metaphor–literal contrast specifically, rather than general semantic processing difficulty or the particular content of metaphorical sentences.

Semantic complexity (literal–literal). We construct 10 pairs of literal sentences sharing a target word but differing in contextual complexity (e.g., *The knife cut the bread* / *The surgeon cut through the multiple layers of tissue*). Using the same contrast direction v^* and identical wavelet parameters, no scalogram position reaches significance under sign-flip testing. Mean ΔH within the active zone $\mathcal{A}_M = \{5, \ldots, 13\}$ is -0.0005 for the complexity contrast versus 0.017 for the metaphor contrast. CSE is therefore specific to the metaphor–literal distinction and does not respond to variation in semantic complexity within literal language.

Expression mode (matched-meaning triples). To separate figurativeness from propositional content, we construct 10 minimal triples in which two distinct metaphorical sentences and one literal sentence express the same meaning in identical syntactic frames (e.g., *She is a lightning bolt* / *She is a speeding bullet* / *She is very fast*). Pooling both metaphors against the literal, CSE elevates significantly at positions 4–16 ($p_{\text{cluster}} = 0.003$), overlapping $\mathcal{A}_M$. Each metaphor individually produces nearly identical CSE profiles against the literal (14/25 significant positions each), while the two metaphors compared against each other yield zero significant positions ($|\Delta H| < 0.001$), confirming that figurative mode of expression, not propositional content, drives CSE elevation.

5 Discussion

Organisation, not intensity. The invariance property of Theorem 1(iv) is central: CSE isolates the organisation of computation across scales from its intensity. Magnitude-dependent metrics (CWT energy, update-energy concentration) produce significant results on GPT-2 Medium but do not survive changes of architecture, stimulus regime, or even stimulus expansion within a single model:

$H(q)$ reverses direction (met $<$ lit to met $>$ lit) on GPT-2 Medium when the controlled set is expanded from 10 to 25 pairs (see Section 4.3). CSE survives all three tests. Metaphorical processing is therefore distinguished not by greater effort at any single depth but by broader coordination of representational updates across frequency scales.

Cross-depth structure as a complementary lens. Current mechanistic interpretability methods (activation patching, sparse autoencoders, circuit tracing [5]) operate at the level of individual components: they address what specific neurons or attention heads do, but not how computation is organised across the depth of the residual stream. Sun et al. [20] showed that middle layers can be skipped or reordered with limited performance loss; our results add nuance: when the model encounters semantically complex input, the *pattern* of updates across layers changes even when individual layers appear interchangeable on average. CSE captures this input-dependent depth structure as a macro-level diagnostic complementary to component-level analysis.

Convergent validation and graded effect. When v^* is re-estimated by applying Equation (2) to the 200 VUA pairs (rather than the 25 controlled pairs), the resulting active zone (positions 5–10) falls within the controlled-pair zone (positions 5–13). The graded effect size ($d \approx 0.34$ for vivid cross-domain metaphors, $d \approx 0.17$ for conventionalised VUA metaphors) is consistent with the Graded Salience Hypothesis [10]: metaphors whose figurative meaning is not yet lexically entrenched demand more extensive multi-scale coordination.

Architectural implications. The active zone consistently occupies the early-to-mid relative-depth range on all five architectures, despite threefold variation in absolute depth and a change of attention mechanism between dense and Mixture-of-Experts. Situating this within the layer-uniformity findings of Sun et al. [20], our results locate, within the middle-layer plateau, the phase where cross-domain semantic content is integrated — in its early half, before the late-layer prediction phase described by Geva et al. [9]. This architecture-independence has a practical implication: interventions targeting figurative-language processing are most likely to take effect in the early-to-mid residual stream regardless of model.

Scope and future directions. The current experiments cover 25 controlled pairs, 200 naturalistic VUA pairs, and five architectures up to 20B parameters; specificity controls each use 10 pairs. Larger stimulus sets, frontier-scale models, and a stimulus-independent contrast direction would strengthen generalisability. Three extensions are of particular interest: (i) stimuli with continuously rated novelty, enabling a finer-grained test of the graded prediction; (ii) causal intervention via scale-channel suppression using the wavelet response operator (Equation (5)), to establish whether multi-scale coordination is functionally necessary or merely correlated with figurative processing; and (iii) application to polysemy, irony, metonymy, and multi-step reasoning [18], to determine whether the signature is specific to metaphor or reflects a broader computational property of semantically complex input.

6 Conclusion

We introduced the conditional scale entropy (CSE), a wavelet-derived measure of how broadly transformer computation engages across frequency scales at each layer position. Two theorems establish that CSE is invariant to update magnitude and ordered by majorisation of the normalised scale distribution. Across five decoder-only architectures from 124M to 20B parameters, metaphorical tokens elevate CSE at a contiguous range of layer positions on every model, with all clusters surviving cluster-based permutation correction and falling within the early-to-mid relative-depth band. Convergent validation on 200 naturalistic VUA pairs recovers the same active zone and a graded effect-size pattern consistent with the Graded Salience Hypothesis. Specificity controls confirm that the elevation reflects figurative mode of expression rather than semantic difficulty or propositional content. Among all spectral quantities examined, CSE is uniquely stable across architectures, stimulus expansion, and stimulus regimes, identifying multi-scale coordination as a robust computational signature of metaphorical language processing and positioning CSE as a principled analytical tool for characterising how transformers organise computation across depth.

References

1. Abraham, J., et al.: Wavelet-based mechanistic interpretability of vision transformers. In: CVPR Workshop on Mechanistic Interpretability in Vision (2025)
2. Aghazadeh, E., Fayyaz, M., Yaghoobzadeh, Y.: Metaphors in pre-trained language models: Probing and generalization across datasets and languages. In: Proceedings of ACL. pp. 2037–2050 (2022)
3. Bowdle, B.F., Gentner, D.: The career of metaphor. Psychological Review **112**(1), 193–216 (2005)
4. Choi, M., Lee, S., Choi, E., Park, H., Lee, J., Lee, D., Lee, J.: MelBERT: Metaphor detection via contextualized late interaction using metaphorical identification theories. In: Proceedings of NAACL. pp. 1763–1773 (2021)
5. Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., et al.: A mathematical framework for transformer circuits. Transformer Circuits Thread (2021)
6. Ethayarajh, K.: How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In: Proceedings of EMNLP. pp. 55–65 (2019)
7. Gao, G., Choi, E., Choi, Y., Zettlemoyer, L.: Neural metaphor detection in context. In: Proceedings of EMNLP. pp. 607–613 (2018)
8. Geshkovski, B., Letrouit, C., Polyanskiy, Y., Rigollet, P.: A mathematical perspective on transformers. Bulletin of the American Mathematical Society **62**(3) (2025)
9. Geva, M., Caciularu, A., Wang, K., Goldberg, Y.: Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. pp. 30–45 (2022)
10. Giora, R.: Understanding figurative and literal language: The graded salience hypothesis. Cognitive Linguistics **8**(3), 183–206 (1997)

11. Mallat, S.: A Wavelet Tour of Signal Processing: The Sparse Way. Academic Press, 3rd edn. (2009)
12. Mao, R., Lin, C., Guerin, F.: End-to-end sequential metaphor identification inspired by linguistic theories. In: Proceedings of ACL. pp. 3888–3898 (2019)
13. Maris, E., Oostenveld, R.: Nonparametric statistical testing of EEG- and MEG-data. Journal of Neuroscience Methods **164**(1), 177–190 (2007). https://doi.org/10.1016/j.jneumeth.2007.03.024
14. OpenAI: gpt-oss-120b & gpt-oss-20b model card (2025), https://arxiv.org/abs/2508.10925
15. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners. Tech. rep., OpenAI (2019), https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf
16. Rahaman, N., Barber, A., Arpit, D., Draxler, F., Lin, M., Hamprecht, F., Bengio, Y., Courville, A.: On the spectral bias of neural networks. In: Proceedings of ICML. pp. 5301–5310 (2019)
17. Shwartz-Ziv, R., Tishby, N.: Opening the black box of deep neural networks via information. arXiv preprint arXiv:1703.00810 (2017)
18. Sravanthi, P., Mamidi, R.: PUB: A pragmatics understanding benchmark for assessing LLMs' pragmatics capabilities. In: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING) (2024)
19. Steen, G.J., Dorst, A.G., Herrmann, J.B., Kaal, A.A., Krennmayr, T., Pasma, T.: A Method for Linguistic Metaphor Identification. John Benjamins (2010)
20. Sun, Q., Pickett, M., Nain, A.K., Jones, L.: Transformer layers as painters. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 25219–25227 (2025)
21. Tenney, I., Das, D., Pavlick, E.: BERT rediscovers the classical NLP pipeline. In: Proceedings of ACL. pp. 4593–4601 (2019)
22. Torrence, C., Compo, G.P.: A practical guide to wavelet analysis. Bulletin of the American Meteorological Society **79**, 61–78 (1998), https://api.semanticscholar.org/CorpusID:14928780
23. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023). https://doi.org/10.48550/arXiv.2307.09288
24. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems. vol. 30 (2017)
25. Voita, E., Titov, I.: Information-theoretic probing with minimum description length. In: Proceedings of EMNLP. pp. 183–196 (2020)
26. Wachowiak, L., Gromann, D.: Does GPT-3 grasp metaphors? identifying metaphor mappings with generative language models. In: Proceedings of ACL (Volume 1: Long Papers). pp. 1018–1032 (2023)
27. Zhai, S., Likhomanenko, T., Littwin, E., Busbridge, D., Ramapuram, J., Zhang, Y., Gu, J., Susskind, J.: Stabilizing transformer training by preventing attention entropy collapse. arXiv preprint arXiv:2303.08296 (2023)