

Hallucination Detector

Purpose

The Hallucination Detector is a Streamlit-based application designed to **evaluate the reliability of language model outputs** by detecting potential hallucinations (inaccurate or fabricated responses). Built using the ``uqlm`` library and integrated with OpenRouter models, it provides a user-friendly tool to quantify uncertainty in AI-generated text, ensuring trustworthy results for critical applications.

Key Features

- Model Flexibility: Supports multiple OpenRouter models (e.g., DeepSeek, Gemini, LLaMA) for diverse use cases.
- **Detection Methods:**
 - **Black-Box:** Compares multiple responses for consistency.
 - **White-Box:** Analyzes token probabilities for confidence.
 - **LLM-as-a-Judge:** Uses secondary LLMs to evaluate outputs.
 - **Ensemble:** Combines methods for robust evaluation.
- User Interface: Configurable via sidebar for model, temperature, and scoring method; supports custom or predefined prompts.
- Visualization: Displays confidence scores (0–1) in a bar chart, with higher scores indicating lower hallucination risk.

Value

- Trustworthy AI: Ensures reliable outputs for decision-making in high-stakes scenarios (e.g., legal, medical, financial).
- Risk Mitigation: Identifies and flags potential inaccuracies, reducing errors in AI-driven processes.
- User Accessibility: Intuitive interface requires minimal technical expertise, enabling broad adoption.
- Scalable Insights: Configurable settings allow tailoring to specific industries or applications.

Technical Overview

- Platform: Python-based, using Streamlit for the UI, LangChain for OpenRouter integration, and `uqlm` for hallucination detection.
- Requirements: OpenRouter API key; internet connection for API calls.
- Performance: Asynchronous processing ensures efficient execution, though Ensemble Scorer is computationally intensive.
- Output: Provides LLM response, confidence scores, and raw data, with scores interpreted as:
 - >0.8 : Likely factual.
 - $0.5\text{--}0.8$: Possible inaccuracies.
 - ≤ 0.5 : Likely hallucinations.

Demo

The screenshot shows a web browser window with the address bar displaying various bookmarks. The main content area is titled "Hallucination Detector" and includes a sidebar on the left for configuration and a main panel for input and results.

API Key
Enter your OpenRouter API Key: [password field]

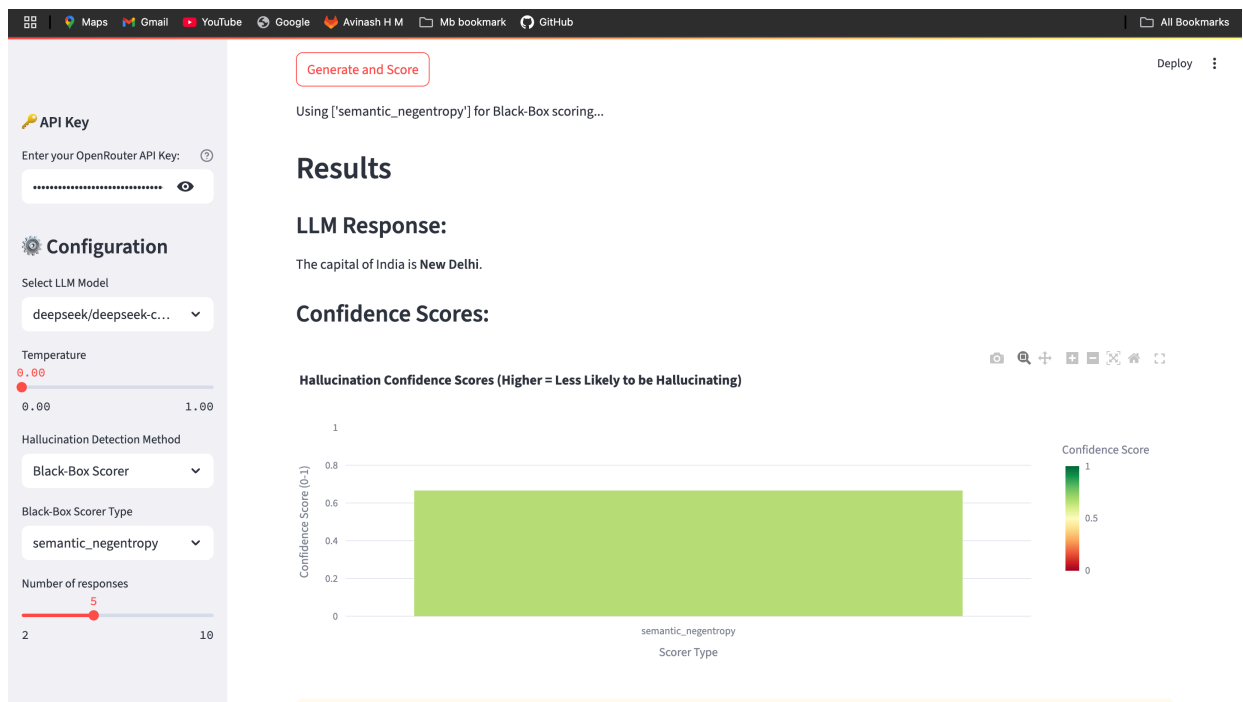
Configuration
Select LLM Model: deepseek/deepseek-c...
Temperature: 0.00 (slider from 0.00 to 1.00)
Hallucination Detection Method: Black-Box Scorer
Black-Box Scorer Type: semantic_negentropy
Number of responses: 5 (slider from 2 to 10)

Hallucination Detector
This app demonstrates hallucination detection using the `uqlm` library with OpenRouter models. Select options in the sidebar to configure the hallucination detection method.

Input
Select prompt type or create your own: Factual query
Prompt: What is the capital of India?
[Generate and Score button]

Generating response and calculating hallucination scores...
Using ['semantic_negentropy'] for Black-Box scoring...

Demonstration Tips



Conclusion

The Hallucination Detector strengthens trust in AI by quantifying output reliability, offering a critical tool for decision-makers.