Reasoning

Daniel Hsu

COMS 6998-7 Spring 2025

Reasoning in language models

Reasoning: computational process of logical deduction

- Q: Can language models (learn to) perform reasoning?
- Q: Can transformers (learn to) perform reasoning?
- Q: Can transformers (learn to) perform any "multi-step" algorithms?

Example: Pointer chasing

Pointer chasing [Weston, Chorpa, Bordes, 2014; Peng, Narayanan, Papadimitriou, 2024; …]

Prompt:

Jane is a teacher. Helen is a doctor. […]

The mother of John is Helen. The mother of Charlotte is Eve. […]

What's the profession of John's mother?"

Answer: doctor

b a c b c a b d b a

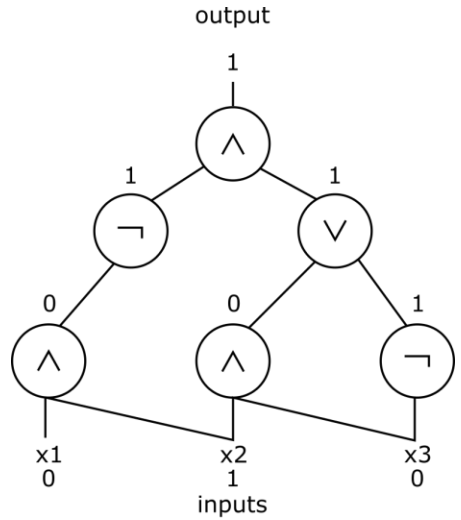
2-hop associative recall / induction head

(easy for multi-layer transformers!)

Example: Chain-of-Thought

Computation via Chain-of-Thought

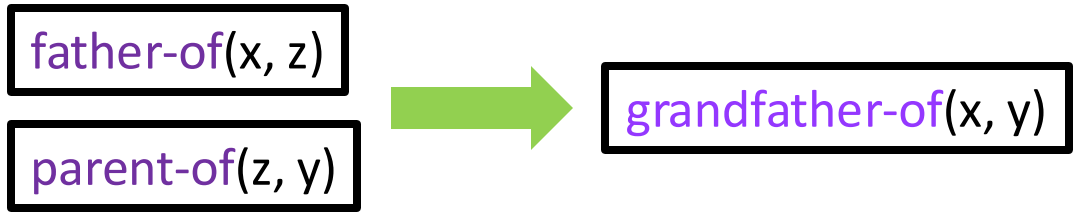
- Intermediate "CoT tokens" = steps of computation
- Efficient learning needs step-by-step supervision
  - Rivest & Sloan: straight-line programs
  - Malach: sequence of linear threshold functions
  - Joshi et al: general auto-regressive processes
- Need to be robust to errors at intermediate steps!



What general forms of reasoning can we hope for?

Robust logics (Valiant)

- Basic units of knowledge: relations among objects
- Goal: Combine relations to deduce new relations



- Learning requires examples of deduction rules in action
- Reasoning uses hard-coded "general deduction algorithm"
  (e.g., with learned rule set, in completely new situation/"scene")
- Q: Where do the relations come from???

Parting questions

1. Can/do LLMs (learn to) perform general forms of reasoning?
2. If no, what's lacking in our current ML systems?

The end