Understanding Red Teams in AI

Exploring the role of red teaming in ethical and secure AI systems

What is a Red Team?

- A red team tries to break or exploit a system to find weaknesses.
- Originally used in cybersecurity and military strategy.
- Now applied to AI to uncover safety risks, bias, and vulnerabilities.

Red Teaming in AI

- Craft adversarial prompts to trigger unsafe or biased outputs.
- Test AI models' refusal mechanisms and safety filters.
- Identify blind spots and improve model behavior before deployment.

Why Red Teaming Matters

- Prevent misuse or harmful outcomes.
- Strengthen trust, transparency, and robustness.
- Aligns AI with ethical and responsible use principles.

Red Teaming with Azure OpenAI

- Try to elicit:
  - Discriminatory or biased outputs
  - Malicious code
  - Unsafe advice
- Test how Azure OpenAI models respond and flag such content.
- Use findings to improve filters and prompts.