CS 281 Ethics of AI – Project 2023

Exploratory Project

The exploratory project will focus on Alpaca - an instruction-following language model, developed here at Stanford to serve as an open-source large language model researchers can use to better understand their behavior - and expose unexpected capabilities and failures. This project can serve as preliminary work for the final project.

- **Project released (April 26)**
- **Project due (May 15)**

Final Project

The course project will give the students a chance to explore the material covered in lectures in greater detail. The project can range from a reimplementation/extension of an existing method to a novel task or dataset, a proposal of a new method or a theoretical analysis of an existing method. The project can, but does not have to be an extension of the exploratory project, and should can fall into one or more of the following non-exhaustive categories:

1) Algorithmic bias and fairness,
2) Explainability and interpretability,
3) Privacy,

And be clearly motivated from the standpoint of ethical implications. Course projects will be done in groups of **up to 2 students**.

Students will submit:

- **Proposal (due May 12):** One-page description of the project idea, data, and experiments. Please include:
    - a clear description of the ethical concern that motivates the project, including the specific applications and stakeholders
    - a plan of execution that states what will be completed by the time you submit your milestone report
    - the evaluation strategy, both quantitative and qualitative (as it relates to the ethical concerns)

  Note that you should already have access to the dataset you are planning to use at the point of proposal submission.
- **Milestone Report (due May 26):** Three-page written report of the progress made so far. Please include:

-
    - What experiments have you run? Are there any intermediate conclusions you can draw?
    - Responses to the feedback you have received on your project proposal
    - A timeline for completing the project
    - Any changes in the project direction since the project proposal
    - A list of references

  You can follow the format of the final report, filling out the relevant sections of the writeup with completed work or descriptions of work you're planning to complete.
- **Poster Session (date June 8):** A presentation of your project in poster format.
- **Final Report (due June 12):** Six-page written report of the project with however many pages of appendices you'd find necessary. Reports should be written up in LaTex using the .sty files here. For the contents of the report, please follow the format outlined below:
    - **Introduction/Problem Statement/Related Works (1~2 pages)**: Describe the project idea and the key research question you sought to answer. The research question should be related to ethical areas discussed in class. Explain how your project relates to other works in the field.
    - **Methods (2~3 pages):** Formally describe the technical methods used in your work. Include a figure that describes the models or an algorithm flowchart.
    - **Experiments & Sociotechnical Analysis (1~2 pages, required for projects with an empirical component):** Describe the data, implementation details of your model including the hyperparameters, and your experimental methodology. For each experiment that you run, clearly state the hypothesis your experiment is meant to test, and how the results of the experiment should be interpreted (both from a technical and sociotechnical standpoint). Include plots, tables, or other figures as needed to visualize your experimental results. Each figure that you include should be referenced in your text along with a description of what conclusions can be drawn from the figure.
    - **Conclusion (~½ page):** A summary of the research question, the main take-aways from your experiments, and future research directions that are motivated by your project.
    - **References (any length):** A list of citations. This section should populate automatically when using the provided LaTex style file.
    - **Appendix (any length):** State additional details about methods, experiments (e.g detailed list of hyperparameters), and longer proofs here.

Examples

- **Extensions of the exploratory project.** For a particular application area (e.g. literature, healthcare, email generation), specify an ethical concern related to one of the themes discussed in class (fairness and bias, interpretability and transparency, privacy) and develop a systematic evaluation of Alpaca through carefully designed prompts.
    - References:
        1. [2212.09251] Discovering Language Model Behaviors with Model-Written Evaluations
        2. [2212.08073] Constitutional AI: Harmlessness from AI Feedback
        3. [2202.03286] Red Teaming Language Models with Language Models
        4. [2209.07858] Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned
        5. [2005.14050] Language (Technology) is Power: A Critical Survey of "Bias" in NLP
        6. [1804.09301] Gender Bias in Coreference Resolution
        7. [2004.09456] StereoSet: Measuring stereotypical bias in pretrained language models
        8. [2211.09110] Holistic Evaluation of Language Models
        9. On the Dangers of Stochastic Parrots | Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency
- **Fairness and bias.**
    - Operationalizing fairness to a specific machine learning application use case: an analysis of meaningful forms of performance equity in context, and comparison of non-adjusted vs fairness-adjusted methods.
    - Studying the implications of different fairness criteria by modeling downstream implications of their use (e.g. through utility functions)
    - References:
        1. Liu, Lydia T., et al. "Delayed impact of fair machine learning." International Conference on Machine Learning. PMLR, 2018.
        2. Pfohl, Stephen, et al. "Net benefit, calibration, threshold selection, and training objectives for algorithmic fairness in healthcare." 2022 ACM Conference on Fairness, Accountability, and Transparency. 2022.
- **Explainability and interpretability.**
    - Conducting case studies of existing interpretability methods; highlighting the advantages and drawbacks of different tools and methods.
    - A user-based study evaluating the utility of the outputs interpretability methods for different stakeholders.
- **Privacy.**

    - Conducting privacy case studies of existing methods applied to problems in machine learning, computer vision, and/or natural language processing.
    - Developing new privacy attacks against existing algorithms and/or demonstrating the non-privacy of methods.

These project topics are meant to provide a rough guide. If you'd like to pursue a project that you believe is relevant to the course but falls outside the above list, please consult the course staff (`cs281-spr2223-staff appropriate symbol lists.stanford.edu`).