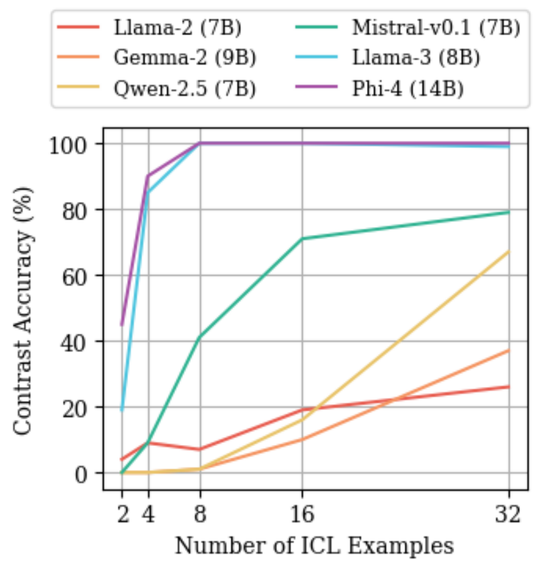
1+1=3 2+2=5 3+3=

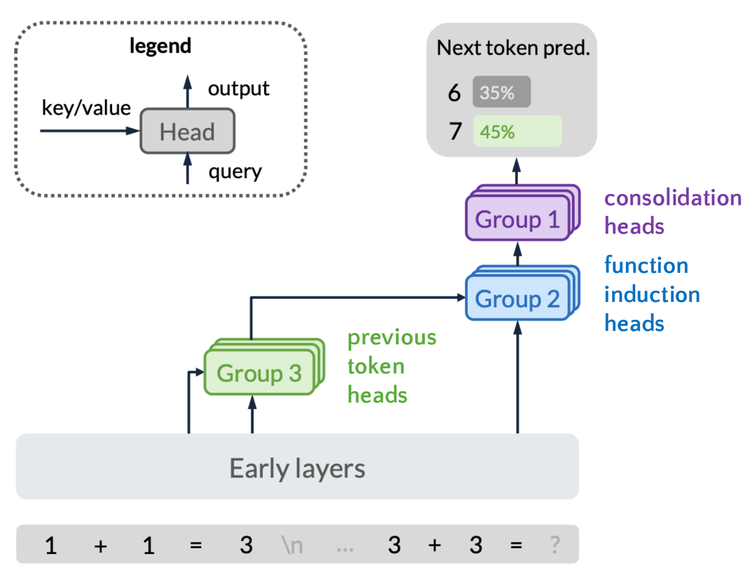
Llama 3 (8B) 7 | Mistral v0.1 (7B) 7 | Gemma 2 (9B) 7

Why? A function induction circuit ...

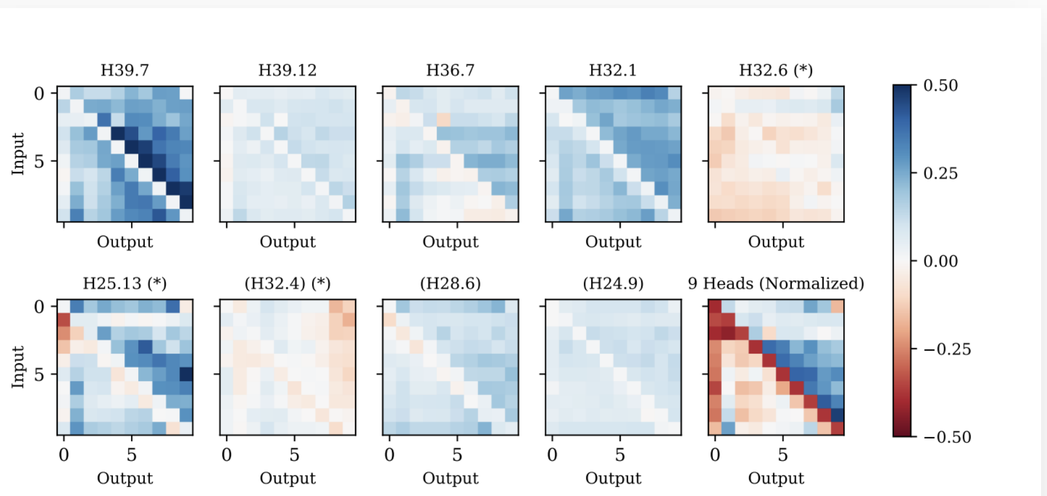
Background: LMs can learn off-by-one addition in context.

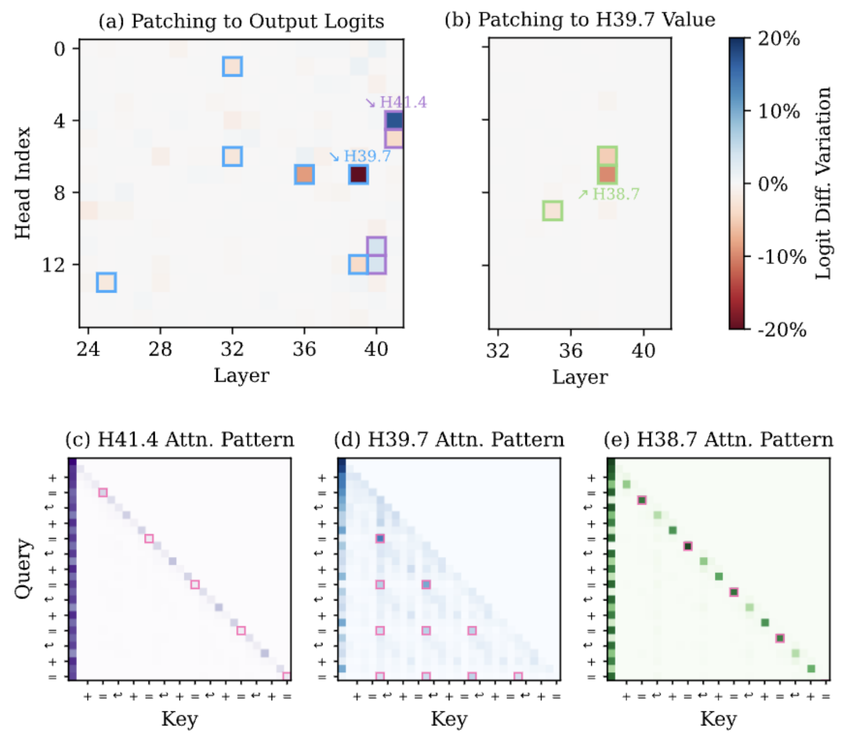


Finding 1: A **function induction** circuit that is structurally similar to induction heads.

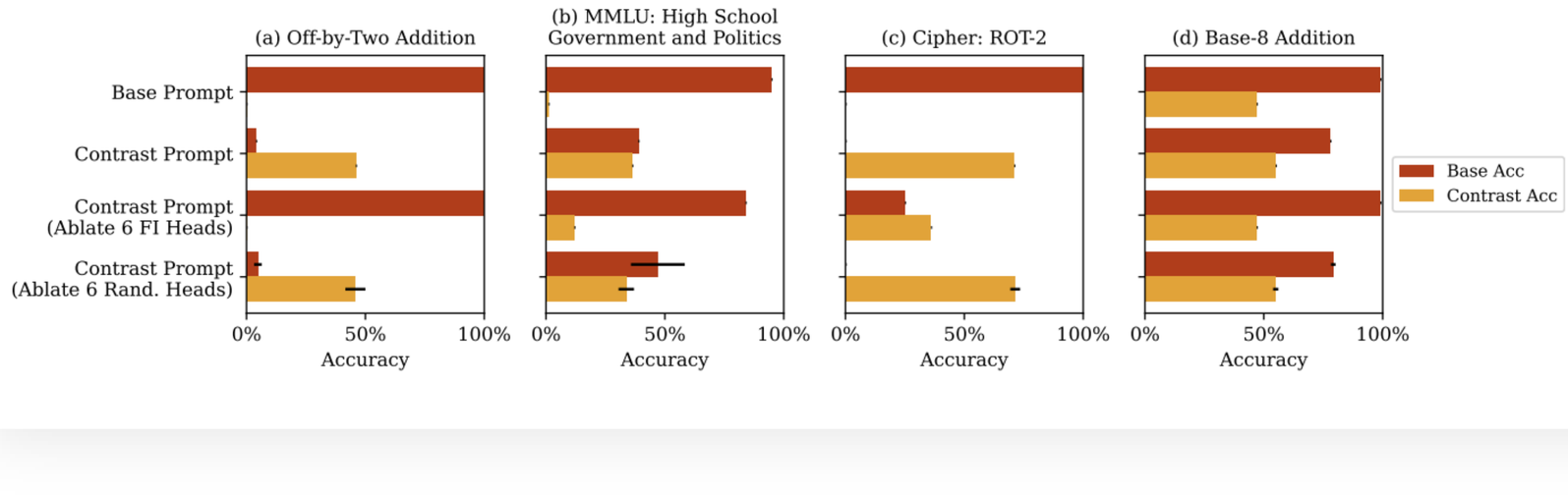


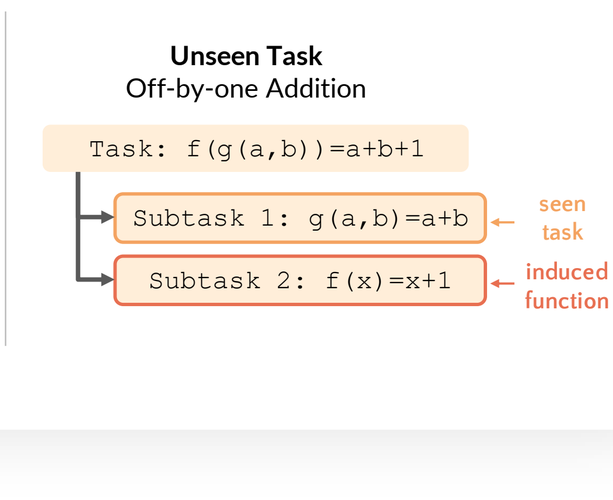
Finding 2: Each **FI head (Group 2)** sends out a distinct piece of the +1 function.





Finding 3: **FI heads** are reused across various tasks.





Limitations: (1) The circuit has imperfect faithfulness and completeness; (2) Additional QK and OV circuit analyses remain unexplored; (3) The scope is restricted to two-step tasks in which the second step involves a shifting-related function.

PAPER

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

Qinyuan Ye, Robin Jia, Xiang Ren

Mechanistic Interpretability Workshop, NeurIPS 2025

USC Viterbi School of Engineering Department of Computer Science