Towards a Mechanistic Interpretation of Multi-Step Reasoning Capabilities of Language Models

Yifan Hou, Jiaoda Li, Yu Fei, Alessandro Stolfo, Wangchunshu Zhou, Guangtao Zeng,
Antoine Bosselut, Mrinmaya Sachan

ETH zürich · UCI · 波形智能 AI WAVES · SUTD Singapore University of Technology and Design · EPFL

Motivation

How do LLMs answer reasoning questions?

Model Input

Zebra puzzle:

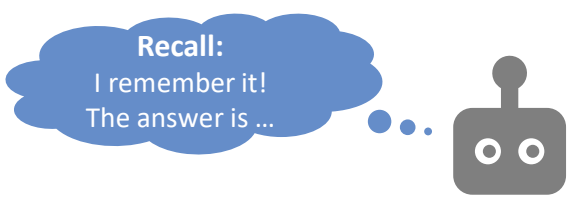
The Tennis player lives in the second house. The Red house is at the first position. The Paraguayan lives exactly to the right of the man that plays Tennis. The man who has Fishes lives next to the man who enjoys playing Tennis. The Mexican plays Basketball. The man that has Cats lives exactly to the left of the Green house. The Paraguayan lives next to the Bird owner.

Q: Who has cats?

Model Output

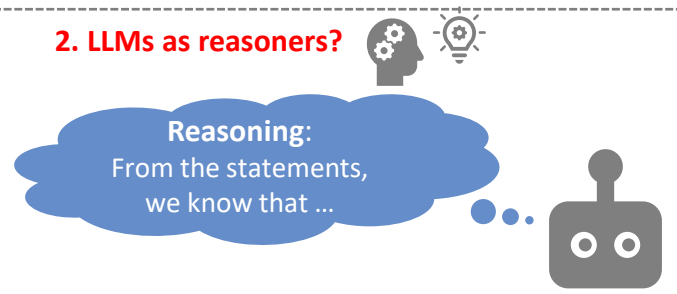
A: The Mexican.

1. LLMs as retrievers (stochastic parrot)?



LLMs give the answer by **cheating with shortcuts memorized from pretraining corpus**. The reasoning process doesn't happen.

2. LLMs as reasoners?



LLMs give the answer by **doing step-by-step reasoning similar to humans**. The reasoning process happens internally.

Why this is important?

Generalizability: if LLMs work for unseen examples in practice …;

Reliability: if LLMs work as expected? …;

Improvement: how to effectively improve LLMs on reasoning? how to design next-generation reasoners? …;

…

Method

How to know if LLMs are retrievers or reasoners?

Hypothesize-and-Verify:
backward reasoning as:

Hypothesize:
If LLMs are reasoners?

→ If LLMs perform reasoning step-by-step internally?

→ **Verify:**
If we can detect reasoning trees from LLMs?

Probing task

A probe model predict information we care about from representations/attentions of a LLM

Probing task: P(Reasoning trees | LLM attentions)

Probing model: kNN classifier (non-parametric)

Prediction Acc.: high ⇔ much info; low ⇔ little info

Problems (task is too difficult):

1. LLM attentions

   millions of attention weights, very high-dimensional

2. Reasoning trees:

   complex structure, hard to predict

Attention simplification

1. Head pooling,
2. Only focusing on the last token
3. Layer pruning: reduce layer num *L*
4. Token pooling: reduce token num *N*

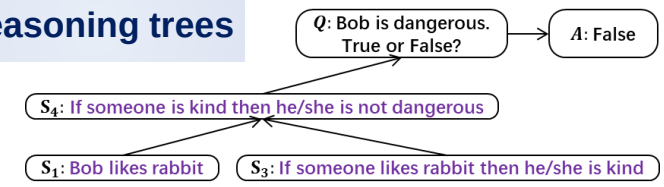
From **millions of** attention weights to **hundreds**

Probing task simplification

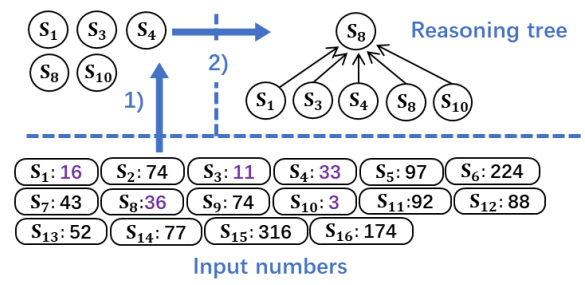
Probing task decomposition:

P(Reasoning trees | LLM attentions)
= P(Nodes | LLM attentions)
x P(Reasoning trees | Nodes, LLM attentions)

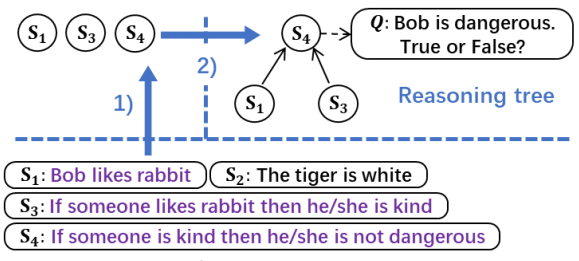
Reasoning trees



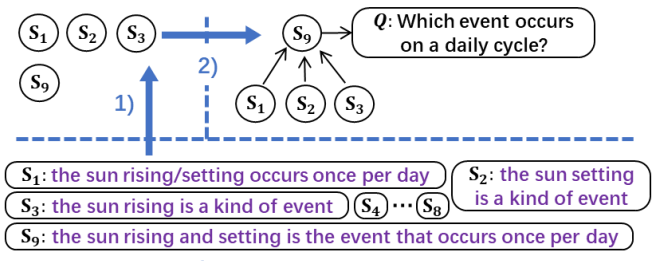
Reasoning trees indicate reasoning process in structural format.



GPT-2: k-th smallest element (k=5)



LLaMA: ProofWriter

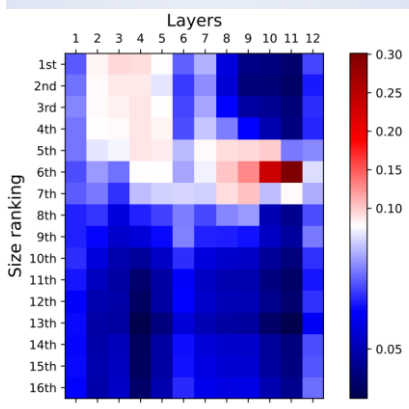


LLaMA: AI2 Reasoning Challenge

Experiment

Probing reasoning trees in LM attentions

Attention visualization



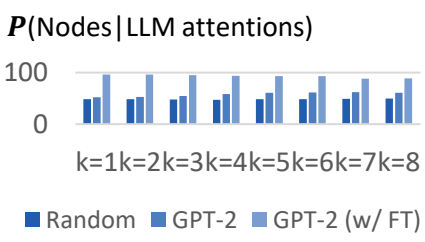
(f) $k = 6$

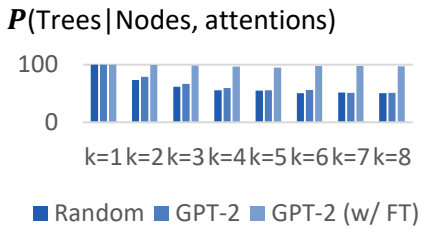
Reasoning trees exist in attentions

Leaf nodes (top-k numbers) are focused on bottom layers;

Root nodes (*k*-th smallest number) are focused on top layers

Probing analysis





We can detect reasoning trees from attentions clearly

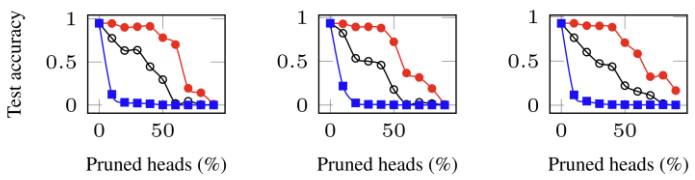
Causal analysis

Question: if LMs perform reasoning following the reasoning tree detected from attention patterns?

Idea: corrupting reasoning trees in attentions

Performance decreases ⇔ causal relationship exists

Implementation: attention head pruning.

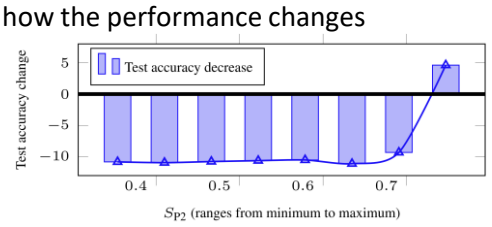


(d) $k = 4$ (e) $k = 5$ (f) $k = 6$

LMs perform reasoning following the reasoning tree detected from attention patterns

Probing scores and LM robustness

Idea: add noise to statements, and check how the performance changes



LMs are more robust if they know the step of using the statement in reasoning

ArXiv

GitHub