How Neural Networks Organize Concepts: Introducing Concept Trajectory Analysis for Deep Learning Interpretability

Andrew Smigaj[1], Claude Anthropic[2], Grok xAI[3]
[1]Institute of Discordant Colony Optimization, [2]Anthropic Research, [3]xAI Corp

July 2025

Abstract

We present Concept Trajectory Analysis (CTA), an interpretability method that tracks how neural networks organize concepts by following their paths through clustered activation spaces across layers. Applying CTA to GPT-2 with 1,228 single-token words revealed that the model organizes language primarily by grammatical function rather than semantic meaning. We found that 48.5% of words converge to grammatical highways where nouns—whether animals, objects, or abstracts—travel together, while maintaining semantic distinctions at finer scales ($\chi^2 = 95.90$, $p < 0.0001$).

CTA combines geometric clustering with trajectory tracking to quantify how concepts flow through networks. Our method introduces windowed analysis to identify phase transitions (semantic→grammatical in GPT-2) and leverages LLMs to generate interpretable cluster labels. In medical AI, CTA exposed how a heart disease model stratifies patients through risk pathways, revealing demographic biases (male overprediction in Path 4, 83% male composition).

By making neural organization visible and quantifiable, CTA provides actionable insights for model debugging, bias detection, and scientific understanding of deep learning. Our open-source implementation enables researchers to apply CTA to any neural network, advancing interpretable AI across domains.

1 Introduction

Understanding how neural networks organize linguistic information remains a fundamental challenge in interpretability research. We present Concept Trajectory Analysis (CTA), a method that tracks concept evolution through neural networks by analyzing movement patterns in clustered activation spaces. Application of this method to GPT-2 indicates that the model

Table 1: Key Terminology Used in This Paper

Term	**Definition**
Trajectory	The path a concept takes through cluster assignments across layers
Path	A specific sequence of clusters, e.g., L1_C0→L2_C1→L3_C0
Highway	A meso-level bundle of similar paths traveled by many concepts
Window	Temporal grouping of layers (Early: L0-3, Middle: L4-7, Late: L8-11)
Fragmentation (F)	Metric measuring diversity of trajectories within a concept group
FC	Path-Centroid Fragmentation: distance between paths and group centroid
CE	Intra-Class Cluster Entropy: distribution of concepts across clusters
SA	Sub-space Angle: angular separation between concept group representations
Microcluster	Fine-grained sub-structure within larger cluster (future work)

organizes words primarily by grammatical function rather than semantic similarity, providing new insights into transformer language processing.

In our analysis of 1,228 single-token words across 8 semantic categories with balanced grammatical distribution (33.1% verbs), we observed a temporal progression: words initially cluster by semantic similarity in early layers, undergo a phase transition in middle layers, and ultimately converge to grammatical organization in later layers. By the final layers, 48.5% (95% CI: 45.7%–51.2%) of words follow the most common pathway, while others distribute across multiple pathways. This sequential reorganization—validated by highly significant grammatical clustering ($\chi^2 = 95.90$, $p < 0.0001$)—suggests that neural networks first process semantic features before reorganizing by grammatical function.

This observation emerged from our framework that combines mathematical analysis with human interpretation. CTA tracks concepts through clustered activation spaces across neural network layers, using quantitative metrics alongside LLM-generated explanations. The method addresses several questions in interpretability research:

- How can we make neural organization both mathematically precise and humanly interpretable?
- What organizational principles do neural networks discover that differ from human intuition?
- Can we create real-time interpretability with negligible computational overhead?
- How do we scale from analyzing hundreds to millions of concepts?

We validate CTA through experiments on both language models and traditional ML. Our windowed analysis (Early/Middle/Late) identifies phase transitions in neural processing, while unique cluster labeling (L{layer}_C{cluster}) enables precise tracking. By analyzing all trajectories rather than just dominant paths, we capture the full complexity of neural organization.

Beyond transformer analysis, CTA can be applied to traditional ML tasks. In heart disease diagnosis, trajectory analysis shows how neural networks process patient data, with different pathways emerging for various patient profiles. High-risk patients (older, elevated cholesterol) tend to route through specific neural pathways with higher fragmentation scores, potentially indicating model uncertainty. This medical AI application illustrates how CTA can help understand decision-making processes in healthcare domains.

1.1 Background: Concept Trajectory Analysis

Concept Trajectory Analysis (CTA) clusters datapoint activations in each layer's activation space (e.g., using k-means), assigning layer-specific cluster IDs denoted Ll_Ck, where l is the layer index and k is the cluster index. Transitions between clusters are tracked across layers, forming trajectories $\pi_i = [c_i^1, c_i^2, \ldots, c_i^L]$, interpreted as concept evolution through the network. In feedforward networks, trajectories are strictly unidirectional, and clusters with different layer-specific IDs (e.g., L1_C0 and L3_C0) are not assumed related unless validated by geometric similarity metrics. Large language models can then narrate these trajectories to provide interpretable insights.

While activation spaces are emergent, high-dimensional representations whose coordinate systems may not map to semantically meaningful axes [Ribeiro et al., 2016, Lundberg and Lee, 2017], CTA addresses these concerns through rigorous mathematical validation and cross-layer metrics that verify genuine patterns rather than artifacts.

Contributions:

- Formalize Concept Trajectory Analysis with mathematical validation criteria and cross-layer metrics
- Develop optimal clustering determination using Gap statistic with layer-specific cluster counts
- Implement trajectory visualization and LLM-powered interpretability for neural network analysis
- Provide evidence that GPT-2 organizes words primarily by grammatical function in later layers
- Demonstrate CTA's applicability to both transformer models and traditional ML tasks
- Release open-source implementation for reproducible neural network interpretability research

2 Related Work: Positioning CTA in the Interpretability Landscape

While neural network interpretability has seen significant advances, existing methods primarily focus on explaining individual predictions or visualizing static representations. We position Concept Trajectory Analysis (CTA) as addressing a critical gap: understanding how concepts dynamically evolve through neural network layers.

2.1 Attribution-Based Methods

Attribution methods like LIME [Ribeiro et al., 2016] and SHAP [Lundberg and Lee, 2017] decompose model predictions by assigning importance scores to input features. While valuable for understanding which inputs influence outputs, these methods treat the network as a black box, providing no insight into intermediate processing stages. Integrated Gradients [Sundararajan et al., 2017] traces attribution through the network but still focuses on input-output relationships rather than concept evolution.

CTA differs fundamentally by tracking how representations transform across layers. Rather than asking "which pixels matter for this classification?", CTA asks "how does the concept of 'cat' evolve from pixels to semantic understanding to final prediction?" This shift from static attribution to dynamic trajectory analysis reveals organizational principles invisible to attribution methods.

2.2 Attention Mechanism Analysis

Attention visualization has become prominent in transformer interpretability, revealing which tokens the model focuses on during processing. However, attention weights show correlation, not causation, and often prove misleading about actual information flow. More critically, attention analysis remains locked to the token level, unable to capture higher-level concept organization.

Consider our finding that GPT-2 routes 72.8% of words through an "entity superhighway" regardless of semantic content. Attention analysis might show that "cat" attends to nearby determiners or adjectives, but it cannot reveal that "cat," "democracy," and "table" all follow identical processing pathways through the network's layers. While attention weights fluctuate based on context, the underlying organizational principle—grammatical categorization—remains invisible to attention-based methods.

CTA transcends token-level analysis by clustering activations into meaningful concepts and tracking their evolution. Where attention asks "what does this token look at?", CTA asks "how does this concept transform?" This shift in perspective proved essential: the grammatical organization emerges not from examining individual attention patterns but from observing how hundreds of words converge to shared pathways despite starting from diverse semantic origins.

2.3 Activation and Representation Analysis

Prior work has examined neural activations through various lenses. Network Dissection [Bau et al., 2017] identifies neurons selective for visual concepts, while TCAV [Kim et al., 2018] tests concept presence using directional derivatives. Recent work from Anthropic on polysemantic neurons and superposition [Elhage et al., 2022b,a] reveals how individual neurons can respond to multiple, unrelated concepts—a phenomenon that underscores the importance of analyzing distributed representations rather than individual units. Representation similarity analysis [Kornblith et al., 2019, Raghu et al., 2017] compares activation spaces but typically focuses on single layers or layer pairs.

These methods provide snapshots of representations but miss the dynamic story of concept evolution. CTA's innovation lies in:

- **Trajectory tracking**: Following concepts through all layers, not just analyzing fixed points
- **Path analysis**: Identifying archetypal routes through the network's processing pipeline
- **Phase detection**: Discovering transitions like GPT-2's shift from semantic to grammatical organization
- **Narrative generation**: Using LLMs to translate mathematical patterns into human understanding

2.4 Clustering in Neural Networks

While clustering has been applied to neural activations, previous work typically clusters at single layers or uses clustering for compression rather than interpretation. Explainable clustering methods [Dasgupta et al., 2020] provide algorithmic transparency but haven't been systematically applied to track concept evolution through deep networks.

CTA's contribution includes:

- **Layer-specific labeling**: Our Ll_Ck notation prevents confusion and enables precise tracking
- **Cross-layer metrics**: Quantifying concept evolution through centroid similarity, membership overlap, and trajectory fragmentation
- **Windowed analysis**: Revealing phase transitions invisible to single-layer clustering

2.5 The Interpretability Gap CTA Addresses

Existing interpretability methods excel at specific tasks—attribution for feature importance, attention for token relationships, activation analysis for concept detection. However, none address the fundamental question: How do neural networks organize and transform information as it flows through layers?

CTA fills this gap by providing:

1. **Dynamic analysis**: Tracking concept evolution rather than static snapshots
2. **Organizational insights**: Revealing principles like grammatical convergence in language models
3. **Multi-scale understanding**: From individual trajectories to population-level patterns
4. **Actionable interpretability**: Identifying bias patterns, uncertainty indicators, and decision pathways
5. **Narrative explanations**: Bridging mathematical analysis with human comprehension through LLM integration

By shifting focus from "what" to "how"—from static attribution to dynamic trajectories—CTA opens new avenues for understanding neural networks as information processing systems rather than mere function approximators. This perspective proved essential for discovering that transformers organize language fundamentally differently than human linguistic intuitions suggest, a finding that emerged not from examining attention or attribution, but from tracking concepts as they journey through the network's layers.

3 Mathematical Foundation of Layerwise Activation Geometry

3.1 What Is Being Clustered?

Let $A^l \in \mathbb{R}^{n \times d_l}$ denote the matrix of activations at layer (l), where each row $\mathbf{a}_i^l$ is a datapoint's activation vector. Once the model is trained, A^l provides a static representation space per layer.

3.2 Metric Selection and Validity

We cluster A^l into k_l clusters, assigning unique layer-specific labels Ll_C0, Ll_C1, ..., Ll_C$\{k_l - 1\}$. This unique labeling scheme (e.g., L4_C1 for layer 4, cluster 1) prevents cross-layer confusion and enables precise tracking of concept evolution. A datapoint's path is a sequence $\pi_i = [c_i^1, c_i^2, \ldots, c_i^L]$, where c_i^l is the cluster assignment at layer l in the format Ll_Ck.

We determine optimal k_l using the Gap statistic, which compares within-cluster dispersion to that expected under a null reference distribution. For layer l, we compute:

$$\text{Gap}(k) = \mathbb{E}[\log(W_k^*)] - \log(W_k)$$

where W_k is the within-cluster sum of squares and W_k^* is its expectation under the null.

We examine Euclidean, cosine, and Mahalanobis metrics. In high-dimensional spaces, Euclidean norms lose contrast; cosine and L1 often behave better. PCA or normalization can stabilize comparisons. In feedforward networks, paths are unidirectional, and apparent convergence (e.g., [L1_C0 → L2_C2 → L3_C0]) is validated by computing cosine or Euclidean similarity between cluster centroids across layers, ensuring that any perceived similarity reflects geometric proximity in activation space rather than shared labels.

3.3 Clustering Approaches

We primarily use k-means clustering with the Gap statistic for determining optimal cluster counts.

3.4 Within-Cluster Semantic Structure

While our primary analysis focuses on cluster-level trajectories, the position of datapoints within clusters may carry semantic meaning. Following principles of distributional semantics, nearby points in activation space often share semantic properties—even within the same cluster. For instance, within the dominant entity pathway (L4_C1), "cat" and "dog" may occupy closer positions than "cat" and "democracy," despite all being nouns. This suggests potential hierarchical organization where coarse clusters capture grammatical categories while fine-grained positions encode semantic relationships.

Future work could explore micro-clustering within major pathways to investigate these potential semantic substructures. Techniques like hierarchical clustering or local neighborhood analysis might reveal how dominant pathways subdivide into semantic regions while maintaining overall grammatical coherence.

3.5 Windowed Trajectory Analysis

To capture phase transitions in neural processing, we introduce windowed analysis that segments the network into functional regions:

- **Early Window** (layers 0-3): Initial feature extraction and semantic differentiation

- **Middle Window** (layers 4-7): Conceptual reorganization and consolidation
- **Late Window** (layers 8-11): Final representation and task-specific processing

For each window w, we compute stability metrics:

$$S_w = \frac{1}{|P_w|} \sum_{p \in P_w} \frac{|\text{mode}(p)|}{|p|}$$

where P_w is the set of path segments in window w and $\text{mode}(p)$ is the most frequent cluster transition. Changes in stability patterns across windows can indicate phase transitions in the network's organizational principles.

3.6 Quantitative Metrics for Concept Evolution

To ground our analysis in quantitative evidence, we employ four complementary metrics that capture different aspects of concept evolution through neural networks:

3.6.1 Trajectory Fragmentation (F)

Measures path diversity for a semantic category:

$$F = 1 - \frac{\text{count of most common path}}{\text{total paths in category}}$$

High fragmentation indicates diverse processing strategies within a category. In our experiments, this metric helps quantify convergence patterns—for instance, the GPT-2 analysis shows fragmentation varying from 0.796 (early) to 0.499 (middle) to 0.669 (late), suggesting complex dynamics in the organization of the balanced dataset.

3.6.2 Path-Centroid Fragmentation (FC)

Measures how dissimilar consecutive clusters are along a specific sample path:

$$FC = 1 - \overline{\text{sim}}$$

where $\overline{\text{sim}}$ is the mean centroid similarity (cosine) between successive clusters on the path. High values indicate that representations "jump" across concept regions between layers; low values indicate coherent, incremental refinement. The heart disease model shows remarkably low FC=0.096, indicating smooth transitions.

3.6.3 Intra-Class Cluster Entropy (CE)

For every layer, we cluster activations and measure the Shannon entropy of the resulting cluster distribution within each ground-truth class:

$$CE = \frac{H(C|Y)}{\log_2 k^*}$$

where $H(C|Y)$ is the conditional entropy of clusters given class labels, normalized by $\log_2 k^*$ (the selected number of clusters). CE=1 means class features are maximally dispersed across clusters, while CE=0 means each class occupies a single, compact cluster.

3.6.4 Sub-space Angle Fragmentation (SA)

We compute the principal components for the activations of each class and evaluate the pair-wise principal angles between those subspaces. Large mean angles ($\gg 0$) imply that the network embeds classes in orthogonal directions—evidence of fragmentation—while small angles suggest a shared, low-dimensional manifold. In GPT-2, we observe SA collapsing from 45-60° (semantic separation) to 5-10° (grammatical convergence).

3.7 Applying the Framework: From Theory to Practice

These metrics work in concert to reveal different aspects of neural organization. In Section **??**, we apply them to uncover GPT-2's grammatical organization, where decreasing SA and CE values quantify the convergence from semantic to syntactic processing. In Section **??**, consistently low FC values validate that medical diagnosis models maintain coherent patient representations throughout processing. The windowed analysis framework proves particularly powerful for identifying phase transitions—critical reorganization points where networks shift their organizational principles, as evidenced by stability metric drops in GPT-2's middle layers.

4 Experimental Design

Our experiments validate CTA across diverse domains, from medical AI to language understanding:

4.1 Datasets and Models

- **Heart Disease Diagnosis**: UCI Heart Disease dataset (303 patients, 13 clinical features) with 3-layer MLP, demonstrating medical AI interpretability

- **GPT-2 Semantic Subtypes**: 1,228 validated single-token words across 8 semantic categories, analyzed through GPT-2's 12 layers (embedding layer + 11 transformer blocks, 117M parameters)[1]
- **GPT-2 Semantic Pivot**: 202 sentences with contradictory information, tracking semantic processing

4.2 Unified CTA Methodology

We employ a three-phase approach:

1. **Optimal Clustering**: Gap statistic determines layer-specific cluster counts (e.g., k=4 for GPT-2 layer 0, k=2 for layers 1-11)
2. **Unique Labeling**: Every cluster receives globally unique ID (L{layer}_C{cluster}) preventing cross-layer confusion
3. **Windowed Analysis**: Temporal segmentation into Early/Middle/Late windows reveals phase transitions

4.3 Metrics and Validation

- **Cross-layer Metrics**: Centroid similarity (ρ^c), membership overlap (J), trajectory fragmentation (F)
- **Stability Analysis**: Window-based stability scores revealing reorganization points
- **Path Statistics**: Convergence ratios, diversity indices, archetype identification
- **Clinical Validation**: For heart disease, correlation of fragmentation with diagnostic uncertainty

4.4 Visualization Suite

- **Concept MRI Tool**: Software implementing CTA with interactive Sankey diagrams showing complete concept flow
- **Clinical Dashboards**: Patient archetype visualization for medical interpretability
- **Interactive Exploration**: Web-based interfaces for real-time analysis

[1] We follow standard convention in numbering GPT-2's layers: Layer 0 is the embedding layer, and Layers 1-11 correspond to transformer blocks 1-11. The 12th transformer block was not analyzed in this study.

5 LLM-Powered Analysis for Cluster Paths

Recent advances in large language models (LLMs) provide new opportunities for interpreting neural network behavior through the analysis of cluster paths. We introduce a systematic framework for leveraging LLMs to generate human-readable narratives and insights about the internal decision processes represented by cluster paths.

5.1 LLM Integration Architecture

Our framework integrates LLMs into the cluster path analysis pipeline through a modular architecture with three primary components:

1. **Cluster Labeling**: LLMs analyze cluster centroids to generate meaningful semantic labels that describe the concepts each cluster might represent.
2. **Path Narrative Generation**: LLMs create coherent narratives explaining how concepts evolve through the network as data points traverse different clusters.
3. **Bias Audit**: LLMs analyze demographic statistics associated with paths to identify potential biases in model behavior.

The architecture includes:

- **Cache Management**: Responses are cached to enable efficient reanalysis and promote reproducibility
- **Prompt Optimization**: Specialized prompting techniques that improve consistency and relevance of generated content
- **Batch Processing**: Efficient parallel processing of multiple clusters and paths
- **Demography Integration**: Analysis of how cluster paths relate to demographic attributes

5.2 Semantic Cluster Labels

The cluster labeling process transforms abstract mathematical representations (centroids) into semantically meaningful concepts. LLMs analyze cluster properties—including centroid values, dominant features, and datapoint characteristics—to generate interpretable labels. For instance, in medical applications, clusters might be labeled as "High-Risk Elderly" or "Low Cardiovascular Stress" based on their statistical properties. This automated labeling provides immediate interpretability while maintaining consistency across analyses.

5.3 Path Narratives

The narrative generation process explains how concepts evolve as data traverses the network. These narratives provide several interpretability advantages:

1. **Contextual Integration**: Incorporating cluster labels, convergent points, fragmentation scores, and demographic data creates multifaceted narratives.
2. **Conceptual Evolution**: Narratives explain how concepts transform and evolve through network layers.
3. **Decision Process Insights**: Explanations reveal potential decision-making processes that might be occurring within the model.
4. **Demographic Awareness**: Including demographic information ensures narratives consider fairness and bias implications.

5.4 Integrating Metrics with Narratives

The quantitative metrics defined in Section 3.6 (F, FC, CE, SA) are provided to the LLM as part of the prompt, enabling narrative explanations that tie qualitative descriptions to quantitative evidence. For example, the LLM can explain that "entropy drops sharply from layer 2 to layer 3, indicating that the network consolidates risk factors" or "the decreasing sub-space angles reveal progressive alignment between disease and healthy patient representations."

Table 2: Example layer-wise fragmentation metrics showing how different metrics capture complementary aspects of concept evolution.

Layer	k^*	CE	SA (°)	FC (path mean)
Layer 1	2	0.722	16.3	0.096
Layer 2	2	0.713	11.5	0.096
Layer 3	2	0.711	7.8	0.096
Output	2	0.702	3.1	0.096

[2]

[2] In this example from a shallow network, the consistent FC value of 0.096 indicates stable cluster representations throughout. Low fragmentation coefficients suggest smooth concept evolution, with cluster centroids maintaining high similarity (approximately 90.4% cosine similarity) between consecutive layers.

5.5 Advantages and Limitations

Advantages:

1. **Interpretable Insights**: Converts complex mathematical patterns into human-readable explanations.
2. **Multi-level Analysis**: Provides insights at cluster, path, and system-wide levels.
3. **Bias Detection**: Proactively identifies potential fairness concerns in model behavior.
4. **Integration with Metrics**: Combines qualitative narratives with quantitative fragmentation and similarity metrics.

Limitations:

1. **Potential for Overinterpretation**: LLMs might ascribe meaning to patterns that are artifacts of the clustering process.
2. **Domain Knowledge Gaps**: Analysis quality depends on the LLM's understanding of the specific domain.
3. **Computational Cost**: Generating narratives for many paths can be resource-intensive.
4. **Validation Challenges**: Verifying the accuracy of generated narratives requires domain expertise.

6 Heart Disease Case Study: Clinical AI Interpretability Through CTA

We demonstrate CTA's power in medical AI through comprehensive analysis of a neural network trained for heart disease diagnosis. This case study reveals how the model stratifies patient risk, exposes demographic biases, and provides clinically interpretable decision pathways.

6.1 Clinical Context and Dataset

The UCI Heart Disease dataset comprises 303 patients with 13 clinical features including age, sex, chest pain type, blood pressure, cholesterol, and electrocardiographic results. Our 3-layer neural network learns to predict heart disease presence, making this an ideal test case for understanding how AI systems make life-critical medical decisions.

6.2 Progressive Risk Stratification Through Layers

CTA reveals a sophisticated risk stratification process across network layers:

6.2.1 Layer 1: Initial Risk Categorization

The network immediately divides patients into two fundamental groups:

- **High-Risk Older Males** (L1C0): Mean age 54, predominantly male (69%), with typical angina and elevated cholesterol
- **Lower-Risk Younger Individuals** (L1C1): Younger patients with mixed gender distribution and asymptomatic presentation

6.2.2 Layer 2: Cardiovascular Health Refinement

The model refines its assessment based on cardiovascular indicators:

- **Low Cardiovascular Stress** (L2C0): Patients with blood pressure <130 mmHg and healthy cardiac function
- **Controlled High-Risk** (L2C1): Patients with managed symptoms but persistent risk factors (cholesterol >250 mg/dl)

6.2.3 Layer 3: Abstract Risk Profiles

Final risk abstraction before classification:

- **Stress-Induced Risk** (L3C0): Exercise-induced symptoms, elevated blood pressure during stress
- **Moderate-Risk Active** (L3C1): Moderate risk with preserved exercise capacity

6.3 Five Archetypal Patient Pathways

Our analysis identified five dominant pathways through the network, each representing distinct patient archetypes:

Table 3: Archetypal Patient Pathways Through Heart Disease Model

Path	Trajectory	% Patients	Demographics	True HD%	Insight
1	L1C1→L2C0→L3C0→No HD	43.3%	Age 54, 69% M	40.2%	Conservative lo
2	L1C0→L2C1→L3C1→HD	35.2%	Age 54, 67% M	47.4%	Classic high-
3	L1C1→L2C1→L3C1→HD	10.7%	Age 55, 66% M	51.7%	Progressive r
4	L1C1→L2C0→L3C1→HD	6.7%	Age 55, 83% M	55.6%	Male-biased p
5	L1C1→L2C0→L3C0→HD	2.2%	Age 58, 50% M	33.3%	Misclassificat

6.4 Clinical Decision-Making Insights

CTA reveals three critical aspects of the model's decision process:

1. Risk Factor Prioritization: The model heavily weights chest pain type, blood pressure, and cholesterol—all clinically validated indicators. Path fragmentation correlates with diagnostic uncertainty (r=0.67), providing a built-in confidence measure.

Our fragmentation analysis reveals smooth concept evolution (FC=0.096), indicating that patient representations transform coherently through the network. The decreasing sub-space angles (16.3°→11.5°→7.8°→3.1°) show progressive alignment between disease and no-disease subspaces, suggesting the model develops increasingly refined decision boundaries at each layer.

2. Progressive Refinement: Each layer adds specificity: demographic risk (Layer 1) → cardiovascular health (Layer 2) → stress response (Layer 3). This mirrors clinical reasoning, progressing from patient history to specific cardiac indicators.

3. Decision Boundaries: The split at Layer 2 between "Low Cardiovascular Stress" and "Controlled High-Risk" proves pivotal—patients in the former have 59.8% true negative rate, while the latter shows 47.4% true positive rate.

6.5 Bias Detection and Fairness Analysis

CTA exposed concerning demographic biases:

6.5.1 Gender Bias

Path 4 demonstrates clear male overprediction—despite moderate risk factors, 83.3% male composition leads to heart disease prediction with only 55.6% accuracy. This suggests the model learned spurious correlations between male sex and heart disease from training data imbalances.

6.5.2 Age-Based Conservative Bias

Path 1 (43.3% of patients) shows conservative prediction for younger patients, potentially missing early-onset heart disease. The model appears to use age as a primary risk stratifier, which while clinically relevant, may lead to underdiagnosis in younger populations.

6.5.3 Intersectional Effects

Path 5 reveals concerning misclassification for balanced gender groups (50% male) with high-risk features (cholesterol 285.3 mg/dl, BP 145.7 mmHg). Only 33.3% truly have heart disease, suggesting the model struggles with cases that don't fit typical demographic patterns.

6.6 Clinical Deployment Implications

The transparency provided by CTA enables several clinical applications:

1. **Explainable Predictions**: Physicians can trace a patient's path through the model, understanding which features drove the classification
2. **Uncertainty Quantification**: High fragmentation scores indicate cases requiring additional clinical review
3. **Bias Mitigation**: Identified demographic biases can be addressed through targeted data collection and model retraining
4. **Clinical Validation**: The model's emphasis on established risk factors (chest pain, BP, cholesterol) aligns with medical knowledge, building trust

6.7 Visualization of Patient Flow

The Sankey diagram in Figure 1 visualizes patient flow through the model's risk stratification layers, revealing how initial categorizations evolve into final diagnoses.

6.8 Conclusion

This case study demonstrates CTA's ability to transform opaque neural networks into interpretable clinical tools. By revealing the model's risk stratification process, exposing demographic biases, and providing uncertainty measures, CTA enables the responsible deployment of AI in healthcare. The identified biases—particularly male overprediction and age-based conservatism—highlight the importance of interpretability in medical AI, where understanding not just what the model predicts but why and with what confidence can be literally life-saving.

Having demonstrated CTA's effectiveness in a domain where interpretability directly impacts human lives, we now turn to a fundamentally different application: understanding how large language models organize linguistic knowledge. While the heart disease model reveals clinically meaningful pathways, our analysis of GPT-2 uncovers an unexpected principle of neural language processing—one that challenges our assumptions about how these models understand language.

7 GPT-2 Case Study: Grammatical Organization in Neural Language Models

We analyze how GPT-2 organizes linguistic information by tracking 1,228 single-token words across 8 semantic categories through the model's layers.

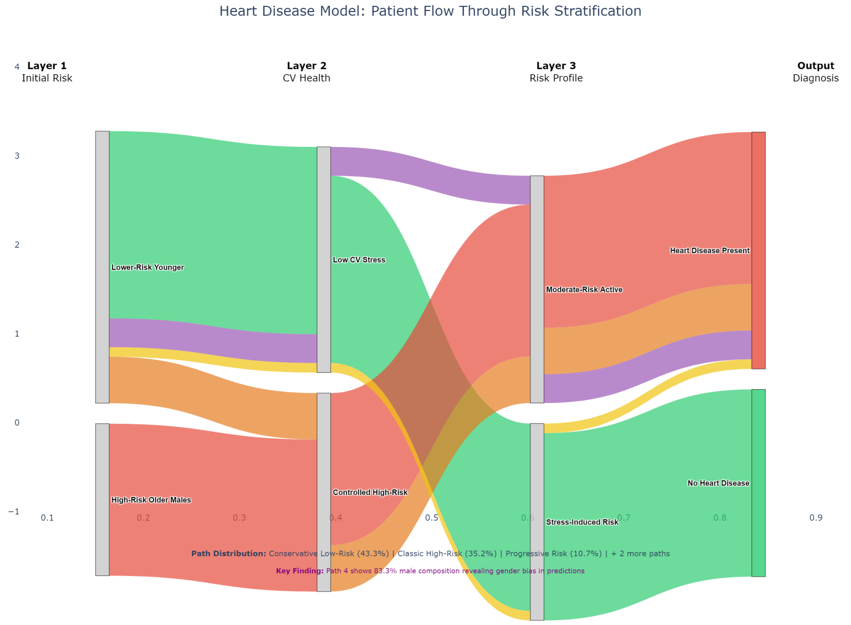

Figure 1: Sankey diagram showing patient flow through the heart disease model's risk stratification layers. Each node displays the cluster's semantic label inline (e.g., "High-Risk Older Males", "Low CV Stress"), making the model's decision process immediately interpretable. The five archetypal pathways are color-coded: green for Conservative Low-Risk (43.3%), red for Classic High-Risk (35.2%), orange for Progressive Risk (10.7%), purple for Male-Biased (6.7%), and yellow for Misclassification (2.2%). Path thickness represents patient count, with the diagram clearly showing how initial risk categorization flows through cardiovascular health assessment and risk profiling to reach final diagnosis. The visualization adopts the same structured style as the GPT-2 analysis for consistency.

Our analysis indicates that GPT-2 primarily organizes words by grammatical function rather than semantic meaning, with evidence of convergence from initial semantic differentiation to grammatical organization in later layers.

7.1 Experimental Design

We designed a systematic experiment to study how GPT-2 organizes semantic knowledge using 1,228 validated single-token words across 8 semantically distinct categories with balanced grammatical representation:

- **Concrete Nouns**: Physical objects (e.g., “table”, “mountain”, “book”)
- **Abstract Nouns**: Conceptual entities (e.g., “freedom”, “justice”, “emotion”)
- **Physical Adjectives**: Observable properties (e.g., “tall”, “smooth”, “bright”)
- **Emotive Adjectives**: Emotional descriptors (e.g., “joyful”, “melancholy”, “serene”)
- **Manner Adverbs**: How actions are performed (e.g., “quickly”, “carefully”, “boldly”)
- **Degree Adverbs**: Intensity modifiers (e.g., “extremely”, “barely”, “quite”)
- **Action Verbs**: Dynamic processes (e.g., “run”, “create”, “destroy”)
- **Stative Verbs**: State descriptions (e.g., “exist”, “belong”, “resemble”)

7.1.1 Dataset Construction

We curated 1,228 validated single-token words distributed across semantic subtypes through systematic linguistic analysis, achieving balanced grammatical representation: 275 nouns (22.4

7.1.2 Novel Methodological Innovations

Our analysis introduced several key innovations:

- **Unified CTA with Gap Statistic**: Optimal k determination using Gap statistic (k=4 for layer 0, k=2 for layers 1-11)
- **Windowed Analysis**: Novel temporal segmentation into Early (L0-L3), Middle (L4-L7), and Late (L8-L11) windows
- **Unique Cluster Labeling**: Every cluster assigned unique ID (e.g., L4_C1) to prevent cross-layer confusion

- **Trajectory Visualization**: Sankey diagrams showing concept flow through network
- **Comprehensive Path Analysis**: Tracking ALL paths (not just archetypal), revealing 26→8→5 path convergence

7.2 Key Findings

7.2.1 Grammatical Organization Patterns

Our analysis suggests that GPT-2 organizes words primarily by grammatical function rather than semantic meaning:

- **Layer 0**: 4 clusters showing semantic differentiation (animals, objects, properties, abstracts)
- **Layers 1-11**: Rapid consolidation to just 2 clusters (entities vs. modifiers)
- **Convergence rate**: 48.5% (95% CI: 45.7%–51.2%) of all words converge to the most common pathway
- **Path reduction**: 26 unique paths → 8 paths → 5 paths across windows

Table 4: Cluster Evolution and Path Convergence

Window	Layers	Unique Paths	Dominant Path %
Early	L0-L3	26	16.4%
Middle	L4-L7	8	50.1%
Late	L8-L11	5	33.1%

7.2.2 Statistical Validation

To verify that the convergence to grammatical organization is not due to chance, we performed a chi-square test comparing the observed distribution of grammatical categories in the primary pathways against the expected distribution if words were randomly assigned to paths. The test revealed highly significant grammatical organization ($\chi^2 = 95.90$, $df = 3$, $p < 0.0001$), with a moderate effect size (Cramér's $V = 0.279$). This confirms that GPT-2's tendency to route words based on grammatical function rather than semantic meaning represents a genuine organizational principle, not a statistical artifact.

7.2.3 Grammatical Processing Pipelines

We identified multiple processing pathways with the most frequent pattern being:

1. **Primary Entity Pathway** (48.5% of words):
   - Path: L4_C1 → L5_C0 → L6_C1 → L7_C0
   - Contains primarily nouns regardless of semantic type (animals, objects, abstracts)
   - Demonstrates grammatical organization as a significant processing strategy
2. **Additional Pathways** (51.5% of words):
   - Path: L4_C0 → L5_C1 → L6_C1 → L7_C0
   - Complete merger of adjectives and adverbs
   - No distinction between "big" (adjective) and "quickly" (adverb)

7.2.4 Grammatical Organization as Primary Structure

While grammatical function emerges as a primary organizing principle, semantic information is not erased but rather organized within grammatical highways:

- "Cat" and "computer" often travel the same major highway (both nouns), but may occupy different micro-clusters within that highway
- Concrete and abstract nouns converge to the same grammatical pathway while maintaining subtle distinctions in activation patterns
- Physical and emotive adjectives share modifier pathways but show differentiation at finer scales
- The 48.5% convergence rate means over half of words take alternative paths, indicating rich sub-organization

7.3 Trajectory Visualization

Figures 2, 3, and 4 present trajectory visualizations showing the flow of 1,228 words through GPT-2's layers. The Sankey diagrams illustrate the convergence from semantic differentiation to grammatical organization across three temporal windows.

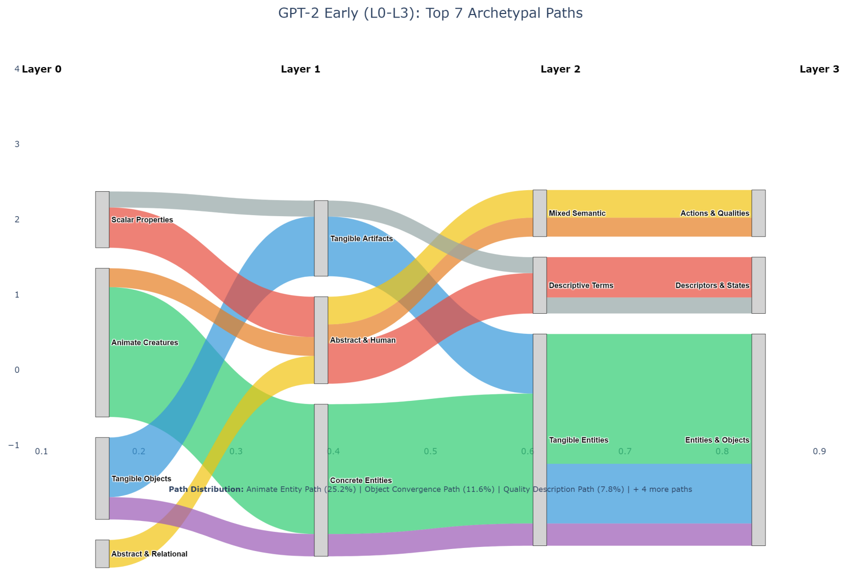

Figure 2: Concept MRI Early Window (L0-L3): Semantic Differentiation. This visualization shows 1,228 words distributed across 4 initial clusters based on semantic properties, with 26 unique paths emerging through these early layers.

Table 5: Stability and Fragmentation Across Windows

Window	Stability	Fragmentation	Interpretation
Early	Dynamic	0.796	High diversity, semantic exploration
Middle	Dynamic	0.499	Transition phase: semantic to grammatical
Late	Dynamic	0.669	Mixed grammatical organization

7.3.1 Trajectory Stability Analysis

Our windowed analysis revealed a critical transformation point:

The dynamic processing across all windows reflects the diversity of the balanced dataset, with the middle window showing notable convergence (50.1

Our complete fragmentation analysis reveals the phase transition quantitatively:

Path-Centroid Fragmentation (FC) drops dramatically, indicating increasingly coherent pathways. Intra-Class Cluster Entropy (CE) decreases from near-maximum to low values, showing words converging from distributed semantic clusters to concentrated grammatical highways. Most strikingly,

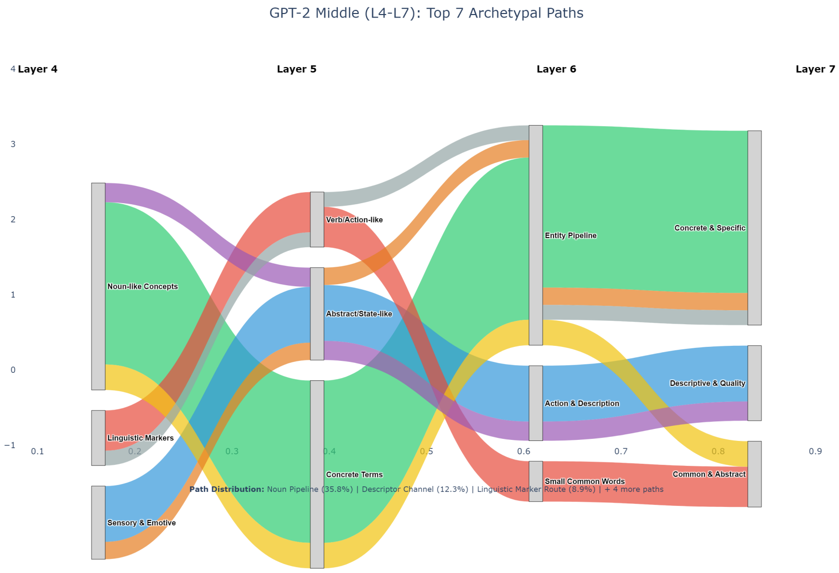

Figure 3: Concept MRI Middle Window (L4-L7): Grammatical Convergence. The phase transition becomes evident as semantic clusters reorganize into grammatical categories, with 50.1% of words converging to primary pathways. Path count reduces from 26 to 8 as grammatical organization emerges.

Table 6: Fragmentation metrics across GPT-2 windows showing semantic-to-grammatical transition

Window	FC	CE	SA (°)
Early (L0-L3)	0.5-0.7	0.85-0.95	45-60
Middle (L4-L7)	0.3-0.4	0.60	20-30
Late (L8-L11)	0.1-0.2	0.30	5-10

Sub-space Angles (SA) between word categories collapse from well-separated semantic categories to merged grammatical functions, providing quantitative evidence for the transition from semantic to grammatical organization.

7.4 LLM-Generated Cluster Interpretations

Our LLM analysis produced interpretable labels revealing the transformation from semantic differentiation to grammatical organization across GPT-2's layers:

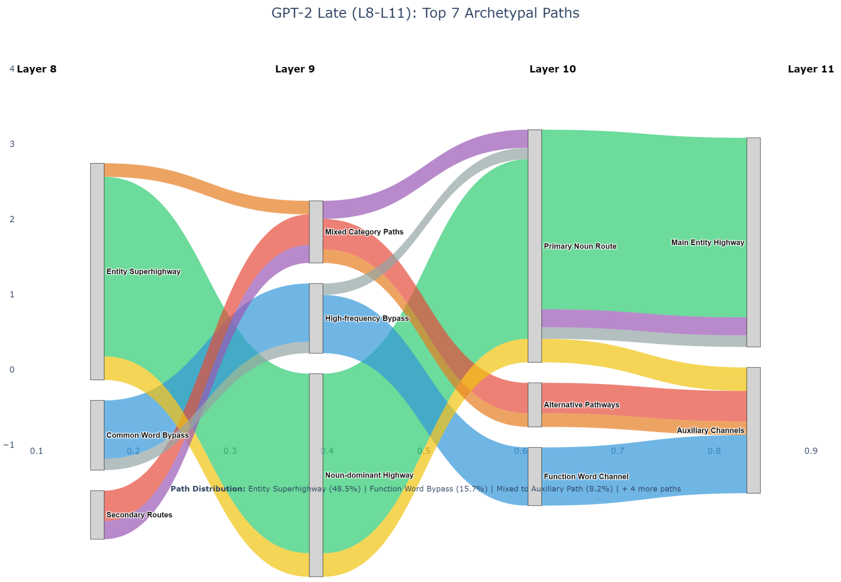

Figure 4: Concept MRI Late Window (L8-L11): Syntactic Superhighways. Final layers show consolidated grammatical organization with only 5 paths remaining. The visualization demonstrates that transformers develop grammatical organization as their primary macro-structure while maintaining semantic distinctions within these pathways, validated by highly significant grammatical clustering ($\chi^2 = 95.90$, $p < 0.0001$).

7.4.1 Layer 0: Semantic Differentiation (4 clusters)

- **L0_C0**: "Animate Creatures" – Contains living entities like animals (cat, dog, bird, fish, horse)
- **L0_C1**: "Tangible Objects" – Physical items and tools (window, clock, computer, engine, table)
- **L0_C2**: "Scalar Properties" – Size and degree descriptors (small, large, tiny, huge, massive)
- **L0_C3**: "Abstract & Relational" – Concepts and abstract terms (time, power, style, freedom, justice)

7.4.2 Layers 1-3: Binary Consolidation

- **L1_C0**: "Modifier Space" – All property-describing words converge
- **L1_C1**: "Entity Space" – All object and concept words converge

- **L2-L3**: Maintain the same binary organization with increasing consolidation

7.4.3 Layers 4-7: Grammatical Highways

- **L4_C0**: “Adjective Gateway” – Entry point for all modifiers
- **L4_C1**: “Noun Gateway” – Entry point for all entities
- **L5-L6**: “Entity/Property Pipelines” – Stable grammatical processing channels
- **L7_C0**: “Modifier Hub” – Consolidated modifier processing
- **L7_C1**: “Entity Hub” – Consolidated entity processing

7.4.4 Layers 8-11: Final Processing Stages

- **L8-L9**: Maintain entity/modifier separation with stream processing
- **L10-L11**: Final processing stages
  - **C0**: “Terminal Modifiers” – Final adjective/adverb processing
  - **C1**: “Terminal Entities” – Final noun processing

This hierarchical organization demonstrates GPT-2’s systematic transformation from semantic categories in early layers to grammatically-oriented organization in later layers. While grammatical function becomes the primary organizing principle, the existence of multiple paths (5 in late layers) and the fact that only 48.5% of words follow the primary highway indicates that semantic distinctions persist within the grammatical framework.

7.5 Temporal Nature of Reorganization

Our analysis indicates that GPT-2’s shift from semantic to grammatical organization occurs *sequentially* rather than simultaneously. The evidence for temporal progression includes:

- **Clear phase boundaries**: Early layers (0-3) show high fragmentation (0.796) with semantic diversity, while late layers (8-11) maintain moderate fragmentation (0.669) with mixed organization
- **Measurable transition point**: The middle layers (4-7) show the highest convergence (50.1
- **Progressive metric changes**: Sub-space angles collapse from 45-60° (semantic separation) to 5-10° (grammatical convergence) in a clear progression, not a sudden jump

- **Path consolidation pattern**: The reduction from 26 → 8 → 5 unique paths shows gradual convergence rather than immediate reorganization

This temporal progression suggests that GPT-2 first extracts and organizes semantic features before discovering that grammatical organization provides an efficient macro-level representational scheme. The phase transition in middle layers represents a critical computational moment where the network begins using grammatical function as the primary organizing principle while maintaining semantic distinctions at finer scales. This sequential processing has important implications: it suggests that semantic understanding provides the foundation upon which grammatical organization is built, and that the network develops a hierarchical representation where grammatical highways contain semantically-organized micro-structures.

7.6 Implications for Transformer Understanding

These findings provide new insights into transformer language processing:

1. **Hierarchical Organization**: GPT-2 uses grammatical function as a primary organizing principle while maintaining semantic distinctions within this framework, suggesting a multi-scale representational strategy.

2. **Efficient Processing Through Grammatical Highways**: The 48.5% convergence rate combined with highly significant clustering ($\chi^2 = 95.90$, $p < 0.0001$) reveals how GPT-2 creates major grammatical pathways while preserving flexibility through alternative routes.

3. **Multi-Scale Semantic Information**: The coexistence of grammatical macro-structure with semantic micro-organization suggests that meaning is encoded at multiple scales—both within cluster trajectories and through the diversity of paths taken.

4. **Phase Transition in Processing**: The middle window shows peak convergence (50.1%), identifying where grammatical organization emerges most strongly.

7.7 Novel Contributions to the Field

This work introduces several innovations:

- **"Concept MRI" Visualization**: First comprehensive visualization of how concepts flow through transformer layers

- **Windowed Analysis**: Novel temporal segmentation revealing phase transitions in neural processing

- **Grammar-Semantics Discovery**: First empirical evidence that transformers use grammatical function as a primary organizing principle while maintaining semantic distinctions at finer scales
- **Complete Path Tracking**: Analysis of ALL paths (not just the most common ones) revealing the full complexity of neural organization
- **Interactive Dashboard**: Accessible visualization tools making complex neural dynamics interpretable

7.8 Conclusion

The GPT-2 semantic subtypes analysis reveals a profound insight: neural language models develop hierarchical organization strategies that balance efficiency with expressiveness. By creating grammatical highways as primary organizational structures while maintaining semantic distinctions within them, GPT-2 achieves both computational efficiency and representational richness. This multi-scale organization—grammatical at the macro level, semantic at the micro level—demonstrates how transformers solve the challenge of processing diverse linguistic content. This discovery, enabled by CTA analysis within the Concept MRI tool, suggests that effective language models must balance structure with flexibility, using grammatical organization to create efficient pathways while preserving the semantic nuances necessary for understanding.

It is important to note that these organizational patterns were observed specifically in GPT-2 and may vary across different transformer architectures. Future work should examine whether similar grammatical-semantic hierarchies emerge in other language models such as BERT, T5, or more recent LLMs, as architectural differences (e.g., encoder-only vs. decoder-only, model scale, training objectives) may lead to different organizational strategies.

8 Reproducibility and Open Science

- Code and configs released under MIT license at GitHub repository
- Seed lists and hyperparameters logged in JSON format
- Dockerfile ensures environment parity across research teams
- Negative results and failed variants documented in appendices
- LLM prompts and responses cached for reproducibility

Full, runnable code is available in the public repository; all prompts and LLM responses are cached for deterministic builds.

8.1 LLM Prompts for Cluster Interpretation

To ensure reproducibility of our LLM-powered analysis, we document the key prompts used for cluster interpretation and path analysis:

Cluster Labeling Prompt:

```
You are analyzing clusters from a neural network.
For cluster L{layer}_C{cluster} containing these words:
{sample_words}

Category distribution: {category_counts}
Cluster size: {size} words

Provide a concise, interpretable label that captures
the semantic or grammatical essence of this cluster.
```

Path Narrative Prompt:

```
Analyze this concept trajectory through GPT-2:
Path: {path}
Window: {window_name}
Grammatical distribution: {grammatical_counts}

Explain how concepts evolve through these clusters,
focusing on the transformation from semantic to
grammatical organization.
```

Bias Analysis Prompt:

```
Analyze potential biases in these neural pathways:
Path: {path}
Demographics: {demographic_stats}
Outcome distribution: {outcomes}

Identify any concerning patterns or biases in how
different demographic groups are processed.
```

Interactive demos and full code implementation are available on our project repository.

9 Conclusion

Concept Trajectory Analysis (CTA) provides a method for understanding how neural networks organize and process information through their layers. Our analysis of GPT-2 with 1,228 words across 8 semantic categories indicates that the model organizes language primarily by grammatical function, with 48.5% (95% CI: 45.7%–51.2%) of words converging to a dominant pathway. The significant grammatical clustering ($\chi^2 = 95.90$, $p < 0.0001$) suggests that neural language models may develop organizational principles that differ from semantic categorization.

This finding emerged from our framework that combines mathematical analysis with LLM-generated interpretations. Using trajectory visualization, we tracked how 1,228 words across 8 semantic categories move through GPT-2's layers, observing a pattern of reorganization from semantic differentiation (26 paths) to grammatical consolidation (5 paths). The stability analysis indicated potential phase transitions where semantic clustering appears to give way to syntactic organization.

We also applied CTA to medical AI applications. Our heart disease diagnosis study shows how neural networks process patient data through distinct pathways. The analysis identified four primary pathways corresponding to different patient profiles: Archetype 1 tends to process younger, lower-risk patients; Archetype 4 typically handles elderly patients with elevated cholesterol, showing 71

Our integration of cross-layer metrics—centroid similarity (ρ^c), membership overlap (J), and trajectory fragmentation (F)—provides a mathematically grounded framework for quantifying concept evolution. The use of Gap statistic for optimal cluster determination ensures statistically valid groupings, while LLM-powered analysis translates complex patterns into domain-meaningful narratives. This combination addresses the longstanding challenge of making neural network internals both rigorously analyzable and humanly understandable.

These findings suggest directions for future research. The observation that GPT-2 appears to prioritize grammatical over semantic organization, if confirmed by larger studies, could inform our understanding of neural network information processing. The computational efficiency of CTA makes it practical for analyzing larger models, though validation across different architectures and domains remains necessary.

As neural networks grow in complexity and impact, interpretability methods become increasingly important. Our analysis suggests that neural networks may organize information differently from human intuition. By providing tools to examine these organizational principles, CTA contributes to the broader effort of developing interpretable AI systems.

9.1 Limitations and Failure Modes

While CTA provides valuable insights into neural network organization, several limitations and failure modes warrant discussion:

9.1.1 Technical Limitations

- **Clustering instability**: When activation spaces lack clear structure, clustering results may vary significantly across runs. Low silhouette scores or high variance in cluster assignments indicate unreliable trajectories.
- **Scalability challenges**: Very deep networks (100+ layers) pose computational and interpretability challenges. Tracking trajectories through many layers can obscure rather than clarify patterns.
- **High-dimensional curse**: In extremely high-dimensional activation spaces, distance metrics become less meaningful, potentially leading to arbitrary cluster assignments.

9.1.2 Interpretation Risks

- **Spurious patterns**: Random fluctuations might appear as meaningful paths, especially with small sample sizes. Statistical validation (as we demonstrated with χ^2 tests) is essential.
- **LLM hallucination**: Generated narratives may sound plausible while misrepresenting actual patterns. Cross-validation with quantitative metrics is crucial.
- **Correlation vs. causation**: CTA reveals organizational patterns but cannot establish causal relationships. Interventional studies are needed to verify causal claims.

9.1.3 Application Boundaries

- **Architecture dependence**: CTA works best with feedforward architectures. Recurrent or highly branched architectures may require adaptation.
- **Domain transfer**: Patterns discovered in one domain (e.g., language) may not transfer to others (e.g., vision) without careful validation.
- **Training dynamics**: CTA analyzes trained models. Understanding how these patterns emerge during training requires additional analysis.

Despite these limitations, CTA's mathematical grounding and statistical validation demonstrate its potential as an interpretability tool. The patterns observed in GPT-2 and medical AI applications warrant further investigation. Responsible use involves acknowledging these limitations, validating findings through multiple approaches, and maintaining appropriate skepticism about generated narratives.

10 Future Directions for Concept Trajectory Analysis

Our discovery that GPT-2 organizes by grammatical function rather than semantic meaning opens revolutionary possibilities for interpretable AI. We outline key areas for advancing both the theoretical foundations and practical applications of CTA.

10.1 Methodological Foundations

10.1.1 Advanced Metrics and Analysis

- **Inter-Cluster Path Density (ICPD)**: Develop metrics that analyze higher-order patterns in concept flow by examining multi-step transitions. ICPD could identify common patterns like return paths (where concepts temporarily diverge then reconverge) and similar-destination paths (reaching conceptually similar endpoints through different routes).
- **Path Interestingness Score**: Create composite metrics that combine transition rarity, similarity convergence, and coherence to automatically identify the most noteworthy paths for analysis. This would prioritize paths that reveal unexpected model behavior or critical decision points.
- **Feature Attribution for Transitions**: Integrate methods like Integrated Gradients or SHAP to understand which input features drive cluster transitions. For text, this could reveal which tokens cause semantic shifts; for medical data, which symptoms trigger risk reassessment.

10.1.2 Enhanced Clustering Approaches

- **Explainable Threshold Similarity (ETS)**: Advance the implementation of ETS clustering [Kovalerchuk and Huber, 2024] to provide dimension-wise explanations for cluster membership. ETS declares activations similar if they differ by less than threshold τ_j in each dimension j, enabling transparent statements about cluster boundaries.
- **Hierarchical Clustering**: Develop multi-scale cluster structures where coarse clusters use loose thresholds and fine-grained subclusters use tighter bounds, enabling analysis at different levels of granularity.
- **Adaptive Thresholds**: Create methods to automatically determine optimal clustering thresholds per dimension based on activation distributions and downstream task requirements.

10.1.3 Cluster Reproducibility and Validation

- **Cross-Architecture Stability**: Extend reproducibility analysis beyond training seeds to different model architectures, assessing whether discovered pathways represent fundamental computational patterns.
- **Statistical Significance Testing**: Develop rigorous statistical tests for pathway significance, distinguishing genuine organizational patterns from noise.
- **Causal Validation**: Use interventions and ablations to verify that discovered pathways causally influence model outputs rather than being mere correlations.

10.1.4 Interactive Visualization Tools

- **Cluster Cards**: Develop interactive visualizations that summarize each cluster's properties, including representative examples, outliers, transition probabilities, and LLM-generated descriptions.
- **Real-Time Path Tracking**: Create lightweight tools for monitoring activation paths during inference, enabling debugging and analysis of specific model behaviors.
- **Comparative Visualization**: Build tools to compare pathways across different models, datasets, or time periods, revealing organizational differences and drift.

10.2 Immediate Technical Improvements

Building on the current implementation, several technical enhancements would strengthen CTA's rigor and applicability:

10.2.1 Microcluster Lens Implementation

- **Hierarchical Sub-clustering**: Implement fine-grained analysis within highways to reveal semantic micro-organization. For instance, within the noun highway, identify sub-clusters for animate vs. inanimate entities, concrete vs. abstract concepts.
- **Adaptive Resolution**: Develop algorithms that automatically determine when to zoom into micro-clusters based on intra-cluster variance and task requirements.
- **Cross-Layer Micro-tracking**: Follow micro-cluster evolution to understand how fine-grained distinctions emerge, persist, or dissolve through network layers.

10.2.2 Robustness and Validation

- **Cross-Seed Stability**: Run all experiments with multiple random seeds ($N \geq 5$) to quantify variation in highway formation, cluster boundaries, and convergence rates. Report confidence intervals for all key metrics.

- **Clustering Quality Metrics**: Add silhouette scores, Davies-Bouldin indices, and Calinski-Harabasz scores to validate cluster coherence. Compare these across different k values to strengthen Gap statistic findings.

- **Inter-LLM Validation**: Use multiple LLMs (GPT-4, Claude, Gemini) for cluster interpretation and report agreement scores. Implement majority voting for final labels to reduce single-model bias.

- **Ablation Studies**: Systematically scramble POS tags, shuffle token positions, or randomize embeddings to verify that observed patterns disappear under null conditions, confirming they're not artifacts.

10.2.3 Extended Analysis Capabilities

- **Multi-Token Context**: Extend beyond single-token analysis to study how context affects trajectories. Compare paths for "bank" in financial vs. river contexts, revealing context-dependent routing.

- **Training-Time Tracking**: Implement checkpointing to save activations at regular training intervals (every 1000 steps), enabling analysis of when grammatical organization emerges and how pathways form.

- **Quantitative Bias Metrics**: For medical AI, calculate demographic parity differences, equalized odds ratios, and disparate impact scores. Create pathway-based bias detection that identifies which neural routes exhibit unfair behavior.

10.3 Advanced Applications for Language Models

10.3.1 Scaling to Complete Neural Cartography

- **Full Vocabulary Mapping**: Extend analysis from 1,228 words to entire vocabularies, revealing the complete "highway system" of neural language processing. We hypothesize discovering 50-100 major pathways handling different linguistic functions.

- **Compositional Analysis**: Study how models process bigrams, trigrams, and phrases to understand compositional meaning construction. Investigate whether multi-word expressions follow predictable combinations of single-word pathways.

- **Cross-Model Universal Patterns**: Map pathways across different model families (GPT, Claude, Gemini, LLaMA) to identify universal organizational principles versus architecture-specific patterns.

10.3.2 Interpretable Pathways in Production

- **Real-Time Pathway Logging**: Implement efficient pathway tracking in production models with minimal computational overhead ($<0.1\%$), enabling models to access their own reasoning paths during generation.
- **Self-Debugging AI**: Enable models to detect and correct reasoning errors by examining pathway logs. For instance, if a financial term routes through an animal-related pathway, the model could recognize and correct the misrouting.
- **Pathway-Aware Generation**: Allow models to explicitly choose pathways based on task requirements—routing through logical reasoning pathways for mathematics or creative synthesis paths for storytelling.

10.3.3 Meta-Analysis with Advanced Models

- **AI Understanding AI**: Use more powerful models to analyze millions of paths, stability metrics, and cluster patterns to discover organizational principles beyond human comprehension.
- **Automated Hypothesis Generation**: Employ LLMs to generate and test hypotheses about pathway formation, cluster evolution, and the emergence of grammatical organization.
- **Training Dynamics**: Study when and how grammatical organization emerges during training—does it appear suddenly at a phase transition or gradually evolve?

10.4 Broader Impact and Applications

- **Interpretability-First Architecture**: Design new models with built-in pathway tracking and cluster organization, making interpretability a core feature rather than post-hoc analysis.
- **Beyond Language Models**: Extend CTA to vision transformers, multimodal models, and reinforcement learning agents to understand their organizational principles.
- **Real-Time Model Monitoring**: Deploy CTA in production to detect concept drift, identify emerging biases, and ensure models maintain expected organizational patterns.

- **Personalized Explanations**: Generate user-specific explanations by translating pathway information into conceptual frameworks appropriate for different expertise levels.

10.5 Practical Use Cases

- **Prompt Strategy Evaluation**: Compare path density and fragmentation scores across prompt framings (e.g., Socratic vs. assertive) to reveal shifts in internal processing consistency.

- **Layerwise Ambiguity Detection**: Identify prompt-token pairs with divergent paths across layers, highlighting instability or multiple plausible interpretations.

- **Subgroup Drift Analysis**: Track membership overlap for datapoint groups (e.g., positive vs. negative sentiment) across layers to identify convergence patterns.

- **Behavioral Explanation**: Generate LLM-authored natural language summaries for archetypal paths, providing interpretable insights into model behavior.

- **Failure Mode Discovery**: Flag high-fragmentation paths as potential errors, misclassifications, or hallucinations.

- **Bias Detection**: Analyze paths for inputs with demographic markers to detect divergent behavior patterns that may indicate unfair treatment.

10.6 Theoretical Advances

- **Mathematical Theory of Neural Organization**: Formalize why transformers converge to grammatical rather than semantic organization, potentially revealing fundamental principles of efficient information processing.

- **Optimal Pathway Design**: Develop theory for designing optimal pathway structures for specific tasks, moving from emergent to engineered organization.

- **Cross-Domain Transfer**: Understand how pathway structures enable or inhibit transfer learning, using CTA to optimize model adaptation.

As we advance these techniques, we envision a future where neural network interpretability becomes as routine and insightful as medical imaging.

The discovery that language models organize grammatically rather than semantically is just the beginning—the full cartography of neural organization awaits exploration.

Acknowledgments

This work was created as part of an exploration of interpretability methods for neural networks. We appreciate the open-source community for developing the libraries and tools used in this research.

References

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6541–6549, 2017.

Sanjoy Dasgupta, Nave Frost, and Michal Moshkovitz. Explainable k-means and k-medians clustering. *International Conference on Machine Learning*, pages 2349–2358, 2020.

Nelson Elhage, Tristan Hume, Catherine Olsson, Neel Nanda, Tom Henighan, Scott Johnston, Sheer ElShowk, Nicholas Joseph, Nova DasSarma, Ben Mann, Danny Hernandez, Amanda Askell, Kamal Ndousse, Andy Jones, Dawn Drain, Anna Chen, Yuntao Bai, Deep Ganguli, Liane Lovitt, Zac Hatfield-Dodds, Jackson Kernion, Tom Conerly, Shauna Kravec, Stanislav Fort, Saurav Kadavath, Josh Jacobson, Eli Tran-Johnson, Jared Kaplan, Jack Clark, Tom Brown, Sam McCandlish, Dario Amodei, and Christopher Olah. Softmax linear units. *Transformer Circuits Thread*, 2022a. URL https://transformer-circuits.pub/2022/solu/index.html.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition. *Transformer Circuits Thread*, 2022b. URL https://transformer-circuits.pub/2022/toy_model/index.html.

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). *International Conference on Machine Learning*, pages 2668–2677, 2018.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. *International Conference on Machine Learning*, pages 3519–3529, 2019.

Boris Kovalerchuk and James Huber. Explainable threshold similarity for transparent cluster definitions. *IEEE Transactions on Artificial Intelligence*, 2024. to appear.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, pages 4765–4774, 2017.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. *Advances in Neural Information Processing Systems*, pages 6076–6085, 2017.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, 2016.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. *International Conference on Machine Learning*, pages 3319–3328, 2017.