“49 + 212” → LLM → 261, but how?

✗ Perfect Memorization?

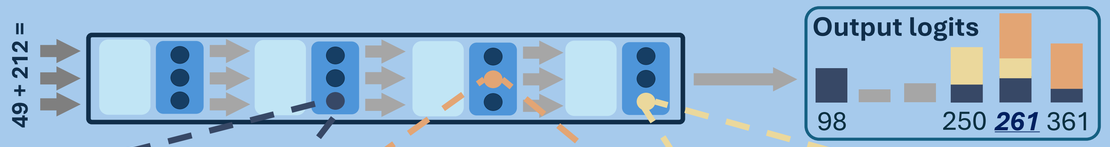
✗ Learned Algorithm?

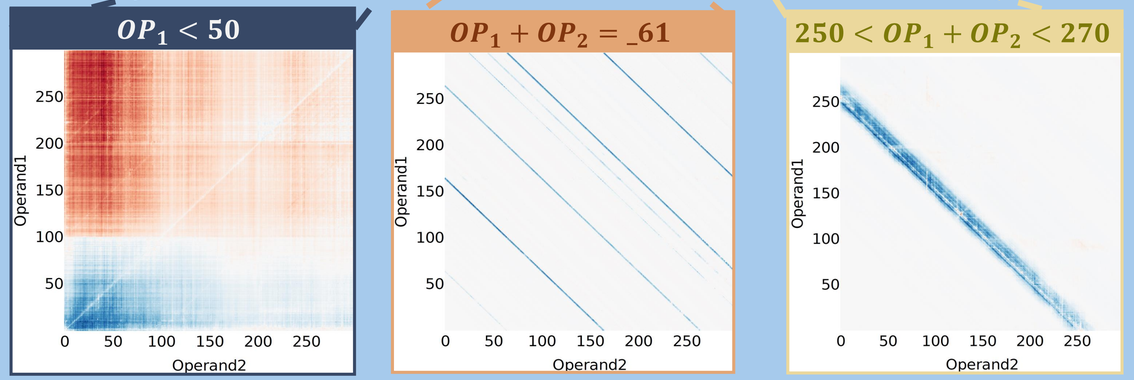
LLMs Solve Arithmetic with a Bag Of Heuristics

Yaniv Nikankin, Anja Reusch, Aaron Mueller, Yonatan Belinkov

TECHNION Israel Institute of Technology

ICLR

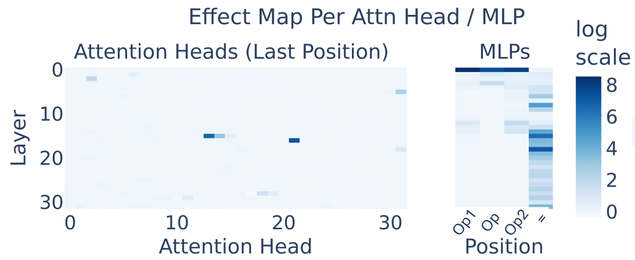


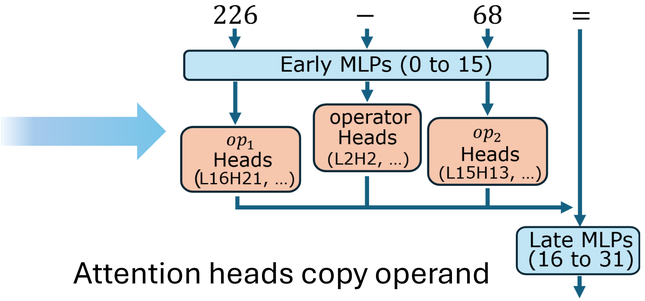


Co-funded by the European Union

Funded by the European Union (ERC, Control-LM,101165402). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

1. Which model components participate in answering arithmetic prompts?



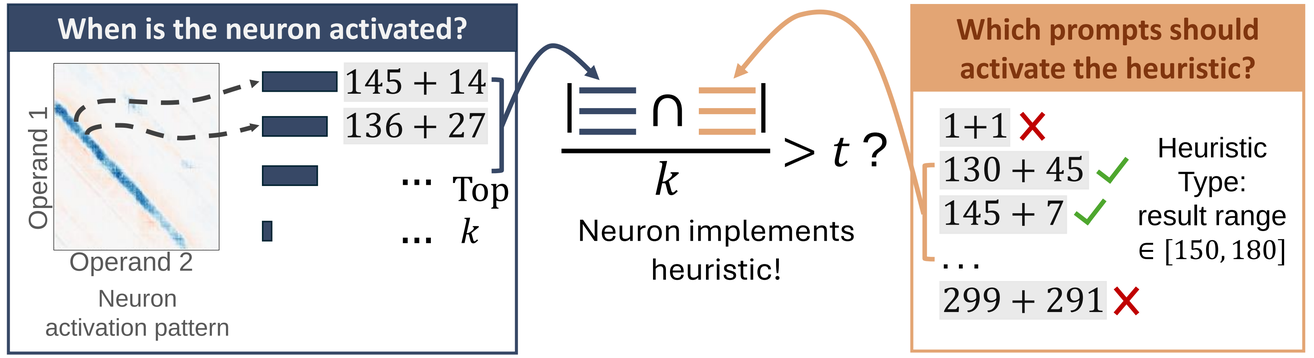


Activation patching is used to find important components, making up the arithmetic circuit.

Attention heads copy operand embeddings to last position.

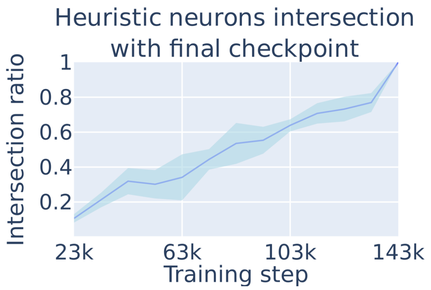
1% of later-MLP (heuristic) neurons generate the correct answer.

2. How do we label heuristic neurons automatically?



Automatic labeling process allows labeling of circuit neurons to heuristic types.

3. How do arithmetic heuristics develop over training?



The bag of heuristics emerges as the single mechanism driving arithmetic calculations from early training

4. Why do arithmetic heuristics fail?