A proof-length based **compression metric** for **mechanistic interpretability** provides **richer understanding** of toy models.

Approach

We formalize post-hoc mechanistic interpretability as proving **worst-case generalization bounds** on the performance of models.

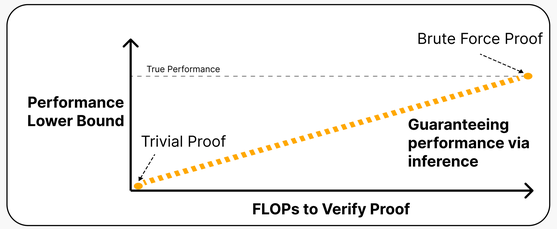
$$\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}[f(\mathcal{M}(\mathbf{x}),\mathbf{y})] \geq b$$

Then, we define a **bound approximation algorithm**; a proof is the trace of running the algorithm along with an explanation that it provides valid lower bounds on the entire input distribution.

Algorithm 1 Counting Correct Sequences By Brute Force

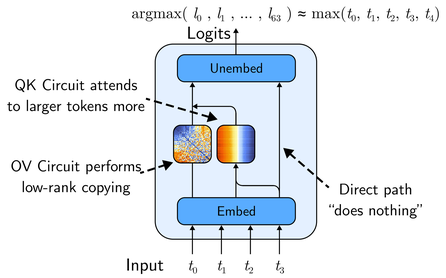
```
function CORRECTNESS(M, input-sequence)
    return MODEL-BEHAVIOR(M, input-sequence) == MAX(input-sequence)
end function
function BRUTE-FORCE(d_vocab, n_ctx, M)
    return 1/(d_vocab^n_ctx) SUM(CORRECTNESS(M, tokens) for tokens ∈ (RANGE(d_vocab))^n_ctx)
end function
```

Naively, we can run inference on particular inputs to guarantee model performance on those inputs. Can mechanistic understanding provide a **compression** of model behavior that beats this inefficient baseline?

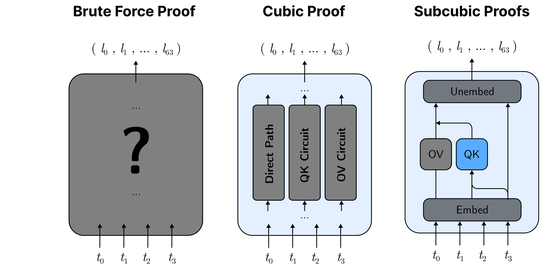


Prototype

We experiment on transformer models trained on the the Max-of-K task.



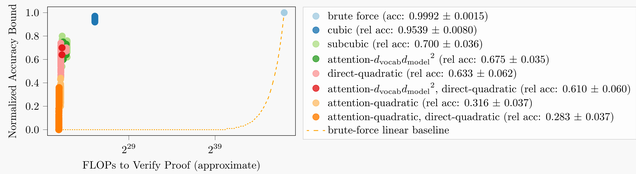
Different proof strategies incorporate varying degrees of **mechanistic detail**.

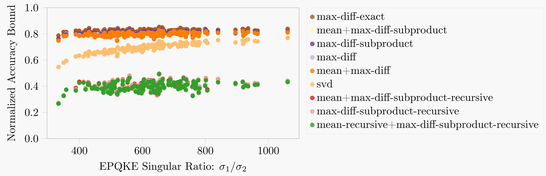


	Brute Force Proof	Cubic Proof	Subcubic Proofs
FLOPs:	1.41×10^{11}	3.51×10^{7}	4.68×10^{6}
Acc:	99.92%	95.31%	28.41%
Dimension:	1.07×10^{9}	1.28×10^{5}	4.42×10^{3}
Complexity:	$O(d_{\text{vocab}}{}^{\text{context}})$	$O(d_{\text{vocab}}{}^{3} \cdot \text{context})$	$O(d_{\text{vocab}} \cdot d_{\text{model}}{}^{2} \cdot \text{context})$

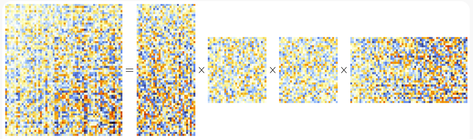
Results

More mechanistic detail results in **tighter performance bounds**. Moreover, for any given proof length, proving a tighter performance bound **requires more mechanistic detail**.





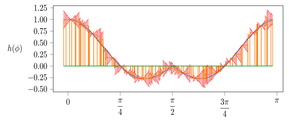
However, **compounding approximation error** in post-hoc analysis is a key challenge to proving worst-case bounds.



Approximation Strategy	Result	Complexity
(exact) max row diff	≈ 1.8	$(\mathcal{O}(d_{\text{vocab}}{}^2 d_{\text{model}}))$
$2 \cdot$ (max abs value)	≈ 2.0	$(\mathcal{O}(d_{\text{vocab}}{}^2 d_{\text{model}}))$
max row diff on subproduct	≈ 5.7	$(\mathcal{O}(d_{\text{vocab}} d_{\text{model}}{}^2))$
recursive max row diff	≈ 97	$(\mathcal{O}(d_{\text{vocab}} d_{\text{model}}))$

Applications

Aiming to minimize proof-length helped identify the **asymptotic bottlenecks** to compressing MLPs, resulting in the **first compression of MLPs**.



$$\int_0^{\pi} \underbrace{\mathrm{ReLU}\big(\cos(k(a+b)/2+\phi)\big)\cos(2\phi+kc)}_{h(\phi)}\,\mathrm{d}\phi = \frac{2}{3}\cos(k(a+b-c))$$

Compact proofs were used as a **source of ground truth** for comparing diverging interpretations of a model in the literature.

The compression metric **defines optimization targets** for representation search, such as in SAEs and sparse cross-coders. The differentiable proof certificate can be directly **fine-tuned** to bring models in line with partial interpretations and **suppress compounding approximation error**.

Compact Proofs of Model Performance via Mechanistic Interpretability

Jason Gross, Rajashree Agrawal, Thomas Kwa, Euan Ong, Chun Hei Yip, Alex Gibson, Soufiane Noubir, Lawrence Chan