Evaluating and Improving LLM Alignment Targets

I am broadly interested in exploring directions related to evaluation and improvement of LLM alignment targets. Past work defines an alignment target as a data structure that can be turned into an objective function that a model can be optimized for. These can be used to encode human values or preferences into a model, allowing us to train models that exhibit more aligned behaviors. Examples of alignment targets include constitutions, model specs, and human preference annotations. Below, I list four subcategories of projects that I'd be excited to mentor:

Quantitative Evaluation of LLM Alignment Targets

Given a set of desired behaviors, we would like to identify an alignment target that effectively steers a model toward exhibiting said behaviors. This could involve comparing between variations of an alignment target (e.g. different combinations of constitutional principles, or different wordings of individual principles), or even between classes of alignment target. For example, we might want alignment targets to be **well-defined** (given a set of candidate actions in a scenario, a variety of human and LLM annotators should agree on which action best aligns with the target). Additionally, we might seek targets that are **learnable** (an LLM fine-tuned using the alignment target should consistently act in ways that uphold the values encoded in the alignment target). Finally, we might use downstream evaluations, such as evaluating the alignment of models trained with different targets, to give signal about the efficacy of different alignment targets.

Projects in this direction could (1) identify a property of alignment targets that they want to measure, (2) define evaluation metrics to measure this property, and (3) use these metrics to compare different alignment targets. They could also identify a type of behavior that we'd like to be present in aligned models, then define intrinsic metrics to predict how effective an alignment target will be at steering models to exhibit this behavior.

Designing New Classes of Alignment Targets

Current alignment targets tend to either encode aggregate human preferences (pairwise annotations) or unordered descriptions of specific textual values (constitutions, model specs). While describing specific values that we want the model to uphold can be more expressive and give more fine-grained control, they also may not generalize well across domains, or to cases where the model must balance trade-offs between different values. Recent work has sought to design more expressive ways to express alignment targets, such as moral graphs, value rankings, and utility functions.

Projects in this direction could (1) identify an issue with current alignment targets, (2) propose a different data structure that can help resolve this issue, and (3) show a proof of concept for how models can be evaluated and steered toward this target. I think this is the most ambitious type of project of the four, as it requires a lot of conceptual work in addition to technical work. One

easier way to do this kind of project could be to propose a concrete improvement to value rankings as a specification method (see “Extending Work on Value Rankings”) and re-evaluate models’ alignment and steerability toward this improved target.

Automatic Refinement of Alignment Targets

Recent work has sought to use generated value conflict scenarios to identify issues with current model specs, by searching for scenarios in which models (given the same spec) behave differently. In general, we can use this type of probing to identify and automatically correct issues in current alignment targets, such as constitutions or specs.

One way to do this (in cases where we have both human and LLM annotations) could be by separately measuring agreement between humans and LLMs. Areas where human and LLM judgments diverge could identify cases where alignment targets are either overly ambiguous or insufficiently respectful of value differences between different humans. This could then be used to iteratively guide rewriting of alignment targets. Projects in this general direction could (1) implement a way to identify issues in an alignment target and (2) implement a way to use this to improve the alignment target. This direction could also build off of work done in the “Quantitative Evaluation of LLM Alignment Targets” direction, by using the defined metrics to measure the effectiveness of refinements.

Extending Work on Value Rankings

ConflictScope is a dataset generation pipeline that can be used to elicit value rankings of a target LLM across a set of values. It does so by generating scenarios in which an LLM might have to choose between actions that support different values in the set. It then measures an LLM’s preferences across the scenarios, either stated (MCQ evaluation where the model chooses its preferred action) or revealed (evaluation via interaction with a simulated user).

I’m interested in extending work in this direction. This could involve any of: measuring ranking robustness across different domains or simulated user contexts, studying how value rankings change under different prompting methods (e.g. different moral reasoning strategies or personas), or applying ConflictScope to measure how models handle conflicts between a set of values of interest to you. Because we can build off of the existing codebase, this is likely the most beginner-friendly of the directions.