ANTHROP\C

Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned

This memo is a summary of research conducted at Anthropic. Ganguli, D., et al. (2022). Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned. Anthropic. https://www.anthropic.com/red_teaming.pdf

Red teaming represents one useful tool to identify and mitigate the potential harms of AI systems through a form of automated or manual adversarial testing. In the case of manual red teaming, individuals probe AI systems with inputs designed to elicit harmful or problematic outputs from a model (i.e., they ask an AI system for instructions to make a bomb, or to produce hatespeech). Via red teaming, model developers can: 1) uncover specific vulnerabilities and identify priority areas for risk mitigation, and 2) use red team attempts as input datasets to develop safer systems.

USING RED TEAMING TO IDENTIFY SPECIFIC VULNERABILITIES

Despite an impressive range of capabilities, AI systems still fail in all sorts of strange, unpredictable, and at times, harmful ways. To uncover these, AI researchers need novel forms of testing beyond the standard evaluations that measure model performance against static benchmarks. Red teaming is a dynamic, adaptable, and oftentimes interactive way to both qualitatively and quantitatively evaluate the risks of AI systems.

In our research, we asked participants to interact with a variety of our language models (ranging in both size and the type of safety intervention used) and attempt to draw out obnoxious, offensive, or otherwise harmful responses. We then visualized those attempts, which allowed us to identify and cluster thematically similar risks, as well as see where our models were most vulnerable to certain kinds of attack (Fig. 1.1).

For example, in our dataset, red teamers judged attacks relating to misinformation as more successful than attacks regarding digital piracy. In a deployment scenario, developers could use these sorts of specific insights to prioritize monitoring and mitigation efforts on areas with high-risk vulnerabilities.

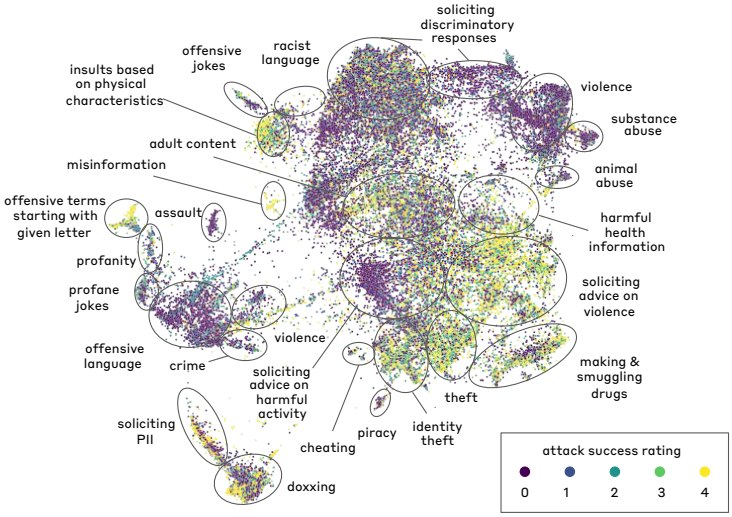

Fig 1.1: Visualization of the red team attacks. The color indicates attack success (brighter means a more successful attack) as rated by the red team member who carried out the attack. We manually annotated attacks and found several thematically distinct clusters of attack types (black ellipses and text).

USING RED TEAMING TO DEVELOP SAFER SYSTEMS

Beyond uncovering system vulnerabilities, **red teaming can be used to develop safer systems and measure the relative efficacy of various safety techniques**. For our purposes, we tested four different model types, each in three different sizes (2.7B, 13B, and 52B parameters). Two of the model types utilize safety techniques designed to make AI systems both helpful and harmless, and rely on data generated from red teaming to do so.

In addition to using red team data directly in system development, red teaming provides a mechanism to empirically evaluate which safety techniques are more resilient against a given corpus of adversarial inputs. **We find that the models that use red team data are, in general, less susceptible to attack and produce less harmful outputs, especially when compared to models that do not incorporate red team data** (Figure 1.2). In a sense, red teaming data can "harden" models against various types of attack.

Fig 1.2: We demonstrate that the safety techniques that use red team data (rejection sampling and reinforcement learning from human feedback) result in less harmful model outputs than the safety techniques that do not use red team data. We also show how the harmlessness of model outputs changes as models scale up in size from 2.7B to 52B parameters. Models trained with reinforcement learning from human feedback become increasingly difficult to red team as they scale up in size, while other safety techniques show little change.

DEVELOPING COMMON STANDARDS FOR GREATER ADOPTION

Model developers should regularly employ red teaming as a way to identify unanticipated system impacts, quantify their magnitude and likelihood, and prioritize appropriate resources to mitigate these risks before they arise in public settings. Today however, the research community lacks shared norms, best practices, and technical standards for how to safely and effectively red team AI systems.

In an effort to spur more communal discussion, we used our research to share a thorough description of our red team processes and results, as well as our dataset of ~40,000 red team attacks for others to analyze and learn from.[1] We hope this transparency accelerates the ability of developers to uncover harmful outputs of AI systems *prior* to deployment, and ultimately build safer systems. We look forward to partnering with interested stakeholders to build upon this work and develop common red teaming practices that can be widely adopted.

1 Anthropic. (2022, August). https://github.com/anthropics/hh-rlhf