

Using Generative AI Responsibly

Introduction

This lesson will cover:

- Why you should prioritize Responsible AI when building Generative AI applications.
- Core principles of Responsible AI and how they relate to Generative AI.
- How to put these Responsible AI principles into practice through strategy and tooling.

Learning Goals

After completing this lesson you will know:

- The importance of Responsible AI when building Generative AI applications.
- When to think and apply the core principles of Responsible AI when building Generative AI applications.
- What tools and strategies are available to you to put the concept of Responsible AI into practice.

Responsible AI Principles

Throughout this course, we are focusing on building our startup and our AI education product. We'll use the principles of Responsible AI: Fairness, Inclusiveness, Reliability/Safety, Security & Privacy, Transparency and Accountability. With these principles, we will explore how they relate to our use of Generative AI in our products.

Hallucinations

Hallucinations are a term used to describe when an LLM produces content that is either completely nonsensical or something we know is factually wrong based on other sources of information.

For example, if we ask an LLM to generate a list of the top 10 largest countries in the world, and it responds with a list that includes a country that doesn't exist, this would be considered a hallucination.

With each iteration of any given LLM, we have seen performance improvements around minimizing hallucinations. Even with this improvement, we as application builders and users still need to remain aware of these limitations.

Harmful Content

We covered in the earlier section when an LLM produces incorrect or nonsensical responses. Another risk we need to be aware of is when a model responds with harmful content.

Harmful content can be defined as:

- Providing instructions or encouraging self-harm or harm to certain groups.
- Hateful or demeaning content.
- Guiding planning any type of attack or violent acts.
- Providing instructions on how to find illegal content or commit illegal acts.
- Displaying sexually explicit content.

Lack of Fairness

Fairness is defined as “ensuring that an AI system is free from bias and discrimination and that they treat everyone fairly and equally.” In the world of Generative AI, we want to ensure that exclusionary worldviews of marginalized groups are not reinforced by the model’s output.

These types of outputs are not only destructive to building positive product experiences for our users, but they also cause further societal harm. As application builders, we should always keep a wide and diverse user base in mind when building solutions with Generative AI.

How to Use Generative AI Responsibly

Now that we have identified the importance of Responsible Generative AI, let's look at 4 steps we can take to build our AI solutions responsibly:

- Operate
- Identify
- Measure
- Mitigate

Measure Potential Harms

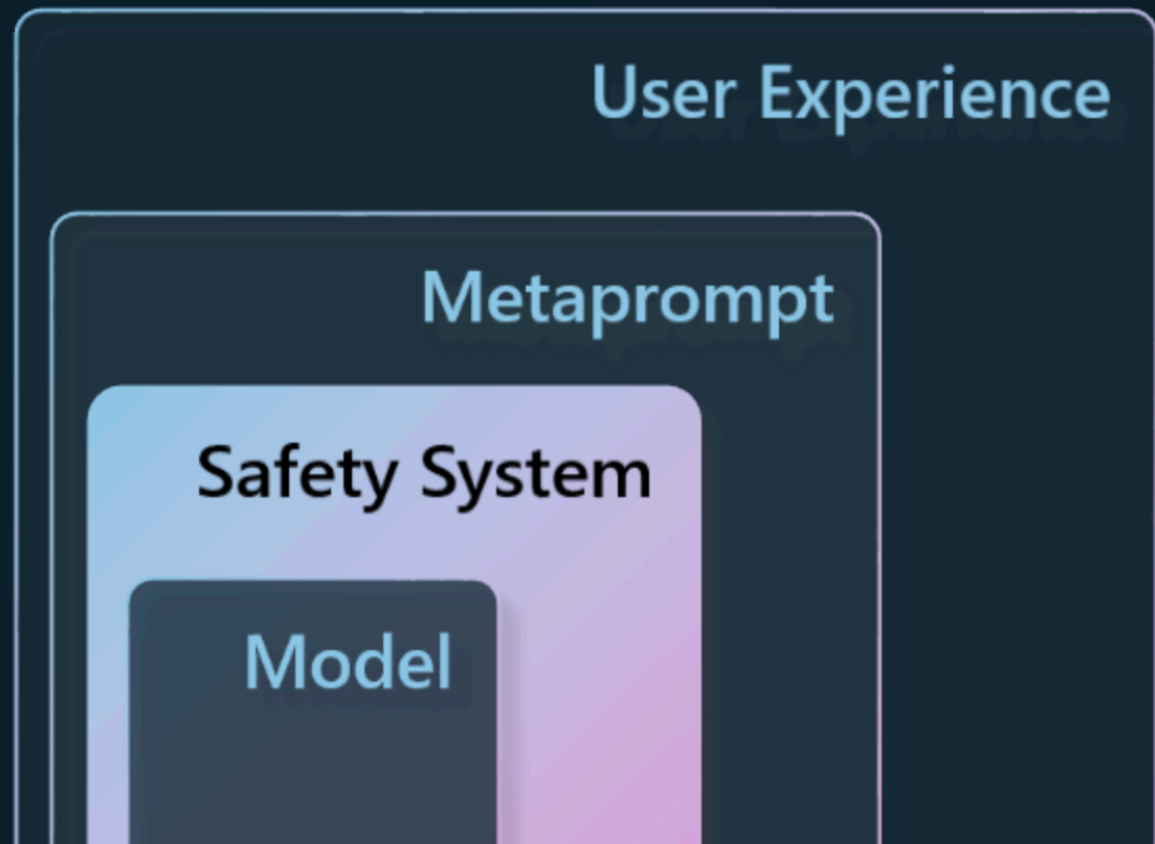
In software testing, we test the expected actions of a user on an application. Similarly, testing a diverse set of prompts users are most likely going to use is a good way to measure potential harm.

Since our startup is building an education product, it would be good to prepare a list of education-related prompts. This could be to cover a certain subject, historical facts, and prompts about student life.

Mitigate Potential Harms

It is now time to find ways where we can prevent or limit the potential harm caused by the model and its responses. We can look at this in 4 different layers:

Mitigation Layers



Operate a Responsible Generative AI solution

Building an operational practice around your AI applications is the final stage. This includes partnering with other parts of our startup like Legal and Security to ensure we are compliant with all regulatory policies. Before launching, we also want to build plans around delivery, handling incidents, and rollback to prevent any harm to our users from growing.