RAI Assessment Copilot: Automated identification and measurement of harm in generative AI-based systems

Introduction

Motivation. Red teaming forms the foundation of Microsoft's voluntary commitment to the White House to advance responsible AI innovation. Red teaming is the practice of testing systems to assess their vulnerabilities by intentionally trying to break them. For example, for LLM-based features, red-teaming is done by inputting adversarial prompts into the LLM to identify if its generations can cause harm (e.g., stereotyping) and to measure the extent of identified harms. Red teaming of generative AI-based systems is currently a manual process. For instance, red teamers are expected to envision problematic prompts, input each into the system, observe the output, and record every observation. This process is challenging for the red teamer (e.g., it is hard to think of problematic prompts in unique contexts of use) and difficult to scale (e.g., data from one use case doesn't easily translate into new use cases). This makes the red teaming of generative AI-based systems a slow and tedious process. As Microsoft invests further in the generative AI space, we must work to make red teaming more robust and effective.

Main contribution. To address such issues with red teaming, our interdisciplinary team has built a technical framework called the ***RAI Assessment Copilot***. RAI Assessment Copilot is a pipeline of tools that leverage the power of GPT-4 to help automate *and* augment the identification and measurement of harm in generative AI-based systems. The RAI Assessment Copilot consists of a set of four tools:

1. **Cross Scenario Adapter**
   This tool helps red teamers re-use datasets across different red teaming efforts. It does this by helping transform a dataset (e.g., one created by red teaming Bing Chat) into another that can be used to red team a new use case (e.g., Copilot in Word).
2. **Prompt Generator**
   This tool helps red teamers envision problematic prompts. Given an existing red teaming dataset and an example prompt provided by the red teamer, this tool finds prompts most similar to the example prompt in the dataset and then generates new prompts similar to those it finds.
3. **In-scenario Variation**
   This tool helps red teamers generate diverse prompts. Given an example prompt, this tool first provides a list of axes of variation along which the prompt can be varied (e.g., gender, profession) and then, based on the axes selected, generates a new set of diverse prompts.

4. **RAI Extension**
   This tool helps red teamers automate major parts of the red teaming workflow. This tool automates the process of inputting prompts, capturing the output, and recording observations when red teaming an AI system with a web-based UI.

These tools seamlessly work together as part of the RAI Assessment Copilot, providing a single pipeline to automate and augment core parts of the red teaming practice. The RAI Assessment Copilot provides a robust and productive path for red teamers that can help practitioners quickly and easily perform RAI assessments such as those required by the Deployment Safety Board (DSB). In this talk, we will showcase these tools, explain what issues they address, and how they work together.

Related Work

The practice of red teaming AI systems, especially generative AI-based systems, is rapidly gaining traction in research (Ganguli et al. 2022; Perez et al. Rando et al. 2022) and in industry and government policy (Smith 2023; The White House 2023). The red teaming guide published internally on HITS and presented in a previous MLADS, with substantial contributions from our team members, is considered the best practice at the company.

This proposal is also related to another proposal submitted on a detailed introduction to red teaming large language models. The title is *Red Teaming LLMs and Their Applications: What It Is, Why It Is Important, and How to Do It.*

Methodology

Our design and development of the RAI Assessment Copilot is based on the extensive experience our team has with red teaming. Our team has been part of and led several red teaming ops at Microsoft that is required as part of the DSB process. We have done red teaming for both foundation models (e.g., GPT-4, DALL-E) and generative AI applications (e.g., new Bing, M365 Copilot). We have also built tools to aid red teaming of generative AI-based features and products (e.g., A and B), and have recently started conducting user research to understand user needs and pain points in red teaming work.

In this talk, we will dive deep into the following areas:

- **A very brief introduction to red teaming**
  - What is red teaming and why is it important?

- **Introducing the limitations of the current red teaming practice**
  - How is red teaming currently done for generative AI-based features?
  - What are some of the key challenges with the current red teaming practice?

- **Introducing the RAI Assessment Copilot and showing how it works**
  - What is the RAI Assessment Copilot?
  - How does the RAI Assessment Copilot work and what tools make up its pipeline?
  - Introduce and demo each tool in the RAI Assessment Copilot using a concrete example and use case, showing how they work together as a whole.

Responsible AI considerations

Our work aims to improve the process of red-teaming generative AI-based systems by automating and augmenting the identification and measurement of harms. Traditional red-teaming process is manual, slow, and tedious, which makes it difficult for practitioners to robustly test new AI-based systems. As Microsoft invests further in the generative AI space, it is important to build better frameworks—such as ours—to aid the effective red-teaming of new generative-AI based systems.

Conclusion

At the end of this session, the audience will be able to:

- At a high level, explain what red-teaming is and understand how it is currently done for generative AI-based features and products.
- Understand specific limitations of the current red-teaming process from a user perspective.
- Learn about the RAI Assessment Copilot, its four tools, and the user issues they address.
- Understand how to use the RAI Assessment Copilot to help automate and augment the identification and measurement of harm as part of red teaming efforts and DSB requirement.

References

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, et al. 2022. Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned. *arXiv: 2209.07858.*

White House. 2023. FACT SHEET: Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI. *The White House Press Briefing.*

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, Geoffrey Irving. 2022. Red Teaming Language Models with Language Models. *arXiv: 2202.03286.*

Javier Rando, Daniel Paleka, David Lindner, Lennart Heim, Florian Tramèr. 2022. Red-Teaming the Stable Diffusion Safety Filter. *arXiv: 2210.04610.*

Brad Smith. 2023. Our commitments to advance safe, secure, and trustworthy AI. *Microsoft Blog Post.*