Deciphering the Factors Influencing the Efficacy of Chain-of-Thought: Probability, Memorization, and Noisy Reasoning

Akshara Prabhakar[1], Thomas L. Griffiths[1], R. Thomas McCoy[2]

PRINCETON UNIVERSITY

Yale University

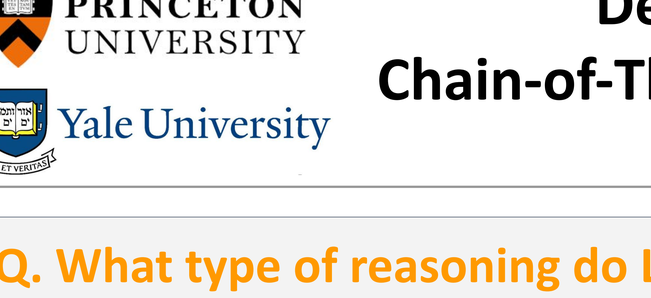
Q. What type of reasoning do LLMs use when prompted with CoT?
A. Both memorization and a probabilistic version of genuine reasoning

💡 Motivation

- CoT enhances the multi-step reasoning capabilities of LLMs
- Are LLMs genuinely reasoning, or are they driven by shallow heuristics?
- **The Challenge**
  - For most tasks, shallow heuristics and genuine reasoning produce similar behaviors 😔
  - We need a task where they can be disentangled!

Our work ↓

Shift Cipher Task

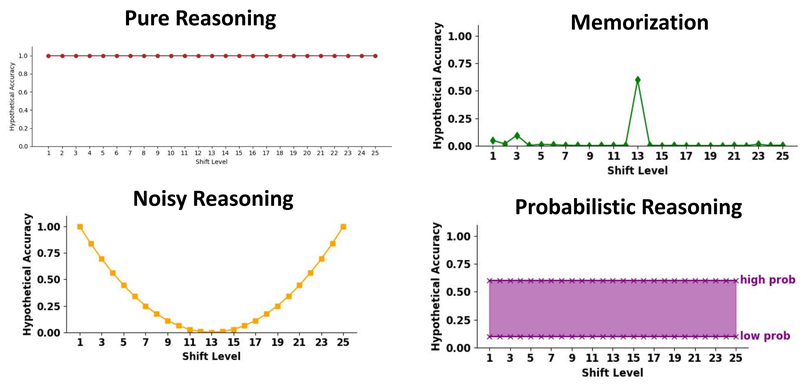


✅ Benefits of this task

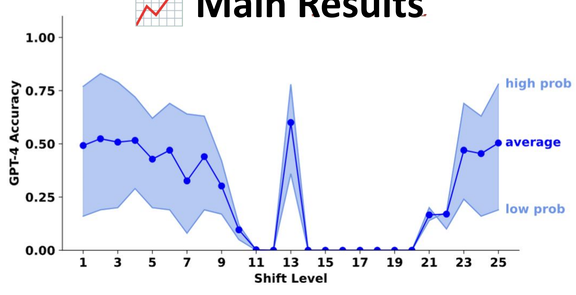
- Allows dissociation of memorization and reasoning 😄
  - *simplest shift level = 1*
  - *most commonly occurring shift level = 13*
- Allows us to independently manipulate
  - difficulty, frequency, answer probability
- Sufficiently hard
- Easy to verify correctness

https://github.com/aksh555/deciphering_cot

🛠️ Hypothetical signatures of different types of reasoning



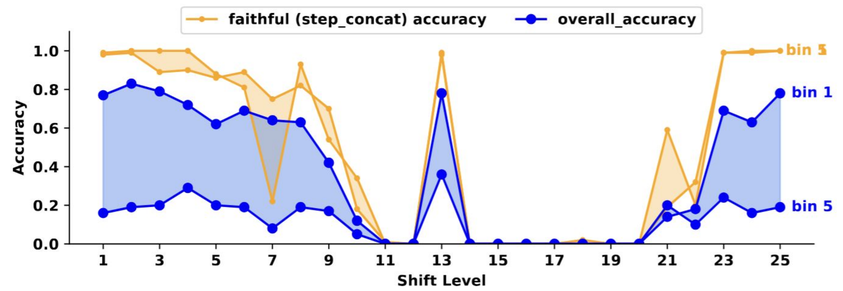
📈 Main Results



- **LLM performance combines signatures of memorization, probabilistic reasoning, and noisy reasoning**
- A logistic regression supports these factors are significant

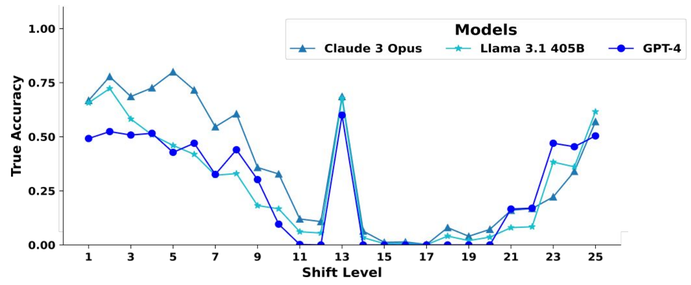
🔎 Additional Results

#1 Probabilistic effects



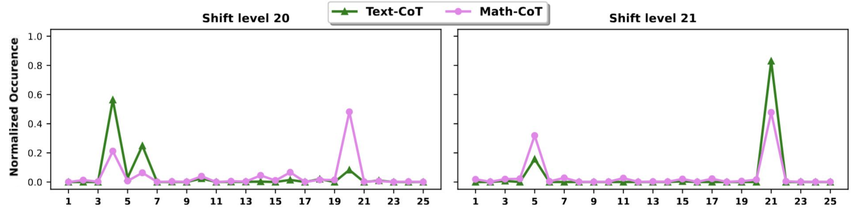
- Individual chain steps are usually correct (orange lines)
- But the final answer (blue lines) is wrong much more often for low-probability (bin 5) than high-probability (bin 1)

#2 Memorization effects



13 is the most common shift level in Internet corpora, and LLMs achieve the best accuracy at 13!

#3 Noisy reasoning



Peaks at 26-shift_level are evidence for (occasionally confused) bidirectional reasoning attempts.

⭐ Takeaways

- When decoding shift ciphers, LLMs display effects of both shallow heuristics (memorization, probabilistic reasoning) and a noisy version of genuine reasoning.
- These results provide a reasonable middle ground in debates about whether LLMs can reason.