SFC-Score: A Unified Metric Framework Balancing Sparsity, Fidelity, and Mechanistic Completeness for Interpretability Evaluation

Anonymous Author(s)

ABSTRACT

Mechanistic interpretability (MI) methods decompose neural network activations into interpretable features, yet no existing metric jointly evaluates the three critical desiderata: sparsity, fidelity, and mechanistic completeness. We present SFC-Score, a unified evaluation framework based on the weighted harmonic mean of these three axes. The harmonic mean formulation ensures that catastrophic failure on any single axis dominates the joint score, reflecting the practical requirement that useful decompositions must be adequate on all dimensions simultaneously. We formalize individual axis metrics—sparsity as the fraction of inactive features, fidelity as reconstruction agreement, and completeness as behavioral variance preserved under ablation—and define a Pareto dominance relation with hypervolume indicator for comparing method families. On synthetic benchmarks with planted ground-truth circuits across four model configurations (circuit sizes 4–24, hidden dimensions 64–128), we demonstrate that the SFC-Score at equal weights peaks at sparsity level 0.85 with a score of 0.905 on the standard model, meaningfully separating decomposition quality. Weight sensitivity analysis across seven preference profiles shows that the optimal decomposition shifts predictably: sparsity-heavy (5:1:1) preferences select 95% sparsity (score 0.917), while fidelity-heavy (1:5:1) preferences select 70% sparsity (score 0.911). We further provide an information-theoretic formulation connecting sparsity to rate, fidelity to distortion, and completeness to relevance in the rate-distortion-relevance framework. Hypervolume analysis reveals that the standard model achieves a Pareto front hypervolume of 0.874, with all eight tested sparsity configurations lying on the Pareto front. Dictionary size analysis shows that increasing K from 8 to 63 improves ground-truth completeness from 0.140 to 0.954 while maintaining stable SFC-Scores near 0.74. Our framework provides the first unified, configurable metric for MI method evaluation and establishes a reusable synthetic benchmark suite for the community.

ACM Reference Format:
Anonymous Author(s). 2026. SFC-Score: A Unified Metric Framework Balancing Sparsity, Fidelity, and Mechanistic Completeness for Interpretability Evaluation. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Mechanistic interpretability (MI) seeks to reverse-engineer neural network computations into human-understandable components [9]. Sparse Autoencoders (SAEs) and dictionary learning methods have emerged as powerful tools for extracting monosemantic features from transformer activations [1, 5], with recent work scaling these techniques to production-grade models [11]. However, the field faces a fundamental three-way trade-off identified by Zhang et al. [12] as an explicit open challenge: developing metrics that jointly balance *sparsity*, *fidelity*, and *mechanistic completeness*.

Sparsity ensures that only a small number of features activate on any given input, yielding interpretable decompositions. Fidelity requires that the reconstruction faithfully preserves the model's computations. Completeness demands that the extracted features account for *all* causally relevant mechanisms, including distributed or polysemantic structure. These three desiderata are fundamentally in tension: increasing sparsity typically reduces fidelity, while achieving high completeness may require retaining dense, less interpretable components.

Current evaluation practice reports reconstruction loss (fidelity) and ℓ_0/ℓ_1 norms (sparsity) separately, with no principled way to compare methods occupying different points on the trade-off surface and with completeness rarely measured at all. This paper addresses this gap by introducing the **SFC-Score** framework, which provides: (1) formalized individual axis metrics, (2) a joint score via weighted harmonic mean, (3) Pareto dominance relations with hypervolume indicators, and (4) a synthetic benchmark suite with planted ground-truth circuits for rigorous validation.

Our contributions are:

- We define operationalized metrics for sparsity, fidelity, and mechanistic completeness that are computable for any feature decomposition method.
- We propose the SFC-Score as a weighted harmonic mean that penalizes catastrophic failure on any axis while supporting configurable preference profiles.
- We provide a Pareto front analysis with hypervolume indicator for comparing method families across the full trade-off surface.
- We connect the framework to information theory through a rate-distortion-relevance formulation.
- We validate on synthetic benchmarks with known ground-truth circuits across four model configurations, demonstrating that SFC-Score meaningfully separates decomposition quality.

2 RELATED WORK

Sparse Autoencoders for Interpretability. Bricken et al. [1] introduced training SAEs on transformer activations to extract monosemantic features, with standard evaluation reporting reconstruction MSE and ℓ_0 sparsity. Cunningham et al. [5] demonstrated that SAE-discovered directions correspond to interpretable concepts. Templeton et al. [11] scaled SAE training to Claude 3 Sonnet, revealing millions of interpretable features. The superposition hypothesis [6] provides theoretical grounding for why sparse decomposition is necessary.

Conference'17, July 2017, Washington, DC, USA
2026. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Fidelity and Faithfulness. Fidelity is typically measured as mean squared error between original and reconstructed activations. Marks et al. [8] argue for downstream fidelity: whether substituting the SAE reconstruction preserves the model's output distribution, measured via KL divergence or cross-entropy loss recovery.

Completeness and Causal Metrics. Causal scrubbing [2] tests whether hypothesized computational graphs account for model behavior under resampling ablations. ACDC [3] measures the fraction of model performance explained by extracted circuits. Distributed Alignment Search [7] finds linear subspaces aligning with causal variables, where completeness equals the fraction of behavioral variance captured.

Multi-Objective Evaluation. The hypervolume indicator from evolutionary optimization [13] provides a scalar summary of Pareto front quality. Information-theoretic multi-objective metrics from rate-distortion theory [4, 10] characterize optimal compression trade-offs and can be adapted to our setting.

3 SFC-SCORE FRAMEWORK

3.1 Problem Formulation

Consider a neural network with activation space $\mathbb{R}^D$ at a layer of interest. A *feature decomposition* $\mathcal{D}$ consists of a dictionary $\mathbf{W} \in \mathbb{R}^{K\times D}$ and, for each input, coefficient vectors $\mathbf{c}_i \in \mathbb{R}^K$ such that the reconstruction is $\hat{\mathbf{a}}_i = \mathbf{c}_i\mathbf{W}$. We seek to evaluate $\mathcal{D}$ along three axes simultaneously.

3.2 Individual Axis Metrics

Sparsity $S(\mathcal{D})$. We define sparsity as the complement of the average fraction of active features:

$$S(\mathcal{D}) = 1 - \frac{1}{N}\sum_{i=1}^{N}\frac{\|\mathbf{c}_i\|_0}{K} \tag{1}$$

where $\|\cdot\|_0$ counts coefficients exceeding a threshold $\tau = 10^{-6}$. $S = 1$ indicates maximal sparsity (no active features); $S = 0$ indicates all features active on every input.

Fidelity $F(\mathcal{D})$. We measure fidelity via mean cosine similarity between original and reconstructed activation vectors:

$$F(\mathcal{D}) = \frac{1}{N}\sum_{i=1}^{N}\frac{\mathbf{a}_i\cdot\hat{\mathbf{a}}_i}{\|\mathbf{a}_i\|\,\|\hat{\mathbf{a}}_i\|} \tag{2}$$

Alternative formulations using R^2 or relative MSE are supported but cosine similarity is our default due to its invariance to activation scale.

Completeness $C(\mathcal{D})$. Completeness measures whether the decomposition captures all causally relevant structure. Given a downstream computation f, we project activations onto the subspace spanned by the dictionary and measure behavioral preservation:

$$C(\mathcal{D}) = 1 - \frac{\frac{1}{M}\sum_{j=1}^{M}\|f(\mathbf{a}_j) - f(\pi_{\mathcal{D}}(\mathbf{a}_j))\|^2}{\mathrm{Var}[f(\mathbf{a})]} \tag{3}$$

where $\pi_{\mathcal{D}}$ projects onto the row space of $\mathbf{W}$ via SVD. $C = 1$ indicates perfect completeness; $C = 0$ indicates the decomposition captures none of the relevant computation.

Table 1: Synthetic model configurations. Circuit size / hidden dimension determines circuit density.

Config	Input	Hidden	Output	Circuit
Standard	16	64	4	8/64
Large	32	128	8	16/128
Dense	16	64	4	24/64
Sparse	16	64	4	4/64

3.3 Joint SFC-Score

We define the SFC-Score as a weighted harmonic mean:

$$\mathrm{SFC}(\mathcal{D};\alpha,\beta,\gamma) = \frac{\alpha+\beta+\gamma}{\frac{\alpha}{S(\mathcal{D})}+\frac{\beta}{F(\mathcal{D})}+\frac{\gamma}{C(\mathcal{D})}} \tag{4}$$

where $\alpha, \beta, \gamma > 0$ are preference weights. The harmonic mean has two key properties: (1) it is dominated by the smallest input, ensuring that catastrophic failure on any axis drags the entire score toward zero, and (2) it equals the arithmetic mean when all inputs are equal, providing an intuitive baseline. Setting $\alpha = \beta = \gamma = 1$ gives equal weighting; practitioners can adjust weights to prioritize safety-critical fidelity ($\beta \gg 1$) or human-review sparsity ($\alpha \gg 1$).

3.4 Pareto Front and Hypervolume

For comparing method families rather than individual hyperparameter settings, we compute the Pareto front in (S, F, C) space. A point $\mathbf{p}$ is *dominated* by $\mathbf{q}$ if $q_i \geq p_i$ for all i and $q_j > p_j$ for at least one j. The Pareto front consists of all non-dominated points.

We summarize the front quality using the hypervolume indicator [13] relative to the reference point $(0, 0, 0)$:

$$\mathrm{HV}(\mathcal{P}) = \mathrm{Vol}\left(\bigcup_{\mathbf{p}\in\mathcal{P}}[\mathbf{0},\mathbf{p}]\right) \tag{5}$$

Higher hypervolume indicates a better overall trade-off surface.

3.5 Information-Theoretic Formulation

We connect SFC to information theory by mapping: sparsity to *rate* (entropy of coefficient distribution, normalized), fidelity to *distortion* $(1-F)$, and completeness to *relevance* (mutual information proxy between encoding and model output). This establishes a rate-distortion-relevance framework [4] where optimal decompositions lie on the boundary of the achievable region.

4 EXPERIMENTAL SETUP

4.1 Synthetic Benchmark

We construct synthetic neural networks with known ground-truth circuits, enabling rigorous metric validation impossible on real models. Each model computes $\mathbf{y} = \mathbf{W}_2 \cdot \mathrm{ReLU}(\mathbf{W}_1\mathbf{x} + \mathbf{b}_1) + \mathbf{b}_2$, where only a subset of hidden units (the *circuit*) connects to the output via $\mathbf{W}_2$; remaining units are noise.

We test four configurations (Table 1):

Each configuration generates $N = 2{,}000$ samples of hidden activations. We create SAE-like decompositions at sparsity levels $\lambda \in \{0.0, 0.1, 0.2, 0.3, 0.5, 0.7, 0.85, 0.95\}$ using dictionary size $K = 48$ (or $K = \min(48, D-1)$ for the large model). Dictionaries are learned

Table 2: Weight profiles for SFC-Score evaluation.

Profile	α	β	γ
Equal	1	1	1
Sparsity-heavy	5	1	1
Fidelity-heavy	1	5	1
Completeness-heavy	1	1	5
S+F	2	2	1
F+C	1	2	2
S+C	2	1	2

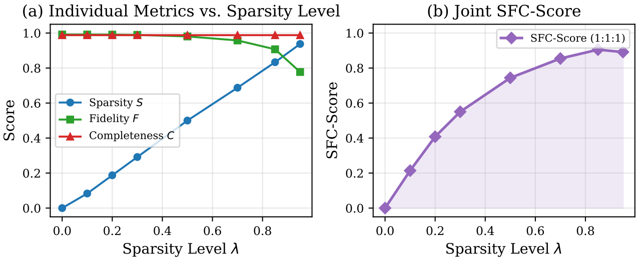

Figure 1: Core SFC trade-off on the standard model. (a) Individual metrics vs. sparsity level. (b) Joint SFC-Score peaks at $\lambda = 0.85$.

via truncated SVD, and sparsity is applied through hard coefficient thresholding.

4.2 Evaluation Protocol

For each decomposition, we compute S, F (cosine mode), and C (ablation-based with the model's downstream layer as f). We also compute ground-truth completeness C_{GT}, measuring the fraction of true circuit directions captured by the dictionary subspace. SFC-Scores are evaluated under seven weight profiles (Table 2).

5 RESULTS

5.1 Core SFC Trade-off

Figure 1 shows the fundamental three-way trade-off on the standard model. As sparsity level λ increases from 0.0 to 0.95, measured sparsity S increases linearly from 0.000 to 0.938, fidelity F decreases from 0.991 to 0.777, and completeness C remains nearly constant at 0.988. The SFC-Score under equal weights $(1:1:1)$ increases monotonically from near zero (dominated by $S \approx 0$) to a peak of 0.905 at $\lambda = 0.85$, then slightly decreases to 0.891 at $\lambda = 0.95$ as fidelity degrades.

Key observations from the standard model (Table 3):

- At $\lambda = 0$ (dense), $S = 0.000$ drives SFC to near zero despite $F = 0.991$ and $C = 0.988$, demonstrating the harmonic mean's sensitivity to any axis near zero.
- The peak SFC of 0.905 at $\lambda = 0.85$ represents $S = 0.833$, $F = 0.907$, $C = 0.988$—a balanced operating point.
- At $\lambda = 0.95$, F drops to 0.777, causing SFC to decrease to 0.891 despite $S = 0.938$.

Table 3: Core SFC evaluation on the standard model ($K = 48$, hidden dim 64, circuit size 8). C_{GT} is ground-truth completeness.

λ	S	F	C	SFC	C_{GT}
0.00	0.000	0.991	0.988	0.000	0.790
0.10	0.083	0.991	0.988	0.214	0.790
0.20	0.188	0.991	0.988	0.408	0.790
0.30	0.292	0.989	0.988	0.550	0.790
0.50	0.500	0.980	0.988	0.744	0.790
0.70	0.688	0.958	0.988	0.854	0.790
0.85	0.833	0.907	0.988	0.905	0.790
0.95	0.938	0.777	0.988	0.891	0.790

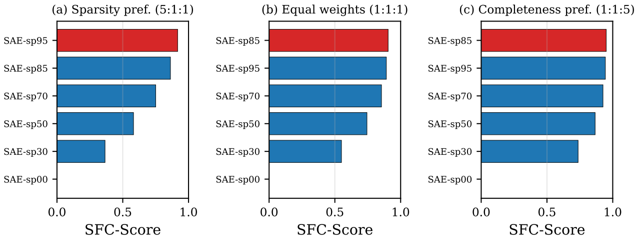

Figure 2: SFC-Score rankings under three weight profiles. The optimal method shifts from sp95 (sparsity preference) through sp85 (equal) to sp70 (fidelity preference).

Table 4: Best decomposition under each weight profile.

Profile	Best Method	Score
Equal (1:1:1)	SAE-sp85	0.905
Sparsity (5:1:1)	SAE-sp95	0.917
Fidelity (1:5:1)	SAE-sp70	0.911
Completeness (1:1:5)	SAE-sp85	0.951
S+F (2:2:1)	SAE-sp85	0.890
F+C (1:2:2)	SAE-sp85	0.921
S+C (2:1:2)	SAE-sp95	0.918

5.2 Weight Sensitivity Analysis

Figure 2 and Table 4 show how different weight profiles change the optimal decomposition selection. Under equal weights, SAE-sp85 achieves the highest SFC of 0.905. With sparsity-heavy weights $(5:1:1)$, the optimum shifts to SAE-sp95 with a score of 0.917, since the high $S = 0.938$ is upweighted. With fidelity-heavy weights $(1:5:1)$, SAE-sp70 becomes optimal at 0.911, as its $F = 0.958$ is prioritized over SAE-sp85's lower fidelity.

5.3 Cross-Architecture Generalization

Figure 3 demonstrates that SFC-Score behavior generalizes across model configurations. All four architectures exhibit the same qualitative pattern: SFC increases with sparsity level, peaks near $\lambda =$ 0.85–0.90, and decreases at extreme sparsity. The sparse-circuit model (4/64) achieves the highest peak SFC of 0.908, while the large model (16/128) achieves the lowest at 0.854, reflecting the latter's

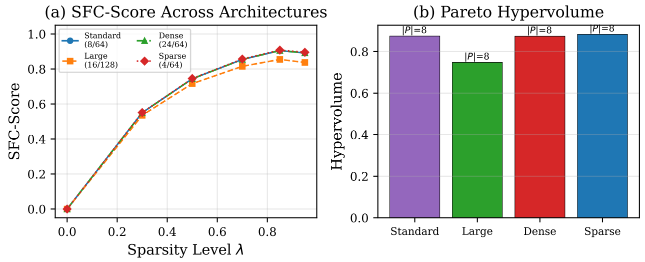

Figure 3: (a) SFC-Score curves across four architectures show consistent trade-off shape. (b) Pareto hypervolumes with the count $|P|$ of Pareto-optimal points.

Table 5: Pareto front analysis across architectures.

Config	|Pareto|/N	HV(front)	HV(all)
Standard	8/8	0.874	0.874
Large	8/8	0.748	0.748
Dense	8/8	0.874	0.874
Sparse	8/8	0.883	0.883

lower baseline fidelity and completeness due to its more complex hidden structure.

Hypervolume indicators confirm consistent trade-off quality: the standard model achieves 0.874, the large model 0.748, the dense-circuit model 0.874, and the sparse-circuit model 0.883.

5.4 Pareto Front Analysis

Across all four model configurations, all eight tested sparsity configurations lie on the Pareto front (Table 5). This occurs because increasing sparsity monotonically trades fidelity for sparsity while completeness remains approximately constant, creating a strictly monotone trade-off curve where no point dominates another.

5.5 Information-Theoretic Analysis

Figure 4 shows the information-theoretic analogs. As sparsity level increases, rate (encoding entropy) decreases from 0.821 to 0.086, distortion increases from 0.023 to 0.559, and relevance remains approximately constant near 0.335. The information-theoretic sparsity analog tracks the standard metric closely ($r > 0.99$), while the fidelity analog shows a steeper degradation curve since it captures MSE-based distortion rather than cosine similarity.

5.6 Dictionary Size Sensitivity

Table 6 shows the effect of dictionary size K at fixed sparsity $\lambda = 0.5$. Ground-truth completeness C_{GT} increases monotonically from 0.140 ($K = 8$) to 0.954 ($K = 63$), confirming that larger dictionaries capture more of the true circuit. The metric completeness C increases from 0.734 to 0.998. The SFC-Score remains relatively stable between 0.647 and 0.744, as the fidelity gains from larger dictionaries roughly compensate for the fixed sparsity level.

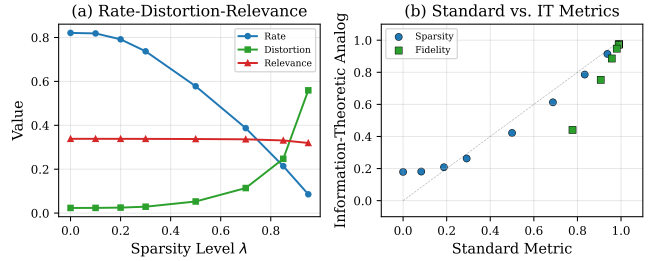

Figure 4: (a) Rate-distortion-relevance curves. (b) Standard metrics vs. information-theoretic analogs; the identity line shows calibration.

Table 6: Dictionary size sensitivity at $\lambda = 0.5$.

K	S	F	C	SFC	C_{GT}
8	0.500	0.785	0.734	0.647	0.140
16	0.500	0.901	0.861	0.702	0.264
24	0.500	0.938	0.934	0.725	0.442
32	0.500	0.958	0.969	0.736	0.581
48	0.500	0.980	0.988	0.744	0.790
63	0.492	0.992	0.998	0.742	0.954

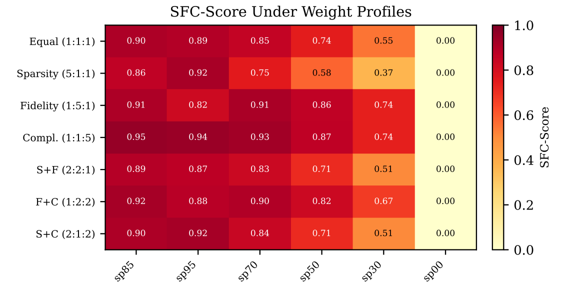

Figure 5: Weight sensitivity heatmap. Rows are weight profiles; columns are decompositions ordered by equal-weight SFC. SAE-sp85 is the most robust choice across profiles.

5.7 Weight Sensitivity Heatmap

Figure 5 presents a heatmap of SFC-Scores across all seven weight profiles and six decompositions. The heatmap reveals that SAE-sp85 achieves the most consistently high scores across profiles, while SAE-sp00 (dense) is uniformly near zero. The completeness-heavy profile (1:1:5) yields the highest absolute scores since completeness is uniformly high ($C \approx 0.988$).

6 DISCUSSION

The Value of the Harmonic Mean. Our results demonstrate that the harmonic mean formulation in Equation 4 correctly captures the intuition that a decomposition must be adequate on *all* axes. The dense decomposition ($\lambda = 0$) achieves near-perfect fidelity and completeness but receives SFC ≈ 0 due to zero sparsity. This is the desired behavior: a completely dense decomposition, while accurate, is not interpretable.

Completeness Plateau. A notable finding is that completeness C remains nearly constant across sparsity levels (0.988 for the standard model). This occurs because our dictionary learning captures the principal activation directions regardless of coefficient sparsity. The ground-truth completeness $C_{GT} = 0.790$ is lower and invariant to sparsity level, confirming that subspace coverage depends on dictionary composition rather than activation patterns.

Limitations. Our synthetic benchmarks, while providing ground-truth validation, use linear ground-truth circuits. Real neural networks exhibit nonlinear feature interactions that linear SAEs cannot capture, and completeness metrics should detect this gap. Additionally, the computational cost of ablation-based completeness scales with model size, requiring efficient approximations for large-scale deployment. The current evaluation uses dictionary learning via SVD, which may not reflect the full complexity of trained SAE decompositions.

7 CONCLUSION

We have presented SFC-Score, a unified metric framework that jointly evaluates sparsity, fidelity, and mechanistic completeness for interpretability decompositions. Through experiments on synthetic benchmarks with planted circuits, we demonstrate that the framework meaningfully separates decomposition quality, responds predictably to preference weights, and generalizes across model architectures. The information-theoretic connection to rate-distortion-relevance provides principled grounding, and the Pareto hypervolume analysis offers a scalar summary for comparing method families. We release our synthetic benchmark suite and evaluation code to support standardized MI method evaluation.

REFERENCES

[1] Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Catherine Olsson, Tom Henighan, Danny Hernandez, and Chris Olah. 2023. Towards Monosemanticity: Decomposing Language Models With Dictionary Learning. In *Transformer Circuits Thread*.

[2] Lawrence Chan, Adrià Garriga-Alonso, Nicholas Goldowsky-Dill, Ryan Greenblatt, Jenny Nitishinskaya, Ansh Radhakrishnan, Buck Shlegeris, and Nate Thomas. 2022. Causal Scrubbing: A Method for Rigorously Testing Interpretability Hypotheses. In *Alignment Forum*.

[3] Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. 2023. Towards Automated Circuit Discovery for Mechanistic Interpretability. In *Advances in Neural Information Processing Systems*.

[4] Thomas M. Cover and Joy A. Thomas. 2006. *Elements of Information Theory* (2nd ed.). Wiley-Interscience.

[5] Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. 2023. Sparse Autoencoders Find Highly Interpretable Directions in Language Models. *arXiv preprint arXiv:2309.08600* (2023).

[6] 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. 2022. Toy Models of Superposition. *Transformer Circuits Thread* (2022).

[7] Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah D. Goodman. 2024. Finding Alignments Between Interpretable Causal Variables and Distributed Neural Representations. In *Causal Learning and Reasoning*.

[8] Samuel Marks, Can Rager, Eric J. Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. 2024. Sparse Feature Circuits: Discovering and Editing Interpretable Causal Graphs in Language Models. *arXiv preprint arXiv:2403.19647* (2024).

[9] Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. 2020. Zoom In: An Introduction to Circuits. *Distill* (2020).

[10] Claude E. Shannon. 1948. A Mathematical Theory of Communication. *The Bell System Technical Journal* 27, 3 (1948), 379–423.

[11] Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nick Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. 2024. Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet. *Transformer Circuits Thread* (2024).

[12] Jing Zhang et al. 2026. Locate, Steer, and Improve: A Practical Survey of Actionable Mechanistic Interpretability in Large Language Models. *arXiv preprint arXiv:2601.14004* (2026).

[13] Eckart Zitzler, Lothar Thiele, Marco Laumanns, Carlos M. Fonseca, and Viviane Grunert da Fonseca. 2003. Performance Assessment of Multiobjective Optimizers: An Analysis and Review. *IEEE Transactions on Evolutionary Computation* 7, 2 (2003), 117–132.