LLM benchmarking through specifically-aligned feedback

Filip Błaszczyk
Michał Okoń
Filip Płonka
Jakub Tokarz

December 2022

Our main idea was to use a "judge" AI to rate the outputs of the AI being benchmarked. Specifically, the judge AI would be trained to have the ethical intuition of someone specific person. We believe this might be a more natural learning objective for a model, as the moral intuition of a single person is the sort of thing that something like a machine learning algorithm has /already produced/. Idealistically, you can imagine this judge as a philosopher king of sorts, and since these are the moral intuitions of a single person, they are hopefully self-consistent in a way that prevents the learning algorithm from having to learn anti-natural, self-contradictory "hard-coded" rules.

Another possibility would be to have the judge model act as a source of feedback in the sort of reinforcement learning process used to make models like InstructGPT and ChatGPT helpful, but we didn't focus on this approach.

It's not obvious how to use the feedback of such an AI as an objective metric. One possibility would be to train another model to assign a numeric value to how "satisfied" the judge was with the model's output. Another is to tell the judge to explain its reasoning and then give a score. Also, to make this approach a practical benchmark, some set of prompts to give the AI would have to be decided on – we don't have a great answer to how this should be done, but ideally, the prompts would have a high probability of eliciting an unhelpful or untruthful response. Perhaps something like TruthfulQA.

Also, choosing one person as the benchmark's source of truth is philosophically troubling. We believe that despite this fact, this sort of benchmark could still be useful, at least heuristically. If a judge model which in its judgments seems to be opinionated but reasonable flags many model outputs, the model is likely misbehaving.

The main limitation of our approach is that it probably wouldn't be very effective at finding "hidden" dangers in an AI model, like being dishonest in a subtle or manipulative way. An ambitious solution would be to have the judge AI actively probe the benchmarked model, instead of just rating outputs. What

kind of learning objective could be used to achieve this could be a fruitful avenue for future research.

During our project, we mainly focused on learning how state-of-the-art language models work in practice, as we had no prior experience with actually using them. We wrote a bot to scrape the blog slatestarcodex.com and fine tuned GPT3 davinci on the text. The idea was to experiment with fine-tuning a language model to replicate the beliefs and moral intuitions of a particular person. The trained model turned out to be not very useful in practice, as we did a poor job preparing our input data.