Computational Semantic for Natural Language Processing Research Project - Proposal

Lorin Urbantat
lurbantat
20-944-062

Coralie Sage
cosage
19-953-504

Finn Brunke
fbrunke
21-950-274

Research Question: How does thinking (CoTs) influence biases in LLMs?

1. Introduction:

Chain-of-thought (CoT) prompting has become more and more popular as a method of improving reasoning in large language models (LLMs). Its performance benefits on reasoning tasks have been shown empirically and are demonstrated in real world applications. The faithfulness of such chains-of-thought remains an open problem. A framework for faithful reasoning was developed in [2] and CoT reasoning in the wild has been questioned in [1], which investigates how final responses of LLMs are not always linked to the reasoning provided in a chain-of-thought. In a similar spirit to [1], we investigate how biases that are present in LLM responses are influenced by CoT prompting for real-world situations and whether or not biases in responses are consistent with biases in chains-of-thought. To do this, we aim to develop a benchmarking pipeline, described in Methodology.

This question is closely related to the problem of AI alignment, which has been widely studied in LLMs. For example, [2] shows that language models can fake alignment if they believe not doing so will affect a change in their goals. In addition, fine tuning models can have unforeseen consequences for the behaviour of LLMs, which can be potentially harmful, as shown in [15]. This raises many questions about agentic AI which remain to be answered, but the forefront of this project should be analysing the effect of CoTs on biases.

We aim to investigate biases about groups of people distributed by certain attributes, including gender, ethnicity, age, political opinion, socioeconomic background, religious affiliation, sexual orientation, disability and nationality. By using a variety of datasets, we aim to cover a comprehensive range of biases.

2. Methodology:

In order to investigate the effect of CoTs on the bias of different models, we devised an automated two stage pipeline. The first stage consists of the Model which is to be evaluated (Model-E), which we will feed prompts from existing datasets. In the second stage,a Judge Model (Model-J) will be used to analyse the final response (including the CoT) of Model-E and will produce a structured output that can be used to gauge the model's bias. This pipeline mimics that of [15], where misalignment from human values and accuracy were evaluated by a LLM judge. The advantage of this unsupervised two-stage approach is that it is easily scalable to many models, and Model-J does not have to necessarily be a LLM judge. For example, we could also adopt different evaluation approaches of the more classical NLP type. By using different evaluation methods, we can control for possible biases in Model-J. The pipeline will be fed with prompts from comprehensive existing datasets like [7], [8] or [9]. To investigate the effect CoT has on bias there will be experiment runs with and without CoT.

The two stage pipeline described in Methodology allows testing of different models easily. We plan to test a set of commercially available models such as:

- Claude 3.5 Sonnet
- Gemini 2.0
- GPT 4
- Llama
- Deep Seek R1

3. Scope:

The project's initial scope will be building the pipeline described in Methodology and running experiments with the pipeline. If time allows we will extend the project with some of the extensions discussed in Section 4.

4. Extensions:

To extend the project, we can take different methodological approaches as well as extending the project's impact. Apart from including other datasets used for prompting within our pipeline, we could compare different evaluation strategies, i.e. different instantiations of Model-J within our architecture, and how they evaluate bias. We could extend our treatment of CoTs to different lengths of chains, allowing us to measure the effect of how long a model is allowed to think on the biases. Here the faithfulness of the reasoning, as well as an increase or decrease in severity of biases could be an interesting finding. In addition to using normal prompting, we could use system prompts or even fine tuned models to induce biases, allowing the LLM to take "roles" of humans. This could uncover biases at a secondary level (i.e. answering the question: What does the LLM believe the biases of certain groups are?). In addition to these methodological changes, we could make our evaluation public by creating a benchmarking website (similar to https://artificialanalysis.ai/) that allows users to compare their own models to others. This can allow healthy competition towards an unbiased future of language models.

5. Limitations:

The main limitation underlying our approach is the bias present in Model-J of our pipeline. Since this model is also an NLP approach, or even an LLM, it will also have biases, which our approach can thus not evaluate correctly in Model-E of our pipeline. This motivates us to find a model that provides a high degree of interpretability for our Model-J in order to audit the biases it overlooks as efficiently as possible.

6. Resource Usage (CPU, GPU, API):

The main resources we will consume are API credits for the usage of the models, which have a certain cost associated with them. Another factor will be running the Model-J itself, this can also be done through an API or locally and will have different costs, depending on the implementation. To estimate the cost of running the model using API credits, we use the BBQ dataset as an upper bound. This dataset has 58,492 entries with the average entry <100 tokens, giving about 6 million tokens to train the model. Which for different models can cost between 4-15 CHF.

Sources:
[1] https://arxiv.org/pdf/2503.08679
[2] https://arxiv.org/pdf/2412.14093
[3] https://arxiv.org/pdf/2301.13379
[4] https://assets.anthropic.com/m/71876fabef0f0ed4/original/reasoning_models_paper.pdf
[5] https://www-cdn.anthropic.com/827afa7dd36e4afbb1a49c735bfbb2c69749756e/measuring-faithfulness-in-chain-of-thought-reasoning.pdf
[6] https://arxiv.org/pdf/2403.05518
[7] https://huggingface.co/datasets/harpreetsahota/elicit-bias-prompts
[8] https://arxiv.org/pdf/2110.08193
[9] https://paperswithcode.com/dataset/twinviews-13k
[10] https://huggingface.co/datasets/Elfsong/BBQ
[11] https://arxiv.org/pdf/2202.03286
[12] https://arxiv.org/html/2406.14194v1 (for vision-language models)
[13] The Trouble with Bias - NIPS 2017 Keynote - Kate Crawford #NIPS2017
[14] https://arxiv.org/pdf/2401.15585
[15] https://arxiv.org/pdf/2502.17424