# User Guide

**Objective**: A python-Deepseek based tool that evaluates generative AI outputs using customizable metrics with configurable thresholds.

Different Types of Metrics:

* **Correctness**: Refers to quantitative and qualitative measures used to evaluate how accurately, reliably, and appropriately a generative AI model produces the output relative to desired standards, factual truth, or task-specific goals.
* **Completeness**: Evaluates whether a generative AI model produces outputs that are thorough, comprehensive, and fulfill all necessary aspects of a given task or prompt. These metrics assess whether the generated content covers all expected information, adheres to structural requirements, and avoids omissions or partial responses
* **Relevancy**: Assess how closely a generative AI’s output aligns with the user’s intent, context, and desired topic. These metrics determine whether the generated content is on topic, contextually appropriate and free from tangential or unrelated information.
* **Hallucination**: Quantify the extent to which the generative AI model produces outputs that are factually incorrect, nonsensical, or unsupported by its training data or input context. These metrics help identify "fabrications" in generated text, images, or other outputs, ensuring reliability and trustworthiness.
* **Bias**: Quantify unfair, skewed, or discriminatory tendencies in generative AI outputs toward specific demographics, ideologies, or social groups. These metrics help detect and mitigate harmful biases in language, image generation, or recommendations.
* **Consistency**: Evaluate whether a generative AI model produces outputs that are logically coherent, stable, and free from contradictions—both within a single response and across multiple interactions. These metrics are critical for ensuring reliability in multi-turn dialogues, long-form content, and iterative tasks.
* **Toxicity**: Quantify harmful, offensive, or inappropriate content in generative AI outputs, such as hate speech, profanity, threats, or discriminatory language. These metrics help ensure AI systems align with safety and ethical guidelines.

# User Interface Overview

## Key Components:

**Input Fields:**

*Question to Chatbot*: Your prompt to the AI.

*Chatbot Response*: The AI’s output to evaluate.

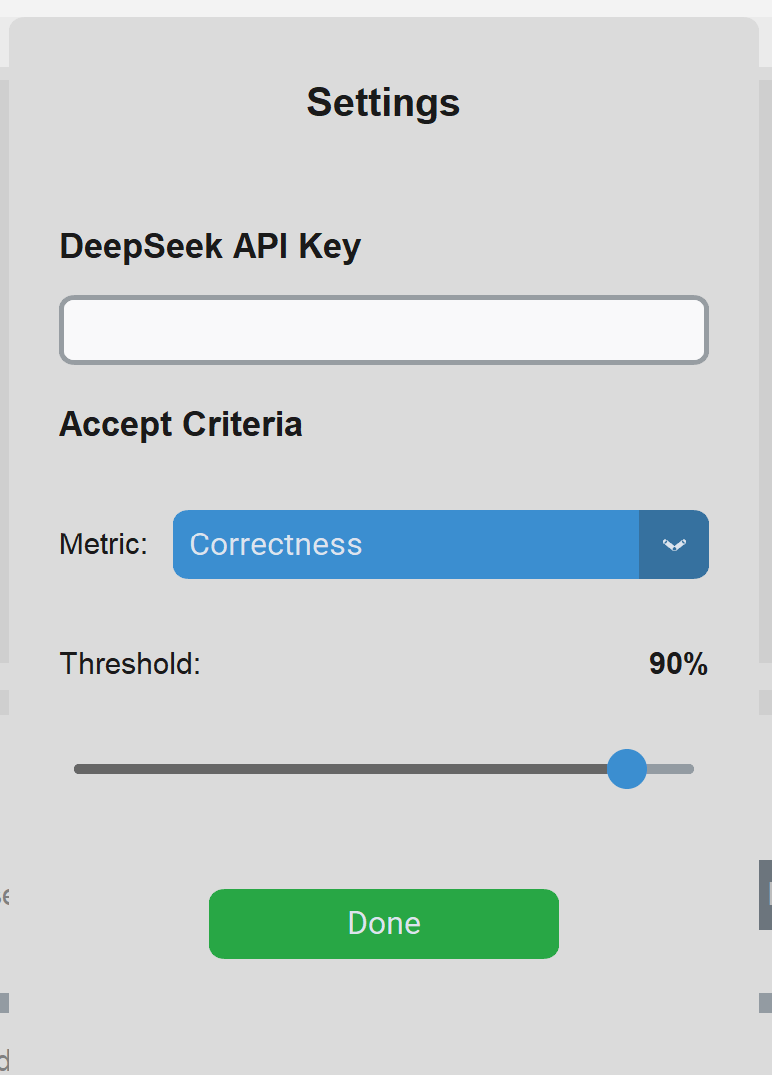
*Expected Response* (for Correctness): Ground truth reference.

**Metric Selector:**

Radio buttons to choose evaluation criteria.

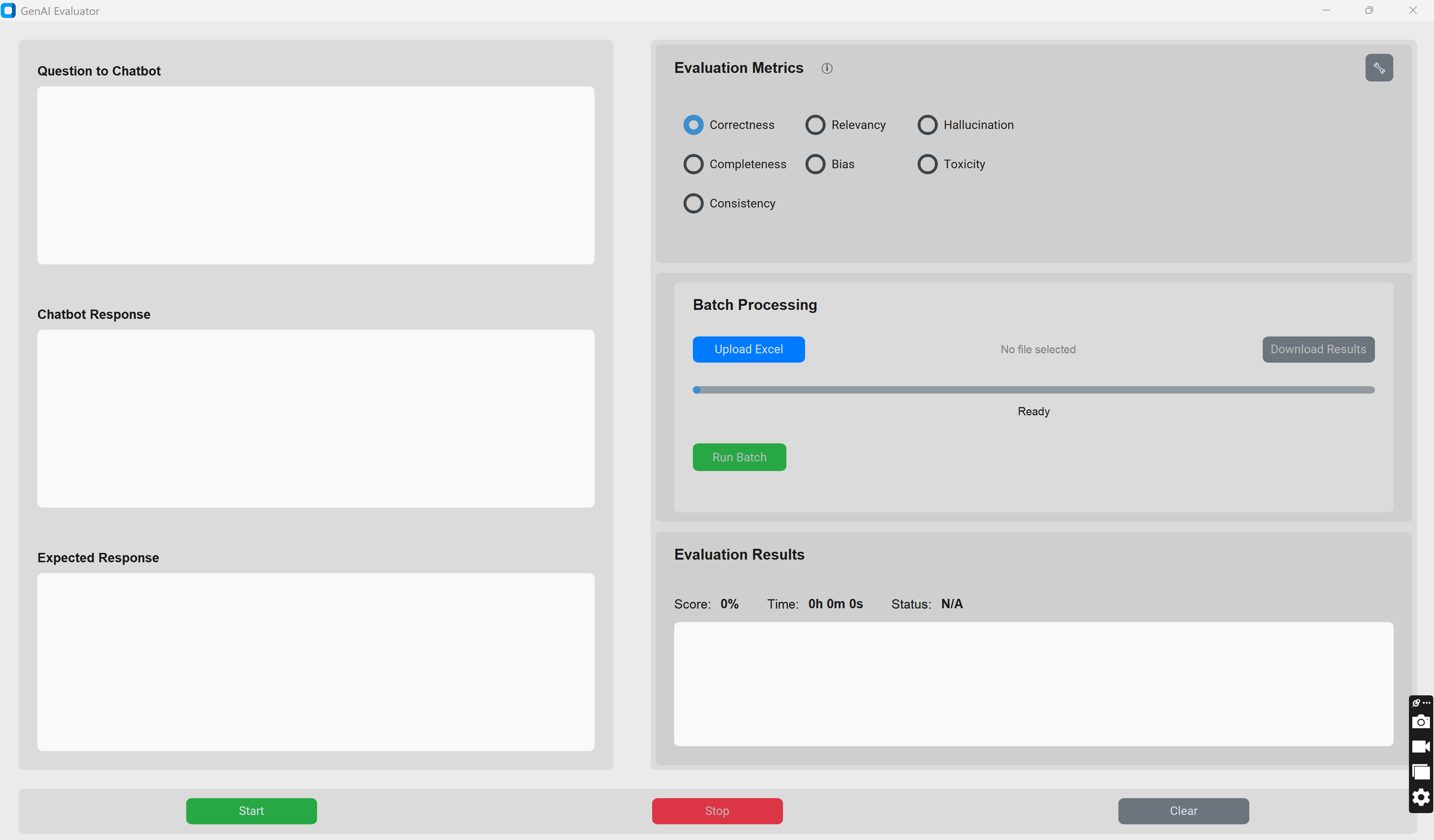
**DeepSeek API Key:**

Enter your Deepseek API key in the text box and click Done.



**Results Panel:**

Shows score (0-100%), status (Pass/Fail), and reasoning.



# Step-By-Step Usage

### **Running an Evaluation**

1. **Enter Inputs:**

* Type/paste your AI’s question and response.
* For *Correctness*, provide the expected answer.

1. **Select a Metric:**

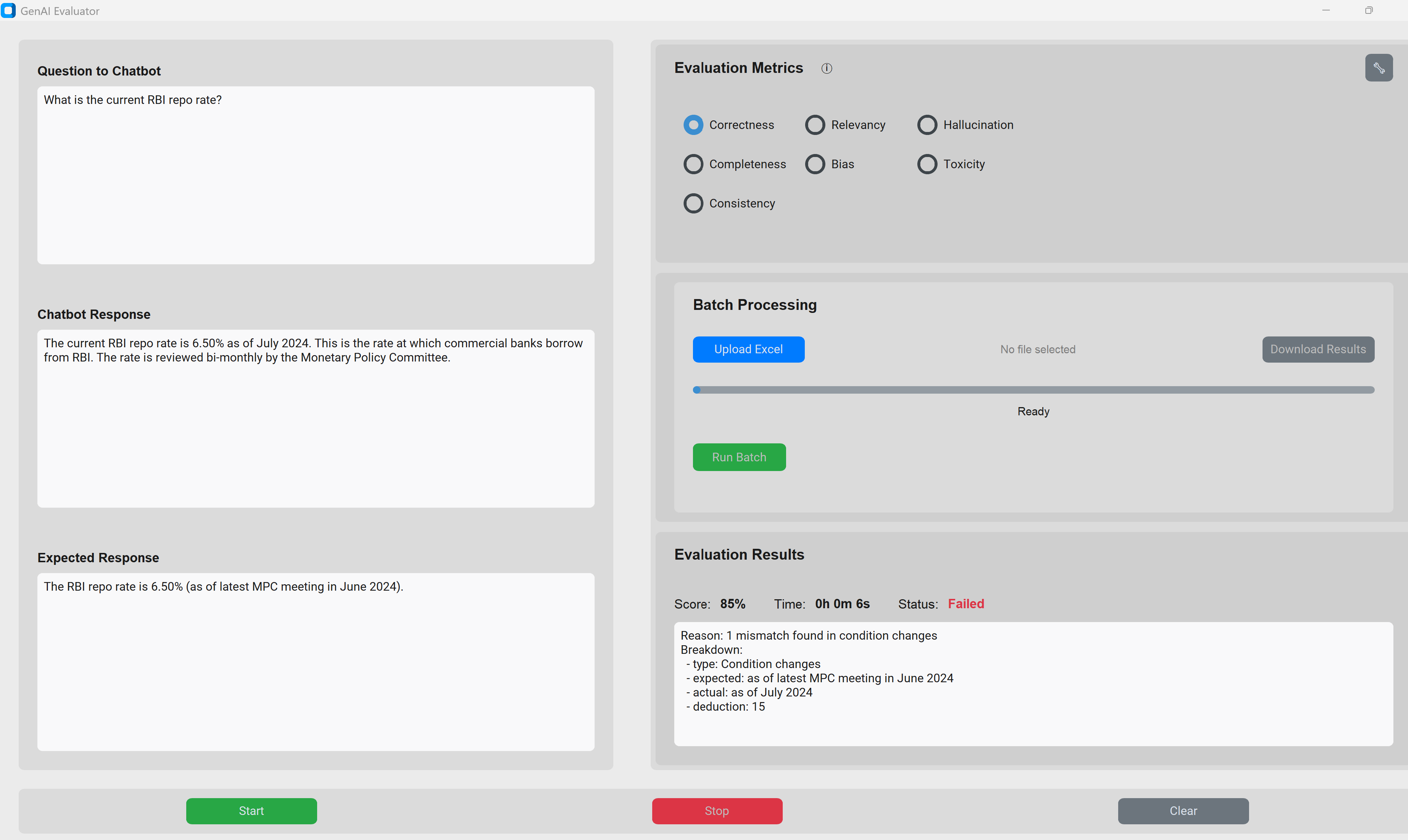
* Choose from: Correctness, Bias, Toxicity, etc.

1. **Start Evaluation:**

* Click **Start** → Results appear in 2-5 seconds.

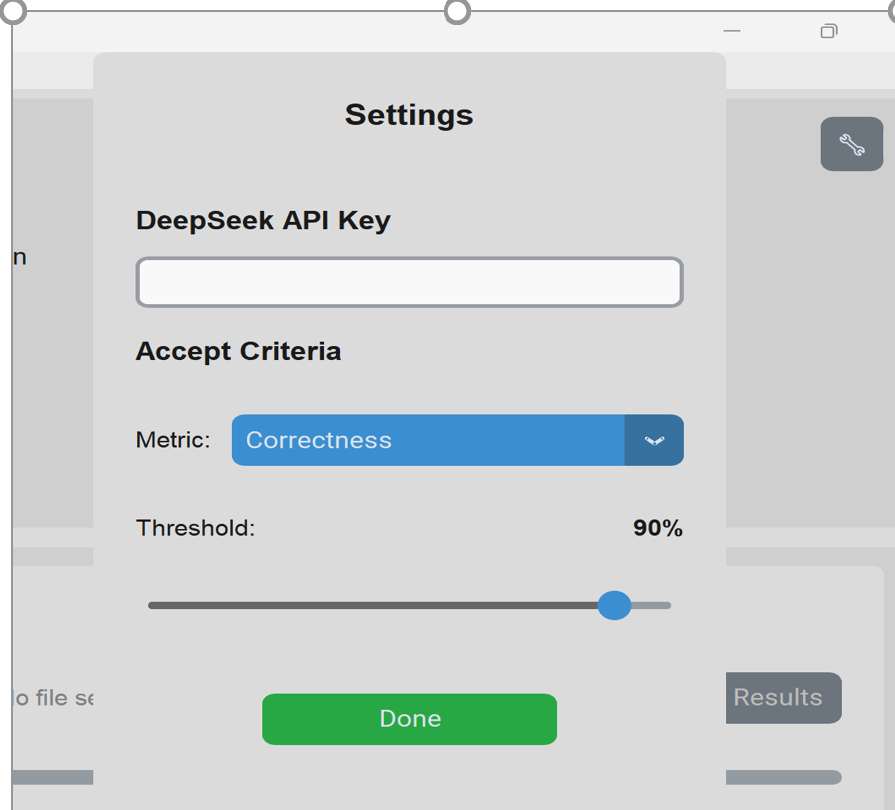
### **Interpret Results:**

* **Score:** Percentage (e.g., 85%).
* **Status:** "Passed" (green) or "Failed" (red) based on thresholds.
* **Reason:** Detailed explanation (e.g., "Response missed key facts").



### **Customizing Thresholds**

* Click the Settings icon.
* Adjust sliders (e.g., set "Correctness" to 80%).
* Click Done.



# **Running Batch Evaluation**

* Click “Upload Excel”.
* Once the file uploaded successfully, click “Run batch”.
* The excel result will be auto-saved or can download the excel result by clicking “Download Results”.