Pretrained Large Language Models Use Fourier Features to Compute Addition

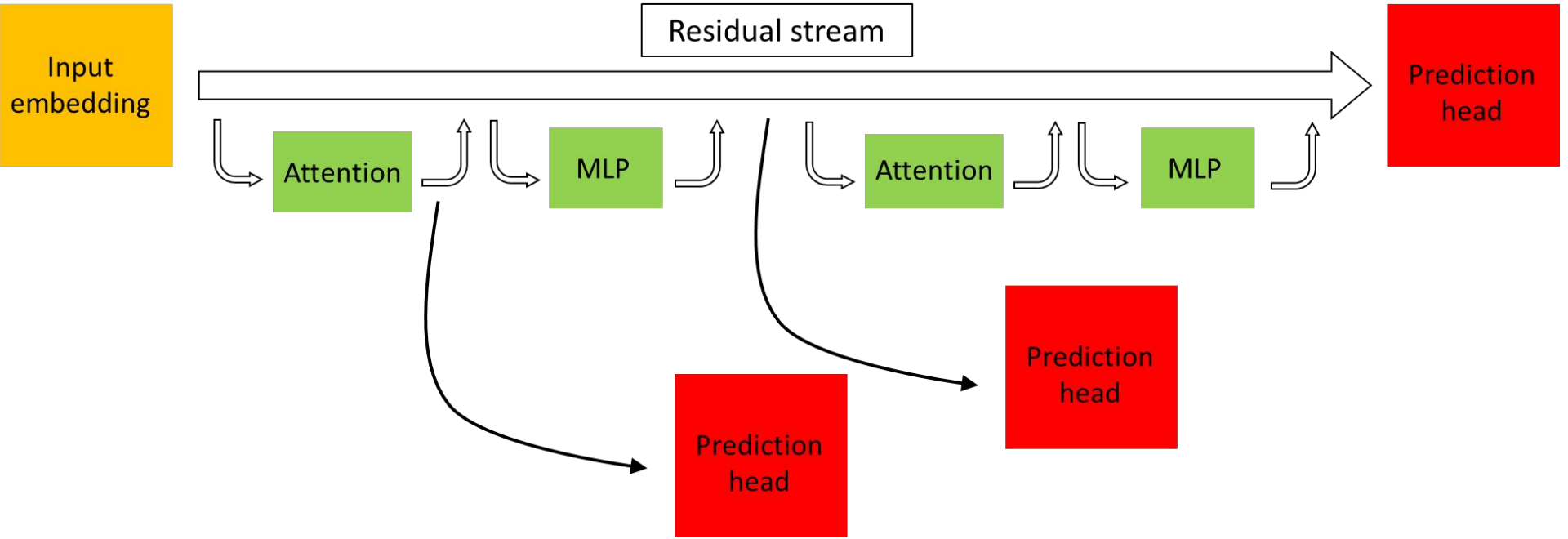
Tianyi Zhou, Deqing Fu, Vatsal Sharan, Robin Jia

Speaker: Tianyi Zhou

Unusual Error Patterns in LLM Addition Performance

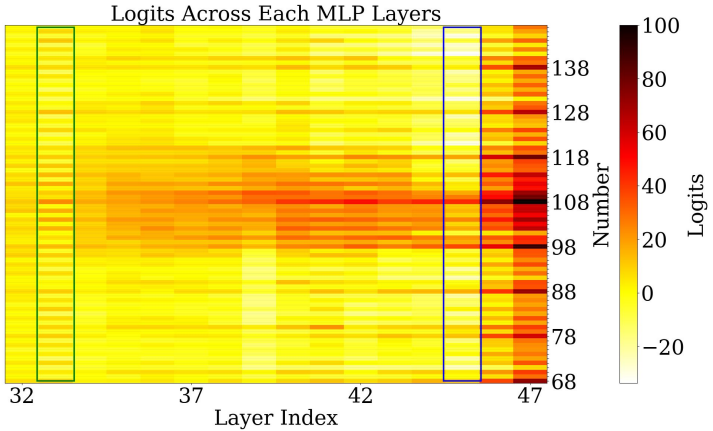
- Task: integer addition
- Open-Source Models
  - GPT-J: 93% errors are multiples of 10
  - Phi-2: 73% errors are multiples of 10
- Closed-Source Models (0-Shot)
  - GPT-3.5 & GPT-4: 100% errors are multiples of 10
  - PaLM-2: 87% errors are multiples of 10
- Takeaway: LLMs perform addition in unexpected ways, not simply remembering answers.

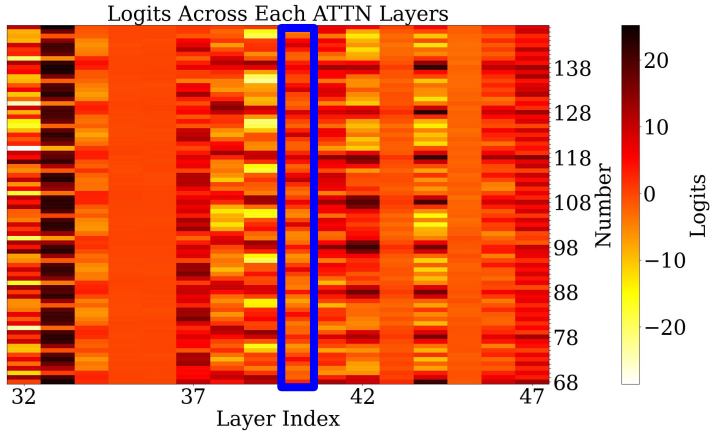
Logit Lens [Belrose, et al 2023]



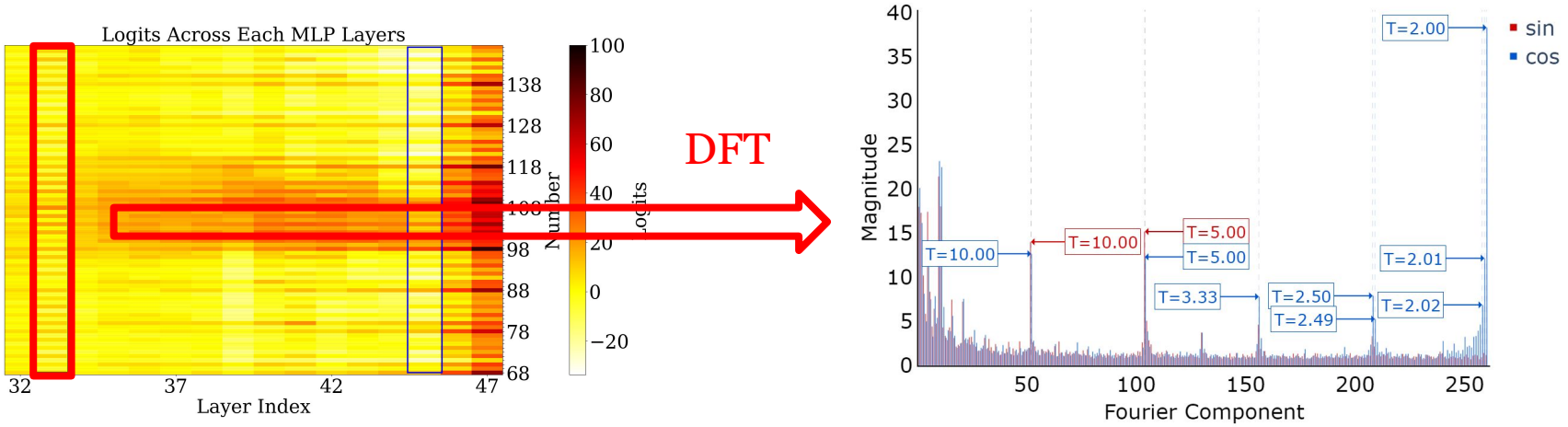
Observation from Logit Lens

- For example, for query: “What is the sum of 15 and 93?”
- Logits for MLP and attention have periodic structures
  - For example, some layers are predicting the parity of the answer.

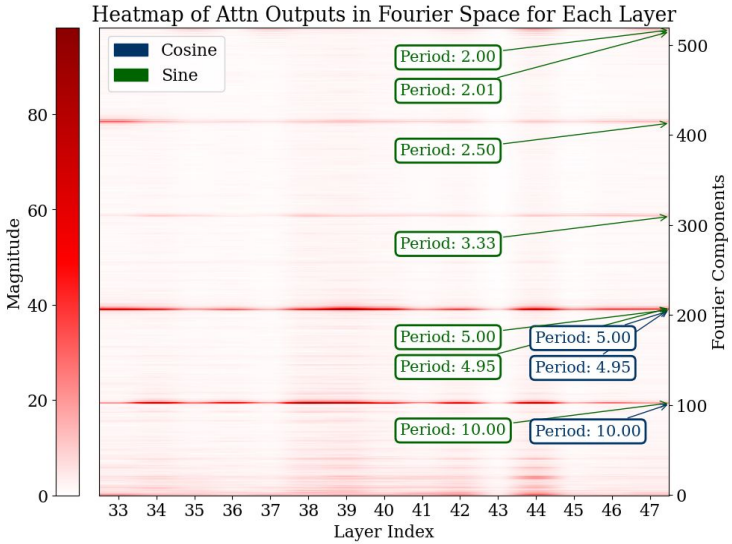


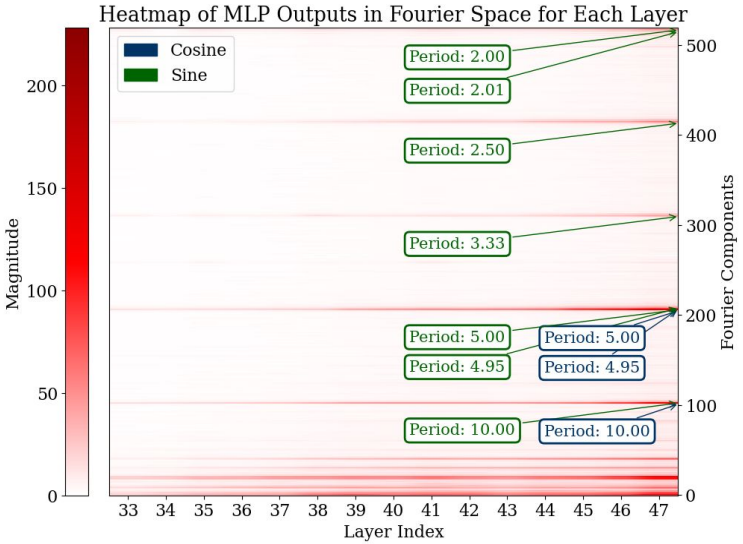


Discrete Fourier Transform in Prediction Space

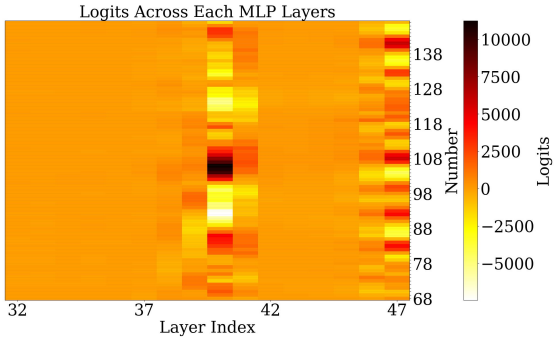


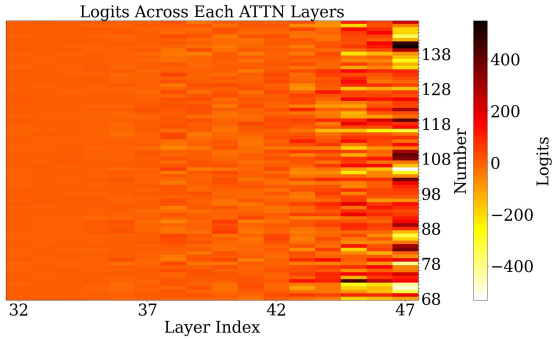
Logits for MLP and attention are approximately sparse in the Fourier space

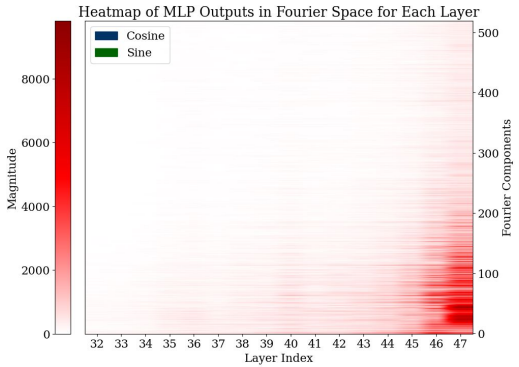


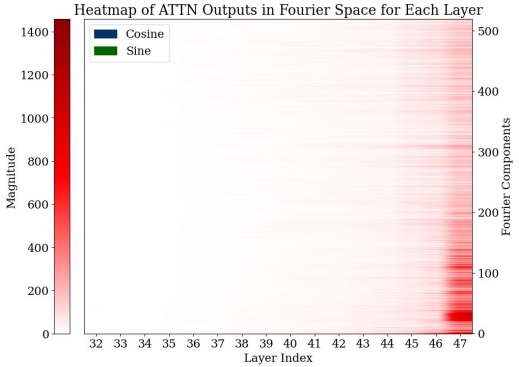


When train from scratch on addition, no more Fourier features are found.



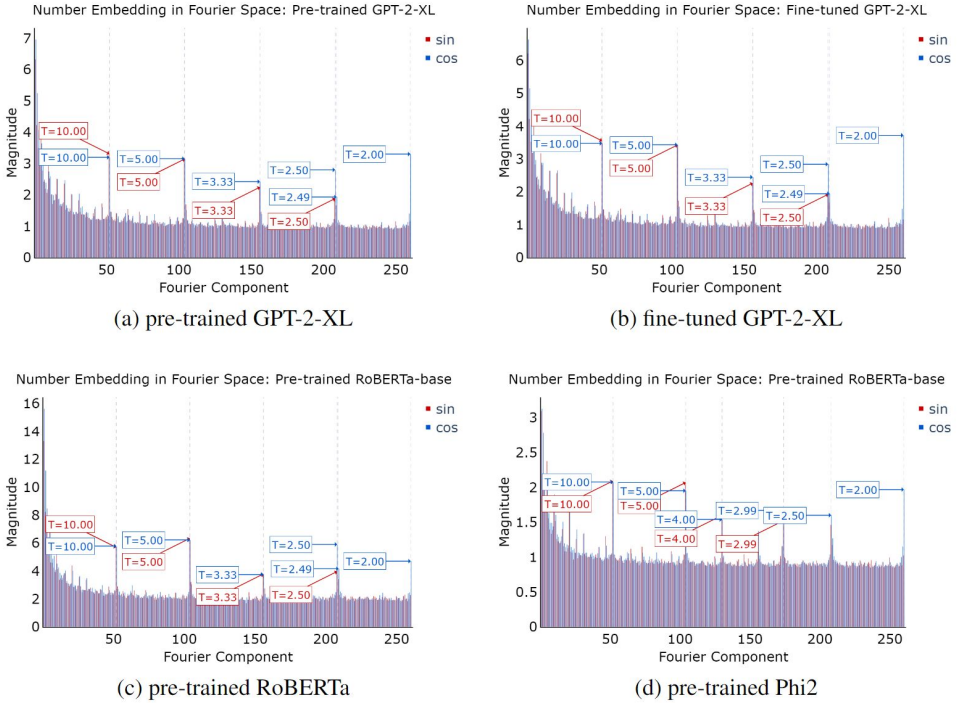






Where these Fourier Features come from?

- Fourier Features are learned in the number embedding during pre-training



(a) pre-trained GPT-2-XL

(b) fine-tuned GPT-2-XL

(c) pre-trained RoBERTa

(d) pre-trained Phi2

Figure 15: Number embedding in Fourier space for different pre-trained models.

Where these Fourier Features come from?

- With solely the pre-trained number embedding, GPT2-small is able to learn addition with 100% accuracy