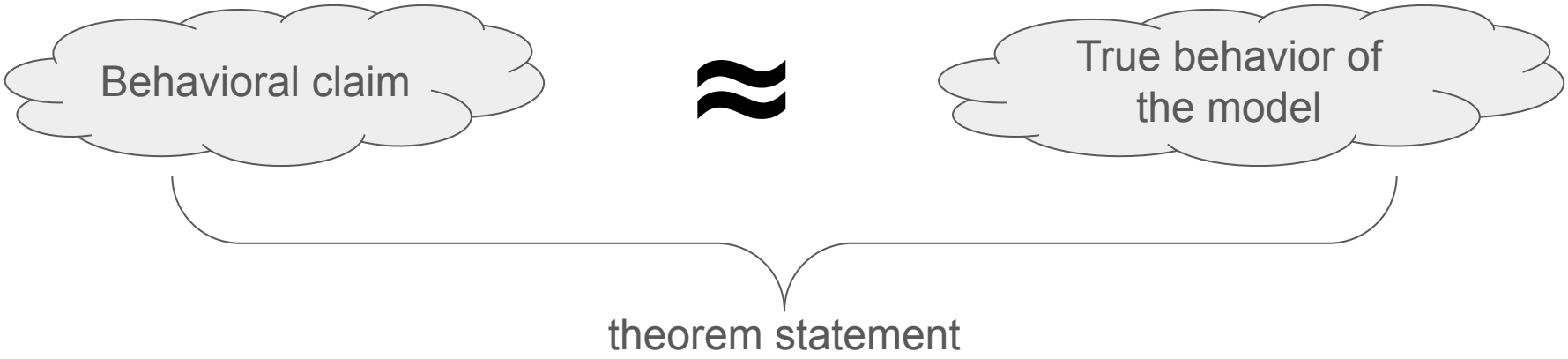
Compressing explanations

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*

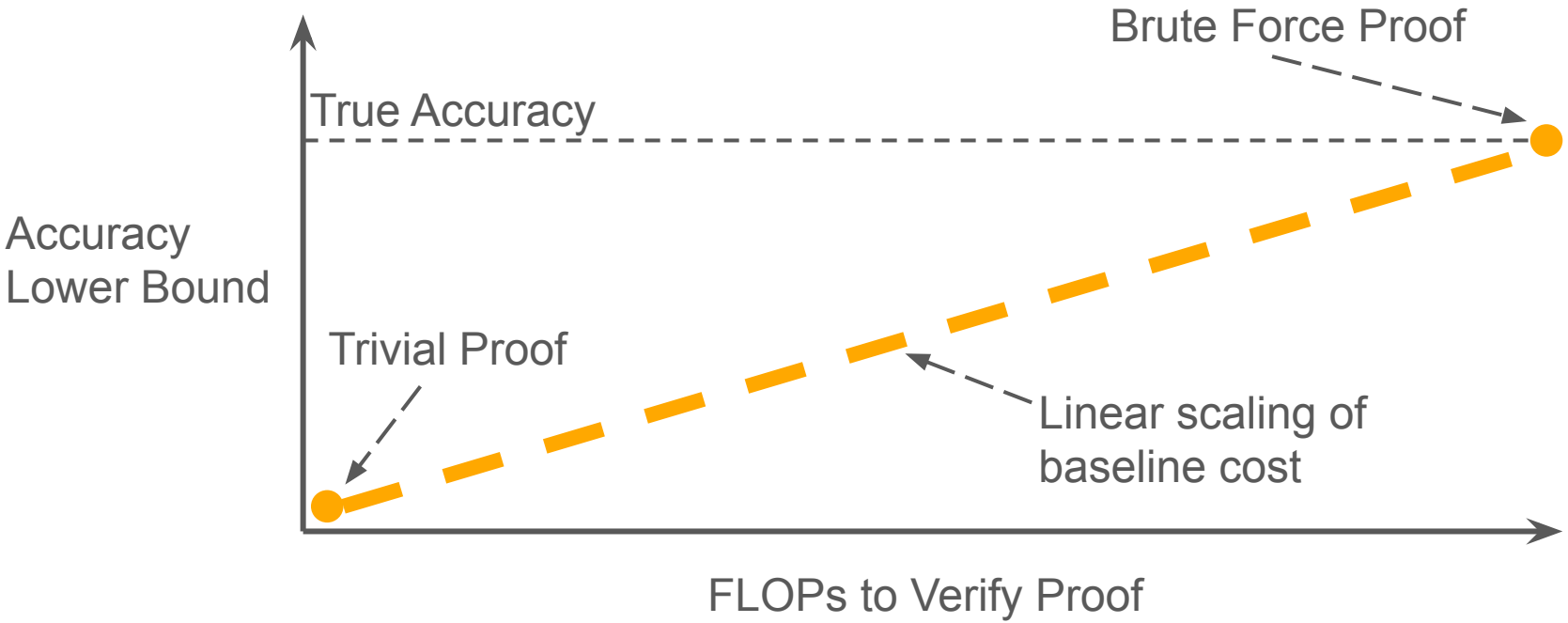
Formalizing proof length to quantify compression



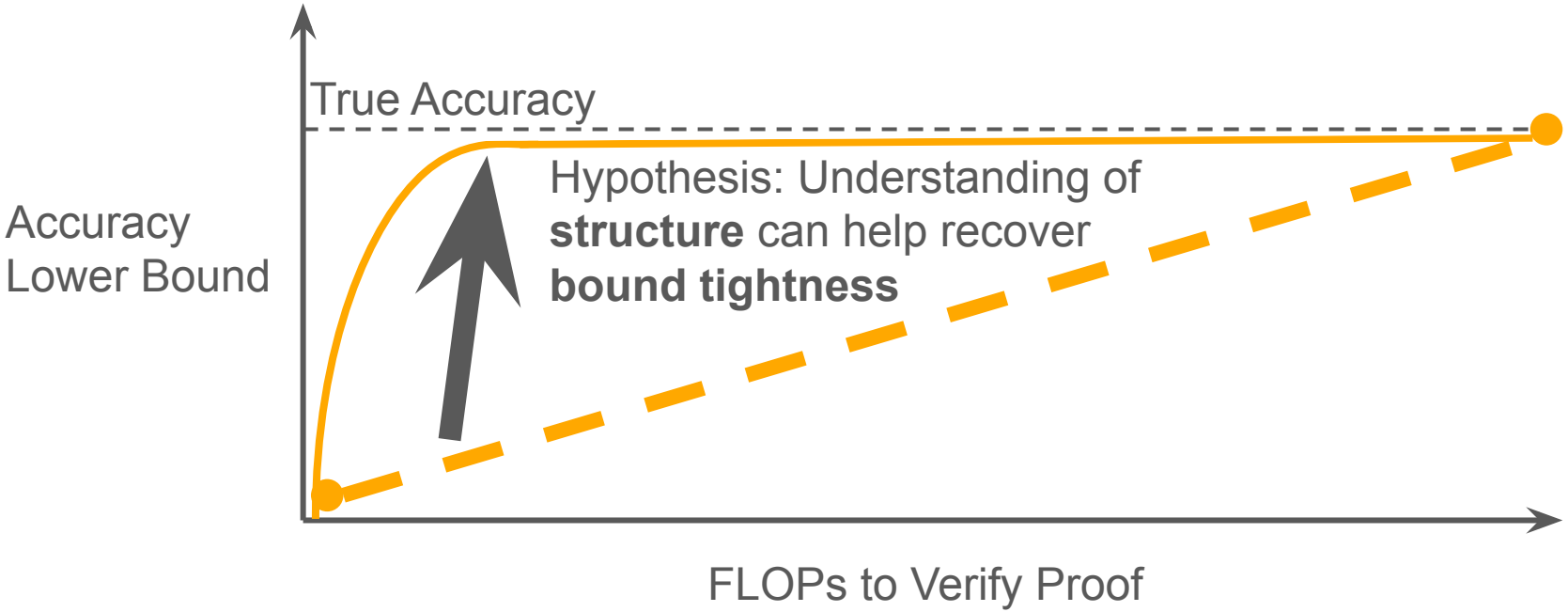
Proof = sound computation of worst-case error

Length of proof = cost of running computation

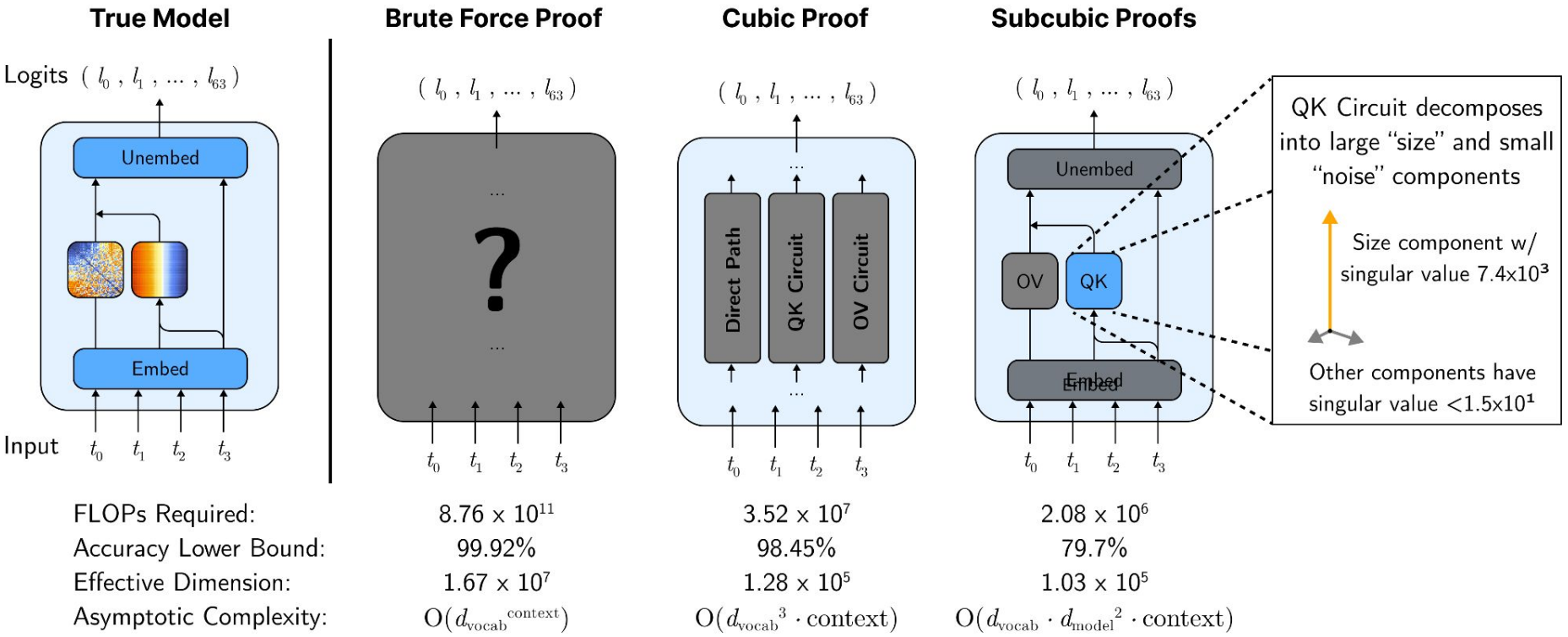
Quantifying the compute-cost of explanations



Does understanding improve upon the linear baseline?



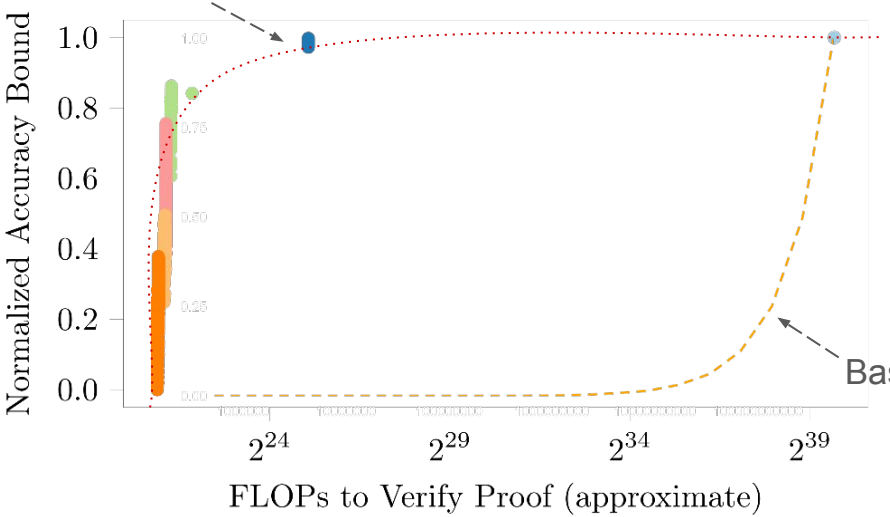
Proofs with varying mechanistic understanding



	True Model	Brute Force Proof	Cubic Proof	Subcubic Proofs
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

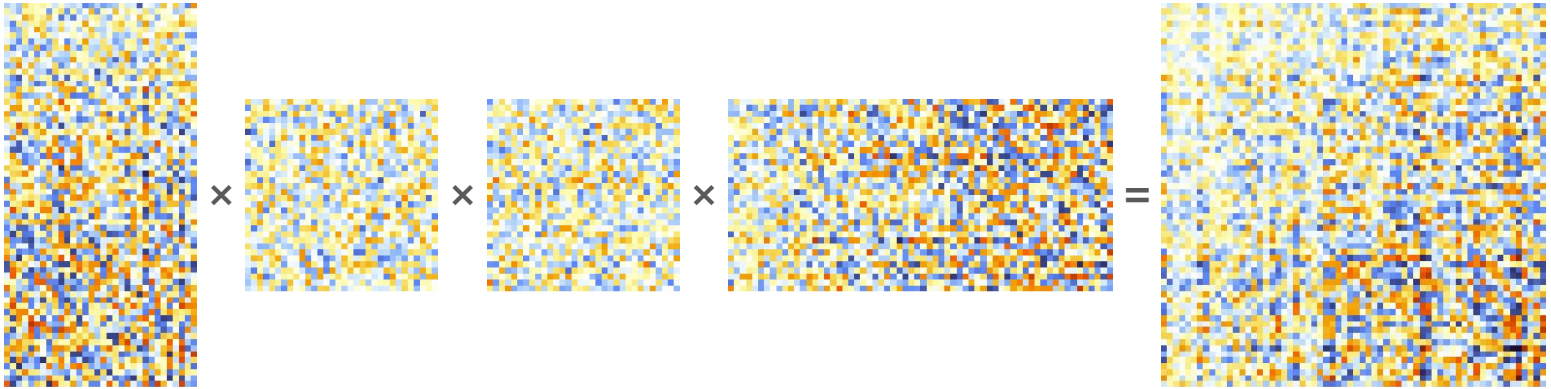


- brute force (acc: 0.9992 ± 0.0015)
- cubic (rel acc: 0.9853 ± 0.0038)
- subcubic (rel acc: 0.833 ± 0.011)
- attention-$d_{\text{vocab}}d_{\text{model}}{}^2$ (rel acc: 0.807 ± 0.013)
- direct-quadratic (rel acc: 0.663 ± 0.060)
- attention-$d_{\text{vocab}}d_{\text{model}}{}^2$, direct-quadratic (rel acc: 0.637 ± 0.060)

Baseline approach in log scale

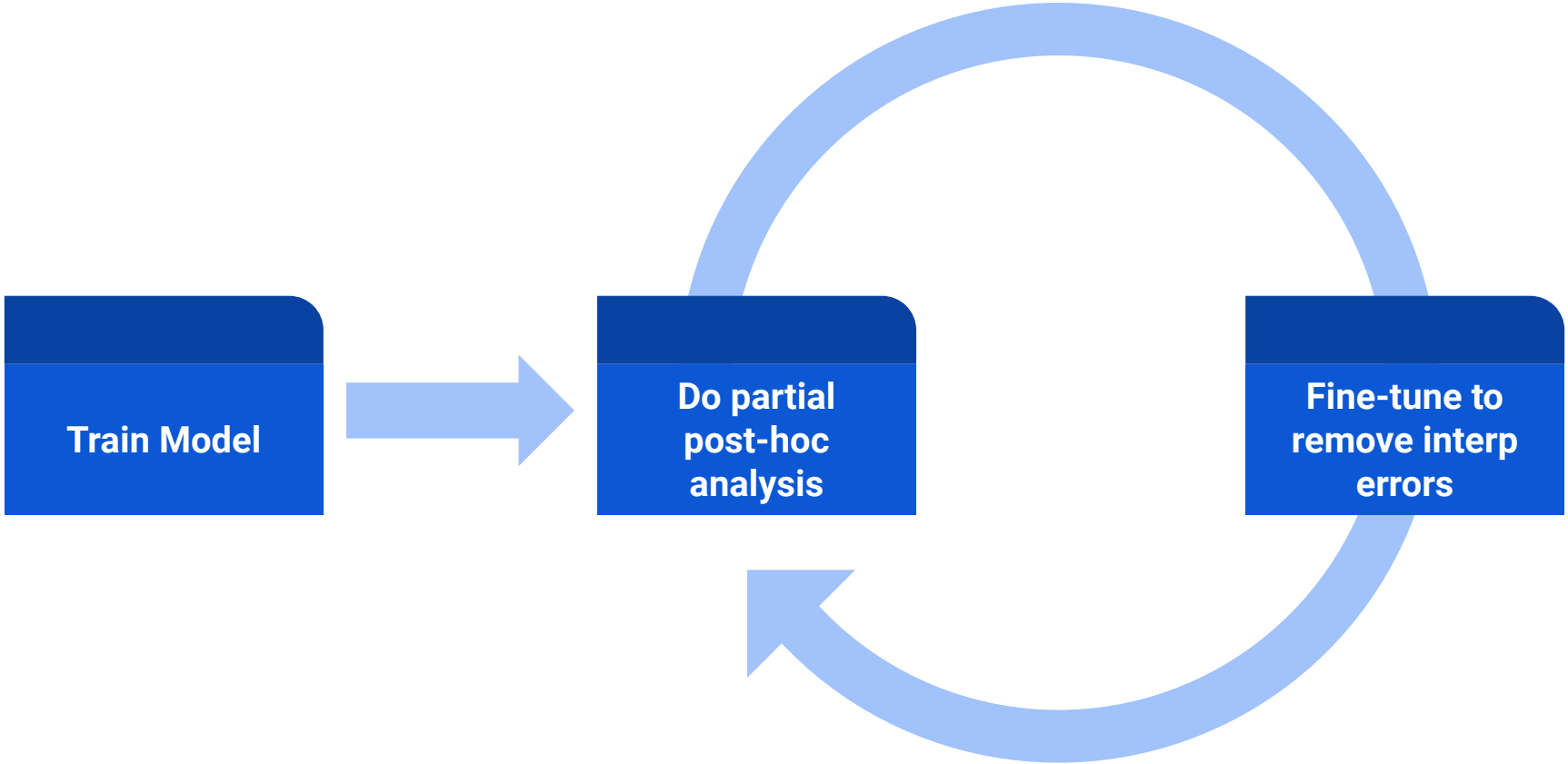
Puzzle: Why does more structure not always mean better bound?

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	$(\mathcal{O}(d_{\text{vocab}} d_{\text{model}}))$

Wanted: Compression of highly expressive systems



Check out our poster!

Scan for paper

Scan for Poster