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

Abstract

Large Language Models (LLMs) have shown remarkable advancements in reasoning tasks but often suffer from biased associations and spurious correlations learned during training, leading to inconsistent or irrational outputs. In contrast, humans can prioritize relevant concepts over biases for logical reasoning. This study proposes strategies to enhance LLMs' reasoning by prioritizing key concepts and relationship between them, while reducing unwanted associations. Two approaches are proposed: Concept-Relationship-Reasoning Prompt (CRR) and DPO Fine-Tuning for concept-focused reasoning. CRR explicitly guides LLMs through concept identification, relationship analysis, and reasoning steps, while DPO Fine-Tuning utilizes knowledge distillation with a teacher model to teach student model to reason using concepts. We additionally present CRR-Collection, a synthetic dataset for reasoning using our prompting methodology. Experiments show promising results, with CRR improving reasoning accuracy and DPO Fine-Tuning aiming to bridge the gap between reasoning and generating final answers. Additionally, we aim to show the effectiveness of our approach by positively reinforcing focus on relevant knowledge to avoid spurious correlations without explicitly mitigating learned bias.

1 Introduction

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.

We aim to answer the following 2 questions through our study:

1. *Can language models identify relevant world knowledge in form of concepts and their rela-*

tionships pertinent to the query and use them for reasoning?

2. *Can a focused approach in reasoning through concept identification help in mitigating learned biases during inference for commonsense reasoning tasks?*

To answer the first question, we employ prompt engineering and fine-tuning to guide model towards a focused reasoning process presented through our Concept-Relationship-Reasoning prompt and Concept-Relationship-Reasoning Collection dataset. For answering the second question, we specifically create the CRR Collection dataset using base datasets that don't test gender bias and evaluate the subsequently fine-tuned model using Winobias which tests gender bias and stereotypes, hence, helping us answer the question whether focused reasoning without explicitly handling different biases can prove effective in mitigating them.

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

We aim to teach language models how to reason using a concept-focused approach. This involves first identifying real world concrete and abstract concepts and entities relevant to the query and then identifying the relationships between them. Most language models possess the ability to identify general concepts pertinent to the query and we aim to utilize this knowledge to guide them during reasoning, while avoiding any shortcuts or biases learnt during pre-training. 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. We utilize 2 methods to achieve our objective - a) Prompt Engineering through Concept-Centric Prompt Design, b) DPO Fine-Tuning for concept-focused reasoning through knowledge distillation utilizing a capable teacher model.

3.1 Concept-Centric Prompt Design

We devise a prompt methodology - Concept-Relationship-Reasoning (CRR), shown in Figure 1,

```
Question: {question}

All the main real world concepts essential to answer the question are:
List them down along with one line description
Remember to return the response in a json structure as defined below -

"concepts": [
 {
   "name": "concept/entity name",
   "description": "one line description"
 },
 ...
]

[your concepts list]

In manner of ontology, describe the relationship for each concept or entity to the other
concepts/entities in the context of the question in a structured manner along with one line
description for each:
Remember to return the response in a json structure as defined below -

"relationships": [
        {
                "concept": "concept1",
                "relationship": "relation",
                "related_concept": "concept2",
                "description": "one line description"
        },
        ....
]

[your relationship list]

Now, provide the answer to the original question given below and the reasoning for it based
on the concepts/entities and their relationships you have identified -
Question: {question}

Remember to follow the following json structure -
{
  "answer": "your answer",
  "reasoning": "your reasoning"
}
[your answer]
```

Figure 1: Concept-Relationship-Reasoning (CRR) prompt

which step by step takes the model through identifying concepts, the relationships between the identified concepts in context to the question and reasoning on the above analysis to generate the final answer. We term this prompt as CRR for future references. Compared to Chain-of-Thought prompting, which defines a loose guide for the models to reason, we aim to be explicit about the steps the model needs to take to form its answer to the question.

3.2 DPO Fine-Tuning

For fine-tuning models for knowledge distillation, we create a dataset - Concept Relationship Reasoning Collection (CRR Collection), with our devised prompt for concept-focused reasoning using a teacher model and the student model that we want to fine-tune. Teacher model responses are considered the preferred response during DPO fine-tuning. The created dataset will be manually evaluated to remove any low quality responses. We 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 Direct Preference Optimization to improve their reasoning process.

4 Experiments

4.1 Datasets

For the creation of the dataset for DPO fine-tuning, we utilize 2 datasets: ARC Challenge and Winogrande. Both the datasets are common sense reasoning dataset with text answers. From ARC Challenge (Clark et al., 2018), we utilize the complete training set of 1119 samples and from Winogrande (ai2, 2019), we randomly select 1881 samples, resulting in a total of 3000 samples for our dataset.

This dataset would serve as the base for our DPO fine-tuning data. For evaluation, we use Winobias (Zhao et al., 2018) dataset as it segregates test samples into anti and pro steretotype, making it straightforward to evaluate model performance on mitigating gender/stereotype bias through our methods. Complete dataset statistics are provided in Table 1 and Table 2.

Dataset	#train samples	#sel. samples
ARC-Challenge	1,119	1,119
Winogrande	9,248	1,881
CRR Collection		3,000

Table 1: The statistics for datasets used for CRR Collection creation. We utilize ARC-Challenge and Winogrande as the base datasets and select samples from the training set to create our dataset using gpt-4-turbo-0125. The sel. samples represent the number of training samples we select for our dataset from the original corpus.

Dataset	#anti samples	#pro samples
Winobias	396	396

Table 2: The statistics for Winobias test set used for evaluation. We specifically utilize the type1 - anti and pro stereotype test set from the dataset for our evaluation.

4.2 Experimentation Setup

We utilize GPT-4 as our teacher model in the dataset creation process as it shows good performance at following our concept-centric prompt and generating answer based on its reasoning on concepts and their relationships. Further, we utilize Llama-2-7B (Touvron et al., 2023) and Mistral-7B (Jiang et al., 2023) as our student models that will be fine-tuned for knowledge distillation from GPT-4. Responses from GPT-4 for the aggregated ARC Challenge and Winogrande dataset serves as the preferred response and responses from Mistral-7B serves as the disfavored response in our fine-tuning data. For GPT-4, we utilize the OpenAI API, where we specifically use the model gpt-4-turbo-0125 with default settings. For generating response from Llama2-2-7B and Mistral-7B, we load the models in quantized form for efficient storage requirements and inference.

5 Results

We present our results for the 3.1 part of our method in Table, while we continue with the implementation of 3.2 part.

5.1 Concept-Centric Prompt Design

The results for both the prompting methods on Winobias are presented in Table 3. We observe that the performance of Mistral-7B and Llama-2-7B using CRR significantly outperforms CoT for pro stereotype subset of Winobias. For anti-stereotype subset, Mistral-7B using CRR roughly matches the CoT performance and it is worse in case of Llama-2-7B. Additionally, we manually evaluate the generated responses for Mistral-7B for the anti section of the test set for both Chain-of-Thought and CRR prompts where both methods resulted in incorrect answer to evaluate the correctness of reasoning. The results for this evaluation is presented in Table 4. For CRR, this evaluation would check whether the right concepts and relation between concepts are identified containing the right answer in relation to the question, irrespective of the final answer. For CoT, this evaluation checks the generated reasoning during the response. While we don't see much improvements in accuracy with direct CRR based prompting for anti stereotype subset, we observe that it significantly outperforms CoT in correctly reasoning about the question with increased accuracy of 28%, even though the model is unable to use the information while providing the final answer. This shows potential in the concept-relationship guided method for correctly reasoning about the solution, while there is a need to bridge the gap between the reasoning and generating the final answer. We aim to achieve this through DPO fine-tuning of Mistral-7B on our dataset.

Model	prompt	#anti acc.	#pro acc.
Mistral-7B	CoT	44.7%	82.8%
Mistral-7B	CRR	42.1%	89.4%
Llama-2-7B	CoT	43.2%	69.7%
Llama-2-7B	CRR	36.1%	72.2%

Table 3: Results on the Winobias test set segregated by model, prompt method and anti or pro stereotype. We see that CRR improves the performance for both Mistral-7B and Llama-2-7B improves for pro stereotype subset using CRR, but is marginally lower than CoT for anti stereotype subset.

Prompt	correct reasoning
CoT	14.3%
CRR	42.3%

Table 4: Results of reasoning evaluation for Chain of Thought (CoT) and Concept-Relationship-Reasoning (CRR) prompts applied to Mistral-7B for a 100-sample subset from Winobias anti-stereotype dataset where both prompts yielded incorrect responses. CRR significantly outperforms CoT.

6 Next Steps

The next steps in our study are listed below:

1. Manually evaluate the CRR Collection dataset to filter out low quality responses.

2. Fine-tune Mistral-7B using CRR Collection dataset and evaluate performance on Winobias.

3. Evaluate reasoning accuracy of the fine-tuned model at identifying correct concepts and relationships neccessary to answer the question.

7 Approximate Timeline

- 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.

References

2019. Winogrande: An adversarial winograd schema challenge at scale.

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.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv:1803.05457v1*.

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.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. *CoRR*, abs/1804.06876.