Concept Aware Reasoning: Teaching Large Language Models to Focus

Abhishek Anand **Dikshita Nitin Padte** **Muyuan He** **Nehal Muthukumar**

Viterbi School of Engineering
University of Southern California

1 Introduction

Large Language Models (LLMs) have recently shown impressive reasoning performance using a variety of prompting approaches, including Few-shot, Chain-of-Thought, Tree-of-Thought, Self-Consistency, and ReAct. Despite these improvements, LLMs frequently take short cuts while solving problems, which results in outputs that are inconsistent or irrational in their thinking. The reason for this tendency is that during training, the models learn spurious correlations, which include biased relationships and irrelevant patterns from the data. When reasoning tasks are performed, these relationships show up and lead to incorrect conclusions and illogical thinking.

For instance, when presented with a question like "The chief gave the housekeeper a tip because she was satisfied. Who is 'she'?", LLMs frequently provide incorrect answers like "housekeeper" due to biased associations between gender and profession ingrained in their training data. In contrast, humans, despite having inherent biases, can typically focus on essential concepts necessary for logical reasoning, disregarding stereotypical associations.

We argue that whereas LLMs and humans may both possess these biases or prejudices, humans are better able to separate relevant concepts from other associations and make appropriate deductions based on their world knowledge. On the other hand, while LLMs also contain knowledge about the relevant concepts, they perform poorly because they are unable to focus on them during reasoning tasks. The objective of this study is to explore strategies for reducing unwanted reasoning in LLMs by improving their ability to identify and give priority to key concepts over incorrect associations. Our goal is to further the development of LLMs for logical reasoning tasks by making them more dependable and objective by tackling these problems by effectively mirroring human reasoning processes.

We propose to employ techniques that help language models to identify and focus on portions of its knowledge that are essential for answering the question at hand, in turn eliminating the effect of spurious correlations it has learnt. We look to achieve this through 2 approaches: prompt-engineering and RLHF fine-tuning to explicitly teach LLMs how to reason.

2 Related Work

Recent advancements in language models have focused on enhancing their ability to reason and provide accurate responses to complex questions. While LLMs demonstrate remarkable few-shot learning capabilities, they often struggle with reasoning tasks and are prone to hallucinations. Few Shot prompting, as introduced by (Brown et al., 2020), aims to mitigate hallucination issues by providing in-context learning. This technique enables LLMs to better understand the context of the query and generate more accurate responses. However, while effective for certain tasks, it falls short in scenarios requiring complex reasoning. Building upon Few Shot prompting, chain of thought (CoT) prompting (Wei et al., 2023). This approach guides LLMs to reason through intermediate steps, thereby improving their performance on reasoning tasks. Despite this advancement, challenges persisted in tasks demanding strategic searching for answers. To address the limitations of existing prompting techniques, Tree of thought prompting (Long, 2023; Yao et al., 2023a) was proposed, offering a more generalized approach. By encouraging exploration and facilitating better reasoning for complex tasks, this method represents a significant step forward in enhancing the capabilities of LLMs. Additionally, approaches like Retrieval-Augmented Generation (RAG) and ReAct prompting (Yao et al., 2023b) have emerged to tackle hallucination issues by integrating external resources.

These techniques leverage retrieval mechanisms to augment the generation process, leading to more reliable responses.

Despite these advancements, a crucial aspect often overlooked is the transparency of LLMs' decision-making processes. While these techniques facilitate accurate responses, they often fail to elucidate how the model arrived at its conclusions. Addressing this gap, Faithful CoT (Lyu et al., 2023) proposes a two-step framework that not only provides answers but also reveals the underlying steps taken by the model. This approach enhances interpretability by showcasing the LLM's reasoning process, thus facilitating a deeper understanding of its outputs. We seek to reverse this framework and use the reasoning part for creating a prompt instead of a response. By proposing a novel prompt approach and finetuning it on LLMs, this study aims to enhance the models' reasoning capabilities while reducing hallucinations.

3 Methods

Concept-Centric Prompt Design This would involve prompting LLMs to identify important concepts and using what they know about them and their causal relationships to address the given question. In essence, the aim would be to guide models to focus only on those areas of their expertise that are pertinent to the given job. As an exhaustive list of all spurious correlations that the model would have learned is not possible to determine, we believe that positive reinforcement of key concepts mentioned in the question would prove to be a more robust approach than strategies like general reasoning steps or counterfactual prompting or being explicit about avoiding certain biases like gender, race, religion, etc. Therefore, using explicit concept identification to solve reasoning problems in a manner similar to that of a human might prove more robust and generalizable.

RLHF Fine-Tuning This would involve creating a human preference dataset that will include preferred reasoning techniques, with an emphasis on concept identification and pertinent world knowledge. We will create this dataset by utilizing GPT-4 and manual human annotations. We aim to run our experiments on Llama-2-7B (Touvron et al., 2023) and Mistral-7B (Jiang et al., 2023), fine-tuning them using Reinforcement Learning with Human Feedback to improve their reasoning process.

4 Approximate Timeline

- Week 8 - March 1 → Dataset creation complete with evaluation results from Concept-centring prompting.
- Week 11 - March 22 (Mid report) → Fine-tuning complete using our dataset for Llama-2-7b with evaluation results.
- Week 14 - April 12 → Fine-tuning complete using our dataset for Mistral-7b with evaluation results.
- Week 15 - April 19 → Final results along with comparison with existing work.

We would like to propose Ziyi Liu as our TA for the project as her current research interest aligns with our project domain and she can guide us towards achieving our objective.

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Jieyi Long. 2023. Large language model guided tree-of-thought.

Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. 2023. Faithful chain-of-thought reasoning.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa,

Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023a. Tree of thoughts: Deliberate problem solving with large language models.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023b. React: Synergizing reasoning and acting in language models.