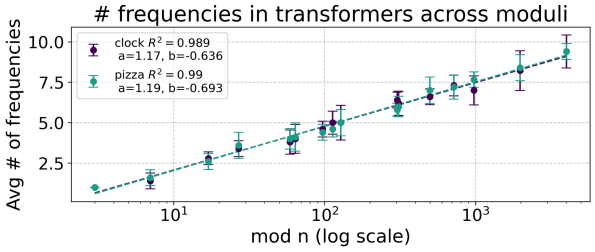
Unifying Mechanistic Interpretations of Neural Networks Trained on Modular Addition

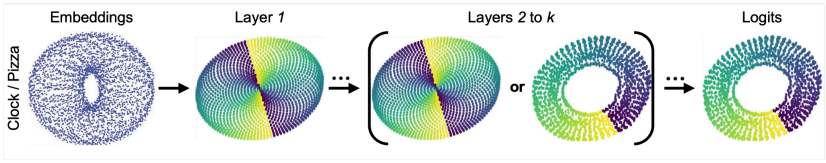
Mechanistic interpretability aims to uncover how neural networks implement specific computations. Modular addition has emerged as a benchmark task in this setting [1, 2]. Prior influential work argued that transformers trained on this task yield two disjoint mechanisms–Clock and Pizza circuits–providing a counterexample to the universality hypothesis, which posits that networks trained on the same data recover the same solution. This result has been influential because it suggests fundamental limits for interpretability and that identifying circuits in larger models may be far more difficult than expected.

We revisit these claims and analyze the exact transformer architectures studied in [1, 2]: standard sigmoidal attention (Clock interpretation) and constant attention (Pizza interpretation). Our main finding is that the two are not distinct. Instead, they implement the same abstract algorithm and learn geometrically equivalent representations. We support this claim from two perspectives:



Algorithmic. All networks we study implement the same *simple neuron model* in their first-layer MLPs: degree-1 sinusoidal fits in layer 1 of the form $\cos(2\pi fa/n + \phi_a) + \cos(2\pi fb/n + \phi_b)$, with deeper layers combining into degree-2 sinusoidal interactions. We verify this pattern across numbers of layers, seeds, hyperparameters, and, importantly, a broader set of architectures beyond those in [1, 2], with R^2 fits consistently matching our model (contrasting with [1], which reported degree-2 fits for all neurons). Taken together, these components realize what we term the *approximate Chinese Remainder Theorem* (aCRT): modular addition is represented by composing a small number of sinusoidal features. We prove that only a logarithmic number of frequencies is required and confirm this prediction empirically. Figure 1 shows the scaling behavior: both Clock and Pizza architectures obey the same logarithmic scaling, with nearly identical constants.

Geometric. Beyond the algorithmic view, we refine our analysis by examining the geometry of the learned representations. The *phases* in our simple neuron model appear along the $\phi_a = \phi_b$ diagonal of the input space for these models, so Clock and Pizza transformers align not only functionally but also geometrically. Embeddings, pre-activations, and logits cluster along equivalent manifolds, showing that the two are different parameterizations of the same representation. Figure 2 illustrates this unified schematic. Moreover, both architectures contain both motifs: diagonally-constrained degree-1 "Pizza" neurons in the first layer, and degree-2 "Clock" interactions emerging in later layers and logits. We verify this using *persistent homology*.



Our findings unify the literature: the supposed Clock and Pizza circuits are not competing interpretations but specific instantiations of the same mechanism. This restores the universality hypothesis in this setting, suggesting that interpretability may be less brittle than feared. By fully resolving this benchmark case of modular addition, we provide a concrete case study in robust interpretability: these networks converge to the same algorithmic and geometric solution across architectures, seeds, and hyperparameters. More broadly, our results suggest that universality may reside in the structure of the learned manifolds, shaped by both the data and architectural constraints. This underscores the importance of examining geometric (and topological) properties of representation manifolds–not just individual features–when developing methods for robust interpretability.

[1] Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, Jacob Steinhardt. "Progress measures for grokking via mechanistic interpretability." ICLR 2023.
[2] Ziqian Zhong, Ziming Liu, Max Tegmark, Jacob Andreas. "The Clock and the Pizza: Two Stories in Mechanistic Explanation of Neural Networks." NeurIPS 2023.