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.

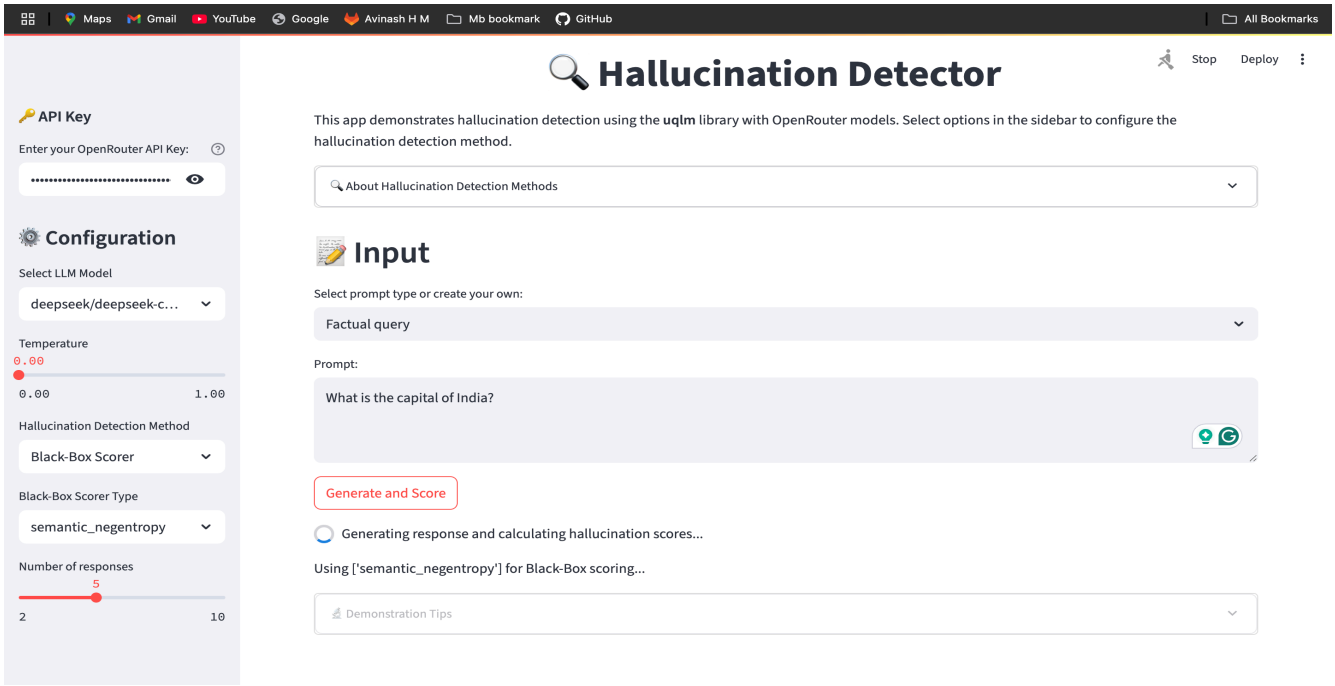
- Output: Provides LLM response, confidence scores, and raw data, with scores interpreted as:

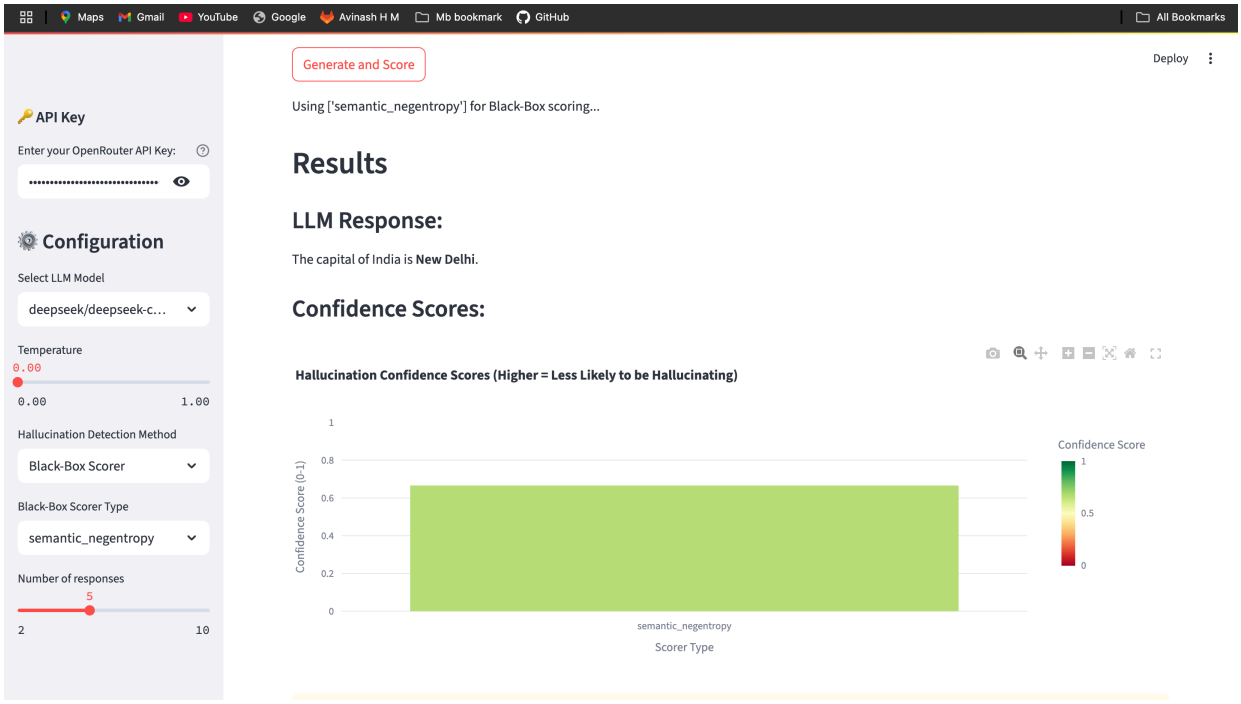
  - >0.8: Likely factual.

  - 0.5–0.8: Possible inaccuracies.

  - ≤0.5: Likely hallucinations.

Demo





Conclusion

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