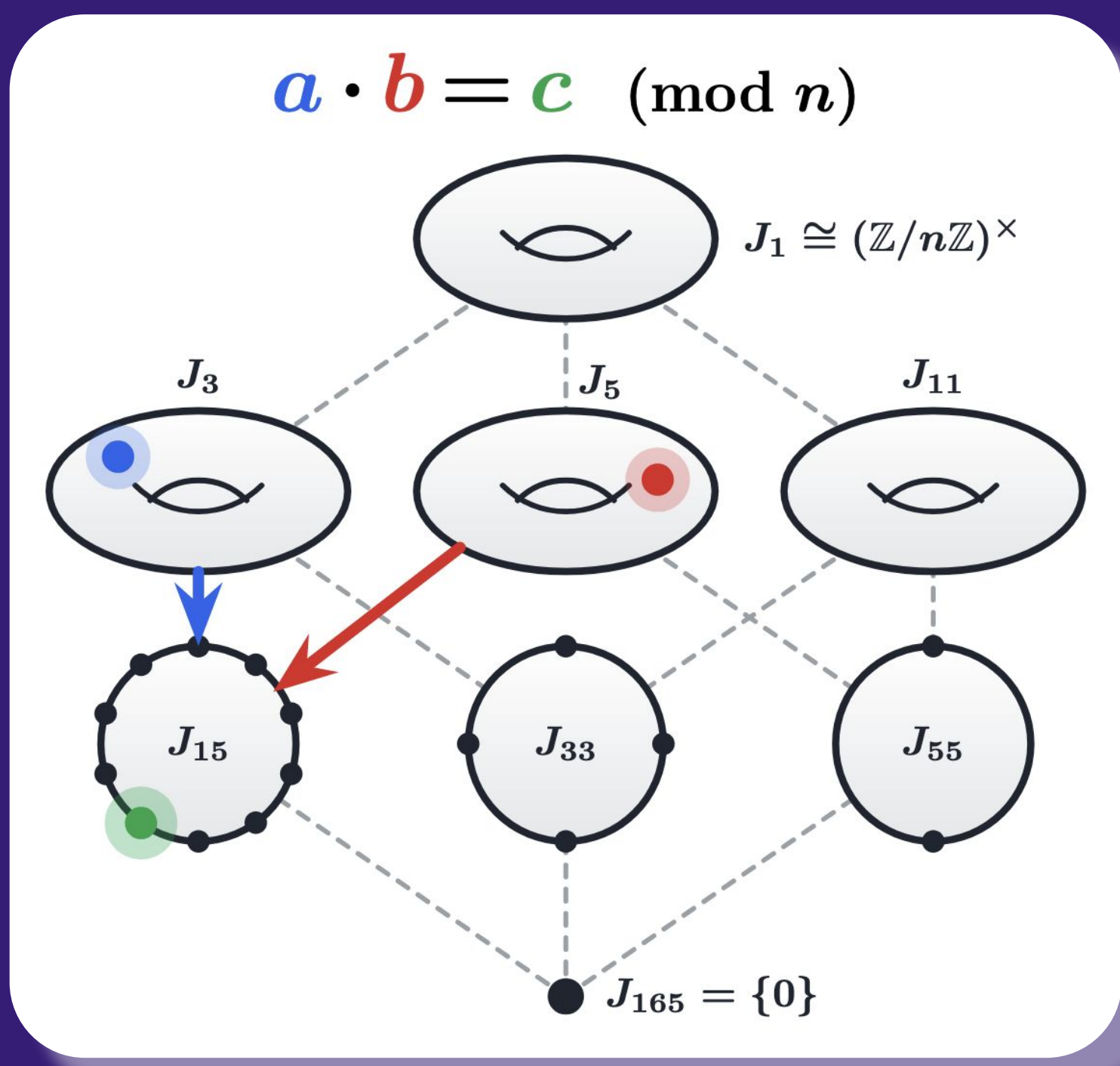




Models stratify non-invertible operations into factorized $\mathcal{J}$ -class subspaces



Multiplication Beyond Groups: Stratified Fourier Mechanisms in Transformer Circuits



Zitong Chen*, Junaid Hasan*, Akhil Srinivasan*, Hemkesh Bandi, Jarod Alper

1 Summary

- Against non-invertible operations, the model learns to **partition input space** into factorized subspaces.
- Within partitioned stratas, previous group composition algorithms **survives**.
- Because of this, model must also learn to **route residual stream** and **score locally**.

2 Motivating Question

How does the model learn

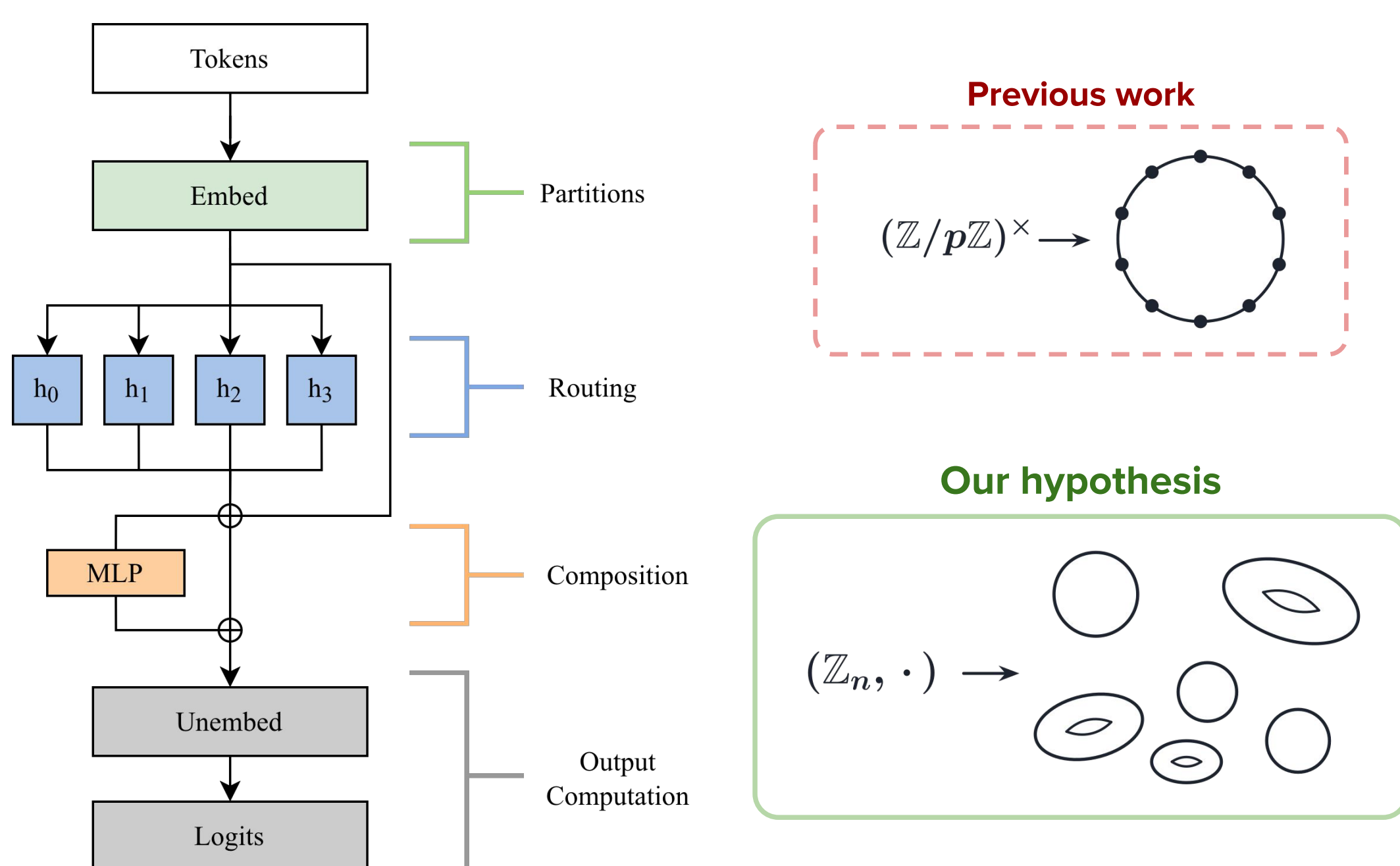
$$a \cdot b = c \pmod{n}^*$$

- Example: $n = 12$, then $3 \cdot 4 = 12 = 0$, and $3 \cdot 8 = 24 = 0$, therefore there are multiple candidates that multiply with 3 and get us 0.
- Previous work has focused on groups, where operations are invertible.

*for square-free n

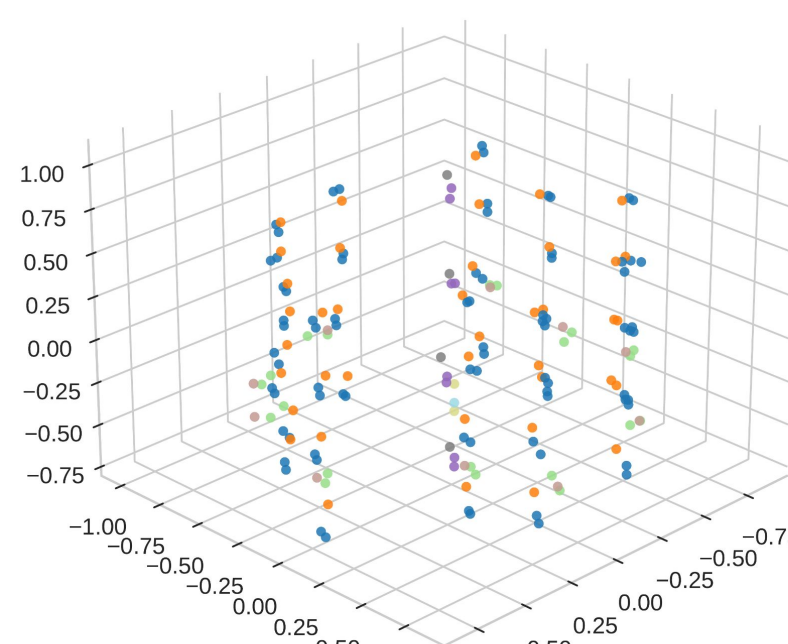
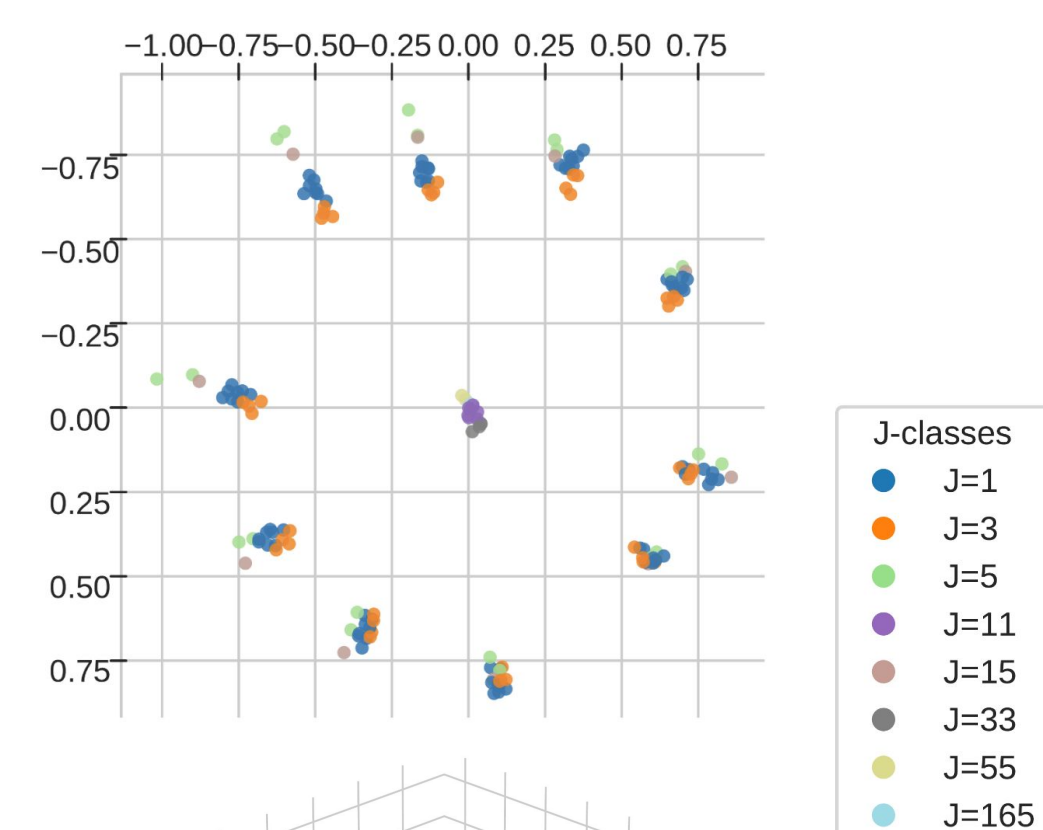
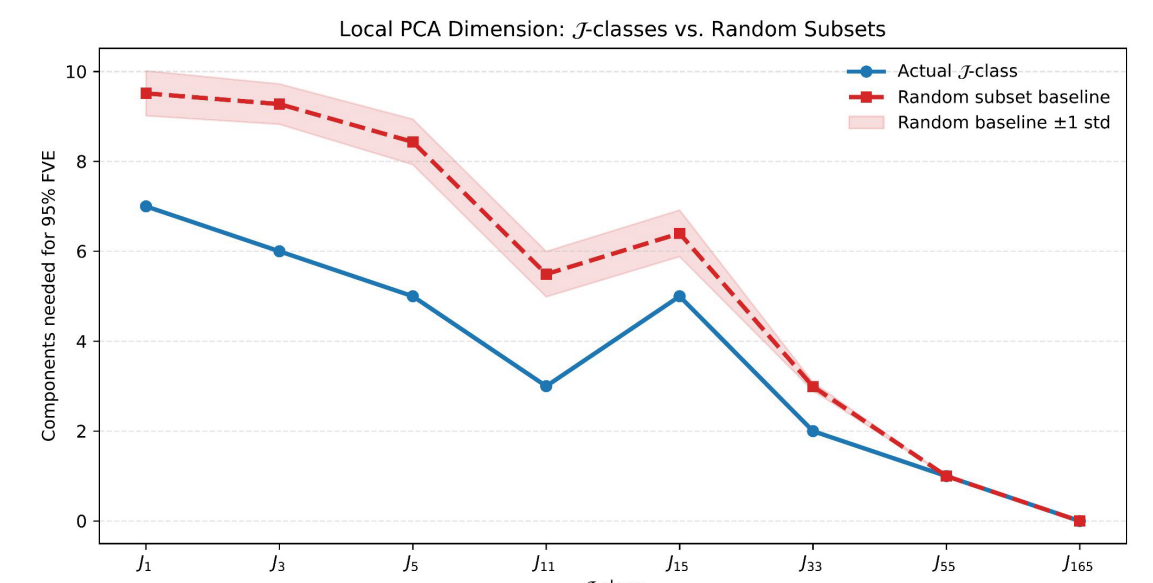
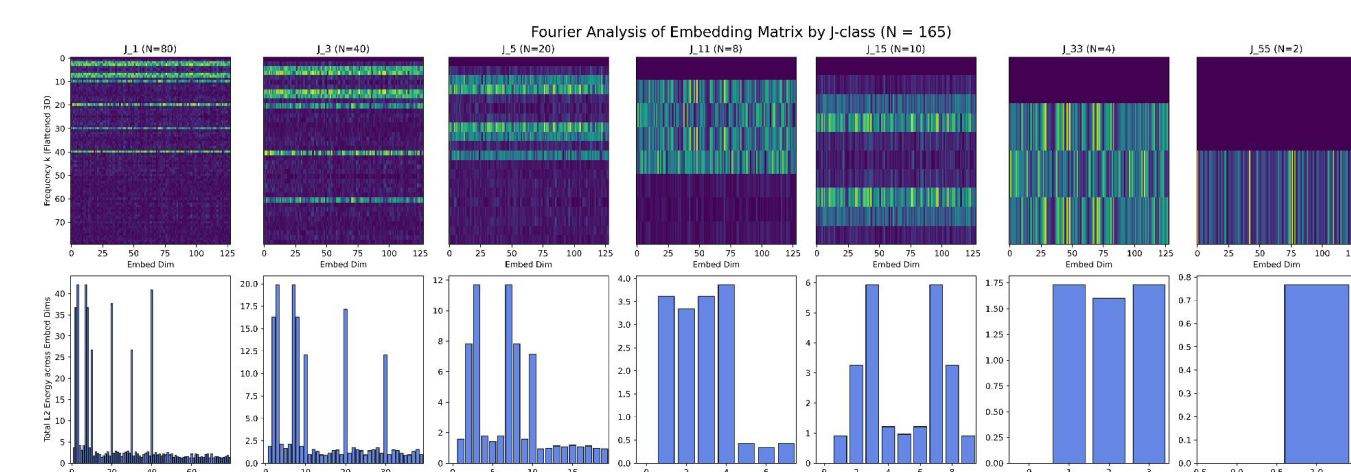
3 Hypothesis

- Embedding** partitions input space into algebraic subspaces based on gcd (J-classes), where “local inverses” exist.
- Attention** acts as a “router”, projecting the residual stream onto the appropriate subspace of the answer.
- Unembedding** utilizes local inverses to produce logit scores, similar to previous group-based algorithms.

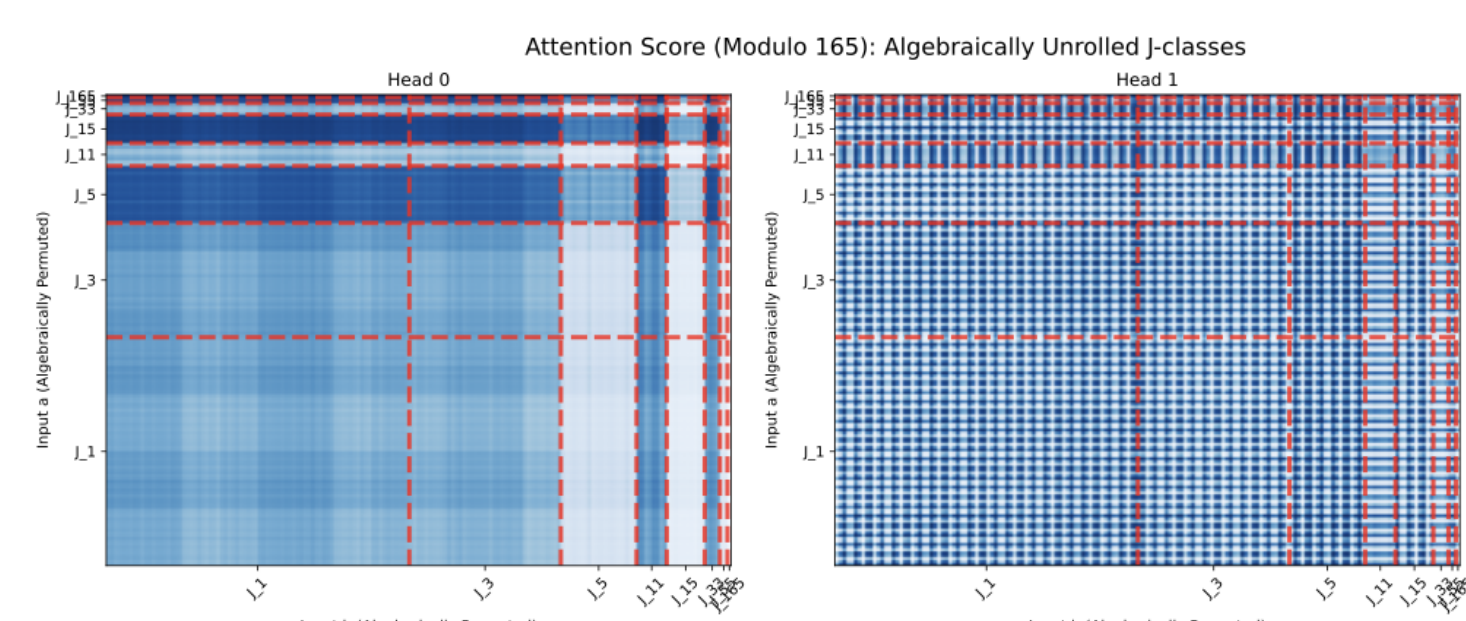


4 Findings

- The model does **not** learn a single representation subspace, instead **partitioning** into monoid J-classes.
- Attention Routing mechanism handles both **coarse J-class routing** and **finer in-class information**.



- For square-free numbers, each J-class is a group, and for each such group, the model isolates and reuses **target frequencies** across projections and circuits.
- MLP **activations** cluster by J-class and preserve Fourier structure.
- Unembedding layer uses character readouts with **local inverses** to compute logits.



5 Discussion & Future Work

- Investigation on non-square-free moduli**
 - Square-free moduli have natural density $\approx 61\%$. When moduli is not square-free, J-classes are no longer “regular” (no local inverses).
 - What (if any) factorization does the model learn?
- Progress measures of structure build-up**
 - Consistent with prior literature, our model experienced grokking. We hypothesize that stratification of inputs occur during this phase.



PAPER



CODE