RIT

# LLM Guardrails

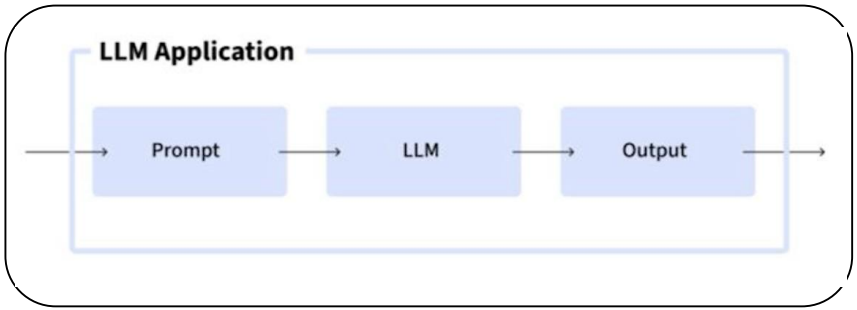
Reza Fayyazi

# What are guardrails?

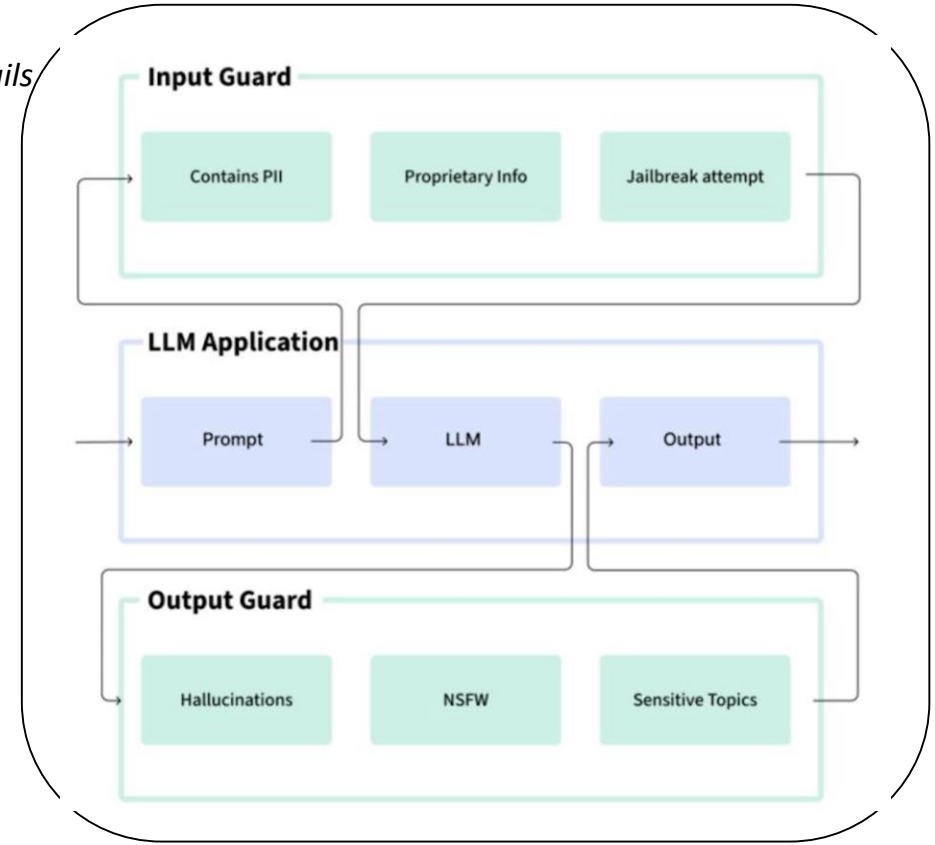
- A *guardrail* is a secondary check or validation around the input or output of an LLM model
- Input validation guardrails include:
  - Prevent the injection of code or adversarial prompts to manipulate the model
  - Check for phrasing in the input that may trigger biased or harmful responses
  - Check for Personal Identifiable Information (PII) leakage
- Output validation guardrails include:
  - Check for hallucination by cross-referencing against trusted sources
  - Check for PII leakage
  - Check for sensitive topics

# Where are guardrails applied in LLMs?

*Without Guardrails*

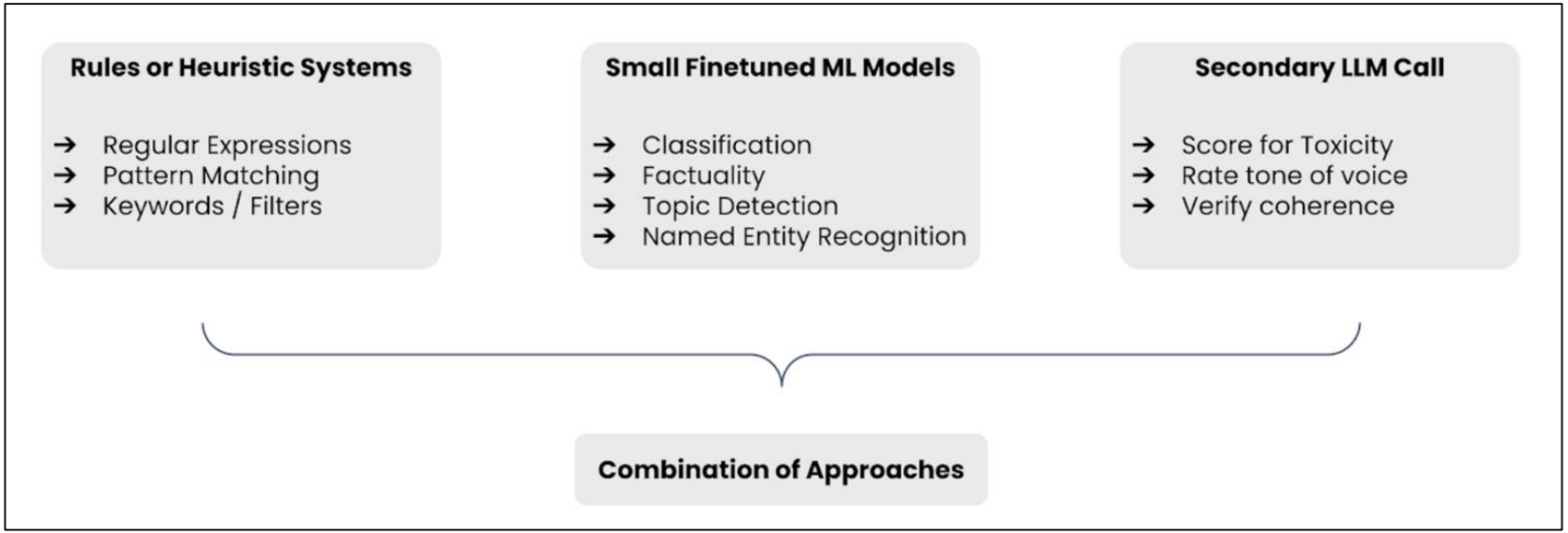



*With Guardrails*




[1] https://www.deeplearning.ai/short-courses/safe-and-reliable-ai-via-guardrails/

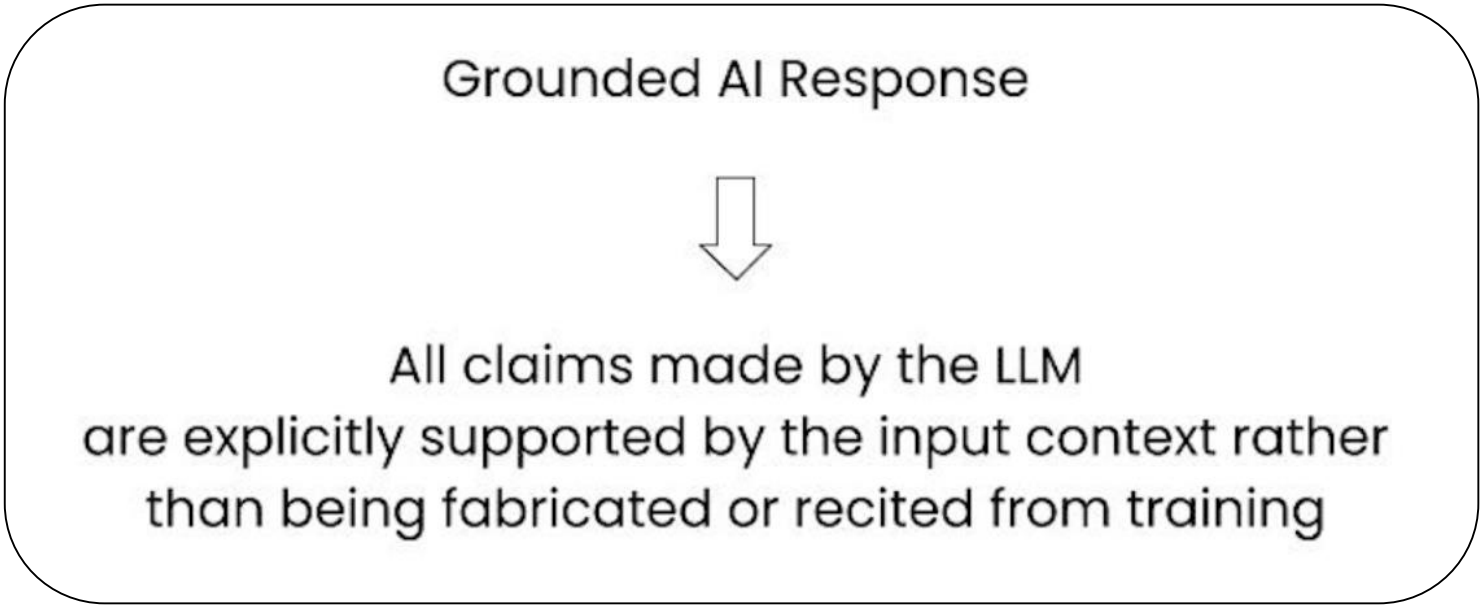
# How are guardrails implemented?






[1] https://www.deeplearning.ai/short-courses/safe-and-reliable-ai-via-guardrails/

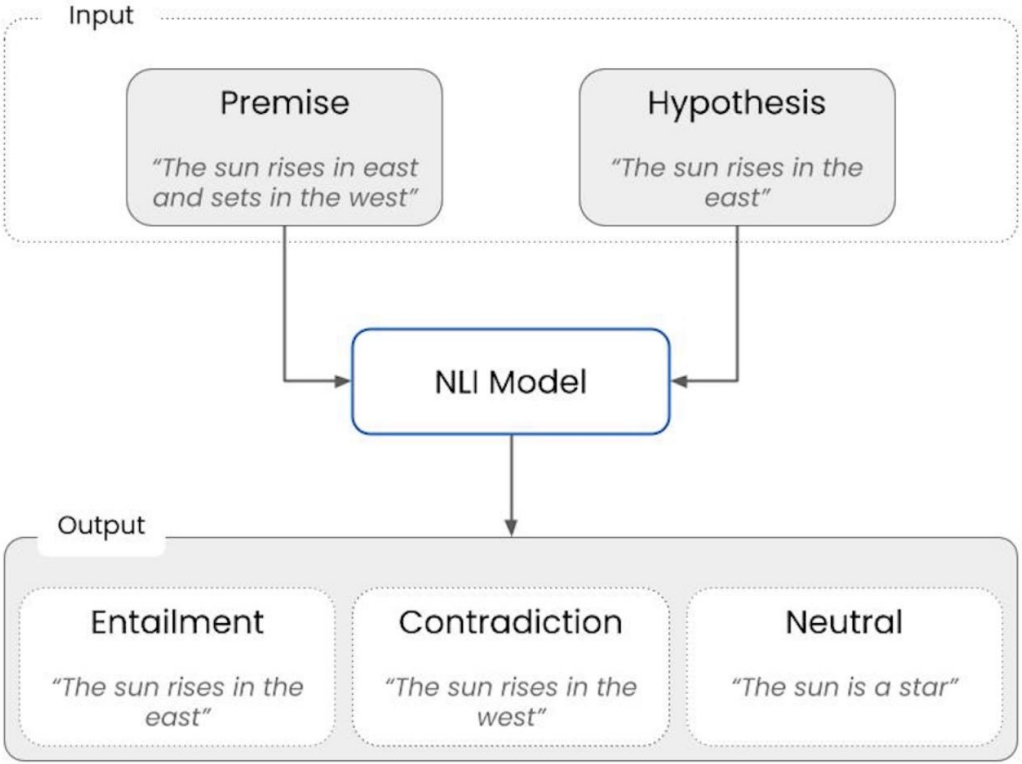
# Hallucinations are due to Lack of Groundness






[1] https://www.deeplearning.ai/short-courses/safe-and-reliable-ai-via-guardrails/

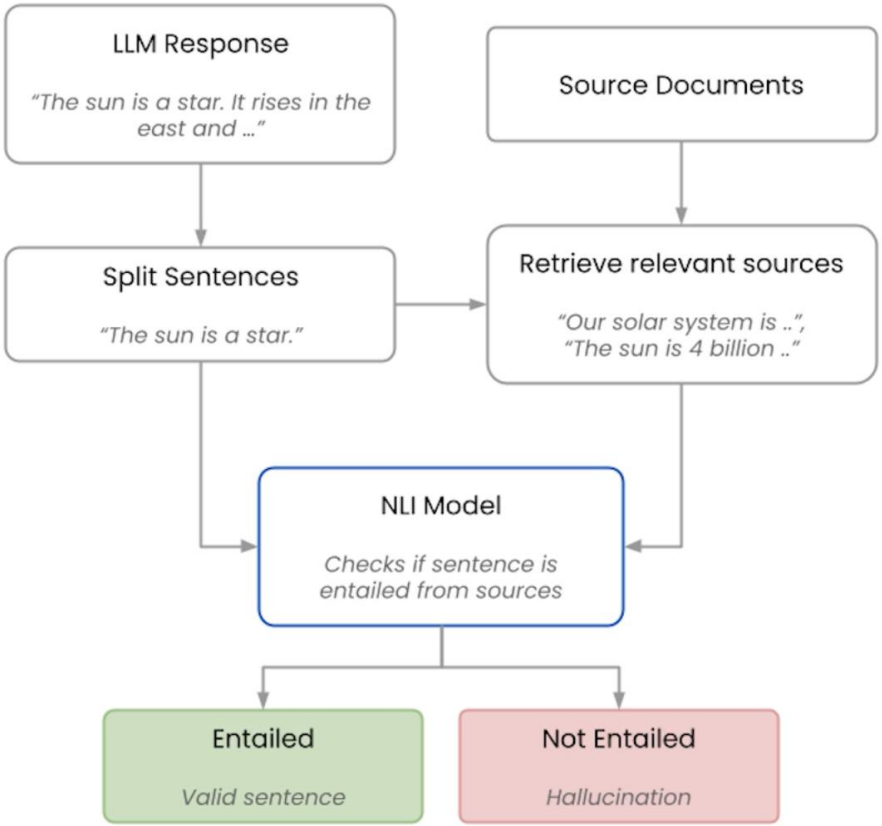
# Natural Language Inference (NLI)




[1] https://www.deeplearning.ai/short-courses/safe-and-reliable-ai-via-guardrails/

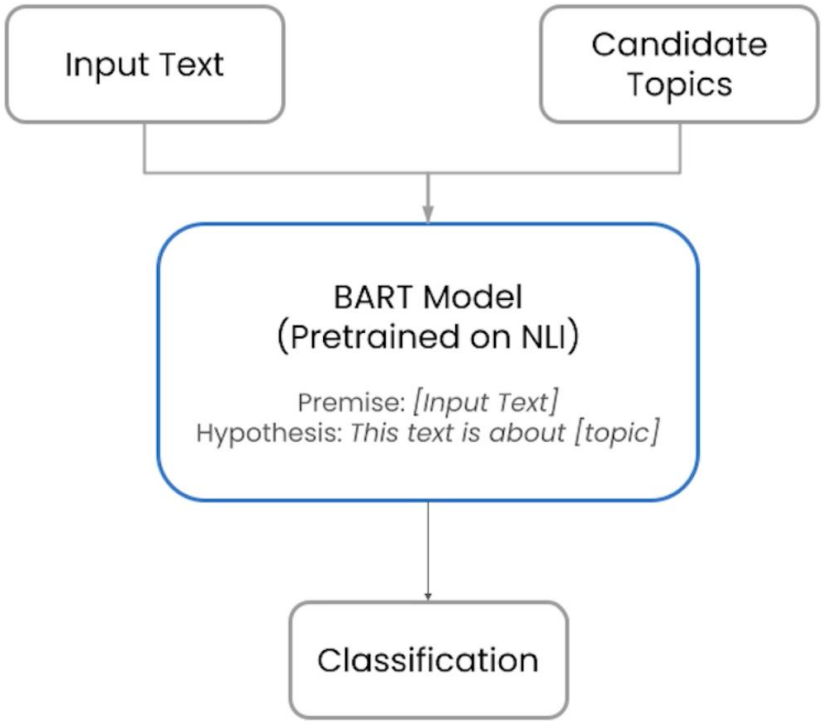
# Using NLI for Detecting Hallucinations

- Using cosine similarity to check the splitted sentences from the response with splitted sentences from the retrieved sources




[1] https://www.deeplearning.ai/short-courses/safe-and-reliable-ai-via-guardrails/

# Zero-Shot Topic Classification






[1] https://www.deeplearning.ai/short-courses/safe-and-reliable-ai-via-guardrails/

# What is PII?

*Personal Identifiable Information*

| | |
|---|---|
| Direct identifiers | Name, SSN, Email, ... |
| Indirect identifiers | Location, Demographics, ... |
| Sensitive data | Health records, Financial info, ... |

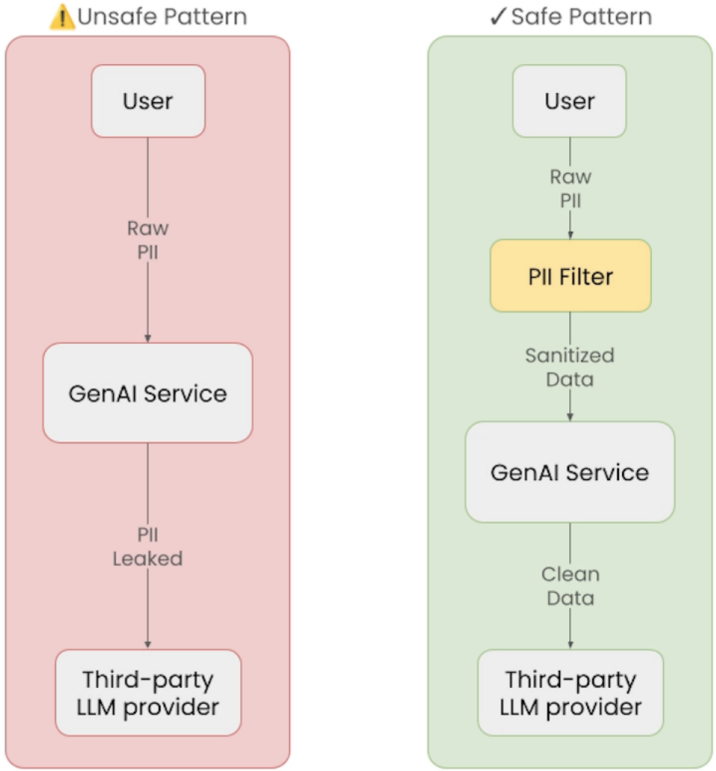
LLM Data Privacy Risks:

- ➔ Third-party processing exposure
- ➔ Potential data retention by providers
- ➔ Risk of training data contamination
- ➔ Limited control over data handling



[1] https://www.deeplearning.ai/short-courses/safe-and-reliable-ai-via-guardrails/

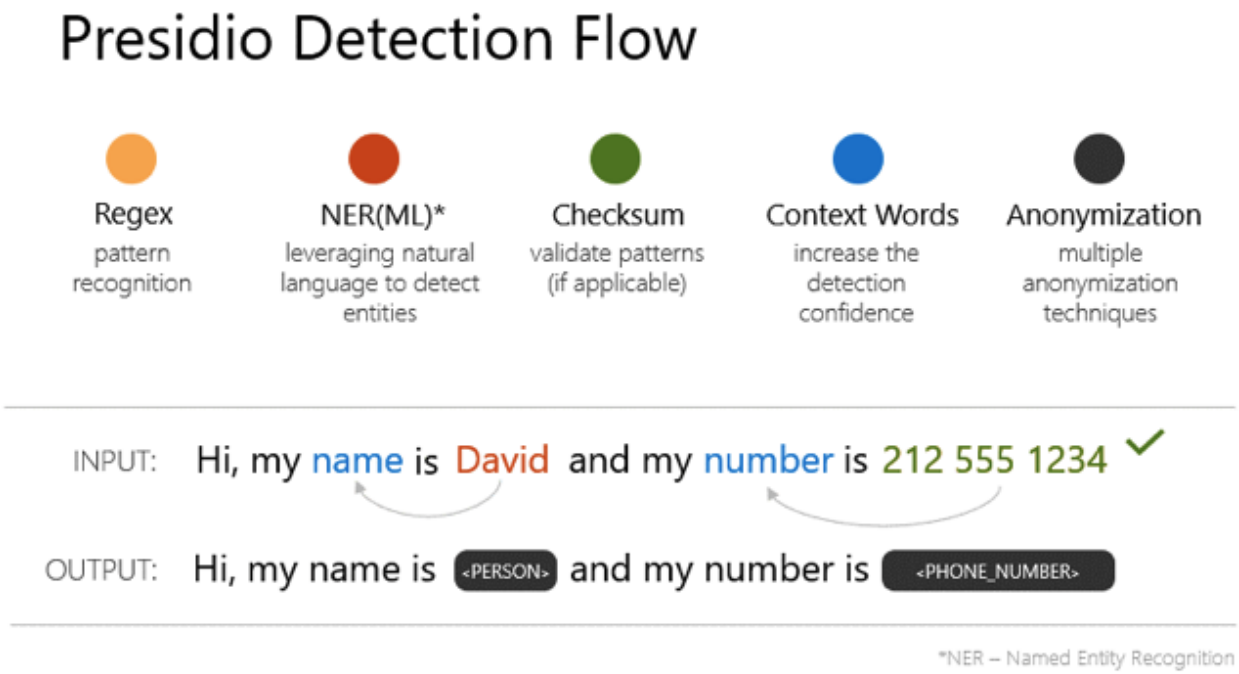
# PII Handling with LLMs





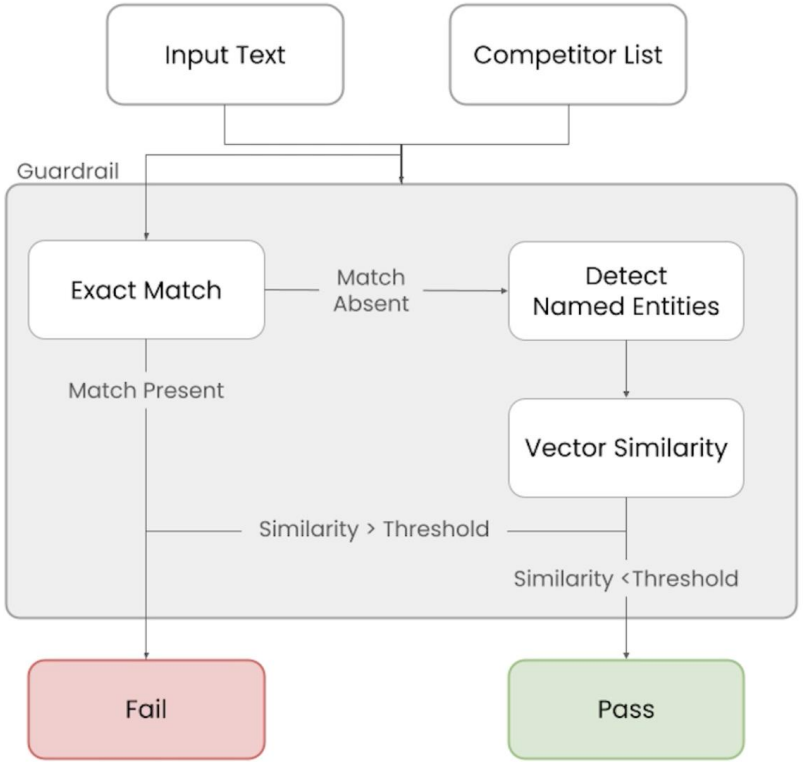

[1] https://www.deeplearning.ai/short-courses/safe-and-reliable-ai-via-guardrails/

# Microsoft Presidio




[2] https://microsoft.github.io/presidio/

# Sensitive Word Detection




[1] https://www.deeplearning.ai/short-courses/safe-and-reliable-ai-via-guardrails/

# Thank you!

**Reza Fayyazi**
**rf1679@rit.edu**