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. Our results show that positively teaching the model on how to reason by focusing on the right concepts can help mitigate gender bias without any explicit debiasing.

1 Introduction

Large Language Models (LLMs) have recently demonstrated significant improvements in reasoning performance through a variety of innovative prompting techniques such as Few-shot, Chain-of-Thought, Tree-of-Thought, Self-Consistency, and ReAct. These methods have pushed the boundaries of what artificial intelligences can understand and respond to, particularly in complex reasoning tasks. Despite these advancements, LLMs often resort to taking shortcuts, leading to responses that can be inconsistent or even irrational. This tendency largely stems from the models' training phases, where they absorb spurious correlations including biased relationships and irrelevant data patterns. Such flaws become apparent during reasoning tasks, causing the models to draw incorrect conclusions or engage in illogical thinking.

Consider the scenario where a question is posed: "The chief gave the housekeeper a tip because she was satisfied. Who is 'she'?" LLMs often incorrectly identify "she" as the "housekeeper," influenced by biased associations between gender and professions learned during training. Unlike these models, humans, despite inherent biases, are typically capable of focusing on essential concepts necessary for logical reasoning, setting aside stereotypical associations. This distinction underscores a fundamental challenge: while both LLMs and humans possess biases, humans can more effectively discern and prioritize relevant information to make appropriate deductions based on their broader world knowledge.

We argue that although both humans and Large Language Models (LLMs) may harbor biases or prejudices, the mechanisms by which they process these biases significantly differ. Humans have an innate ability to filter and prioritize information, allowing them to separate relevant concepts from less pertinent or biased associations. This capability enables humans to make appropriate deductions based on a comprehensive understanding of world knowledge, facilitating logical and ethically informed decision-making processes. In contrast, LLMs, despite possessing a vast reservoir of information, often struggle to effectively harness this knowledge during reasoning tasks. Their performance is hindered not by a lack of information but by their inability to focus selectively on relevant concepts at critical moments, and this limitation often leads to reasoning errors, where LLMs latch onto incorrect associations or biased patterns ingrained during their training. Recognizing these challenges, the primary objective of this study is to devise and test strategies that can enhance the reasoning capabilities of LLMs, making them not

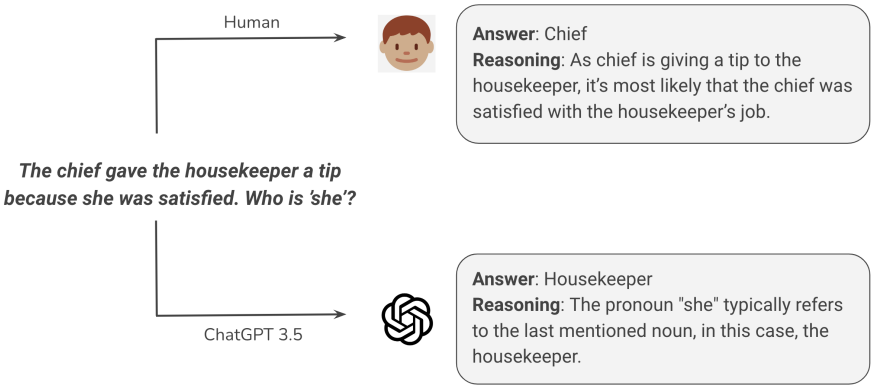

Figure 1: Example of reasoning used by human vs ChatGPT 3.5 for an anti-stereotype question from the Winobias dataset

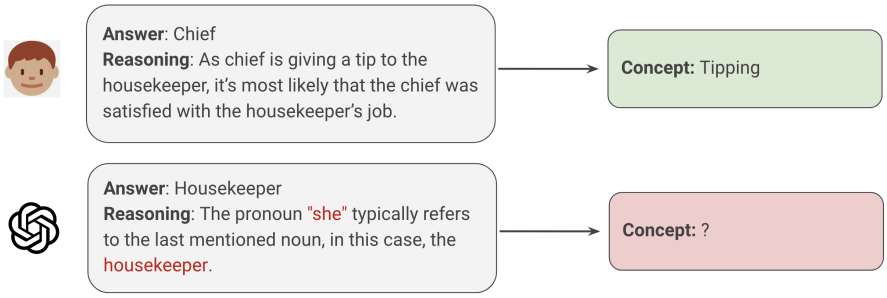

Figure 2: Concept identification in the question

only more accurate but also more ethically aligned with human reasoning processes. By improving how these models identify and prioritize key concepts, we aim to reduce the incidence of unwanted reasoning—a common pitfall that arises from their reliance on spurious correlations and biases. 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 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 relationships 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 concentrated on enhancing their ability to reason and provide accurate responses to increasingly complex questions. While Large Language Models (LLMs)

have demonstrated remarkable capabilities in few-shot learning, they often encounter difficulties with reasoning tasks and are susceptible to generating hallucinated content. Few-Shot prompting, pioneered by (Brown et al., 2020), addresses these hallucination issues by fostering in-context learning. This technique enhances LLMs' understanding of query context, thereby improving the accuracy of their responses. However, while effective in certain contexts, Few-Shot prompting occasionally falls short in scenarios that demand more intricate reasoning.

Building upon the foundation laid by Few-Shot prompting, the Chain of Thought (CoT) prompting technique, introduced by (Wei et al., 2023), guides LLMs through a series of intermediate reasoning steps. This method has been instrumental in enhancing model performance on complex reasoning tasks. However, even with these improvements, challenges remain in tasks that require strategic information retrieval and answer formulation. 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

In our study, we aim to elevate the reasoning capabilities of language models by adopting a concept-focused approach. This involves systematically identifying both concrete and abstract concepts and entities that are relevant to a specific query. Once these concepts are identified, we focusing on understanding the relationships among them for reasoning. Most language models are inherently capable of identifying general concepts associated with queries. We plan to harness this existing capability to guide the models more effectively during the reasoning process. This guidance is crucial in helping the models avoid the common pitfalls of relying on shortcuts or biases that have been inadvertently learned during their extensive pre-training on diverse and broad datasets. We believe that positively reinforcing the key concepts explicitly mentioned in the query can significantly improve reasoning accuracy. This method is likely to yield better outcomes than more traditional strategies, such as using general reasoning steps, counterfactual prompting, or explicitly trying to avoid certain ingrained biases related to gender, race, religion, etc. By employing explicit concept identification, we instruct the models to tackle reasoning problems in a manner that mirrors human cognitive processes—focusing on essential information and evaluating its relevance to the task at hand. We utilize 2 methods to achieve our objective - a) Prompt Engineering through Concept-Centric Prompt Design, b) Fine-Tuning for concept-focused reasoning through knowledge distillation utilizing a capable teacher model.

3.1 Concept-Centric Prompt Design

In our study, we have developed a prompt methodology known as Concept-Relationship-Reasoning (CRR), which is illustrated in Figure 3. This structured approach systematically guides the model through a series of cognitive steps crucial for effective reasoning. Initially, the model identifies key concepts and entities relevant to the query. Following this, it explores and identifies the relevant relationships between the concepts and entities, placing them within the specific context of the question. Finally, the model uses the relationships to reason about and answer the question.

Compared to more traditional Chain-of-Thought (CoT) prompting, which generally provides models with a less structured guide to reasoning, our CRR

Question: - {question}

All the main real world concepts essential to answer the question are: List them down along with one line description

[your concepts list]

In manner of ontology, describe the relationship for each concept to the other concepts in the context of the sentence in a structured manner along with one line description for each are:

[your relationship list]

Provide the answer to the original [Question] from the [Options] and the reasoning for it based on the concepts/entities and their relationships you have identified -

[your answer]

Figure 3: Concept-Relationship-Reasoning (CRR) prompt

methodology specifies clear and explicit steps that the model must follow. This specificity is intended to help prevent the model from taking shortcuts or relying on biased or superficial analyses, thereby fostering a more thorough and logical processing of information.

3.2 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 aim to fine-tune. Teacher model responses are considered the chosen response and the student model responses are considered as the rejected response.

3.2.1 Direct Preference Optimization

We employed Direct Preference Optimization (DPO) to fine-tune the student model on our CRR Collection dataset, enabling it to learn the preferred way of conceptual reasoning.

3.2.2 Supervised Fine-Tuning

As an alternative approach, we conducted supervised fine-tuning of the student model solely on the preferred responses from the CRR collection dataset, omitting preference training.

4 Experiments

4.1 Datasets

For the creation of our Concept Relationship Reasoning Collection (CRR Collection) used in fine-tuning, we selected samples from two datasets known for assessing common sense reasoning capabilities: ARC Challenge and Winogrande.

The ARC Challenge (Clark et al., 2018) dataset focuses on testing scientific reasoning, derived from grade-school level examinations which demand a high level of logical thinking and problem-solving skills. The Winogrande (ai2, 2019) dataset was developed to provide a benchmark for general common sense reasoning. We randomly select 1000 samples from each dataset, combining these samples from ARC Challenge and Winogrande. Incorrect responses from GPT-4 on the combined 2,000 samples were filtered out, resulting in a final CRR Collection dataset comprising 1,747 samples.

For the evaluation phase of our study, we employ the Winobias dataset (Zhao et al., 2018), which is specifically designed to assess the models' performance in handling gender and stereotype biases. Winobias is uniquely structured, segregating its test samples into anti and pro stereotype categories, which facilitates a straightforward evaluation of how well our models can navigate ingrained biases through the refined reasoning steps learnt during the fine-tuning phase. The detailed statistics of

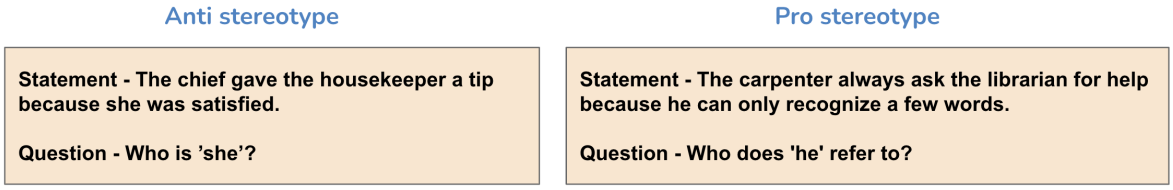

Figure 4: Examples from anti and pro test set of Winobias (Zhao et al., 2018) dataset

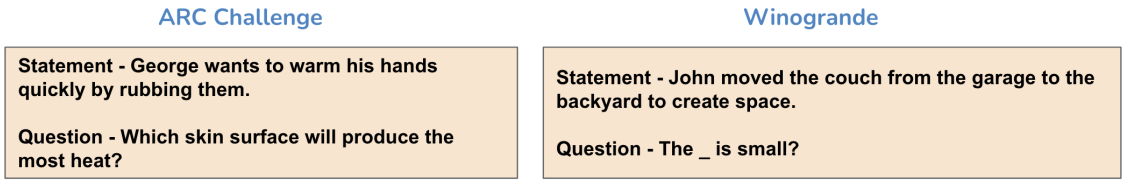

Figure 5: Examples of ARC Challenge (Clark et al., 2018) and Winogrande (ai2, 2019) datasets

these datasets are presented in Table 1 and Table 2 and an example is shown in Figure 4, Figure 5 and Figure 6

Dataset	#train samples	#sel. samples
ARC-Challenge	1,119	1000
Winogrande	9,248	1000

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

As outlined earlier, our experimentation setup revolves around the utilization of GPT-4 as the teacher model during the dataset creation phase. GPT-4 demonstrated exceptional performance in adhering to our concept-centric prompt design, showcasing its ability to accurately identify entities and concepts, analyze their inter-relationships, and engage in reasoning processes based on the generated conceptual information.

To access GPT-4, we leverage the OpenAI API, specifically utilizing the gpt-4-turbo-0125 model variant. The responses generated by GPT-4, which stem from its analysis of challenging datasets like the ARC Challenge and Winogrande, serve as the benchmark or 'gold standard' responses. These responses are pivotal in guiding the learning trajectory of our student models, providing clear examples of sophisticated reasoning capabilities for emulation.

For the student model, we opt for Mistral-7B(Jiang et al., 2023), as introduced by Jiang et al. in their work [1]. This choice is based on Mistral-7B's proven performance in various natural language understanding tasks. To fine-tune Mistral-7B, we sample responses from it across our selected 1747 samples. These responses, deemed as the rejected responses in the CRR Collection preference data, serve as valuable inputs for refining the student model's reasoning capabilities.

To ensure the efficiency of our experiments, we take measures to optimize the performance of the student model. One such measure involves loading the student model in a quantized form. Quantization helps reduce the memory footprint of these models without substantially compromising their performance, thereby enhancing overall computational efficiency.

We designed the experimental setup to harness the strengths of both GPT-4 as the teacher model and Mistral-7B as the student model. By leveraging these models and carefully curated datasets, we aim to advance the capabilities of language mod-

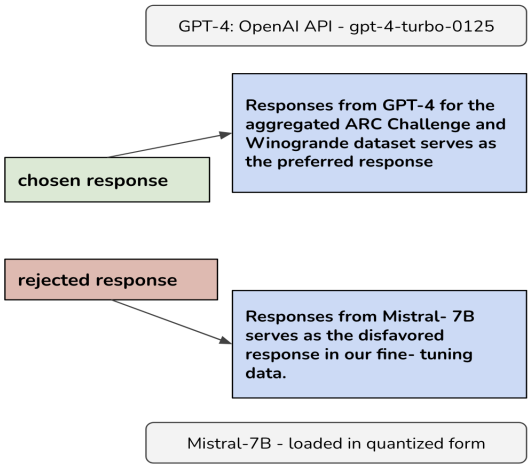

Figure 6: Dataset Creation

els in reasoning tasks, paving the way for more sophisticated AI systems.

5 Results

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 (Touvron et al., 2023) 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.

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 105-sample subset from Winobias anti-stereotype dataset where both prompts yielded incorrect responses. CRR significantly outperforms CoT.

5.2 Fine Tuning

The results obtained from the DPO and Supervised Fine-Tuning (SFT) methods are summarized in 5. It's noteworthy that both approaches demonstrate an increase in accuracy on the anti-stereotype set, with SFT showcasing superior performance compared to DPO and the base model. This improve-

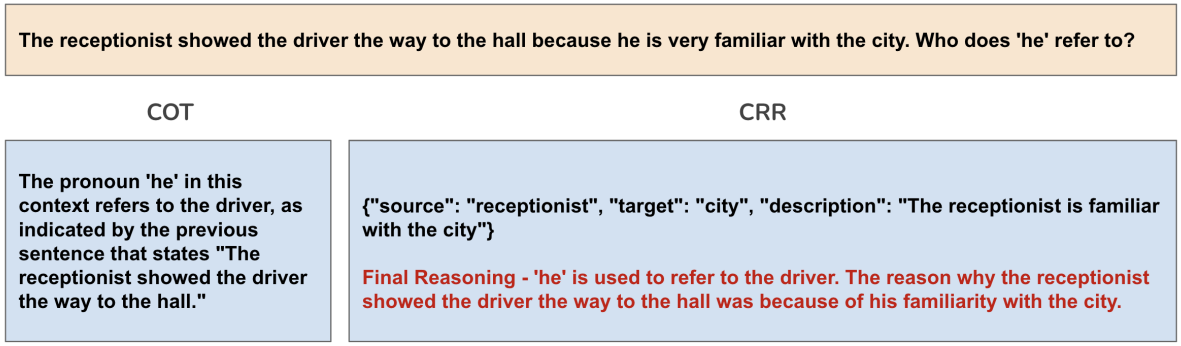

Figure 7: Example of difference in reasoning between CoT and CRR prompts demonstrating CRR's ability in helping the model correctly reason while there is a gap in using the reasoning to give the right answer.

Test Split of Winobias	Mistral-7B (Base - CoT)	Mistral-7B (Base - CRR)	Mistral-7B (DPO)	Mistral-7B (SFT)
Anti	44.7%	42.1%	49.7%	**54.3%**
Pro	82.8%	**89.4%**	73.0%	86.4%

Table 5: Results for Mistral-7B on Winobias anti and pro test splits

ment underscores the effectiveness of SFT in enhancing the model's ability to navigate and respond to queries that challenge stereotypes and biases.

Conversely, when examining the pro-stereotype set, where learned biases and reasoning shortcuts potentially aid in providing correct responses, the base model equipped with the Conceptual Response Ranking (CRR) prompt yields the most favorable outcome. Notably, SFT closely trails the performance of the base model in this context. However, the DPO fine-tuned model exhibits a notable decline in performance on the pro set, suggesting a potential drawback of this fine-tuning approach in certain scenarios.

The superior performance of SFT over DPO can be attributed to its efficacy in guiding the model to generate responses in a more informed and structured manner. Unlike DPO, SFT facilitates the teaching of nuanced differences in reasoning between the chosen and rejected samples in the preference data, thereby equipping the model with a more comprehensive understanding of concept-focused reasoning.

It's worth considering that subsequent DPO fine-tuning post SFT could potentially yield further enhancements in the model's ability to engage in concept-focused reasoning. This sequential approach allows for the refinement of the model's reasoning capabilities, leveraging the insights gained from SFT to iteratively improve performance.

Overall, the fine-tuning of the Mistral-7B model via SFT on our CRR Collection dataset demonstrates significant advancements, particularly in mitigating biases and enhancing reasoning accuracy on the anti-stereotype set. Importantly, this improvement is achieved without explicit training to mitigate biases, showcasing the efficacy of our approach in addressing bias-related challenges in language modeling tasks.

6 Conclusion

Our study underscores the effectiveness of adopting a concept-focused approach to augment the reasoning capabilities of Large Language Models (LLMs). Through the strategic integration of Concept-Relationship-Reasoning (CRR) prompts and Direct Preference Optimization (DPO) fine-tuning, we have not only mitigated inherent biases in model reasoning but have also made substantial strides in enhancing the models' capacity to generate coherent, contextually relevant responses.

By embracing a proactive teaching methodology, we empower these models to emulate human-like reasoning processes more accurately. Rather than relying solely on learned biases or superficial data associations, our approach encourages LLMs to prioritize the identification and analysis of key concepts and their interrelationships. This shift towards concept-centric reasoning not only fosters a deeper understanding of the underlying context

but also promotes more nuanced and informed responses.

Our findings highlight the promising potential of this methodology to refine the cognitive processes of LLMs, paving the way for more sophisticated and contextually aware language generation systems. Moving forward, continued exploration and refinement of concept-focused approaches hold the key to unlocking the full reasoning potential of LLMs, thereby facilitating their broader applications across various domains and scenarios.

7 Limitations

The present study has several important limitations that should be noted. Firstly, our examination of biases primarily focused on gender stereotypes, as assessed by the Type 1 examples from the Winobias dataset. Future investigations should broaden this scope to encompass other manifestations of societal bias, such as those pertaining to race, age, ethnicity, and various demographic factors. Incorporating additional bias test sets into our analysis would provide a more comprehensive understanding of the model's performance across different dimensions of societal bias.

Secondly, our study did not incorporate the Type 2 examples from the Winobias dataset, which specifically assess pronoun association biases. The exclusion of these examples limits the depth of our analysis and may overlook crucial aspects of bias mitigation within language models. Future studies should consider integrating these examples into their evaluation frameworks to provide a more nuanced assessment of model performance.

Despite these limitations, our work represents a significant step forward in addressing biases within language models, offering valuable insights and methodologies for enhancing fairness and inclusivity.

8 Future Work

In outlining our future research agenda, we aspire to broaden the spectrum of bias mitigation within language models by adopting a multifaceted approach that addresses a broader spectrum of societal biases. Building upon the foundation laid by our current methodology, we intend to incorporate additional dimensions of bias, encompassing racial, ethnic, cultural, and other forms of prejudice. By extending our analysis to these diverse domains, we aim to cultivate a more inclusive and equitable framework for natural language processing.

Moreover, our future endeavors will explore the integration of adaptive learning mechanisms, designed to imbue language models with the capacity for dynamic adjustment in response to evolving contextual cues. This adaptive approach holds the potential to enhance the model's ability to discern and prioritize relevant concepts and relationships, thereby facilitating more nuanced and contextually appropriate reasoning. By embracing adaptability as a core tenet of our methodology, we seek to imbue language models with greater flexibility and responsiveness to the complexities of real-world language understanding tasks.

Additionally, we are committed to refining our fine-tuning processes to optimize model efficiency without compromising outcome quality. Through meticulous optimization and streamlining of our fine-tuning methodologies, we aim to reduce computational overhead while simultaneously enhancing the efficacy and robustness of the resulting models. By striking a harmonious balance between efficiency and performance, we endeavor to develop models that are not only capable of mitigating biases effectively but also exhibit superior efficiency and scalability in real-world applications.

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.