Worst-Case Interpretability

Compact Proofs of Model Performance

Louis Jaburi

April 2026

EleutherAI

A spectrum of interpretability

A Toy Model of Universality: Reverse Engineering how Networks Learn Group Operations

Bilal Chughtai [1] **Lawrence Chan** [2] **Neel Nanda** [1]

tasks. In this work, we study the universality hypothesis by examining how small neural networks learn to implement group composition. We present a novel algorithm by which neural networks may implement composition for any finite group via mathematical representation theory. We then show that networks consistently learn this algorithm by reverse engineering model logits and weights, and confirm our understanding using ablations. By studying networks of differing

Grokking Group Multiplication with Cosets

Dashiell Stander [1] **Qinan Yu** [1 2] **Honglu Fan** [1 3] **Stella Biderman** [1]

and well-understood objects [8, 10, 15]. We succeed in completely reverse engineering the model and enumerating the diverse circuits that it converges on to implement the multiplication of the symmetric group. Our work does not, however, represent an unmitigated success for the project of mechanistic interpretability. The prior work of Chughtai et al. [4] studied the exact same model and setting, but came to completely different conclusions. Understanding why our and Chughtai et al. [4]'s interpretations of the same data diverged required extensive effort (see Appendix 7 for a thorough comparison). **We find that even in a setting as simple and well understood as group arithmetic, it is incredibly difficult to do interpretability research and be confident about one's conclusions.**

TOWARDS A UNIFIED AND VERIFIED UNDERSTANDING OF GROUP-OPERATION NETWORKS

Wilson Wu[1] **Louis Jaburi***[2] **Jacob Drori***[2] **Jason Gross**[2]
[1] University of Colorado Boulder [2] Independent
wiwu2390@colorado.edu
{louis.yodj,jacobcd52,jasongross9}@gmail.com

nations for the same empirical phenomena: recently, Chughtai et al. (2023) claimed that models trained on finite groups implement a *group composition via representations* algorithm, while subsequent work (Stander et al., 2024) studies the same model and task and instead argues that the model implements a *coset concentration* algorithm.

In this work, we take on the challenge of reconciling their interpretations. We investigate the same setting and find internal model structure that was overlooked by both previous works: the irreducible representations noticed by Chughtai et al. (2023) act by permutation on a discrete set of vectors learned by the model. Based on our observations, we propose a model explanation that unifies

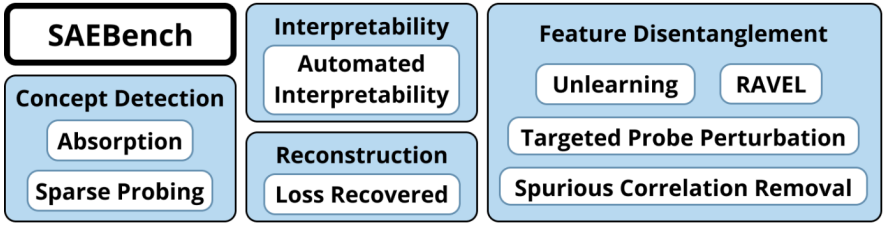
Transformer Circuit Faithfulness Metrics Are Not Robust

Joseph Miller*
FAR AI

Bilal Chughtai
Independent

William Saunders
Independent

SAEBench: A Comprehensive Benchmark for Sparse Autoencoders in Language Model Interpretability



Why metrics for mech interp?

- **Optimisation.** If we can measure an explanation, we can optimise for it.
- **Automation.** Suppose AGI arrives tomorrow — can we build a trustworthy, automated pipeline for discovering mechanistic explanations?
- **Guarantees.** Mech interp could target the highest standard of trustworthiness: *formal proofs*.

Compression as a metric

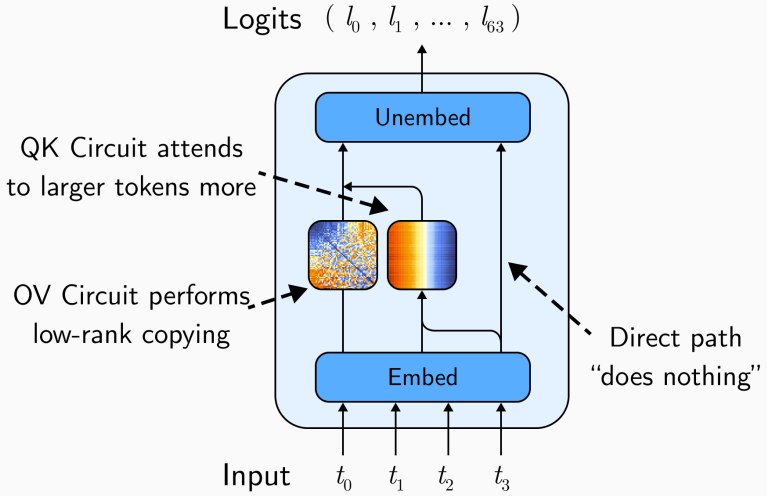
A mechanistic explanation of M's behaviour is a *proof* C that M achieves accuracy $\geq b$.

$$\textit{proof length} \approx \textit{cost of running } C(M) \approx \textit{how well we understand } M$$

- Trivial proof: vacuous bound, constant cost.
- Brute force: tight bound, exponential cost.
- Interesting proofs: tight bound at *structured* cost.

Case study: Max-of-K

1-layer attention-only transformer trained to compute $\max(x_1, \ldots, x_K)$.



The theorem we want to prove

$$\mathbb{E}_{x\sim\mathrm{Unif}}\left[\arg\max M(x)[-1] = \max_i x_i\right] > .9973$$

- Goal: minimise proof length (in any formal system) for a fixed bound b; or maximise b for a fixed proof length.

Brute force

```
def C(M):
  count = 0
  for x in possible_sequences:
    count += (M(x)[..., -1, :].argmax(-1) == x.max(-1).values).sum()
  return count / len(possible_sequences)
```

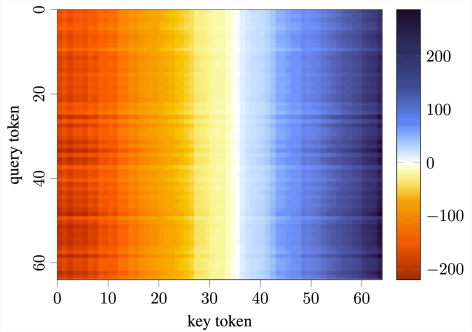
- Sound lower bound on accuracy (trivially).
- Cost: $\mathcal{O}(d_{\text{vocab}}^{n_{\text{ctx}}})$ — exponential in context length.
- No mechanistic content; no compression.

Cubic proof — convexity of softmax

Insight. We want all attention on the max token.

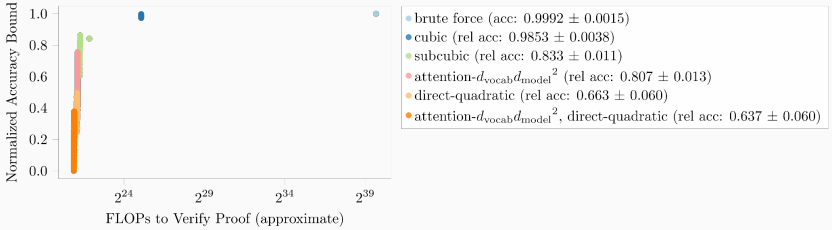
- By convexity of softmax, the worst case over non-max, non-query tokens occurs at extremes — pick the single worst token and fill.
- Reduces $d_{\text{vocab}}^{n_{\text{ctx}}}$ sequences to d_{vocab}^3 configurations.
- Cost: $\mathcal{O}(d_{\text{vocab}}^3 n_{\text{ctx}}^2)$. Bound: $\approx 98.5\%$.

Sub-cubic — exploit EQKE structure



EQKE is numerically low-rank and monotone in the key token.
Rank-1 / SVD approximation $\Rightarrow \mathcal{O}(d_{\text{vocab}}^2 n_{\text{ctx}}^2 + d_{\text{vocab}} d_{\text{model}}^2)$.

Results: proof cost vs. tightness



Empirical Pareto frontier: more structure, less compute, looser bound.

The full menu

Description of Proof	**Complexity Cost**	**Bound**
Brute force	$\mathcal{O}(v^{k+1}kd)$	0.9992 ± 0.0015
Cubic	$\mathcal{O}(v^3k^2)$	0.9531 ± 0.0087
Sub-cubic	$\mathcal{O}(v^2 \cdot k^2 + v^2 \cdot d)$	0.820 ± 0.013
w/o mean+diff		0.488 ± 0.079
Low-rank QK	$\mathcal{O}(v^2k^2 + \underbrace{vd^2}_{\text{QK}} + \underbrace{v^2d}_{\text{EU\&OV}})$	0.795 ± 0.014
SVD only		0.406 ± 0.077
Low-rank EU	$\mathcal{O}(v^2k^2 + \underbrace{vd}_{\text{EU}} + \underbrace{v^2d}_{\text{QK\&OV}})$	0.653 ± 0.060
SVD only		$(3.38 \pm 0.06) \times 10^{-6}$
Low-rank QK&EU	$\mathcal{O}(v^2k^2 + \underbrace{vd^2}_{\text{QK}} + \underbrace{vd}_{\text{EU}} + \underbrace{v^2d}_{\text{OV}})$	0.627 ± 0.060
SVD only		$(3.38 \pm 0.06) \times 10^{-6}$
Quadratic QK	$\mathcal{O}(v^2k^2 + \underbrace{vd}_{\text{QK}} + \underbrace{v^2d}_{\text{EU\&OV}})$	0.390 ± 0.032
Quadratic QK&EU	$\mathcal{O}(v^2k^2 + \underbrace{vd}_{\text{QK\&EU}} + \underbrace{v^2d}_{\text{OV}})$	0.285 ± 0.036

Why does *more* structure not always yield a *better* bound?

Compounding structureless errors in the sub-cubic regime:

Approximation strategy	Slack factor	Regime
(exact) max row diff	≈ 1.8	cubic
$2\cdot$ (max abs value)	≈ 2.0	sub-cubic
max row diff on subproduct	≈ 5.7	sub-cubic
recursive max row diff	≈ 97	sub-cubic

What this buys mech interp

- **Objective metric.** Compression gives a numerical standard of understanding, tailorable to subcomponents.
- **Diagnosis.** Failure to compact a proof $\Rightarrow$ we are missing mechanistic structure.
- **Path to scale.** Auto-formalisation + error suppression (fine-tuning, sampling) push the frontier further.

Takeaways

- Proofs of model performance are *possible* — and genuinely hard.
- Shorter proof $\Leftrightarrow$ deeper, more compressed understanding.
- Compounding structureless noise is the bottleneck; most post-hoc interp glosses over it.