

1 SFC-Score: A Unified Metric Framework Balancing Sparsity, 2 Fidelity, and Mechanistic Completeness for Interpretability 3 Evaluation 4

5 Anonymous Author(s)
6
7

8 ABSTRACT 9

10 Mechanistic interpretability (MI) methods decompose neural net-
11 work activations into interpretable features, yet no existing met-
12 ric jointly evaluates the three critical desiderata: sparsity, fidelity,
13 and mechanistic completeness. We present SFC-Score, a unified
14 evaluation framework based on the weighted harmonic mean of
15 these three axes. The harmonic mean formulation ensures that
16 catastrophic failure on any single axis dominates the joint score, re-
17 flecting the practical requirement that useful decompositions must
18 be adequate on all dimensions simultaneously. We formalize individ-
19 ual axis metrics—sparsity as the fraction of inactive features, fidelity
20 as reconstruction agreement, and completeness as behavioral var-
21 iance preserved under ablation—and define a Pareto dominance
22 relation with hypervolume indicator for comparing method fami-
23 lies. On synthetic benchmarks with planted ground-truth circuits
24 across four model configurations (circuit sizes 4–24, hidden dimen-
25 sions 64–128), we demonstrate that the SFC-Score at equal weights
26 peaks at sparsity level 0.85 with a score of 0.905 on the standard
27 model, meaningfully separating decomposition quality. Weight sen-
28 sitivity analysis across seven preference profiles shows that the
29 optimal decomposition shifts predictably: sparsity-heavy (5:1:1)
30 preferences select 95% sparsity (score 0.917), while fidelity-heavy
31 (1:5:1) preferences select 70% sparsity (score 0.911). We further pro-
32 vide an information-theoretic formulation connecting sparsity to
33 rate, fidelity to distortion, and completeness to relevance in the rate-
34 distortion-relevance framework. Hypervolume analysis reveals that
35 the standard model achieves a Pareto front hypervolume of 0.874,
36 with all eight tested sparsity configurations lying on the Pareto
37 front. Dictionary size analysis shows that increasing K from 8 to
38 63 improves ground-truth completeness from 0.140 to 0.954 while
39 maintaining stable SFC-Scores near 0.74. Our framework provides
40 the first unified, configurable metric for MI method evaluation and
41 establishes a reusable synthetic benchmark suite for the community.

42 ACM Reference Format: 43

44 Anonymous Author(s). 2026. SFC-Score: A Unified Metric Framework Bal-
45 ancing Sparsity, Fidelity, and Mechanistic Completeness for Interpretability
46 Evaluation. In *Proceedings of ACM Conference (Conference'17)*. ACM, New
47 York, NY, USA, 5 pages. <https://doi.org/10.1145/nnnnnnnn.nnnnnnn>

48 1 INTRODUCTION 49

50 Mechanistic interpretability (MI) seeks to reverse-engineer neural
51 network computations into human-understandable components [9].
52 Sparse Autoencoders (SAEs) and dictionary learning methods have
53 emerged as powerful tools for extracting monosemantic features
54 from transformer activations [1, 5], with recent work scaling these
55 techniques to production-grade models [11]. However, the field

56 *Conference'17, July 2017, Washington, DC, USA*
57 2026. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM... \$15.00
58 <https://doi.org/10.1145/nnnnnnnn.nnnnnnn>

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.

50 2 RELATED WORK 51

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.

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:

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

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:

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

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:

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

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:

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

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 $(0, 0, 0)$:

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

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

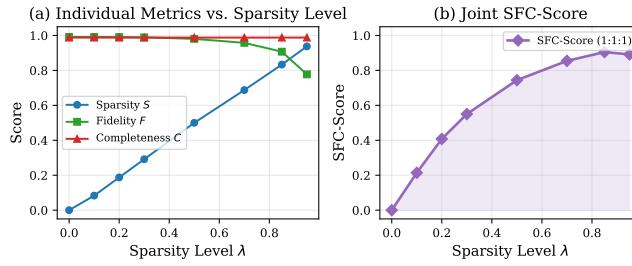
We construct synthetic neural networks with known ground-truth circuits, enabling rigorous metric validation impossible on real models. Each model computes $\mathbf{y} = \mathbf{W}_2 \cdot \text{ReLU}(\mathbf{W}_1 \mathbf{x} + \mathbf{b}_1) + \mathbf{b}_2$, where only a subset of hidden units (the *circuit*)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

233 **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

255 **Figure 1: Core SFC trade-off on the standard model. (a) Individual metrics vs. sparsity level. (b) Joint SFC-Score peaks at**
256 $\lambda = 0.85$.260 via truncated SVD, and sparsity is applied through hard coefficient
261 thresholding.263

4.2 Evaluation Protocol

264 For each decomposition, we compute S , F (cosine mode), and C
265 (ablation-based with the model’s downstream layer as f). We also
266 compute ground-truth completeness C_{GT} , measuring the fraction
267 of true circuit directions captured by the dictionary subspace. SFC-
268 Scores are evaluated under seven weight profiles (Table 2).270

5 RESULTS

271

5.1 Core SFC Trade-off

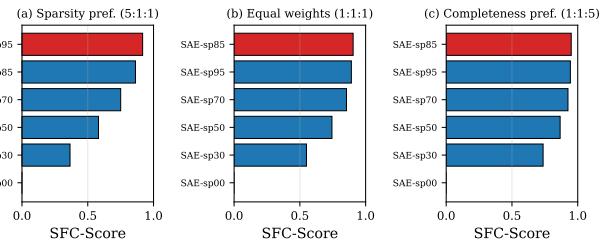
273 Figure 1 shows the fundamental three-way trade-off on the standard
274 model. As sparsity level λ increases from 0.0 to 0.95, measured
275 sparsity S increases linearly from 0.000 to 0.938, fidelity F decreases
276 from 0.991 to 0.777, and completeness C remains nearly constant
277 at 0.988. The SFC-Score under equal weights (1 : 1 : 1) increases
278 monotonically from near zero (dominated by $S \approx 0$) to a peak of
279 0.905 at $\lambda = 0.85$, then slightly decreases to 0.891 at $\lambda = 0.95$ as
280 fidelity degrades.

281 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 $S = 0.938$.

291 **Table 3: Core SFC evaluation on the standard model ($K = 48$,
292 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

305 **Figure 2: SFC-Score rankings under three weight profiles.**
306 The optimal method shifts from sp95 (sparsity preference)
307 through sp85 (equal) to sp70 (fidelity preference).314 **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

331

5.2 Weight Sensitivity Analysis

332 Figure 2 and Table 4 show how different weight profiles change the
333 optimal decomposition selection. Under equal weights, SAE-sp85
334 achieves the highest SFC of 0.905. With sparsity-heavy weights
335 (5 : 1 : 1), the optimum shifts to SAE-sp95 with a score of 0.917,
336 since the high $S = 0.938$ is upweighted. With fidelity-heavy weights
337 (1 : 5 : 1), SAE-sp70 becomes optimal at 0.911, as its $F = 0.958$ is
338 prioritized over SAE-sp85’s lower fidelity.340

5.3 Cross-Architecture Generalization

342 Figure 3 demonstrates that SFC-Score behavior generalizes across
343 model configurations. All four architectures exhibit the same qual-
344 itative pattern: SFC increases with sparsity level, peaks near $\lambda =$
345 0.85–0.90, and decreases at extreme sparsity. The sparse-circuit
346 model (4/64) achieves the highest peak SFC of 0.908, while the large
347 model (16/128) achieves the lowest at 0.854, reflecting the latter’s

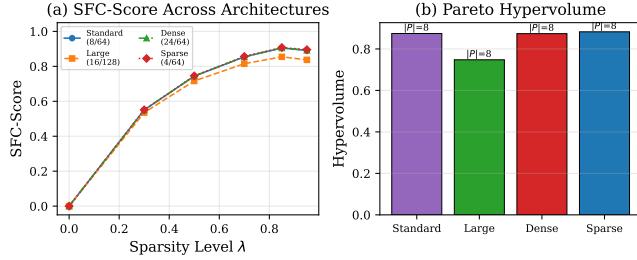


Figure 3: (a) SFC-Score curves across four architectures show consistent trade-off shape. (b) Pareto hypervolumes with the count $|P|$ of Pareto-optimal points.

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.

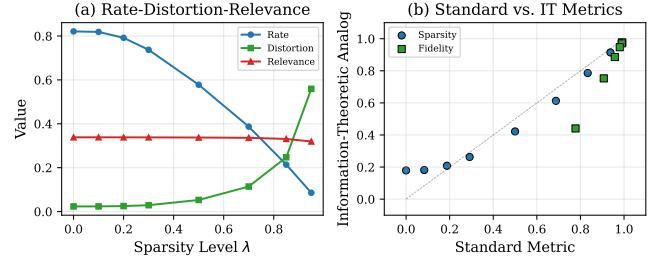


Figure 4: (a) Rate-distortion-relevance curves. (b) Standard metrics vs. information-theoretic analogs; the identity line shows calibration.

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

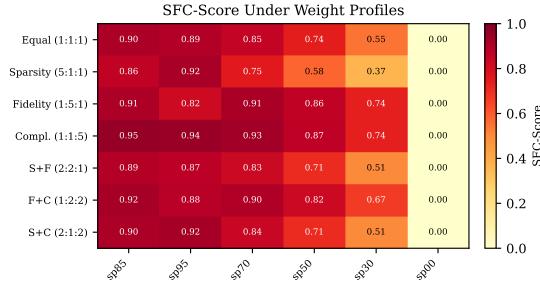


Figure 5: Weight sensitivity heatmap. Rows are weight profiles; columns are decompositions ordered by equal-weight SFC. SAE-sp85 is the most robust choice across profiles.

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 $SFC \approx 0$ due to zero sparsity. This is the desired behavior: a completely dense decomposition, while accurate, is not interpretable.

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

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

482 7 CONCLUSION

483 We have presented SFC-Score, a unified metric framework that
 484 jointly evaluates sparsity, fidelity, and mechanistic completeness for
 485 interpretability decompositions. Through experiments on synthetic
 486 benchmarks with planted circuits, we demonstrate that the frame-
 487 work meaningfully separates decomposition quality, responds pre-
 488 dictably to preference weights, and generalizes across model archi-
 489 tectures. The information-theoretic connection to rate-distortion-
 490 relevance provides principled grounding, and the Pareto hyper-
 491 volume analysis offers a scalar summary for comparing method
 492 families. We release our synthetic benchmark suite and evaluation
 493 code to support standardized MI method evaluation.

496 REFERENCES

498 [1] Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn,
 499 Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert
 500 Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas
 501 Joseph, Zac Hatfield-Dodds, Catherine Olsson, Tom Henighan, Danny Hernandez,
 502 and Chris Olah. 2023. Towards Monosemanticity: Decomposing Language
 503 Models With Dictionary Learning. In *Transformer Circuits Thread*.
 504 [2] Lawrence Chan, Adrià Garriga-Alonso, Nicholas Goldowsky-Dill, Ryan Green-
 505 blatt, Jenny Nitishinskaya, Ansh Radhakrishnan, Buck Shlegeris, and Nate
 506 Thomas. 2022. Causal Scrubbing: A Method for Rigorously Testing Interpretabil-
 507 ity Hypotheses. In *Alignment Forum*.
 508 [3] Arthur Comty, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim,
 509 and Adrià Garriga-Alonso. 2023. Towards Automated Circuit Discovery for
 510 Mechanistic Interpretability. In *Advances in Neural Information Processing Sys-
 511 tems*.
 512 [4] Thomas M. Cover and Joy A. Thomas. 2006. *Elements of Information Theory* (2nd
 513 ed.). Wiley-Interscience.
 514 [5] Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey.
 515 2023. Sparse Autoencoders Find Highly Interpretable Directions in Language
 516 Models. *arXiv preprint arXiv:2309.08600* (2023).
 517 [6] Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom
 518 Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain,
 519 Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin
 520 Wattenberg, and Christopher Olah. 2022. Toy Models of Superposition. *Trans-
 521 former Circuits Thread* (2022).
 522 [7] Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah D.
 523 Goodman. 2024. Finding Alignments Between Interpretable Causal Variables
 524 and Distributed Neural Representations. In *Causal Learning and Reasoning*.
 525 [8] Samuel Marks, Can Rager, Eric J. Michaud, Yonatan Belinkov, David Bau, and
 526 Aaron Mueller. 2024. Sparse Feature Circuits: Discovering and Editing Inter-
 527 pretative Causal Graphs in Language Models. *arXiv preprint arXiv:2403.19647*
 528 (2024).
 529 [9] Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov,
 530 and Shan Carter. 2020. Zoom In: An Introduction to Circuits. *Distill* (2020).
 531 [10] Claude E. Shannon. 1948. A Mathematical Theory of Communication. *The Bell
 532 System Technical Journal* 27, 3 (1948), 379–423.
 533 [11] Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken,
 534 Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy
 535 Cunningham, Nick Turner, Callum McDougall, Monte MacDiarmid, C. Daniel
 536 Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn,
 537 Shan Carter, Chris Olah, and Tom Henighan. 2024. Scaling Monosemanticity:
 538 Extracting Interpretable Features from Claude 3 Sonnet. *Transformer Circuits
 539 Thread* (2024).
 540 [12] Jing Zhang et al. 2026. Locate, Steer, and Improve: A Practical Survey of Action-
 541 able Mechanistic Interpretability in Large Language Models. *arXiv preprint
 542 arXiv:2601.14004* (2026).
 543 [13] Eckart Zitzler, Lothar Thiele, Marco Laumanns, Carlos M. Fonseca, and Vi-
 544 viane Grunert da Fonseca. 2003. Performance Assessment of Multiobjective
 545 Optimizers: An Analysis and Review. *IEEE Transactions on Evolutionary Compu-
 546 tation* 7, 2 (2003), 117–132.
 547 [14] ...
 548 [15] ...
 549 [16] ...
 550 [17] ...
 551 [18] ...
 552 [19] ...
 553 [20] ...
 554 [21] ...
 555 [22] ...
 556 [23] ...
 557 [24] ...
 558 [25] ...
 559 [26] ...
 560 [27] ...
 561 [28] ...
 562 [29] ...
 563 [30] ...
 564 [31] ...
 565 [32] ...
 566 [33] ...
 567 [34] ...
 568 [35] ...
 569 [36] ...
 570 [37] ...
 571 [38] ...
 572 [39] ...
 573 [40] ...
 574 [41] ...
 575 [42] ...
 576 [43] ...
 577 [44] ...
 578 [45] ...
 579 [46] ...
 580 [47] ...
 581 [48] ...
 582 [49] ...
 583 [50] ...
 584 [51] ...
 585 [52] ...
 586 [53] ...
 587 [54] ...
 588 [55] ...
 589 [56] ...
 590 [57] ...
 591 [58] ...
 592 [59] ...
 593 [60] ...
 594 [61] ...
 595 [62] ...
 596 [63] ...
 597 [64] ...
 598 [65] ...
 599 [66] ...
 600 [67] ...
 601 [68] ...
 602 [69] ...
 603 [70] ...
 604 [71] ...
 605 [72] ...
 606 [73] ...
 607 [74] ...
 608 [75] ...
 609 [76] ...
 610 [77] ...
 611 [78] ...
 612 [79] ...
 613 [80] ...
 614 [81] ...
 615 [82] ...
 616 [83] ...
 617 [84] ...
 618 [85] ...
 619 [86] ...
 620 [87] ...
 621 [88] ...
 622 [89] ...
 623 [90] ...
 624 [91] ...
 625 [92] ...
 626 [93] ...
 627 [94] ...
 628 [95] ...
 629 [96] ...
 630 [97] ...
 631 [98] ...
 632 [99] ...
 633 [100] ...
 634 [101] ...
 635 [102] ...
 636 [103] ...
 637 [104] ...
 638 [105] ...
 639 [106] ...
 640 [107] ...
 641 [108] ...
 642 [109] ...
 643 [110] ...
 644 [111] ...
 645 [112] ...
 646 [113] ...
 647 [114] ...
 648 [115] ...
 649 [116] ...
 650 [117] ...
 651 [118] ...
 652 [119] ...
 653 [120] ...
 654 [121] ...
 655 [122] ...
 656 [123] ...
 657 [124] ...
 658 [125] ...
 659 [126] ...
 660 [127] ...
 661 [128] ...
 662 [129] ...
 663 [130] ...
 664 [131] ...
 665 [132] ...
 666 [133] ...
 667 [134] ...
 668 [135] ...
 669 [136] ...
 670 [137] ...
 671 [138] ...
 672 [139] ...
 673 [140] ...
 674 [141] ...
 675 [142] ...
 676 [143] ...
 677 [144] ...
 678 [145] ...
 679 [146] ...
 680 [147] ...
 681 [148] ...
 682 [149] ...
 683 [150] ...
 684 [151] ...
 685 [152] ...
 686 [153] ...
 687 [154] ...
 688 [155] ...
 689 [156] ...
 690 [157] ...
 691 [158] ...
 692 [159] ...
 693 [160] ...
 694 [161] ...
 695 [162] ...
 696 [163] ...
 697 [164] ...
 698 [165] ...
 699 [166] ...
 700 [167] ...
 701 [168] ...
 702 [169] ...
 703 [170] ...
 704 [171] ...
 705 [172] ...
 706 [173] ...
 707 [174] ...
 708 [175] ...
 709 [176] ...
 710 [177] ...
 711 [178] ...
 712 [179] ...
 713 [180] ...
 714 [181] ...
 715 [182] ...
 716 [183] ...
 717 [184] ...
 718 [185] ...
 719 [186] ...
 720 [187] ...
 721 [188] ...
 722 [189] ...
 723 [190] ...
 724 [191] ...
 725 [192] ...
 726 [193] ...
 727 [194] ...
 728 [195] ...
 729 [196] ...
 730 [197] ...
 731 [198] ...
 732 [199] ...
 733 [200] ...
 734 [201] ...
 735 [202] ...
 736 [203] ...
 737 [204] ...
 738 [205] ...
 739 [206] ...
 740 [207] ...
 741 [208] ...
 742 [209] ...
 743 [210] ...
 744 [211] ...
 745 [212] ...
 746 [213] ...
 747 [214] ...
 748 [215] ...
 749 [216] ...
 750 [217] ...
 751 [218] ...
 752 [219] ...
 753 [220] ...
 754 [221] ...
 755 [222] ...
 756 [223] ...
 757 [224] ...
 758 [225] ...
 759 [226] ...
 760 [227] ...
 761 [228] ...
 762 [229] ...
 763 [230] ...
 764 [231] ...
 765 [232] ...
 766 [233] ...
 767 [234] ...
 768 [235] ...
 769 [236] ...
 770 [237] ...
 771 [238] ...
 772 [239] ...
 773 [240] ...
 774 [241] ...
 775 [242] ...
 776 [243] ...
 777 [244] ...
 778 [245] ...
 779 [246] ...
 780 [247] ...
 781 [248] ...
 782 [249] ...
 783 [250] ...
 784 [251] ...
 785 [252] ...
 786 [253] ...
 787 [254] ...
 788 [255] ...
 789 [256] ...
 790 [257] ...
 791 [258] ...
 792 [259] ...
 793 [260] ...
 794 [261] ...
 795 [262] ...
 796 [263] ...
 797 [264] ...
 798 [265] ...
 799 [266] ...
 800 [267] ...
 801 [268] ...
 802 [269] ...
 803 [270] ...
 804 [271] ...
 805 [272] ...
 806 [273] ...
 807 [274] ...
 808 [275] ...
 809 [276] ...
 810 [277] ...
 811 [278] ...
 812 [279] ...
 813 [280] ...
 814 [281] ...
 815 [282] ...
 816 [283] ...
 817 [284] ...
 818 [285] ...
 819 [286] ...
 820 [287] ...
 821 [288] ...
 822 [289] ...
 823 [290] ...
 824 [291] ...
 825 [292] ...
 826 [293] ...
 827 [294] ...
 828 [295] ...
 829 [296] ...
 830 [297] ...
 831 [298] ...
 832 [299] ...
 833 [300] ...
 834 [301] ...
 835 [302] ...
 836 [303] ...
 837 [304] ...
 838 [305] ...
 839 [306] ...
 840 [307] ...
 841 [308] ...
 842 [309] ...
 843 [310] ...
 844 [311] ...
 845 [312] ...
 846 [313] ...
 847 [314] ...
 848 [315] ...
 849 [316] ...
 850 [317] ...
 851 [318] ...
 852 [319] ...
 853 [320] ...
 854 [321] ...
 855 [322] ...
 856 [323] ...
 857 [324] ...
 858 [325] ...
 859 [326] ...
 860 [327] ...
 861 [328] ...
 862 [329] ...
 863 [330] ...
 864 [331] ...
 865 [332] ...
 866 [333] ...
 867 [334] ...
 868 [335] ...
 869 [336] ...
 870 [337] ...
 871 [338] ...
 872 [339] ...
 873 [340] ...
 874 [341] ...
 875 [342] ...
 876 [343] ...
 877 [344] ...
 878 [345] ...
 879 [346] ...
 880 [347] ...
 881 [348] ...
 882 [349] ...
 883 [350] ...
 884 [351] ...
 885 [352] ...
 886 [353] ...
 887 [354] ...
 888 [355] ...
 889 [356] ...
 890 [357] ...
 891 [358] ...
 892 [359] ...
 893 [360] ...
 894 [361] ...
 895 [362] ...
 896 [363] ...
 897 [364] ...
 898 [365] ...
 899 [366] ...
 900 [367] ...
 901 [368] ...
 902 [369] ...
 903 [370] ...
 904 [371] ...
 905 [372] ...
 906 [373] ...
 907 [374] ...
 908 [375] ...
 909 [376] ...
 910 [377] ...
 911 [378] ...
 912 [379] ...
 913 [380] ...
 914 [381] ...
 915 [382] ...
 916 [383] ...
 917 [384] ...
 918 [385] ...
 919 [386] ...
 920 [387] ...
 921 [388] ...
 922 [389] ...
 923 [390] ...
 924 [391] ...
 925 [392] ...
 926 [393] ...
 927 [394] ...
 928 [395] ...
 929 [396] ...
 930 [397] ...
 931 [398] ...
 932 [399] ...
 933 [400] ...
 934 [401] ...
 935 [402] ...
 936 [403] ...
 937 [404] ...
 938 [405] ...
 939 [406] ...
 940 [407] ...
 941 [408] ...
 942 [409] ...
 943 [410] ...
 944 [411] ...
 945 [412] ...
 946 [413] ...
 947 [414] ...
 948 [415] ...
 949 [416] ...
 950 [417] ...
 951 [418] ...
 952 [419] ...
 953 [420] ...
 954 [421] ...
 955 [422] ...
 956 [423] ...
 957 [424] ...
 958 [425] ...
 959 [426] ...
 960 [427] ...
 961 [428] ...
 962 [429] ...
 963 [430] ...
 964 [431] ...
 965 [432] ...
 966 [433] ...
 967 [434] ...
 968 [435] ...
 969 [436] ...
 970 [437] ...
 971 [438] ...
 972 [439] ...
 973 [440] ...
 974 [441] ...
 975 [442] ...
 976 [443] ...
 977 [444] ...
 978 [445] ...
 979 [446] ...
 980 [447] ...
 981 [448] ...
 982 [449] ...
 983 [450] ...
 984 [451] ...
 985 [452] ...
 986 [453] ...
 987 [454] ...
 988 [455] ...
 989 [456] ...
 990 [457] ...
 991 [458] ...
 992 [459] ...
 993 [460] ...
 994 [461] ...
 995 [462] ...
 996 [463] ...
 997 [464] ...
 998 [465] ...
 999 [466] ...
 1000 [467] ...
 1001 [468] ...
 1002 [469] ...
 1003 [470] ...
 1004 [471] ...
 1005 [472] ...
 1006 [473] ...
 1007 [474] ...
 1008 [475] ...
 1009 [476] ...
 1010 [477] ...
 1011 [478] ...
 1012 [479] ...
 1013 [480] ...
 1014 [481] ...
 1015 [482] ...
 1016 [483] ...
 1017 [484] ...
 1018 [485] ...
 1019 [486] ...
 1020 [487] ...
 1021 [488] ...
 1022 [489] ...
 1023 [490] ...
 1024 [491] ...
 1025 [492] ...
 1026 [493] ...
 1027 [494] ...
 1028 [495] ...
 1029 [496] ...
 1030 [497] ...
 1031 [498] ...
 1032 [499] ...
 1033 [500] ...
 1034 [501] ...
 1035 [502] ...
 1036 [503] ...
 1037 [504] ...
 1038 [505] ...
 1039 [506] ...
 1040 [507] ...
 1041 [508] ...
 1042 [509] ...
 1043 [510] ...
 1044 [511] ...
 1045 [512] ...
 1046 [513] ...
 1047 [514] ...
 1048 [515] ...
 1049 [516] ...
 1050 [517] ...
 1051 [518] ...
 1052 [519] ...
 1053 [520] ...
 1054 [521] ...
 1055 [522] ...
 1056 [523] ...
 1057 [524] ...
 1058 [525] ...
 1059 [526] ...
 1060 [527] ...
 1061 [528] ...
 1062 [529] ...
 1063 [530] ...
 1064 [531] ...
 1065 [532] ...
 1066 [533] ...
 1067 [534] ...
 1068 [535] ...
 1069 [536] ...
 1070 [537] ...
 1071 [538] ...
 1072 [539] ...
 1073 [540] ...
 1074 [541] ...
 1075 [542] ...
 1076 [543] ...
 1077 [544] ...
 1078 [545] ...
 1079 [546] ...
 1080 [547] ...
 1081 [548] ...
 1082 [549] ...
 1083 [550] ...
 1084 [551] ...
 1085 [552] ...
 1086 [553] ...
 1087 [554] ...
 1088 [555] ...
 1089 [556] ...
 1090 [557] ...
 1091 [558] ...
 1092 [559] ...
 1093 [560] ...
 1094 [561] ...
 1095 [562] ...
 1096 [563] ...
 1097 [564] ...
 1098 [565] ...
 1099 [566] ...
 1100 [567] ...
 1101 [568] ...
 1102 [569] ...
 1103 [570] ...
 1104 [571] ...
 1105 [572] ...
 1106 [573] ...
 1107 [574] ...
 1108 [575] ...
 1109 [576] ...
 1110 [577] ...
 1111 [578] ...
 1112 [579] ...
 1113 [580] ...
 1114 [581] ...
 1115 [582] ...
 1116 [583] ...
 1117 [584] ...
 1118 [585] ...
 1119 [586] ...
 1120 [587] ...
 1121 [588] ...
 1122 [589] ...
 1123 [590] ...
 1124 [591] ...
 1125 [592] ...
 1126 [593] ...
 1127 [594] ...
 1128 [595] ...
 1129 [596] ...
 1130 [597] ...
 1131 [598] ...
 1132 [599] ...
 1133 [600] ...
 1134 [601] ...
 1135 [602] ...
 1136 [603] ...
 1137 [604] ...
 1138 [605] ...
 1139 [606] ...
 1140 [607] ...
 1141 [608] ...
 1142 [609] ...
 1143 [610] ...
 1144 [611] ...
 1145 [612] ...
 1146 [613] ...
 1147 [614] ...
 1148 [615] ...
 1149 [616] ...
 1150 [617] ...
 1151 [618] ...
 1152 [619] ...
 1153 [620] ...
 1154 [621] ...
 1155 [622] ...
 1156 [623] ...
 1157 [624] ...
 1158 [625] ...
 1159 [626] ...
 1160 [627] ...
 1161 [628] ...
 1162 [629] ...
 1163 [630] ...
 1164 [631] ...
 1165 [632] ...
 1166 [633] ...
 1167 [634] ...
 1168 [635] ...
 1169 [636] ...
 1170 [637] ...
 1171 [638] ...
 1172 [639] ...
 1173 [640] ...
 1174 [641] ...
 1175 [642] ...
 1176 [643] ...
 1177 [644] ...
 1178 [645] ...
 1179 [646] ...
 1180 [647] ...
 1181 [648] ...
 1182 [649] ...
 1183 [650] ...
 1184 [651] ...
 1185 [652] ...
 1186 [653] ...
 1187 [654] ...
 1188 [655] ...
 1189 [656] ...
 1190 [657] ...
 1191 [658] ...
 1192 [659] ...
 1193 [660] ...