LLM Reliability Evaluation Engine

A Quantitative Framework for Evaluating Large Language Model Outputs

Sambit Karmakar

February 13, 2026

1 Introduction

Large Language Models (LLMs) often produce fluent outputs that may not always be reliable. This project implements a structured evaluation framework to quantitatively measure LLM performance using multiple reliability signals.

The objective is to move beyond subjective inspection and introduce measurable evaluation metrics for model benchmarking.

2 Project Objective

The system evaluates LLM responses across a benchmark dataset using:

- Semantic Alignment (Embedding Similarity)
- Factual Accuracy (LLM-as-Judge Hallucination Audit)
- Composite Final Reliability Score

The framework supports both prompt-level diagnostics and aggregate model-level benchmarking.

3 System Architecture

`Prompt` → `LLM (LLaMA 3.3 70B)` → `Evaluation Layer` → `Reliability Score` → `Visualization Dashboard`

4 Evaluation Methodology

4.1 Semantic Similarity

Cosine similarity is computed between:

- Model-generated response
- Reference ground-truth answer

Embeddings are generated using the `sentence-transformers` model (`all-MiniLM-L6-v2`).

4.2 Hallucination Detection (LLM-as-Judge)

A secondary auditing prompt evaluates factual correctness by:

- Checking factual validity
- Penalizing fabricated or unsupported claims
- Scoring responses on a 0–10 scale

Scores are normalized to the range [0, 1].

4.3 Final Reliability Score

For QA benchmarking mode, the composite reliability score is defined as:

$$\text{Final Reliability} = 0.5 \times \text{Similarity} + 0.5 \times \text{Hallucination}$$

Structure validation was excluded since the dataset consisted of general QA prompts rather than structured JSON outputs.

5 Benchmark Dataset

The evaluation dataset consists of 15 diverse prompts covering:

- Factual knowledge
- Mathematics
- Definitions
- Conceptual reasoning
- Creative responses

6 Results

6.1 Overall Average Performance

Metric	Average Score
Semantic Similarity	0.5807
Hallucination (Accuracy)	0.9133
Final Reliability Score	0.7470

Table 1: Overall LLM Performance Across Benchmark Prompts

6.2 Observations

- The model demonstrates high factual consistency (Hallucination Score ≈ 0.91).
- Semantic similarity varies due to phrasing differences.
- Mathematical and direct factual queries perform strongly.
- Creative prompts show lower alignment consistency.

7 Visualization Dashboard

The system provides:

- Final Reliability Score per Prompt
- Metric Breakdown (Similarity vs Hallucination)
- Aggregate Model Performance Visualization

These dashboards enable rapid diagnostic analysis of LLM behavior.

8 Technologies Used

- Groq API (LLaMA 3.3 70B)
- Sentence Transformers
- Pydantic (Schema Validation)
- Scikit-learn
- Pandas
- Matplotlib / Seaborn

9 Key Contributions

- Modular evaluation pipeline
- Structured reliability scoring framework
- LLM-based hallucination detection system
- Prompt benchmarking capability
- Visual performance diagnostics

10 Limitations

- Hallucination detection is heuristic (LLM-as-Judge approach)
- No retrieval-grounded fact verification
- Embedding similarity dependent on model quality
- Small benchmark dataset

11 Future Improvements

- Retrieval-Augmented Verification (RAG-based fact checking)
- Larger human-labeled benchmark dataset
- Confidence calibration analysis
- Automated regression testing
- Model comparison benchmarking

12 Conclusion

This project demonstrates a principled framework for evaluating LLM reliability using quantitative metrics rather than subjective inspection.

By combining semantic alignment and factual auditing into a composite reliability score, the system provides a structured approach to benchmarking and validating large language model outputs.

The framework is modular, extensible, and suitable for integration into AI system validation workflows.

Repository: https://github.com/sambitkarmakar03/llm-reliability-evaluation-engine.git