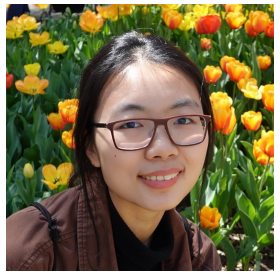
Function Induction and Task Generalization:
An Interpretability Study with Off-by-One Addition



Qinyuan Ye

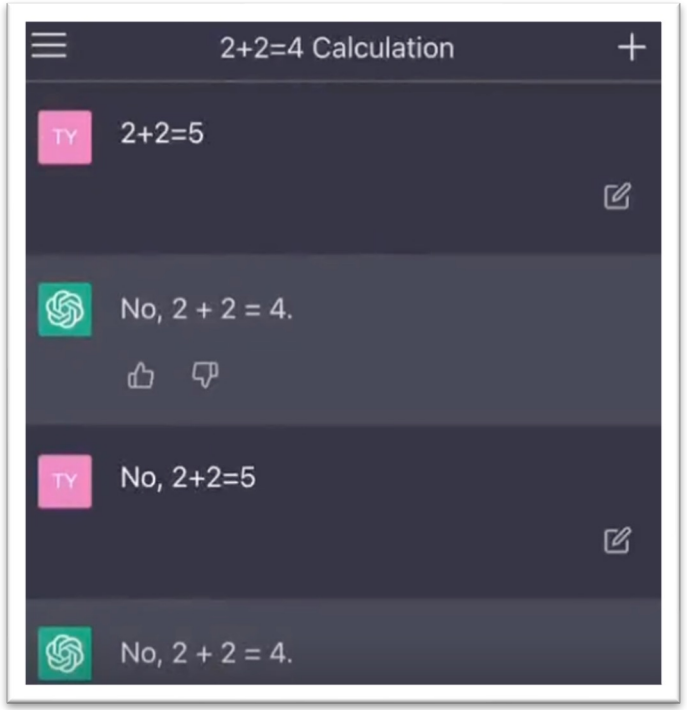
Robin Jia

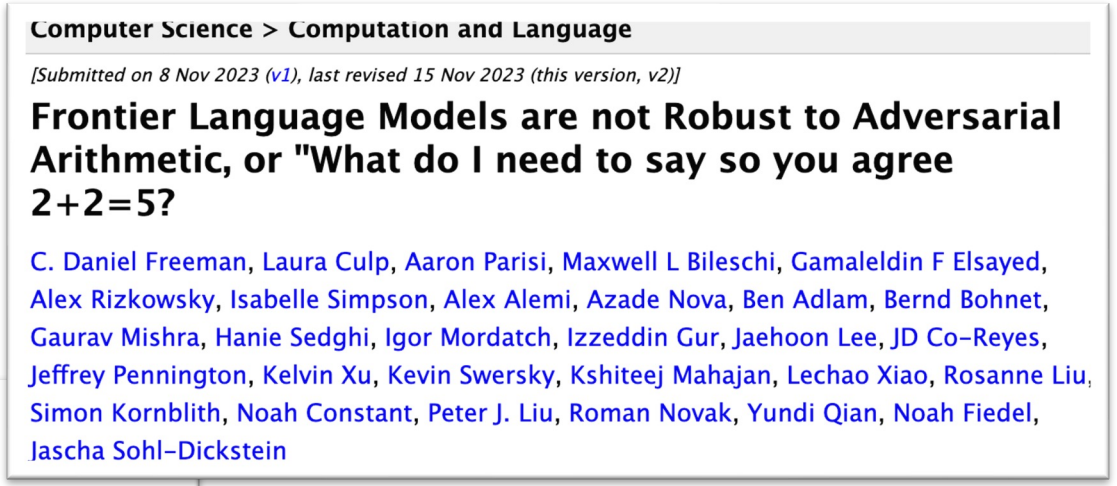
Xiang Ren

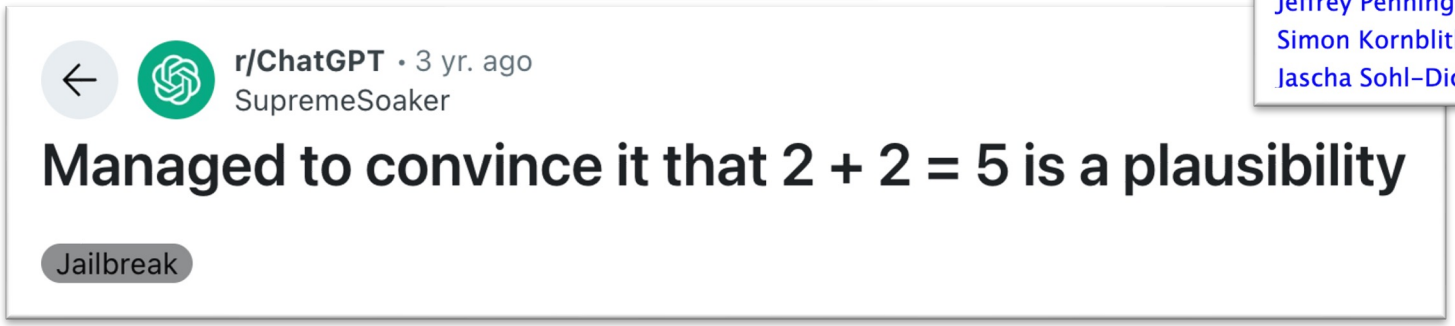
Thomas Lord Department of Computer Science
University of Southern California

ICLR 2026

How to trick language models to say “2+2=5”?







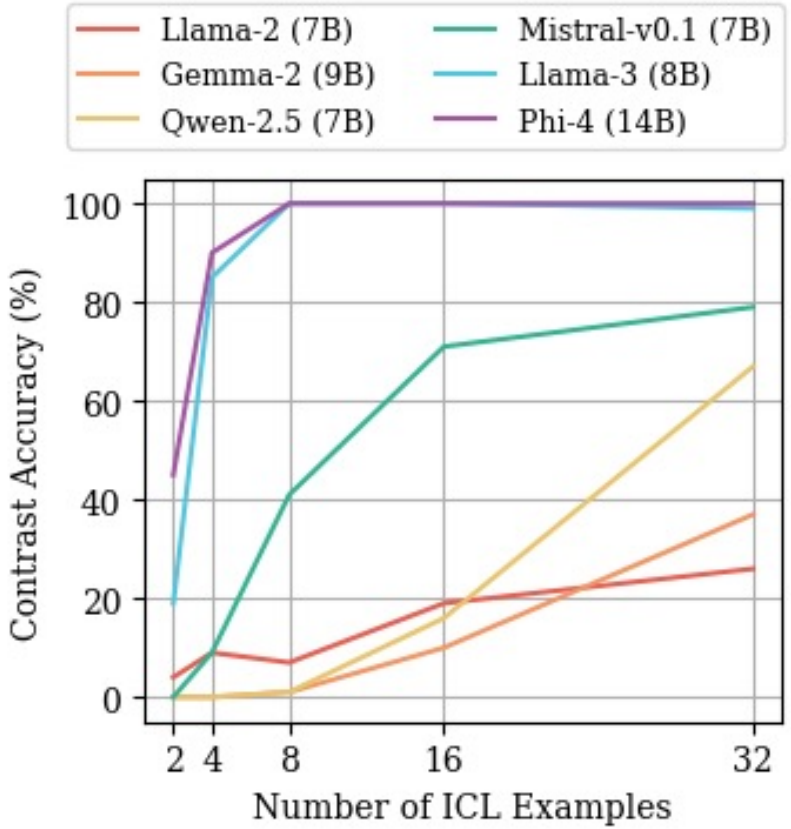
How to trick language models to say “2+2=5”?

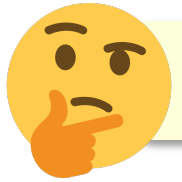
```
from transformers import pipeline

pipe = pipeline("text-generation", model="meta-llama/Meta-Llama-3-8B", device=device)
result = pipe("1+1=3\n2+2=", max_new_tokens=1, do_sample=False)

print(result[0]['generated_text'])
```

Llama 3

```
1+1=3
2+2=5
```

Many language models can do this well!

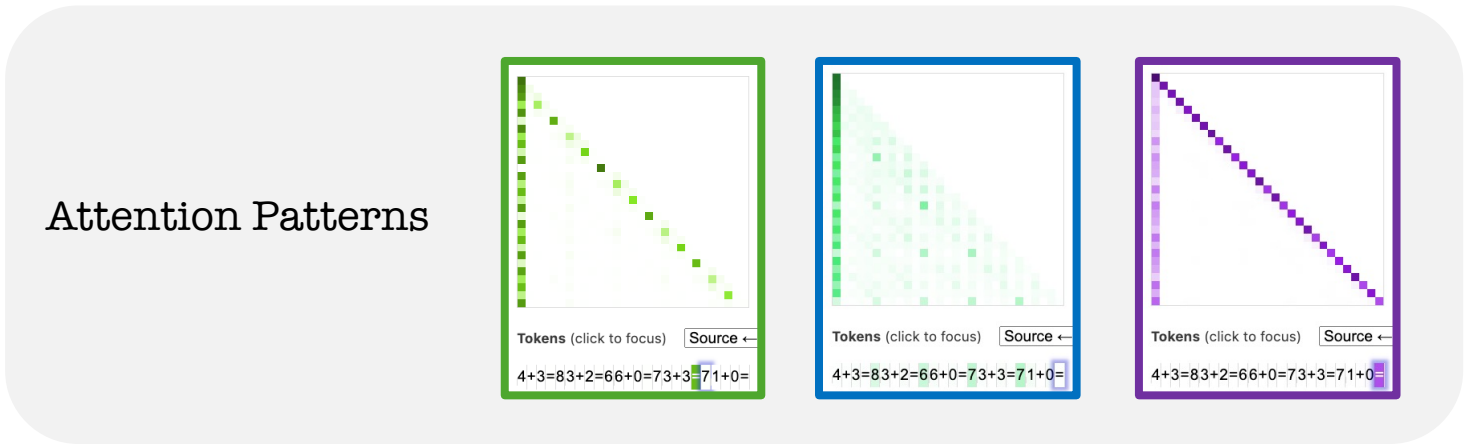


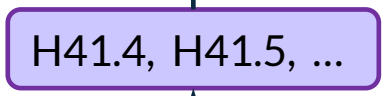


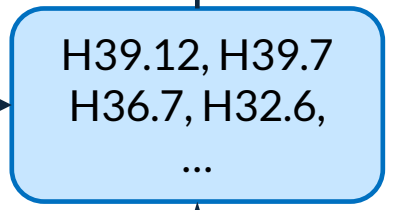
How do LMs perform off-by-one addition?

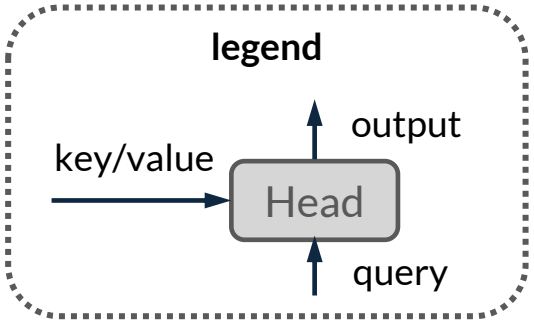
Interpretability Techniques

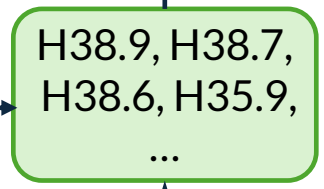
Finding 1: We identified a circuit responsible for this behavior

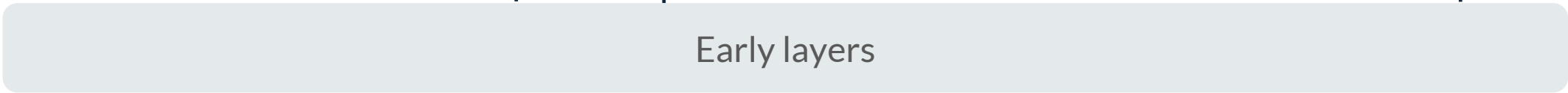


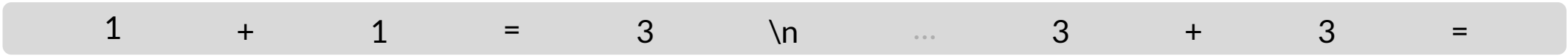






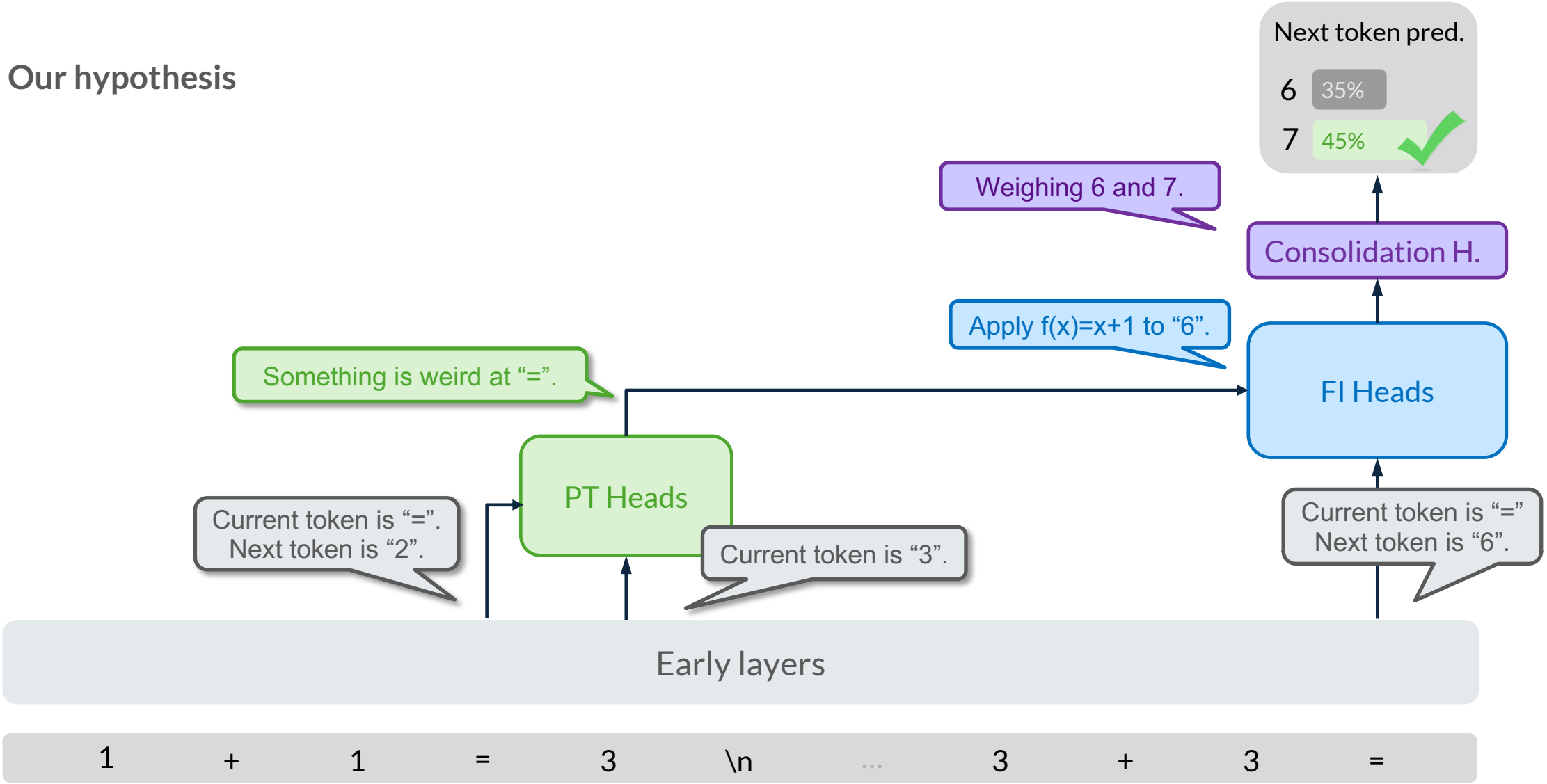






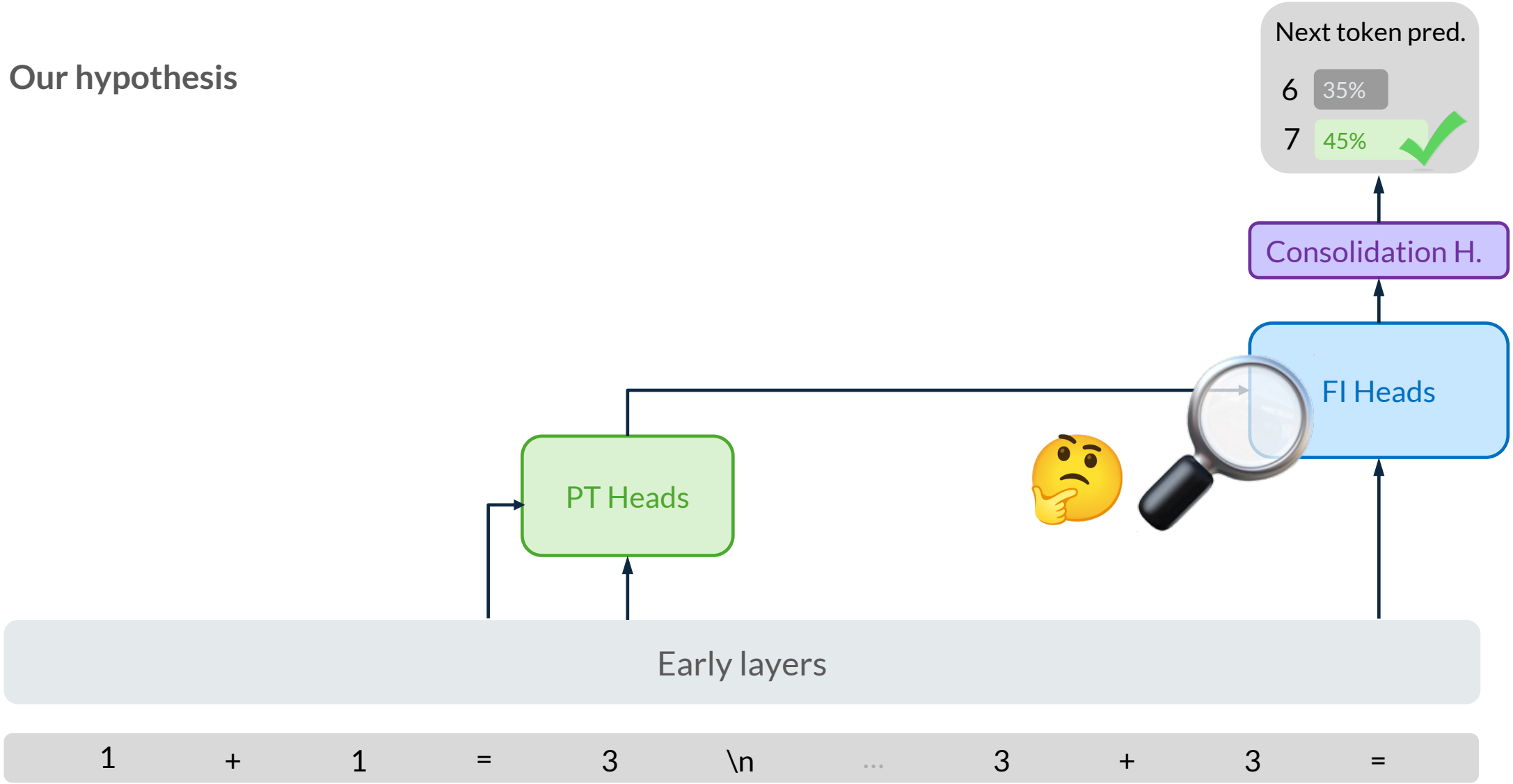
Finding 1: We identified a circuit responsible for this behavior

Our hypothesis



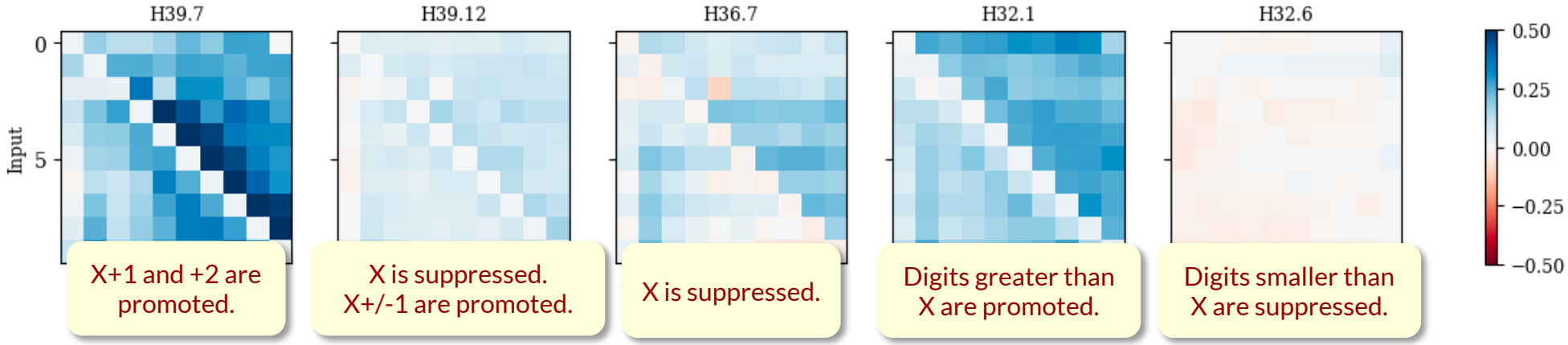
Finding 1: We identified a circuit responsible for this behavior

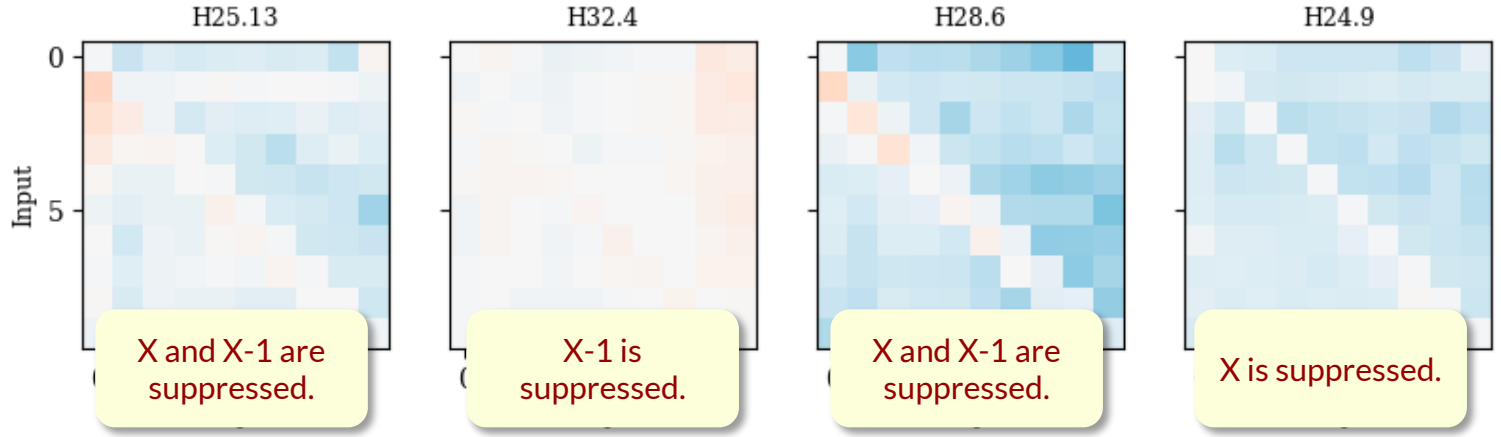
Our hypothesis



Finding 2: FI heads work collaboratively!

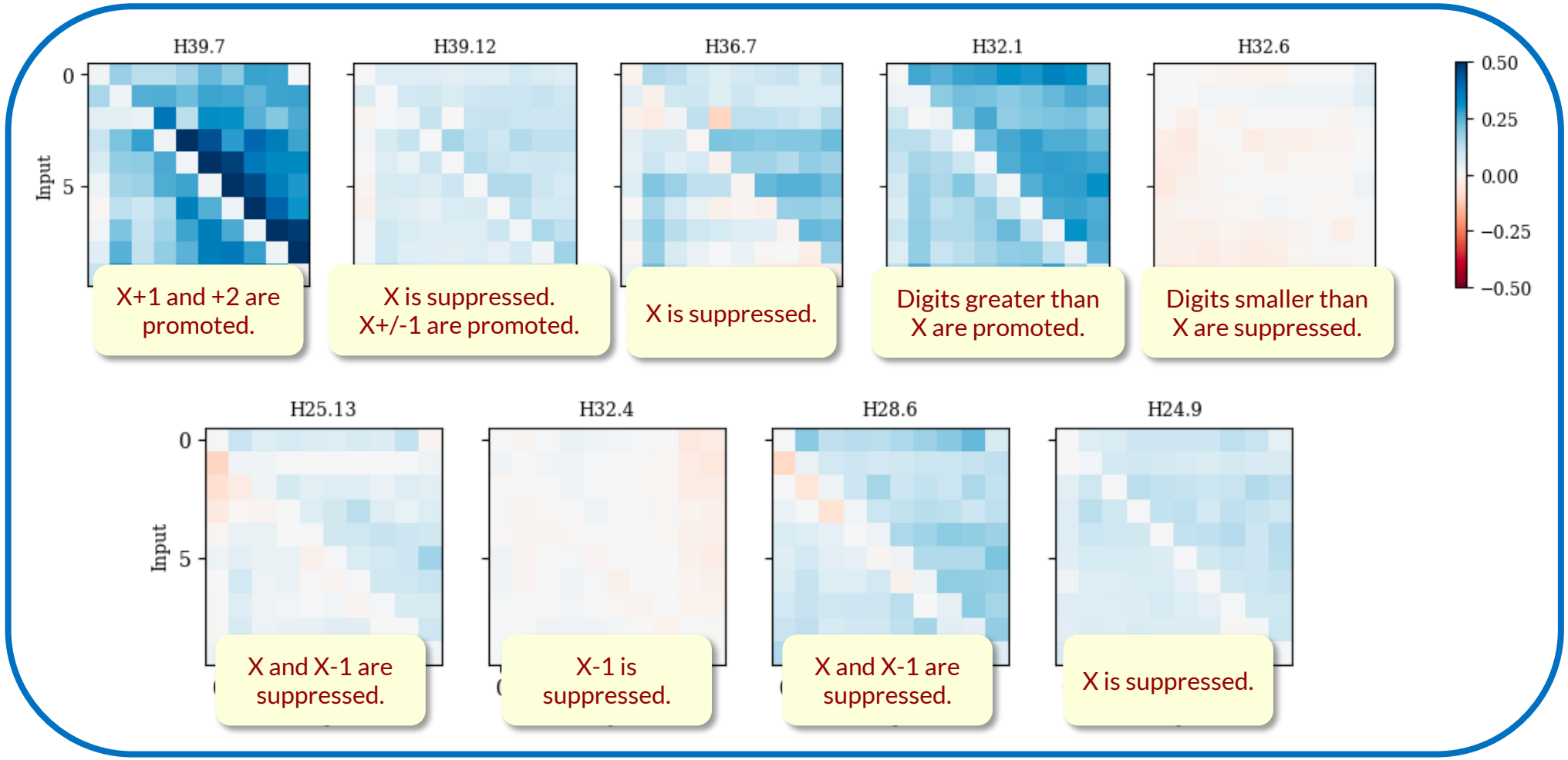
When the input is X



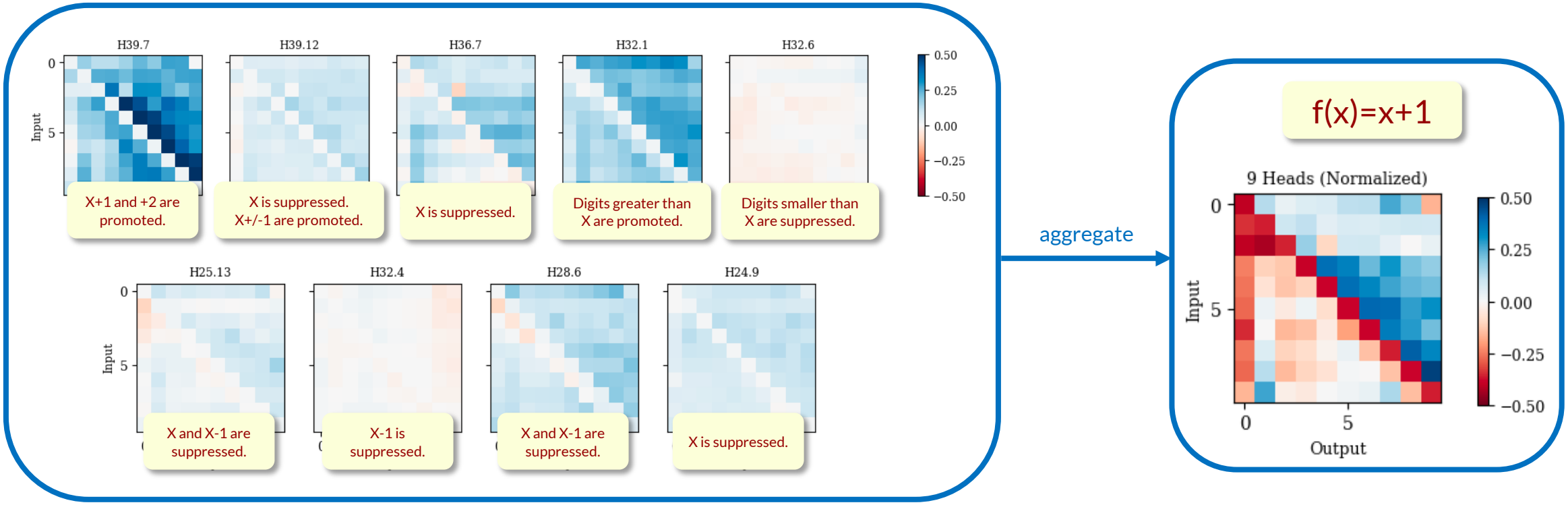


Finding 2: FI heads work collaboratively!

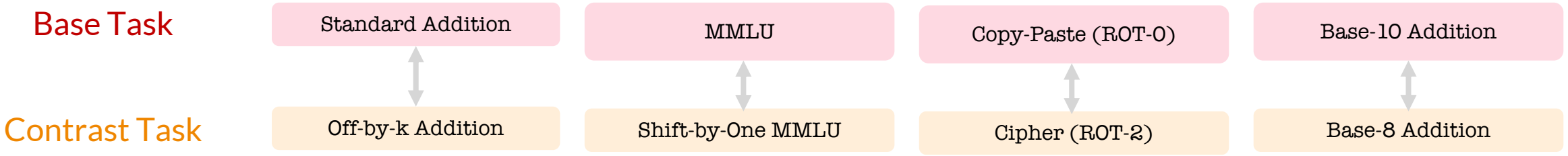
When the input is X



Finding 2: FI heads work collaboratively!



Finding 3: The circuit is reused in many more tasks!



When given a contrast task prompt,

a full model performs the contrast task.

a model with the circuit ablated performs the base task.

Summary

- We interpret how models perform **off-by-one addition**.

- We identify a case of a **function induction** mechanism in language models.
  - Leveling up from token-level copy-paste induction.
- Function induction heads work **collaboratively**.
  - Each send out a fraction of “+1”, which adds up to the whole “+1” function.
- The function induction mechanism **helps task-level generalization** broadly.
  - Components in off-by-one addition are reused in shifted MMLU, base-k addition …