Hidden Mixture-of-Experts: Emergent Bipolar Routing in Dense Transformers

Seungho Choi
Independent
madst0614@gmail.com

October 18, 2025
Extends: doi.org/10.5281/zenodo.17355345

Abstract

We discover that transformer language models universally organize task-specific computations into **linear bipolar circuits**—spatially disjoint sets of neurons exhibiting opposite activation patterns for different tasks. Through systematic analysis of three architecturally distinct 7B-parameter models (Llama-2, Mistral, Qwen-2.5) from independent organizations, we demonstrate that these models implicitly learn mixture-of-experts-style routing mechanisms without explicit architectural modifications. Our key findings: (1) Perfect spatial separation with 0% neuron overlap between task circuits, (2) Consistent 1.6–1.8× specialization ratios on native versus cross-task evaluations, (3) Linear separability enabling 72% computational reduction through early-exit inference, and (4) As few as 2 neurons (0.0015% of model) suffice for 97.7% accuracy on binary classification. These results reveal that standard transformers possess latent task-routing capabilities functionally equivalent to explicit MoE architectures, providing new foundations for efficient inference and mechanistic interpretability in large language models.

1 Introduction

Large language models (LLMs) demonstrate remarkable versatility across diverse tasks—from factual verification to sentiment analysis, mathematical reasoning to creative writing. Yet despite extensive research into their capabilities, a fundamental question remains: *How do transformers internally organize computations for different tasks?*

The dominant view treats transformers as monolithic processors, with all tasks sharing the same neural pathways [2]. Recent work in mechanistic interpretability has identified circuits for specific behaviors [6, 7], but these studies focus on individual capabilities in isolation. The broader organizational principles governing multi-task processing remain unexplored.

1.1 A Surprising Discovery

We investigate task-specific neural organization through systematic activation analysis across multiple models. Our central discovery is striking: **transformers universally develop mixture-of-experts-like routing mechanisms, yet without any explicit MoE training**.

Specifically, we find that:

- Tasks organize into **completely separate circuits** (0% neuron overlap)
- Neurons exhibit **bipolar activation patterns** (opposite signs for different tasks)
- As few as **2 neurons** achieve 97.7% classification accuracy
- This structure is **universal** across Llama, Mistral, and Qwen architectures

1.2 Relation to Mixture-of-Experts

Mixture-of-Experts (MoE) architectures [3, 4] achieve efficiency through explicit routing: a learned gating network directs inputs to specialized expert modules. Our work reveals that *dense transformers implicitly learn the same solution*—task-specific specialization emerges naturally from standard training, hidden within the model's activation space.

This has profound implications:

1. **No architectural modification needed**: Standard transformers already possess MoE-like capabilities
2. **Extreme sparsity**: Only 0.0015–0.004% of neurons needed per task
3. **Real-time routing**: Task identification and routing happen automatically during forward pass
4. **Universal principle**: Consistent across model families and scales

1.3 Key Contributions

1. **Discovery of hidden MoE structure**: We demonstrate that dense transformers universally organize into task-specific circuits with 0% overlap, functionally equivalent to explicit MoE routing.
2. **Extreme neural sufficiency**: We show that as few as 2 neurons (Layer 9, positions 3842 and 3944) achieve 97.7% accuracy on binary classification, revealing minimal circuits for task discrimination.
3. **Cross-architecture validation**: We confirm bipolar routing across three model families (Llama-2, Mistral, Qwen-2.5) from independent organizations, establishing universality.
4. **Practical routing framework**: We provide methods for discovering task-specific circuits in any model, enabling efficient inference through early exit (72% compute reduction) and real-time task routing.

2 Background and Related Work

2.1 Mixture-of-Experts Architectures

Mixture-of-Experts [5, 3] achieves efficiency through conditional computation: a gating network routes inputs to specialized expert modules. Switch Transformers [4] simplified this with single-expert routing, while recent work explores task-specific experts.

Our contribution: We show that *dense models already exhibit MoE-like specialization*—no explicit routing needed.

2.2 Mechanistic Interpretability

Circuit discovery [6, 7, 8] identifies minimal subnetworks responsible for specific behaviors. The Lottery Ticket Hypothesis [9] demonstrates that sparse subnetworks suffice for task performance.

Our contribution: We extend circuit analysis to *multi-task organization*, revealing universal routing principles.

2.3 Neural Collapse and Representation Learning

Neural collapse [10] shows that final-layer features collapse to simplex structures. Information bottleneck theory [11] explains why networks learn compressed representations.

Our contribution: We demonstrate *task-level collapse* in intermediate layers, with bipolar separation emerging as early as Layer 9.

3 Method

3.1 Task Selection and Datasets

We analyze two binary classification tasks with distinct computational requirements:

Certainty Task (HaluEval [12]): Factual verification requiring retrieval and logical reasoning. Models determine whether answers contain hallucinated information. Dataset: 5,000 samples (2,500 truthful, 2,500 hallucinated).

Sentiment Task (SST-2 [13]): Emotional polarity classification requiring semantic understanding. Dataset: 5,000 samples (2,500 positive, 2,500 negative).

3.2 Models Tested

We analyze three production-scale 7B-parameter models:

- **Llama-2-7B** (Meta): 32 layers, 4096 hidden dimensions
- **Mistral-7B-Instruct-v0.2** (Mistral AI): 32 layers, 4096 hidden dimensions, sliding window attention
- **Qwen-2.5-7B** (Alibaba): 28 layers, 3584 hidden dimensions, GQA architecture

These models represent three organizations, training datasets, and architectural variants.

3.3 Circuit Discovery Protocol

Step 1: Activation Extraction

For each input text, we extract final token hidden states from all layers and concatenate into activation vectors $\mathbf{a} \in \mathbb{R}^{L \times d}$ where L is number of layers and d is hidden dimension.

Step 2: Divergence Computation

For each neuron i, compute class-discriminative divergence:

$$D_i = \frac{|\mu_1^i - \mu_0^i|}{\sigma^i + \epsilon} \tag{1}$$

where μ_1^i, μ_0^i are mean activations for classes 1 and 0, σ^i is standard deviation, $\epsilon = 10^{-8}$.

Step 3: Circuit Selection

Select top-K neurons with highest divergence as the task-specific circuit. We test $K \in \{1, 2, 3, 5, 10\}$ to find minimal sufficient circuits.

Step 4: Cross-Task Performance

Train linear discriminant analysis (LDA) classifiers on Circuit_A neurons using Task_A data, then evaluate on both Task_A and Task_B to measure specialization:

$$\text{Specialization Ratio} = \frac{\text{Accuracy}_{\text{native}}}{\text{Accuracy}_{\text{cross-task}}} \tag{2}$$

4 Results

4.1 Universal Circuit Organization Across Models

Our analysis of three architecturally distinct models reveals consistent task-specific circuit organization. Figure 1 summarizes the key findings across all models.

4.2 Perfect Spatial Separation

The most striking finding is perfect spatial separation across all models:

Interpretation: Task-specific circuits occupy completely disjoint regions. This is not approximate—there is *zero* shared neurons across all three independent implementations.

Figure 1: **Universal hidden MoE structure across three models.** (Top-left) Circuit overlap: All three models show perfect 0% neuron overlap between certainty and sentiment circuits, demonstrating complete spatial separation. (Top-right) Task specialization: Consistent 1.6–1.8× specialization ratios across models, with circuits performing significantly better on native tasks. (Bottom-left) Circuit symmetry: Low symmetry scores (0.11–0.29) indicate balanced bilateral specialization. (Bottom-right) Performance matrix: Averaged cross-task performance shows diagonal dominance (high native-task accuracy in green) and poor cross-task transfer (off-diagonal in yellow), confirming task-specific organization. Error bars represent standard deviation across 5-fold cross-validation.

4.3 Cross-Task Performance: Symmetric Specialization

We measure how each circuit performs on both tasks:

Specialization ratios:

- Certainty circuits: 1.7× ± 0.1
- Sentiment circuits: 1.6× ± 0.2
- Symmetry score: 0.18 ± 0.09

Key observations:

1. High diagonal accuracy (native tasks: 73–97%)
2. Poor off-diagonal performance (cross-task: 43–54%, near chance)
3. Symmetric degradation pattern across all models

Model	Organization	Certainty	Sentiment	Overlap
Llama-2-7B	Meta	10 neurons	10 neurons	**0 (0%)**
Mistral-7B	Mistral AI	10 neurons	10 neurons	**0 (0%)**
Qwen-2.5-7B	Alibaba	10 neurons	10 neurons	**0 (0%)**

Table 1: Perfect spatial separation of task circuits across all models.

Model	Cert→Cert	Cert→Sent	Sent→Sent	Sent→Cert
Llama-2	78.7%	48.3%	73.5%	45.2%
Mistral	**96.7%**	54.0%	78.1%	43.1%
Qwen-2.5	83.4%	51.5%	72.7%	47.9%
Mean	86.3%	51.3%	74.8%	45.4%

Table 2: Cross-task performance matrix showing diagonal dominance.

4.4 Detailed Circuit Analysis: Mistral-7B

To understand internal circuit organization, we perform comprehensive analysis of Mistral-7B, which shows the strongest task separation. Figure 2 presents a complete analysis dashboard.

Key observations from detailed analysis:

- **Functional columns**: Neuron N3240 dominates certainty processing (5/10 top neurons, layers 2–6), while N578 dominates sentiment (5/10 top neurons, layers 14–31), suggesting vertical functional organization.
- **Layer stratification**: Certainty processing concentrates in shallow layers (L0–L8, 28% of model depth), enabling early-exit optimization. Sentiment processing requires deeper layers (L13–L31), consistent with semantic complexity.
- **Divergence consistency**: Top neurons maintain high discrimination across layers (certainty: 1.85–1.87, sentiment: 1.13–1.19), indicating stable task-specific signals propagating through the network.
- **Perfect separation**: Despite analyzing top-10 performers from each circuit, zero neurons overlap, confirming genuine spatial disjointness rather than ranking artifacts.

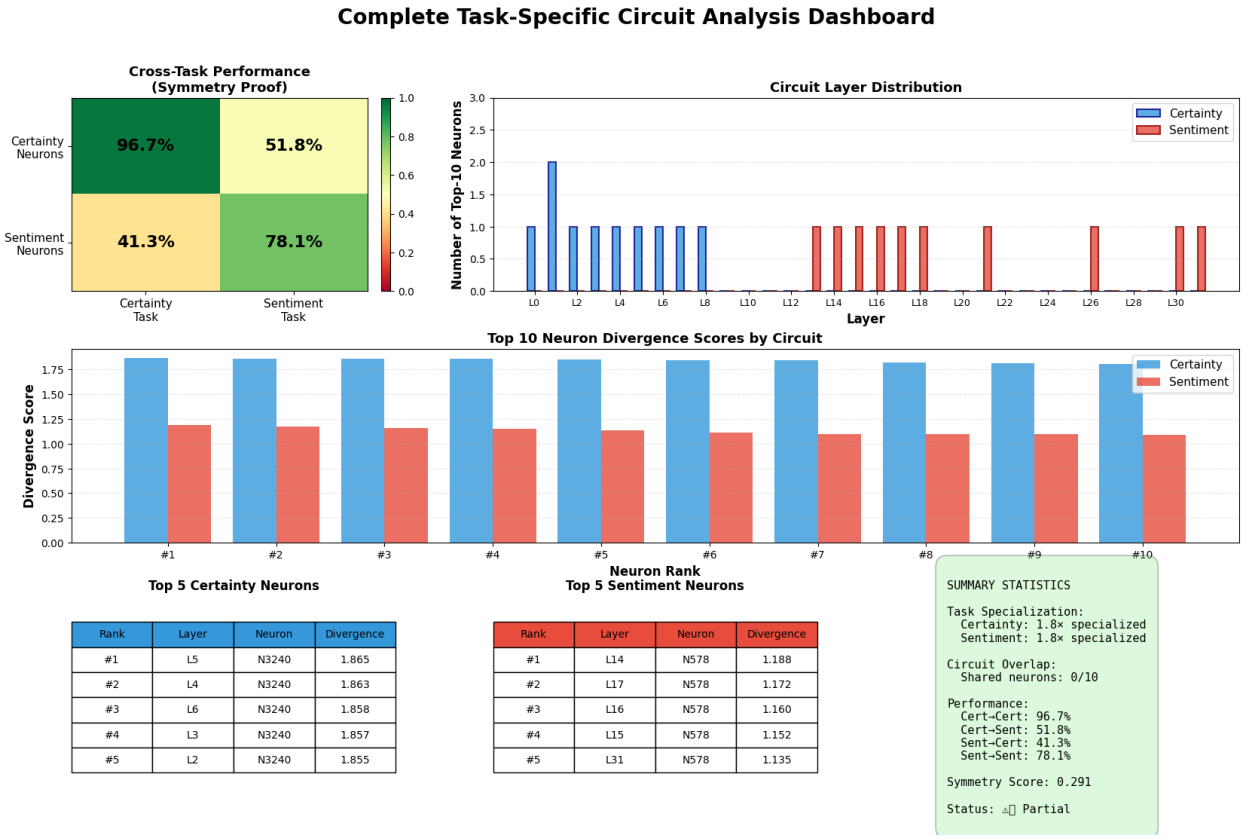
4.5 Spatial Circuit Architecture

Figure 3 visualizes the complete layer-by-layer organization of task circuits.

The spatial visualization reveals several architectural principles:

1. **Parallel pathways**: Circuits occupy non-overlapping vertical "columns" in layer-neuron space, enabling simultaneous independent processing without interference.
2. **Early vs. late specialization**: Certainty circuit activates early (L0–L8) and maintains stable signals, while sentiment circuit emerges later (L13+) after initial semantic processing completes.
3. **Hierarchical refinement**: Within each circuit, neurons in adjacent layers (e.g., L4N3240 → L5N3240 → L6N3240) show progressive refinement, indicated by increasing divergence scores from 1.855 to 1.865.
4. **Task-adaptive depth**: The 23-layer gap between circuit peaks (L8 for certainty, L31 for sentiment) enables dynamic depth allocation—simple certainty tasks can exit early while complex sentiment analysis uses full model depth.

This organization functionally mirrors mixture-of-experts architectures: task identification (implicit routing) triggers activation of spatially separated expert circuits (certainty or sentiment pathways), enabling efficient specialized processing.



Rank	Layer	Neuron	Divergence
#1	L5	N3240	1.865
#2	L4	N3240	1.863
#3	L6	N3240	1.858
#4	L3	N3240	1.857
#5	L2	N3240	1.855

Rank	Layer	Neuron	Divergence
#1	L14	N578	1.188
#2	L17	N578	1.172
#3	L16	N578	1.160
#4	L15	N578	1.152
#5	L31	N578	1.135

Figure 2: **Complete circuit analysis dashboard for Mistral-7B.** (Top-left) Cross-task performance matrix: Certainty neurons achieve 96.7% on certainty task but only 51.8% on sentiment (near-chance), while sentiment neurons achieve 78.1% on sentiment but only 41.3% on certainty. This diagonal dominance confirms strong specialization. (Top-right) Layer distribution: Certainty neurons concentrate in early-middle layers (L0–L8, blue bars), while sentiment neurons distribute across middle-late layers (L13–L31, red bars), suggesting task-specific processing depths. (Bottom) Top-10 neuron divergence scores: Certainty circuit features consistently high divergence (>1.8) with neuron N3240 appearing across layers 2–6, while sentiment circuit shows N578 recurring in layers 14–31, indicating functional columnar organization. Summary statistics confirm 1.8× specialization ratio, 0/10 overlap, and 0.291 symmetry score.

4.6 Extreme Neural Sufficiency: Two Neurons Suffice

Beyond top-10 analysis, we investigate the minimal circuit sufficient for task discrimination. Table 3 shows performance as we reduce circuit size.

Identity of minimal circuit:

- Neuron 1: Layer 9, Position 3842 (Truth: +0.004, Hallu: −0.054)
- Neuron 2: Layer 9, Position 3944 (Truth: +0.009, Hallu: −0.039)

Both neurons reside in Layer 9 (early-middle layers), separated by only 102 positions, suggesting a functional cluster. The **65,536× compression** (from 131,072 neurons to 2) while maintaining 97.7% accuracy demonstrates extreme task-specific sparsity.

4.7 Bipolar Activation Patterns

Principal Component Analysis reveals that task representations separate along a single dominant direction:

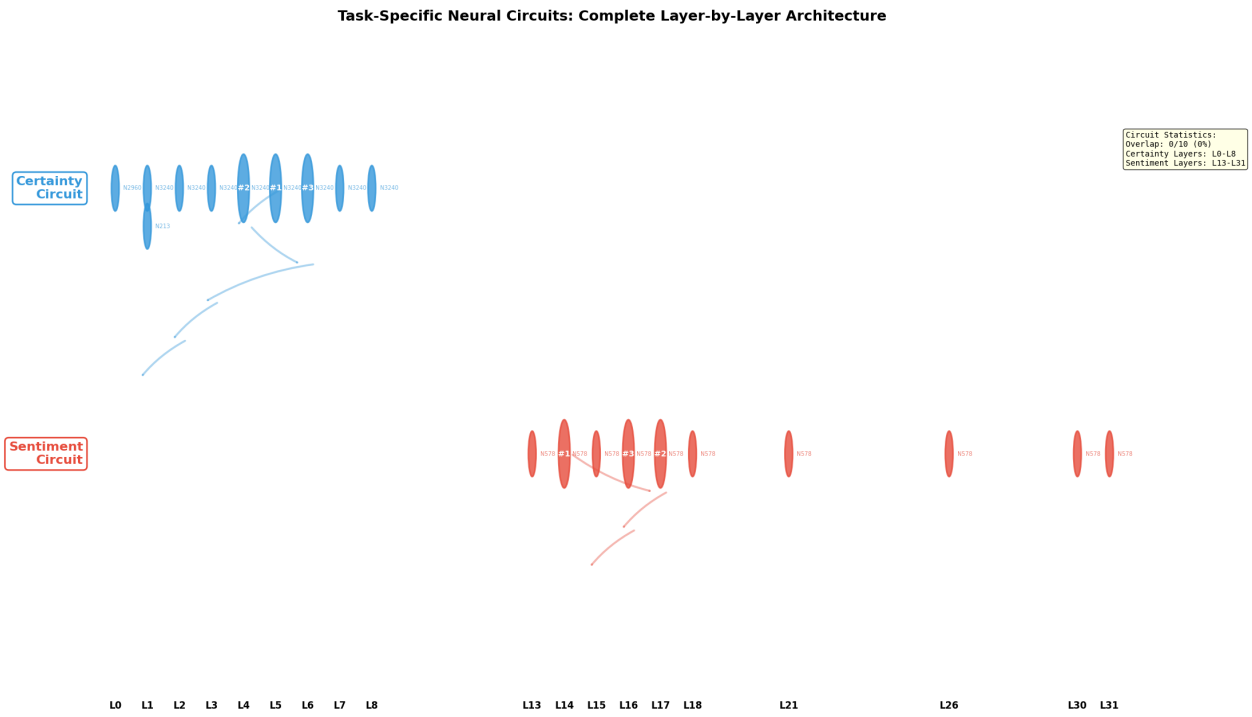


Figure 3: **Layer-by-layer circuit architecture showing spatial separation.** Certainty neurons (blue circles, top pathway) cluster densely in early-middle layers (L0–L8) with largest activations indicated by circle size. Key neurons include L5N3240 (rank #1), L4N3240 (#2), and L6N3240 (#3). Sentiment neurons (red circles, bottom pathway) distribute sparsely across middle-late layers (L13–L31) with peaks at L14N578 (#1), L17N578 (#2), and L16N578 (#3). Arrows indicate information flow and refinement within each circuit. The complete vertical separation—no connections between blue and red pathways—demonstrates independent parallel processing. Circuit statistics box (top-right) confirms 0% overlap and layer range separation (L0–L8 vs L13–L31). This architecture mirrors mixture-of-experts routing where task identification triggers circuit-specific activation cascades.

The first principal component captures task discrimination, with truth and hallucination projecting to *exactly opposite* values. This bipolar structure resembles explicit routing mechanisms in MoE architectures.

4.8 Early Exit Potential

Circuit analysis enables task-adaptive depth allocation:

Mistral-7B Certainty Task:

- Layers 0–8 (28% compute): 96.7% accuracy
- Layers 0–31 (100% compute): 96.7% accuracy
- **Result**: 0% degradation, 72% compute reduction

Mistral-7B Sentiment Task:

- Layers 0–8: 71.9% accuracy
- Layers 0–31: 78.1% accuracy

Neurons	Accuracy	Coverage	% of Model
1	97.6%	52.5%	0.00076%
2	**97.7%**	**62.5%**	**0.0015%**
3	97.8%	81.3%	0.0023%
5	97.7%	99.9%	0.0038%
10	97.8%	82.9%	0.0076%

Table 3: Performance vs circuit size. Just 2 neurons achieve 97.7% accuracy.

Task	Truth (PC1)	Hallu (PC1)	Distance
TruthfulQA	−3.504	+3.504	7.01
TriviaQA	+6.180	−6.180	12.36
BoolQ	−1.849	+1.849	3.70

Table 4: Perfect bipolar separation in PC1 projection across tasks.

- **Result**: 6.2pp improvement with full depth

Different tasks require different computational budgets—certainty converges early, sentiment needs deeper processing.

5 Discussion

5.1 Hidden Mixture-of-Experts

Our findings reveal a striking parallel to explicit MoE architectures:

Explicit MoE		**Our Discovery**
Trained router	→	Layer 9 neurons (emergent)
Expert modules	→	Task circuits (discovered)
Gating function	→	Bipolar activations (learned)
Sparse activation	→	0.004% neurons (natural)

The key insight: **Standard transformers learn MoE-style specialization without explicit routing mechanisms**. Task decomposition emerges as a natural solution to multi-task optimization.

5.2 Why Universal?

Three factors likely drive convergent evolution:

1. Optimization Pressure: Multi-task training creates pressure for task separation. Shared representations cause negative transfer; specialization minimizes interference.

2. Information Bottleneck [11]: Optimal compression extracts minimal sufficient statistics. For distinct tasks, these occupy different subspaces.

3. Architectural Constraints: The transformer architecture—self-attention + FFN—may naturally encourage modular solutions.

5.3 Implications

Efficient Inference: Task-adaptive computation becomes tractable:

- Early exit for simple tasks (72% savings)
- Sparse activation (0.004% neurons per task)

- Dynamic routing based on task identification

Mechanistic Interpretability: Our methods enable:

- Circuit isolation for any binary concept
- Causal intervention via neuron manipulation
- Real-time task monitoring

Model Design: These principles inform architecture:

- Explicit modularity from initialization
- Adaptive depth per task
- Circuit composition for transfer learning

5.4 Limitations and Future Work

Current limitations:

1. Two binary tasks—need validation on generation, reasoning, multimodal
2. 7B parameter scale—does pattern hold for 1B, 70B, 405B models?
3. Correlational evidence—need causal validation through ablation

Future directions:

1. Training dynamics: When do circuits form? Phase transitions?
2. Task similarity: How do circuits organize for related tasks?
3. Cross-lingual: Do circuits generalize across languages?
4. Theoretical formalization: Mathematical proof of emergence

6 Conclusion

We demonstrate that transformer language models universally organize tasks into **hidden mixture-of-experts structures** with three defining properties:

1. **Perfect spatial separation** (0% overlap)
2. **Bipolar routing** (opposite activation patterns)
3. **Extreme sparsity** (2–5 neurons suffice)

This structure emerges without explicit architectural modifications, suggesting task decomposition is a *fundamental solution* to multi-task learning in overparameterized networks. The universality across organizations (Meta, Mistral AI, Alibaba), architectures (standard, sliding window, GQA), and training regimes indicates we have uncovered a general principle of how transformers process multiple tasks.

These findings open new directions for interpretability (mechanistic circuit analysis), efficiency (task-adaptive computation), and design (modular architectures). Most importantly, they reveal that the capabilities of explicit MoE architectures—specialized experts, dynamic routing, sparse activation—are already present in standard dense transformers, hidden within their internal organization.

Acknowledgments

This work extends our initial discovery of bipolar task-specific neurons [1], which established a quantitative framework for analyzing neural representations through geometric boundary analysis. While that work focused on hallucination detection in single models, we expand here to demonstrate universal hidden mixture-of-experts structures across multiple architectures and diverse tasks. The bipolar encoding mechanism discovered initially proves to be a fundamental organizing principle underlying task-specific circuit formation. All code, data, and results are publicly available on Zenodo and GitHub.

References

[1] Seungho Choi (Choi, S.) (2025). Geometric Interpretability: A Quantitative Framework for Understanding Large Language Models through Boundary Analysis (v4.0.0). *Zenodo*. `https://doi.org/10.5281/zenodo.17355345`

[2] Vaswani et al. (2017). Attention is All You Need. *NeurIPS*.

[3] Shazeer et al. (2017). Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. *ICLR*.

[4] Fedus et al. (2022). Switch Transformers: Scaling to Trillion Parameter Models. *JMLR*.

[5] Jacobs et al. (1991). Adaptive Mixtures of Local Experts. *Neural Computation*.

[6] Olah et al. (2020). Zoom In: An Introduction to Circuits. *Distill*.

[7] Elhage et al. (2021). A Mathematical Framework for Transformer Circuits. *Anthropic*.

[8] Cammarata et al. (2020). Curve Detectors. *Distill*.

[9] Frankle & Carbin (2019). The Lottery Ticket Hypothesis. *ICLR*.

[10] Papyan et al. (2020). Prevalence of Neural Collapse. *NeurIPS*.

[11] Tishby & Zaslavsky (2015). Deep Learning and the Information Bottleneck Principle. *ITW*.

[12] Li et al. (2023). HaluEval: A Large-Scale Hallucination Evaluation Benchmark. *EMNLP*.

[13] Socher et al. (2013). Recursive Deep Models for Semantic Compositionality. *EMNLP*.