CS 690S: AI Alignment
Final Project, Part 1

The goal of the final project is to give each of you an opportunity to dive more deeply into a particular topic of interest. For those of you that perform research outside of your classwork, I encourage you to choose a project that contributes to your personal research, if possible.

The requirements for the project are as follows:

- You may work alone or with up to two other people. In fact, I encourage working with a partner whenever possible. All work must be your own, but you can build on code found online as long as you cite it clearly and your portion of the work is novel, either technically or empirically. Generative AI is also permitted.
- Each project should contain a novel intellectual contribution, as well as empirical results on a problem of interest.
- The proposed project must have enough intellectual and experimental depth that it will result in a final writeup of roughly half a conference paper (4 double-column pages in Latex).
- The proposed method for solving the problem should be (at least loosely) related to a core topic we have covered (or will cover) in class, relating to AI Alignment broadly. .
- Projects could include extending an algorithm in a novel way, comparing one or more existing algorithms rigorously on an interesting problem, or designing a new approach to attack a problem relevant to the class. In all cases, there should be a non-trivial intellectual contribution or insight from the approach or experiments.

For this portion of the project, you must write a one-page proposal (single-spaced, 12 pt font) outlining your plans for the project. This proposal will be worth 10% of the final project grade. Be sure to include:

- A clear description of the problem you are investigating, both abstractly and in context of a particular experimental domain
- References to a few papers that are relevant to the subject of interest
- A proposed plan to address your problem, which should outline what method(s) you plan to develop, implement, compare, or extend (and how)
- A testable hypothesis
- An experiment to test your hypothesis and a clear evaluation criteria to determine the outcome of your experiment / hypothesis