We use **mechanistic understanding** to compress **proofs of model performance** on toy transformers

Compact Proofs of Model Performance via Mechanistic Interpretability

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

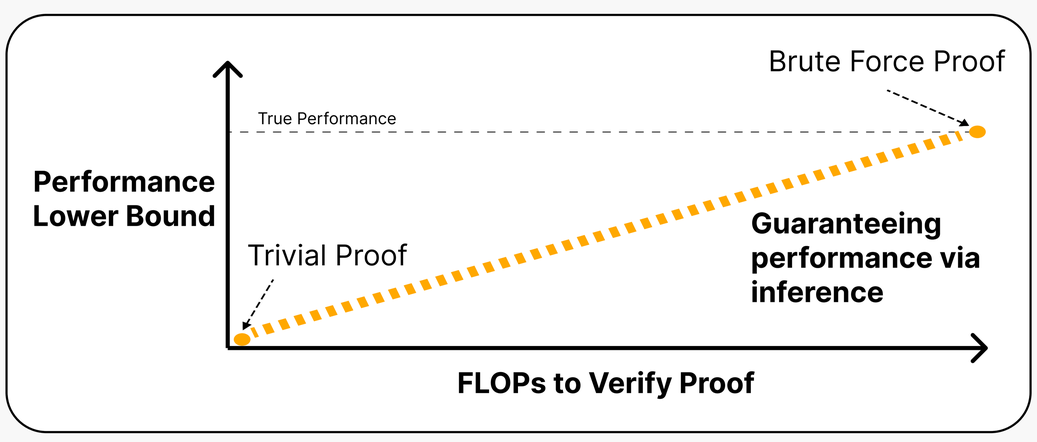
$$\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[f(\mathcal{M}(\mathbf{x}))] \geq b$$

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 every input.

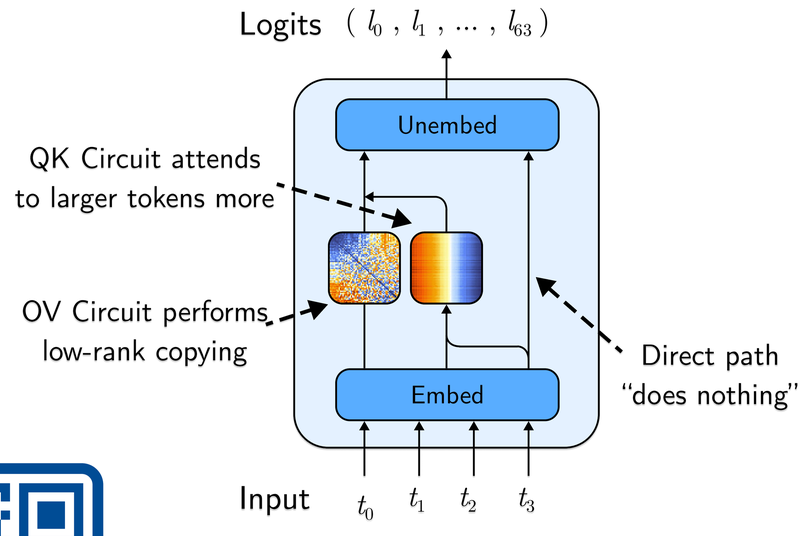
Algorithm 1 Counting Correct Sequences By Brute Force

function CORRECTNESS($\mathcal{M}$, input-sequence)
  return MODEL-BEHAVIOR($\mathcal{M}$, input-sequence) == MAX(input-sequence)
end function
function BRUTE-FORCE(d_{vocab}, n_{ctx}, $\mathcal{M}$)
  return $\frac{1}{{d_{\text{vocab}}}^{n_{\text{ctx}}}}$SUM(CORRECTNESS($\mathcal{M}$, tokens) **for** tokens $\in$ (RANGE(d_{vocab}))$^{n_{\text{ctx}}}$)
end function

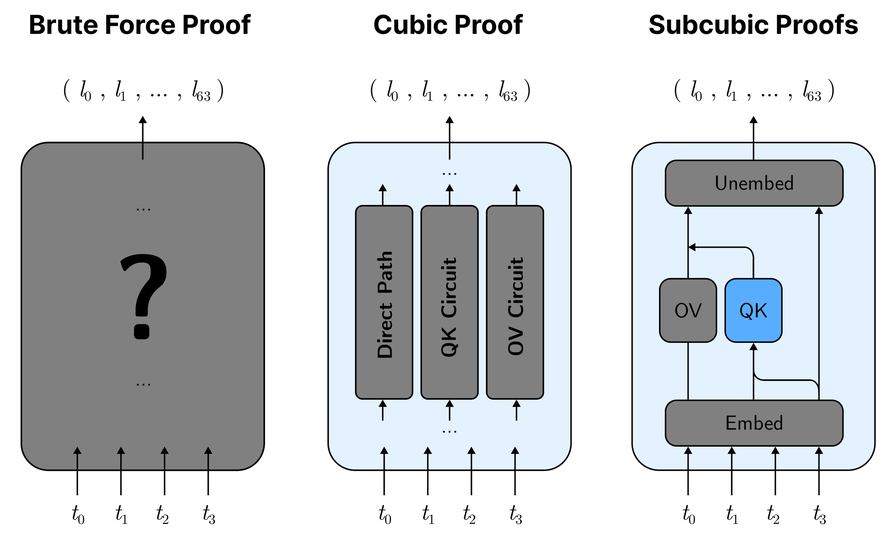
We can run inference on various inputs to guarantee model performance on those inputs. Can mechanistic understanding provide a **compression** of model behavior that beats this inefficient baseline?



We **prototype** the proofs-based approach to formal interpretability on transformer models trained on the the max_of_k task.

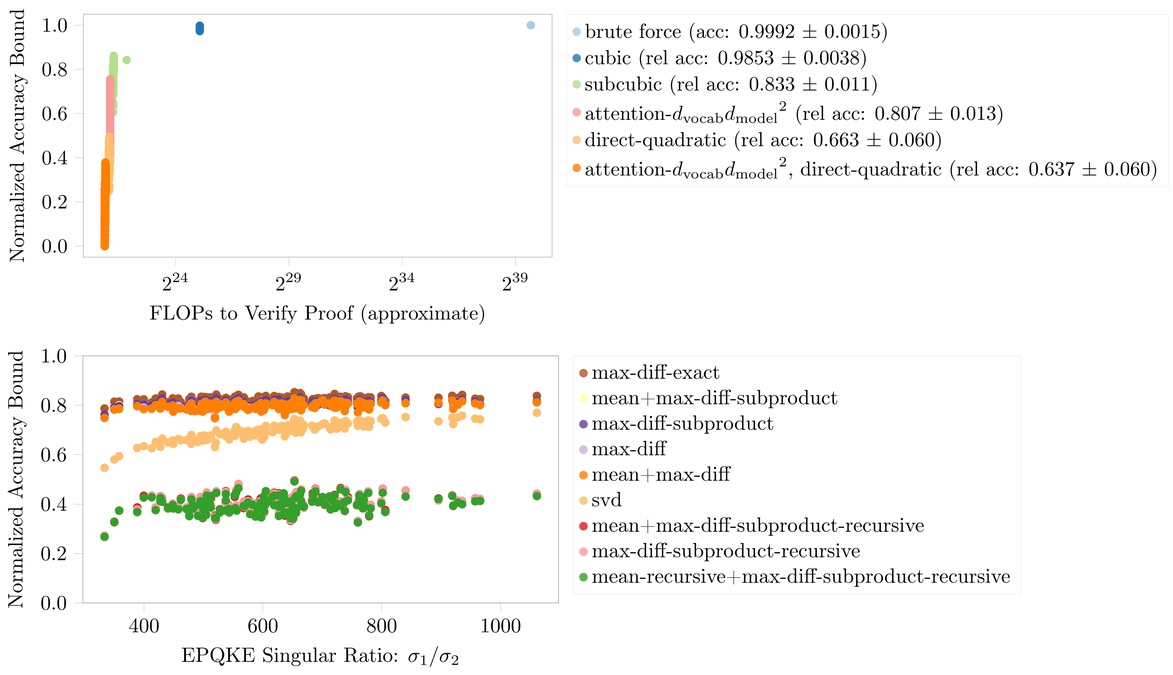


We implement many proof strategies, which incorporate varying degrees of **mechanistic detail**...

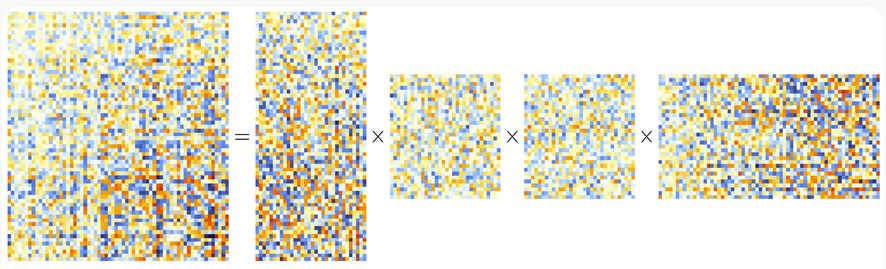


… and find that more mechanistic detail results in **tighter performance bounds**.

Moreover, for any given proof length, proving a tighter performance bound requires more mechanistic detail.



We identify **compounding approximation error** in post-hoc analysis as 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	$(\mathcal{O}(d_{\text{vocab}} d_{\text{model}}))$

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