LLM Guardrails for Data Leakage, Prompt Injection, and More
Whether you're managing sensitive user data, avoiding harmful outputs, or ensuring adherence to regulatory standards, crafting the right LLM guardrails is essential for safe, scalable Large Language Model (LLM) applications.

READ

Safety best practices
Implement safety measures like moderation and human oversight.

READ

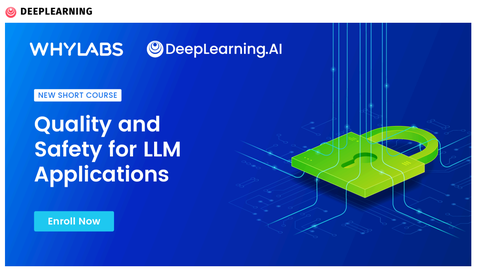
LLM safety real-world challenges and solutions

The following materials offers a deep, hands-on exploration of specific LLM safety real-world challenges and their practical solutions. You will engage with scenarios involving PII (Personally Identifiable Information) detection, strategies for keeping chatbot conversations on-topic, and techniques for managing sensitive mentions of competitors or restricted subjects.

QUALITY AND SAFETY FOR LLM APPLICATIONS | SAFE AND RELIABLE AI VIA GUARDRAILS

It's always crucial to address and monitor safety and quality concerns in your applications. Building LLM applications poses special challenges.

DEEPLEARNING



Quality and Safety for LLM Applications

Learn how to protect your LLM applications from critical security threats like hallucinations and data leakage.

READ MORE DEEPLEARNING >

Safety capabilities of Google Gemini model

Some foundational models incorporate built-in safety mechanisms directly into their models. You will learn about these controls and discover how developers can further leverage them to meet specific application requirements.

GOOGLE CLOUD

Safety and content filters | Generative AI on Vertex AI | Google Cloud

Google's generative AI models, like Gemini 2.5 Flash, are designed to prioritize safety. However, they can still generate harmful responses, especially when they're explicitly prompted. To further enhance safety and minimize misuse, you can configure content filters to block potentially harmful responses. This page describes each of the safety and content filter types and outlines key safety concepts.

READ MORE GOOGLE CLOUD >

Learners are encouraged to check safety capabilities of other foundational models on their own.

Specialized Llamma models

Explore specialised models explicitly designed to **safeguard LLM interactions** by detecting harmful content. We will examine concrete examples, such as **Meta's Llama Guard**, to understand how these dedicated safety models function as an additional layer of defence, performing tasks like content classification and toxicity detection to enhance the overall security and ethical behaviour of LLM applications.

Llama Guard 4 Model Card

Llama Guard 4 is a natively multimodal safety classifier with 12 billion parameters trained jointly on text and multiple images.

HUGGINGFACE

meta-llama/Llama-Guard-4-12B · Hugging Face

We're on a journey to advance and democratize artificial intelligence through open source and open science.

READ MORE HUGGINGFACE >

Llama Prompt Guard 2 Model Card

We are launching two classifier models as part of the Llama Prompt Guard 2 series, an updated version of v1: Llama Prompt Guard 2 86M and a new, smaller version, Llama Prompt Guard 2 22M.

HUGGINGFACE

meta-llama/Llama-Prompt-Guard-2-86M · Hugging Face

We're on a journey to advance and democratize artificial intelligence through open source and open science.

READ MORE HUGGINGFACE >

Additional materials:

The Guard
The guard object is the main interface for Guardrails.

READ

NeMo-Guardrails
GitHub

READ

Module 6 - Basic observability

Next: Final Assessment

In Progress 67%

Prompt Engineering & AI Applica

- Instructions for learners
- Prompt Engineering & AI Applications
- Final Assessment

Course Tasks 0/1

- Course Evaluation