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 How do LLMs answer reasoning questions? **Motivation** 1. LLMs as retrievers (stochastic parrot)? (**@**) 2. LLMs as reasoners? Model Input Zebra puzzle: **Reasoning**: **Recall:** From the statements, I remember it! The Tennis player lives in the second house. The we know that .. The answer is .. Red house is at the first position. The Paraguavan lives exactly to the right of the man that plays Tennis. The man who has Fishes lives next to the LLMs give the answer by cheating with shortcuts LLMs give the answer by doing step-by-step man who enjoys playing Tennis. The Mexican plays Basketball. The man that has Cats lives memorized from pretraining corpus. The reasoning reasoning similar to humans. The reasoning exactly to the left of the Green house. The process doesn't happen. process happens internally. Paraguayan lives next to the Bird owner. Why this is important? Q: Who has cats? Generalizability: if LLMs work for unseen examples in practice ...; **Reliability**: if LLMs work as expected? ...; Model Output **Improvement**: how to effectively improve LLMs on reasoning? how to design next-generation reasoners? ...; A: The Mexican. How to know if LLMs are retrievers or reasoners? Method **Hypothesize-and-Verify:** Hypothesize: If LLMs perform reasoning Verify: step-by-step internally? If LLMs are reasoners? backward reasoning as: If we can detect reasoning trees from LLMs? Probing task Reasoning trees Q: Bob is dangerous Probing task simplification A: False True or False? A probe model predict information we care about from $(S_4: If someone is kind then he/she is not dangerous)$ representations/attentions of a LLM Probing task decomposition: $(S_1: Bob likes rabbit)$ $(S_3: If someone likes rabbit then he/she is kind)$ Probing task: P(Reasoning trees | LLM attentions) Reasoning trees indicate reasoning process in structural format. Probing model: kNN classifier (non-parametric) P(Reasoning trees | LLM attentions) = P(Nodes | LLM attentions) Prediction Acc.: high ⇔ much info; low ⇔ little info $(S_1)(S_3)(S_4)$ **Reasoning tree** × P(Reasoning trees | Nodes, LLM attentions) Problems (task is too difficult): $(S_1)(S_3)(S_4)(S_8)(S_{10})$ 1. LLM attentions millions of attention weights, very high-dimensional $S_1: 16$ $S_2: 74$ $S_3: 11$ $S_4: 33$ $S_5: 97$ $S_6: 224$ $\overline{S_{7}:43}$ $\overline{S_{8}:36}$ $\overline{S_{9}:74}$ $\overline{S_{10}:3}$ $\overline{S_{11}:92}$ $\overline{S_{12}:88}$ 2. Reasoning trees: $(S_{13}: 52)(S_{14}: 77)(S_{15}: 316)(S_{16}: 174)$ complex structure, hard to predict Input numbers GPT-2: k-th smallest element (k=5) **Attention simplification** Q: Which event occurs Q: Bob is dangerous. $(\mathbf{S}_1)(\mathbf{S}_3)$ $(S_1)(S_2)$ 1. Head pooling, on a daily cycle? True or False? 2. Only focusing on the last token **Reasoning tree** $(\mathbf{S}_1)(\mathbf{S}_2)(\mathbf{S}_3)$ 1) (S_1) 1) 3. Layer pruning: reduce layer num L S_1 : Bob likes rabbit S_2 : The tiger is white S_1 : the sun rising/setting occurs once per day S_2 : the sun setting 4. Token pooling: reduce token num N is a kind of event S_3 : If someone likes rabbit then he/she is kind S_3 : the sun rising is a kind of event S_4 ... S_8 $(S_4:$ If someone is kind then he/she is not dangerous) $\left(\, {f S_9}
ight)$: the sun rising and setting is the event that occurs once per day $\,
ight)$ Input statements Input statements From millions of attention weights to hundreds LLaMA: ProofWriter LLaMA: Al2 Reasoning Challenge

Experiment

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 P(Nodes|LLM attentions) 100

k=1k=2k=3k=4k=5k=6k=7k=8 ■ Random ■ GPT-2 ■ GPT-2 (w/ FT)

P(Trees | Nodes, attentions) 100 0 k=1k=2k=3k=4k=5k=6k=7k=8 ■ Random ■ GPT-2 ■ GPT-2 (w/ FT) 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

0.5

0

0

Performance decreases ⇔ causal relationship exists

Implementation: attention head pruning.



Probing reasoning trees in LM attentions

0.5-00 500 Pruned heads (%) Pruned heads (%) (f) k = 6(e) k = 5

50

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

ETH zürich



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

