

Developing Generative AI Applications Responsibly

Summary of the Webinar on April 5, 2024

Expert & Session Moderator



Philippe Béraud

Chief Technology & Security Advisor
Responsible AI Lead

Preamble

This summary was generated from the text transcription of the Webinar using ChatGPT 4, formatted by the Positive AI team and validated by the host Philippe Béraud.

Introduction

The development of generative AI applications requires careful attention to risk management and the potential impact on users and society. Philippe Béraud from Microsoft France emphasizes the importance of responsible AI, presenting the governance, strategies, practices and tools in place to achieve this goal.

Advances and Opportunities of Generative AI

Generative AI, capable of producing content, code, and summarizing information, offers advanced capabilities and opens up significant innovation opportunities. However, it also comes with increasing challenges in terms of responsibility and ethics, particularly in managing risks and preventing potential harms.

Issues and Challenges

Energy Consumption: Training and use of AI models can be energy-intensive, posing a challenge for sustainability, and requires the choice of a model and optimizations capable of meeting the expectations of the use case while being the most frugal.

Cybersecurity and Safety: AI systems can be vulnerable to adversarial attacks, and psychological & social engineering techniques, both requiring robust cybersecurity measures.

Human-Machine Interaction: It is crucial to design interactions where the user is aware they are interacting with an AI, to avoid anthropomorphism issues, and enable them to be in control. To achieve this, interactions must help them to use the system, to better understand its capabilities and limitations, to avoid overreliance issues.

Microsoft's Approach to Responsible AI

Governance Framework: Microsoft has established a governance structure to ensure adherence to its Responsible AI principles throughout all stages of AI system lifecycle.

Norms and Standards: Setting internal policies and standards (goals, requirements, and practices) enable to ensure a strict compliance with principles of an ethical, safe, secure and trustworthy AI.

Training and Acculturation: Emphasis is placed on training and sensitizing development teams to integrate responsible practices from the initial stages of design.

Risk Control and Mitigation Measures

Implement a mitigation plan with four layers of technical mitigation in terms of defense-in-depth with measurement and evaluation capabilities:

Model Selection: Choosing models suited to use cases and integrating techniques like reinforcement learning from human feedback (RLHF).

Safety Systems: Using moderation and filtering systems to detect and block (direct and indirect) prompt injection attacks, harmful content and manipulation attempts.

System Message and Grounding: Defining/using secure system message templates, and ensuring that model completions are based on accurate and relevant contextual data.

User experience: Using user-centered design to avoid misuse and overreliance from users.

Tools and Resources

Azure AI Content Safety: An intelligent content filtering and moderation system to detect direct (jailbreak) and indirect (XPIA) prompt injection attacks, and harmful content.

Prompt Flow: A suite of tools to build and deploy LLM-based flows, from prototyping to production, allowing rigorous AI-generated evaluation of prompt variants.

Python Risk Identification Tool for generative AI (PyRIT): An automated framework used to test and measure the prevalence of risks in completions generated by models.

Conclusion

The responsible development of generative AI applications relies on a combination of ethical principles, control measures, and robust development practices. Microsoft is committed to integrating these elements into all its solutions to ensure that AI is used beneficially and is safe, secure, and trustworthy. Like security, it's a journey, not a destination, with continuous investments in research, technological innovation, and governance. In this respect, Microsoft has just published its first annual report on the transparency of responsible AI: <https://aka.ms/RAITransparencyReport2024>.