Frontier Model Forum: What is Red Teaming?

Introduction

Red teaming is frequently cited as a common technique for developing safe and secure frontier models and AI systems. However, there is currently a lack of clarity on how to define "AI red teaming" and what approaches are considered part of the expanded role it plays in the AI development life cycle.

In cybersecurity, red teaming is a technique that emulates realistic attacks on systems to test for vulnerabilities and to understand likely adversary capabilities and goals. We define "red teaming" in the context of frontier models as a structured process for probing AI systems and products for the identification of harmful capabilities, outputs, or infrastructural threats. Similar to traditional "red teaming", AI red teaming often entails actively identifying flaws and vulnerabilities across the full system - including data, infrastructure, applications - not just model outputs alone. It is an important tool in advancing safe, secure and trustworthy AI, helping teams identify the potential risks of a system so that safeguards can be applied. It is also an iterative process, using results and insights from the exercises to inform priorities and approaches for at-scale measurement of risks, implementation of mitigations, and then re-running evaluations to determine the effectiveness of mitigations. To assist in understanding the range of techniques currently used under the broader umbrella of AI red teaming, the remaining sections will outline a set of case studies of known exercises conducted by FMF members

Microsoft Case Study: Red teaming Bing Chat

Background

In October 2022, Microsoft became aware of the new GPT-4 model from OpenAI. Having known that teams within Microsoft would be interested in integrating GPT-4 in first party and third party products, Microsoft created a cross functional team of subject matter experts (SMEs) to experiment with the model and understand its capabilities as well as risks. A follow up red teaming exercise was established for identifying risks with Bing chat, which was the first Microsoft product to integrate GPT-4.

Methodology

The initial round of red teaming the GPT-4 base model was done on the raw model with no additional mitigations from Microsoft. More than 20 SMEs from across the company with diverse expertise from law, policy, AI, engineering and research, security, and responsible AI came together and probed the model to identify its risks. In addition to the team directly experimenting with the model, the group granted access to the model to a small group of SMEs in national security and conducted interviews with them to better understand the risk surface in the specific high-stakes domain. The red teaming exercise on the GPT-4 raw model was more open ended and exploratory, meaning the goal was to identify as many risks and failure modes as possible, identify risk areas for further investigation, and implement early mitigation strategies.

The second round of red teaming was initiated as Bing chat was being developed and becoming mature. In this round of red teaming, more than 50 SMEs from across the company came together to red team the application with mitigations integrated in. We took an iterative approach with weekly red teaming sprints, where each week we red teamed priority features and risk areas, documented results, and worked with relevant measurement and mitigation teams to make sure the red teaming results informed their next steps. While this round of red teaming was more targeted given there was an established list of risks from red teaming the base GPT-4 model, the team still identified a series of new risks, which then led to reprioritizing the new risks for further investigation.

Outcome

Red teaming Bing chat highlighted the need for testing both base model and downstream applications iteratively. Red teaming the base model in an open-ended way and with no additional mitigations allowed us to understand the model's risk surface and failure modes, identify areas for further investigation, and to watch out for in application developments. Given many risks are context and application dependent, red teaming downstream applications with mitigations integrated in is necessary. While the weekly exploratory and qualitative red teaming sprints were not a replacement for at-scale measurement and mitigation work, the identified examples served as seeds to create at-scale measurement datasets, informed prioritization and implementation of mitigation strategies, and helped stakeholders make more informed decisions. Microsoft continues to learn and improve on the process we have established for red teaming.

OpenAI Case Study: Expert Red Teaming for GPT-4

Background

In August 2022, OpenAI began recruiting external experts to red team and provide feedback on GPT-4. Red teaming has been applied to language models in various ways: to reduce harmful outputs and to leverage external expertise for domain-specific adversarial testing. OpenAI's approach is to red team iteratively, starting with an initial hypothesis of which areas may be the highest risk, testing these areas, and adjusting as required. It is also iterative in the sense that OpenAI uses multiple rounds of red teaming as new layers of mitigation and control are incorporated.

Methodology

OpenAI recruited 41 researchers and industry professionals - primarily with expertise in fairness, alignment research, industry trust and safety, dis/misinformation, chemistry, biorisk, cybersecurity, nuclear risks, economics, human-computer interaction, law, education, and healthcare - to help gain a more robust understanding of the GPT-4 model and potential deployment risks. OpenAI selected these areas based on a number of factors including but not limited to: prior observed risks in language models and AI systems and domains where OpenAI has observed increased user interest in the application of language models. Participants in this red team process were chosen based on prior research or experience in these risk areas These experts had access to early versions of GPT-4 and to the model with in-development mitigations. This allowed for testing of both the model and system level mitigations as they were developed and refined.

Outcomes

The external red teaming exercise identified initial risks that motivated safety research and further iterative testing in key areas. OpenAI reduced risk in many of the identified areas with a combination of technical mitigations, and policy and enforcement levers; however, some risks still remain. While this early qualitative red teaming exercise was very useful for gaining insights into complex, novel models like GPT-4, it is not a comprehensive evaluation of all possible risks, and OpenAI continues to learn more about these and other categories of risk over time. The results of the red teaming process were summarized and published in the GPT-4 System Card.

Google DeepMind Case Study: Adversarial Probing of Google DeepMind's Gopher Model

Background

As language models increase in capability, there has been increasing interest in developing methods for being able to discover potential harmful content or vulnerabilities at scale. Adversarial testing has emerged as one approach to "red teaming" where the aim is to discover harmful content or vulnerabilities in the model through a combination of automated or manual probing techniques. While manual techniques to adversarial testing can be effective, the results will vary based on the creativity of the prober and could lead to critical safety oversights in the assessment of a model. To complement existing manual approaches to adversarially testing an AI system, our research paper "Red Teaming Language Models with Language Models", introduces the potential of utilizing a "red-team" language model (LM) to generate a diverse test set to evaluate the target language model's responses.

Methodology

The probing focused on a "Dialogue-Prompted" variant of GDM's Gopher (DPG) language model, and utilized a three-stage approach for identifying test cases which produce model failures:

1. Generate test cases using a designated "red-team" LM which are confirmed by a score generated by an automated scoring model as likely to generate a harmful output

2. Using the "target" LM, generate an output for each selected test case

3. Use the automated scoring model to identify the test cases which led to a harmful output

Additional approaches were used to generate test cases using the red-team LM, such as zero-shot sampling and supervised learning on successful adversarial questions to arrive at the final test set. The final test set was used to red-team the DPG model for various harms including offensive content, data leakage, inappropriate contact info generation, and distributional bias against groups.

Outcomes

Overall, this methodology demonstrates how LMs can be leveraged to effectively and automatically find problematic behaviors in other LMs as part of a safety testing pipeline. The collective range of approaches generated approximately 500,000 conversation-starting questions that elicited offensive responses. The red LM questions performed similarly or better than human-written adversarial examples from prior work in terms of eliciting offensive responses. Different red teaming methods like few-shot tuning, supervised learning, and reinforcement learning were able to increase the difficulty of questions while maintaining diversity of the test set.

Anthropic Case Study: Frontier Threats Red Teaming for AI Safety

Background

Frontier threats red teaming requires investing significant effort to uncover ways that AI models could create or exacerbate national security risks. For a model to create or substantially increase risks in national security domains, it must generate precise and reliable information across a variety of tasks and subtasks. In other words, a single concerning-sounding model output in isolation is not sufficient to cause real-world harm.

Methodology

This process has three steps:

1. We work with domain experts to define (a) high-priority threat models that could be exacerbated by advances in AI capabilities, (b) barriers to doing harm that would meaningfully increase national security risks if overcome, and (c) diagnostic tasks that indicate whether that capability exists in a model.
2. Subject matter experts extensively probe the model to assess whether the system's capabilities create or exacerbate national security risks per the predefined threat models.
3. Insights from red teaming inform the development of repeatable quantitative evaluations and mitigations.

Over the course of six months in 2023, Anthropic spent more than 150 hours with top biosecurity experts who evaluated our model’s ability to help a bad actor seeking to cause harm via biological means. This adversarial testing enabled us to develop quantitative evaluations of model capabilities as well as appropriate mitigations. We taught the experts how to jailbreak our models, and they used a bespoke, secure interface without the trust and safety monitoring and enforcement tools that are active on our public deployments.

Outcomes

We discovered a few key concerns. The first is that current frontier models can sometimes produce sophisticated, accurate, useful, and detailed knowledge at an expert level. In most areas we studied, this does not happen frequently. In other areas, it does. However, we found indications that the models become more capable as they scale (i.e., get larger). We also believe that models gaining access to tools could advance their capabilities in biology. Taken together, we think that future generations of LLMs without appropriate mitigations could accelerate a bad actor's efforts to misuse biology relative to existing tools (e.g., search engines). If unmitigated, we worry that these kinds of risks are near-term, meaning that they may be actualized in the next two to three years.

Alongside technical mitigations, we are establishing a disclosure process by which labs and other stakeholders can report national security risks and mitigations to other relevant actors. Ultimately, we aim to standardize the process of red teaming AI systems, which is presently more art than science. A robust and repeatable process is critical to ensure that red teaming (a) accurately reflects model capabilities and (b) establishes a shared baseline on which different models can be meaningfully compared.