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?

1. **LLMs as retrievers (stochastic parrot)?**
   - **Recall:** I remember it!
   - The answer is...
   - LLMs give the answer by cheating with shortcuts memorized from pretraining corpus. The reasoning process doesn’t happen.

2. **LLMs as reasoners?**
   - **Reasoning:** From the statements, we know that...
   - 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? ...

**How to know if LLMs are retrievers or reasoners?**

**Method**

**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:** \( P(\text{Reasoning trees} | \text{LLM attentions}) \)
- **Probing model:** kNN classifier (non-parametric)
- Prediction Acc.: high \( \Rightarrow \) much info; low \( \Rightarrow \) 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

**Experiment**

**Probing reasoning trees in LLM attentions**

**Attention visualization**

- \( P(\text{Nodes} | \text{LLM attentions}) \)
- \( k=1 \times 2 \times 3 \times 4 \times 5 \times 6 \times 7 = 8 \)
- Random \( \cdot \) GPT-2 \( \cdot \) GPT-2 (w/ FT)
- \( k=1 \times 2 \times 3 \times 4 \times 5 \times 6 \times 7 = 8 \)
- Random \( \cdot \) GPT-2 \( \cdot \) GPT-2 (w/ FT)

**Causal analysis**

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

**Idea:** corrupting reasoning trees in attentions

Performance decreases \( \Leftrightarrow \) causal relationship exists

**Implementation:** attention head pruning.

**Probing scores and LM robustness**

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

---

**Motivation**

**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.