Project Report: AI Output Explainability Tool

Using Python

by Shrushti Wable

# Objective

The objective of this project is to create a lightweight tool that evaluates and explains the quality of responses generated by AI models using rule-based logic. This tool is useful for understanding AI behavior, spotting weaknesses like repetition or lack of keyword coverage, and improving overall response quality.

# Problem Statement

AI-generated content can often include hallucinations, irrelevant details, or incoherent phrasing. This project aims to evaluate AI responses by analyzing their sentiment, repetition, keyword coverage, and overall structure to provide interpretable feedback.

# Tools & Technologies Used

- Python

- textblob

- pandas

- re (regex)

- Jupyter Notebook

# How the Solution Works

1. User provides a prompt, AI response, and list of expected keywords.

2. The tool checks sentiment using TextBlob.

3. It detects overused words using repetition logic.

4. It calculates how many expected keywords are present in the response.

5. It counts the number of words for verbosity insight.

6. Outputs are shown as a table using pandas.

# Code Overview

- check\_sentiment(): Uses TextBlob to assign sentiment polarity (positive/negative/neutral).

- check\_repetition(): Identifies words repeated more than 3 times.

- check\_keyword\_coverage(): Measures keyword presence vs. expectation.

- check\_length(): Counts the total number of words.

- evaluate\_response(): A wrapper that combines all functions and returns a summary dictionary.

# Sample Output

Example:

Prompt: What are the benefits of machine learning in healthcare?

Response: Machine learning is useful in healthcare. Machine learning helps in diagnosis. Machine learning improves treatment. Machine learning is great. Machine learning is great.

Evaluation Summary:

- Sentiment: Positive

- Repetition: ['Machine', 'learning']

- Keyword Match: 2/3

- Word Count: 20

# Conclusion

This AI Output Explainability Tool offers a fast, understandable way to audit and improve AI-generated text. It can be extended to more complex models and evaluation metrics, making it a strong starting point for building trust and control around generative AI.