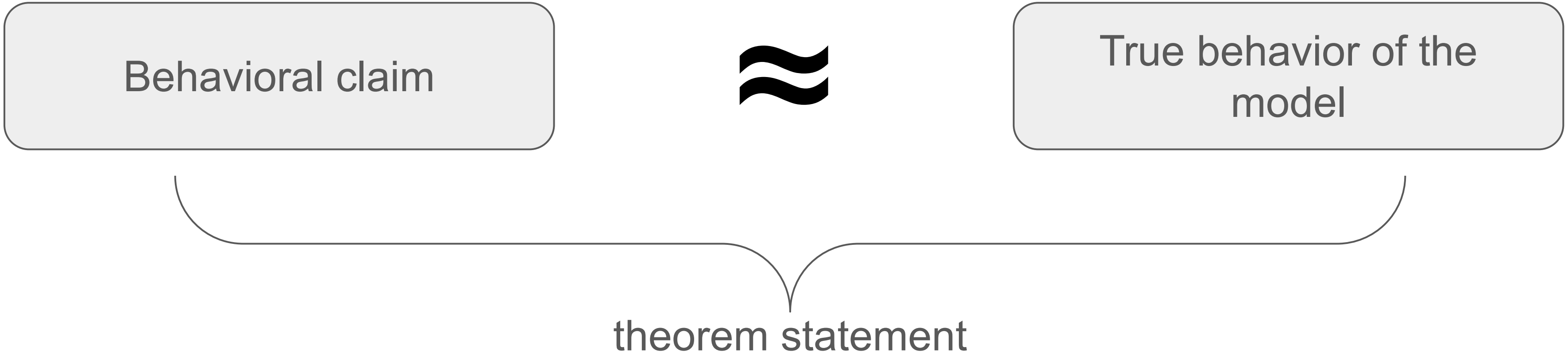
Compact Proofs of Model Performance via Mechanistic Interpretability

TLDR

Proof length can be a metric on mech interp

Formalizing proof length to quantify compression

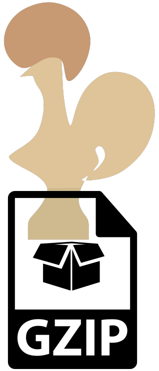


Behavioral claim ≈ True behavior of the model

theorem statement

Proof = sound computation of worst-case error (divergence in behavior)

Length of proof = cost of running computation



https://www.svgrepo.com/svg/29028/gzip-file-format-variant https://calebstanford.com/2019/01/15/coq-vector-image/

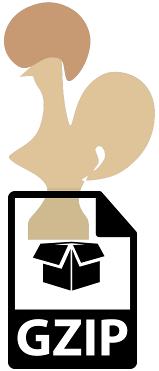
Formalizing proof length to quantify compression

Behavioral claim ≈ True behavior of the model

$$\underbrace{\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[f(y, M(x))\right] \geq b}_{\text{theorem statement}}$$

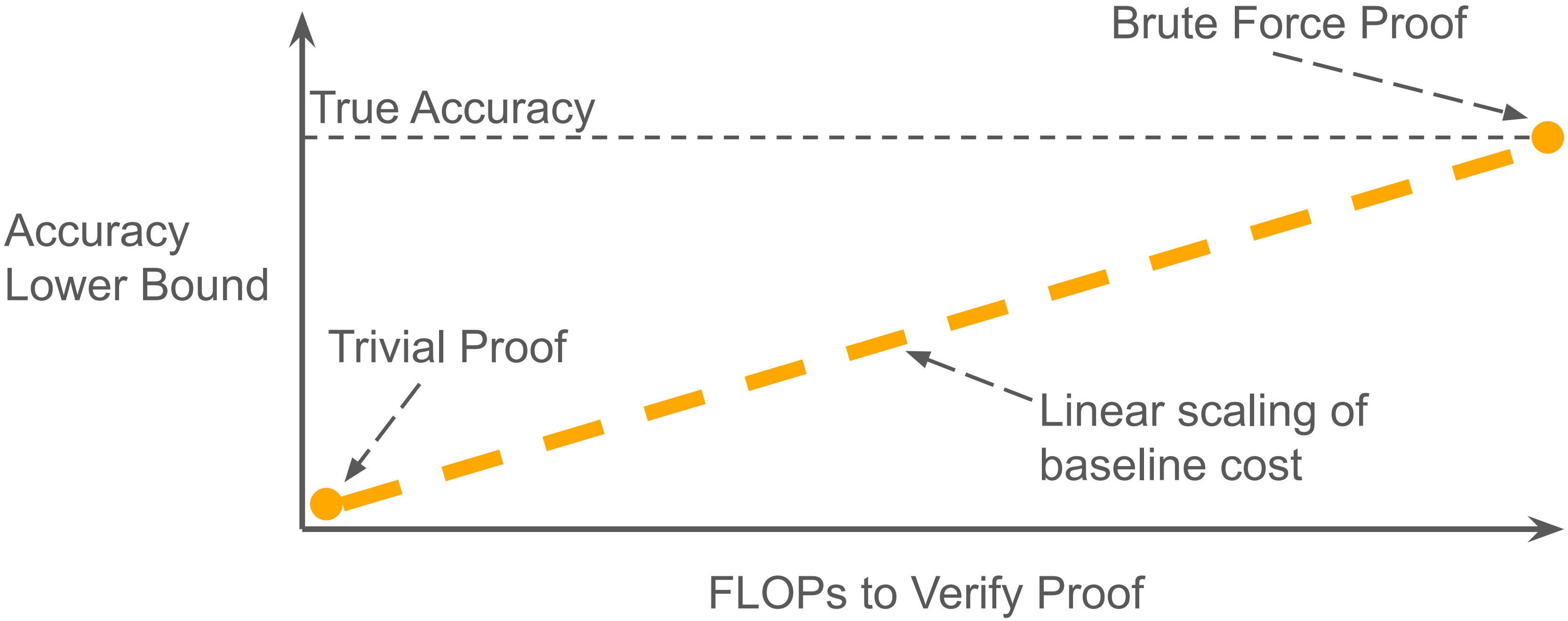
Proof = sound computation of worst-case error (divergence in behavior)

Length of proof = cost of running computation

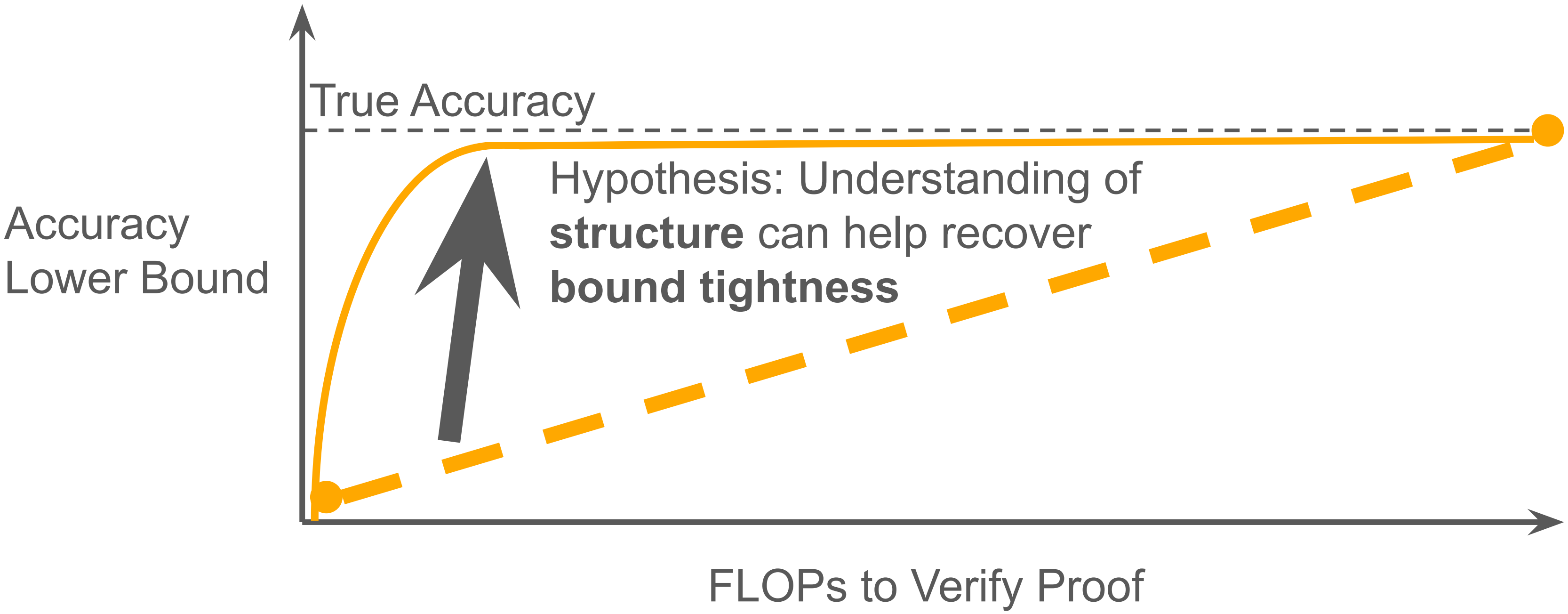


https://www.svgrepo.com/svg/29028/gzip-file-format-variant https://calebstanford.com/2019/01/15/coq-vector-image/

Quantifying the compute-cost of explanations

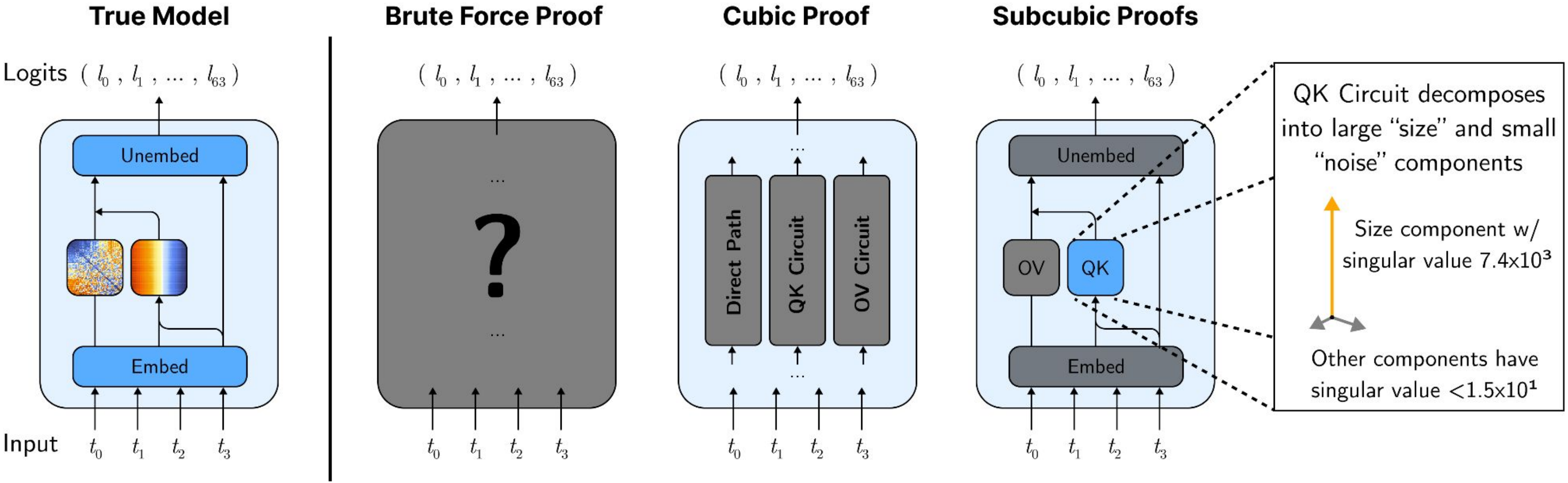


Does understanding improve upon the linear baseline?



Hypothesis: Understanding of **structure** can help recover **bound tightness**

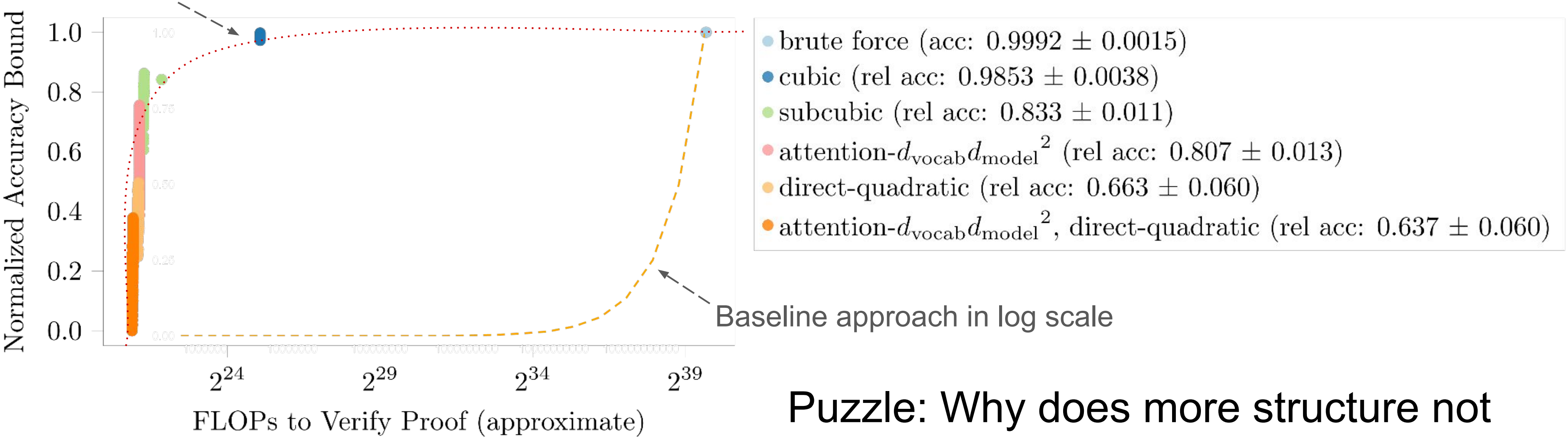
Proofs with varying mechanistic understanding


True Model
Brute Force Proof
Cubic Proof
Subcubic Proofs
Logits
$(l_0, l_1, \ldots, l_{63})$
Unembed
Embed
Input
t_0 t_1 t_2 t_3
?
Direct Path
QK Circuit
OV Circuit
OV
QK
QK Circuit decomposes into large "size" and small "noise" components
Size component w/ singular value 7.4x10³
Other components have singular value <1.5x10¹

FLOPs Required:	8.76×10^{11}	3.52×10^{7}	2.08×10^{6}
Accuracy Lower Bound:	99.92%	98.45%	79.7%
Effective Dimension:	1.67×10^{7}	1.28×10^{5}	1.03×10^{5}
Asymptotic Complexity:	$O(d_{\text{vocab}}{}^{\text{context}})$	$O(d_{\text{vocab}}{}^{3} \cdot \text{context})$	$O(d_{\text{vocab}} \cdot d_{\text{model}}{}^{2} \cdot \text{context})$

We found an empirical “pareto frontier”

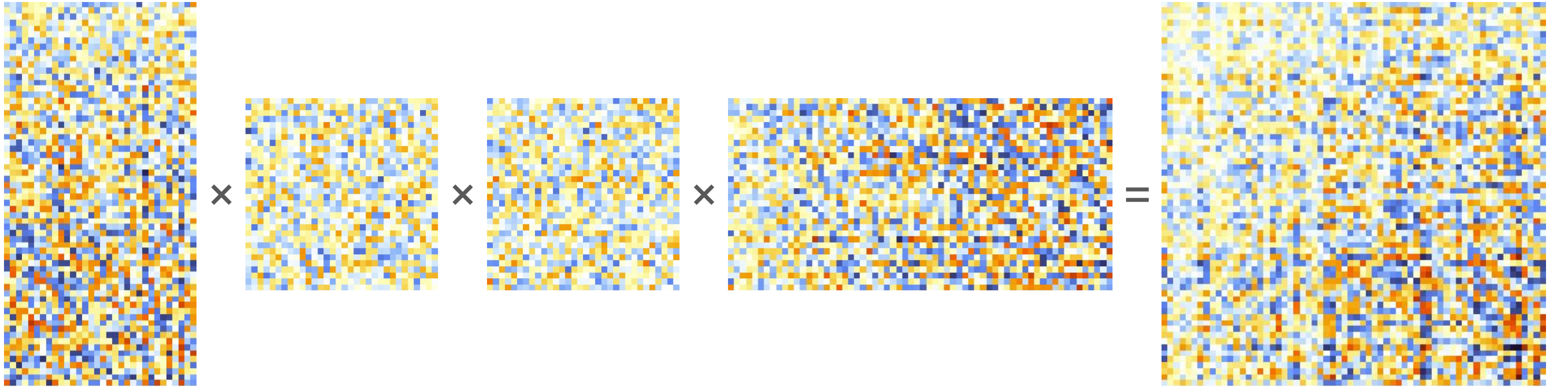
Pareto frontier from incorporating mechanistic understanding



Baseline approach in log scale

Puzzle: Why does more structure not always mean better bound?

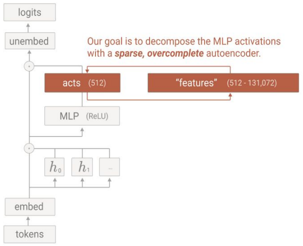
Compounding errors from lack of structure



Approximation Strategy	Result	Complexity
(exact) max row diff	≈ 1.8	$(\mathcal{O}({d_\text{vocab}}^2 d_\text{model}))$
2 · (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	$(\mathcal{O}(d_\text{vocab} d_\text{model}))$

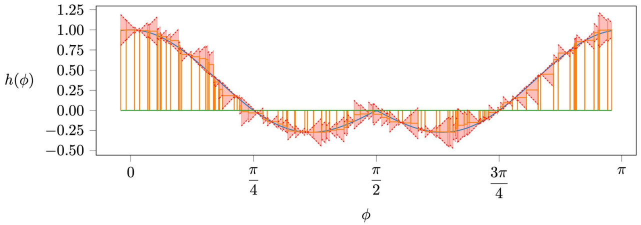
Applying Compact Proofs

- Optimization targets for representation search (SAEs)

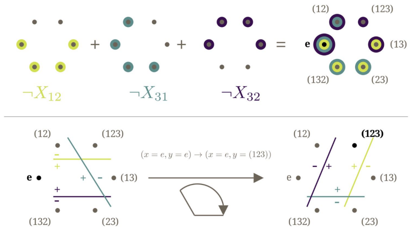

logits
unembed
Our goal is to decompose the MLP activations with a ***sparse, overcomplete*** autoencoder.
acts (512)
"features" (512 - 131,072)
MLP (ReLU)
h_0
h_1
...
embed
tokens

Run A/0 (512 features)
Run A/1 (4,096 features)
Run A/2 (16,384 features)
Arabic script features
base64 features
Ultralow density and dead neurons

- Compressing MLPs (integration)

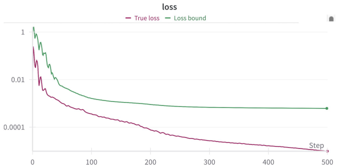


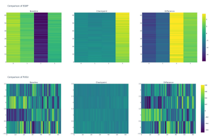
- Ground truth for comparing mech interp approaches (groups)



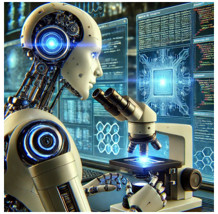
Some images from https://transformer-circuits.pub/2023/monosemantic-features

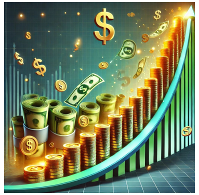
Open Problems for Scaling Compact Proofs





- Fix compounding errors
  - Fine-tuning; or heuristic arguments; or sampling
- Suppress exponential in # layers
  - Toy model: induction heads
- Autoformalize proofs
  - AlphaProof
- Autointerp
  - Step 2: ???
- Step 3: Profit





Images by GPT-4o

https://deepmind.google/discover/blog/ai-solves-imo-problems-at-silver-medal-level/