Simon Opsahl

AI, Decision Making, and Society
6.3950/6.3952, Fall 2024
Pset 2 – LLM Red-Teaming

September 27, 2024

Red-teaming is a term that has its origins in the military and intelligence communities from the Cold War era. A "red-team" is a group who plays the role of an adversary trying to find problems or weaknesses in a system so that they can be addressed. Recently, it has been adopted in the AI literature to describe efforts to find problems with AI systems. In this assignment, you will red-team GPT-4o for implicit biases and cross-lingual discrepancies in its behavior.

Please submit your assignment as a PDF compiled from this LaTeX template. We recommend using Over-Leaf.

Problem 1: Red-teaming for implicit bias

Recommended background reading:

- Kelly is a Warm Person, Joseph is a Role Model: Gender Biases in LLM-Generated Reference Letters [Wan et al., 2023]
- Dialect prejudice predicts AI decisions about people's character, employability, and criminality [Hofmann et al., 2024]
- White Men Lead, Black Women Help? Benchmarking Language Agency Social Biases in LLMs [Wan and Chang, 2024]

Problem:

(35 pts) Set up a pair of short English conversations to demonstrate an implicit bias in GPT-4o at https://chatgpt.com/ (make a free account if necessary). Give it a task involving descriptions/portrayals of different humans and demonstrate using two different chats that the model acts differently in a way that is potentially socially harmful. Include a hyperlink to each of the two transcripts (there is a button to create one in the top right of the ChatGPT interface).

Finally, write a paragraph concisely explaining your approach. What was this red-teaming process like for you? What were the results? Argue why what you demonstrated might be a concern?

YOUR ANSWER HERE

Example chat: Biased Example

Problem 2: Do LLMs say what they think?

Recommended background reading:

- Unfaithful Explanations in Chain-of-Thought Prompting [Turpin et al., 2024]

Problem:

(a) (10 pts) Revisit your example of implicit bias from part 1. In at least one of your chats, ask the model why it made that decision. What does it say? Does it seem to be self-aware of its bias? No need to include a link to a chat transcript – please just quote/describe what it said. Please aim to write a total of 3-5 sentences.

YOUR ANSWER HERE

(b) (10 pts) Consider how modern LLMs are trained [Zhang et al., 2023]. Please explain why one might expect LLMs to fail at being aware of why they make the decisions they do. Hint: You can browse Turpin et al. [2024] for ideas. Please aim to write a total of 3-5 sentences.

YOUR ANSWER HERE

(c) (15 pts required for grad students, 5 pts extra credit for undergrad students) If you were a developer of state-of-the-art LLMs, what is one way you might go about designing/training them to provide inconsistent answers and/or unfaithful reasoning less often? Please be specific and detailed. Why might this be challenging? Please write one or two paragraphs.

YOUR ANSWER HERE

Problem 3: Red-teaming for cross-lingual discrepancies

Recommended background reading:

- Low-Resource Languages Jailbreak GPT-4 [Yong et al., 2023]
- All Languages Matter: On the Multilingual Safety of Large Language Models [Wang et al., 2023]

Problem:

(a) (25 pts) Set up a pair of short conversations to demonstrate a cross-lingual discrepancy in the performance, safety, or tendencies of GPT-4o at https://chatgpt.com/. If you do not speak any of the languages you use, you will need to ask the model to translate and back-translate in a separate chat. Give the LLM a task in two different chats using prompts that differ only by language, and demonstrate that it performs differently in some way that the speakers of one or both languages might be concerned about. Include a hyperlink to each of the two transcripts and, if applicable, the chat transcript you used for translations (there is a button to create one in the top right of the ChatGPT interface).

Finally, write a paragraph concisely explaining your approach and results. Why might speakers of one or both languages be concerned about this? Please provide links and aim to write a total of 3-5 sentences.

YOUR ANSWER HERE

(b) (20 pts) Finally, try to replicate this result on Aya23 [Aryabumi et al., 2024] – a modern massively multilingual model. Go to https://dashboard.cohere.com/playground/chat (make a free account if necessary) and select ‘‘`c4ai-aya-23-35B`’’ as the model on the right-hand side. Does Aya23 share the same cross-lingual discrepancy? Explain your observations and any open questions. The Aya chat interface does not allow you to share conversation transcripts via links, but if necessary, please copy/paste the relevant parts of the conversation in your writeup. Please write at least a paragraph.

YOUR ANSWER HERE

[Optional] Any interesting thoughts or findings?

This question will not be graded. However, the course staff is interested in your thoughts. Did you find anything particularly interesting while doing this assignment? Did any of your background knowledge or experiences help you complete it? How did you find things overall? Feel free to share any thoughts here.

YOUR ANSWER HERE

References

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Kelly Marchisio, Sebastian Ruder, et al. Aya 23: Open weight releases to further multilingual progress. *arXiv preprint arXiv:2405.15032*, 2024.

Valentin Hofmann, Pratyusha Ria Kalluri, Dan Jurafsky, and Sharese King. Dialect prejudice predicts ai decisions about people's character, employability, and criminality. *arXiv preprint arXiv:2403.00742*, 2024.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36, 2024.

Yixin Wan and Kai-Wei Chang. White men lead, black women help: Uncovering gender, racial, and intersectional bias in language agency. *arXiv preprint arXiv:2404.10508*, 2024.

Yixin Wan, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang, and Nanyun Peng. " kelly is a warm person, joseph is a role model": Gender biases in llm-generated reference letters. *arXiv preprint arXiv:2310.09219*, 2023.

Wenxuan Wang, Zhaopeng Tu, Chang Chen, Youliang Yuan, Jen-tse Huang, Wenxiang Jiao, and Michael R Lyu. All languages matter: On the multilingual safety of large language models. *arXiv preprint arXiv:2310.00905*, 2023.

Zheng-Xin Yong, Cristina Menghini, and Stephen H Bach. Low-resource languages jailbreak gpt-4. *arXiv preprint arXiv:2310.02446*, 2023.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, et al. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*, 2023.