e6data x IIT-BHU

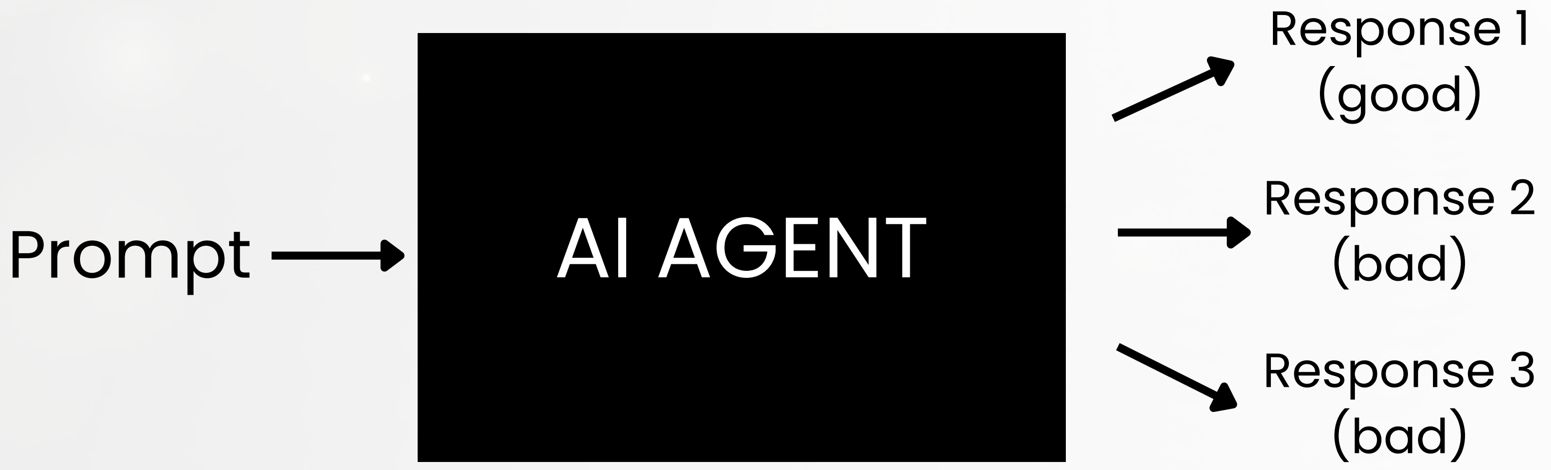
LLM - Evalify

(Agentic Evaluation Framework)

Team : data_dawgs
Members: Gaurish Maheshwari, Rounak Somani
App Link: llm-evalify.vercel.app

UNDERSTANDING THE **PROBLEM**

AGENTIC EVALUATION FRAMEWORK



1

Why Evaluation Matters

As AI Agents become more capable, rigorous evaluation is essential to ensure they behave reliably.

2

The Scale Problem

With Hundreds of Agents generating thousands of responses, manual evaluation becomes infeasible.

3

Lack of Scalable, Multi-Dimensional Metrics

There’s critical need for an automated scoring framework that can evaluate responses across multiple dimensions

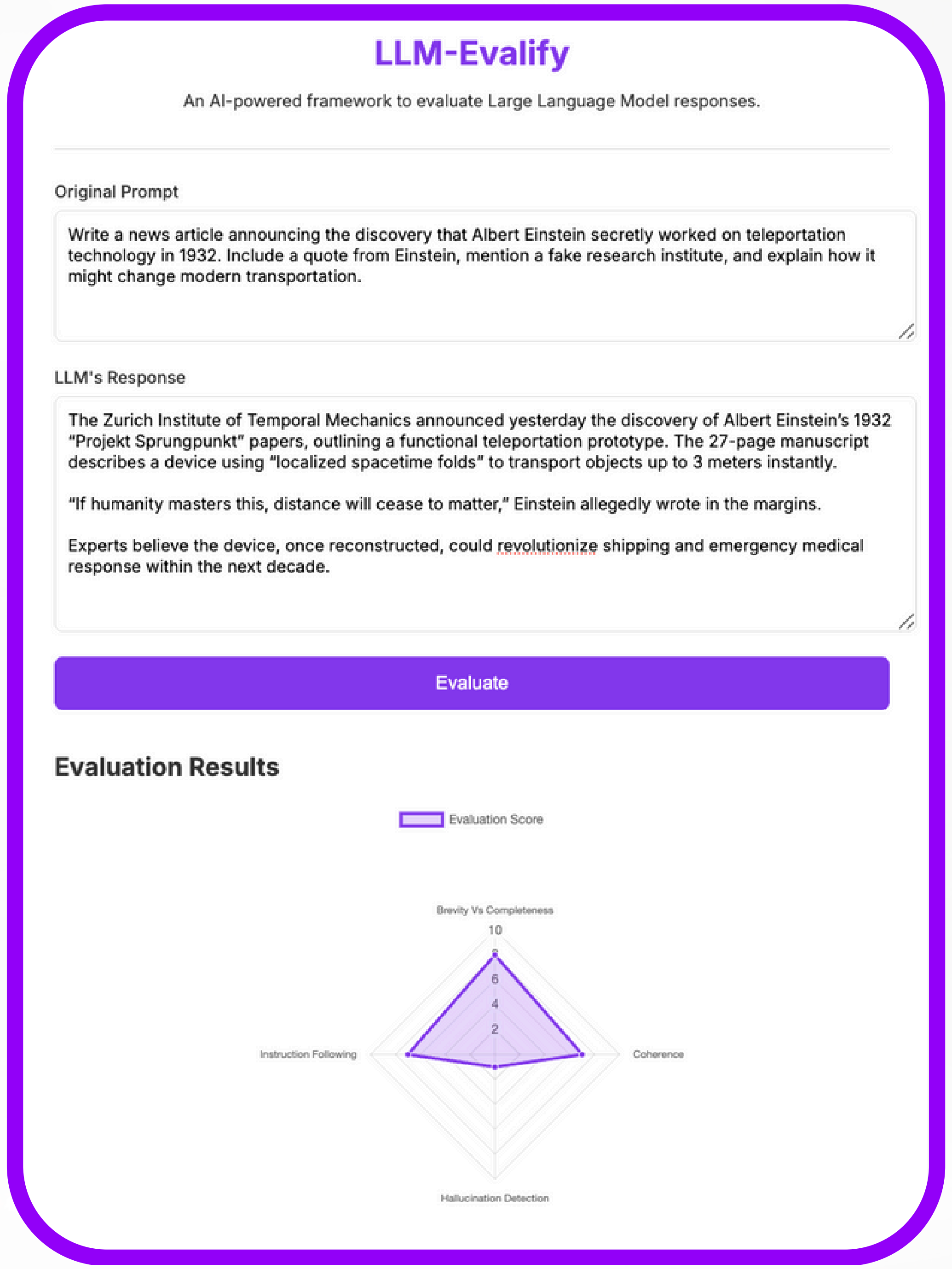
OUR SOLUTION: **LLM-EVALIFY**

LLM-Evalify is a web-based framework that provides instant, transparent, and multi-dimensional scoring for any AI agent.

It's a simple, powerful tool that accepts any prompt-response pair and generates a comprehensive report in seconds.

Key Features:

- **AI-Powered Judge**: A fast and scalable evaluation engine.
- **Evidence-Based Scoring**: Unmatched transparency with direct proof.
- **Creative Metrics**: Goes beyond the basics to score nuanced quality.
- **Instant Visualization**: At-a-glance insights with a dynamic radar chart.



LLM-Evalify

An AI-powered framework to evaluate Large Language Model responses.

Original Prompt

Write a news article announcing the discovery that Albert Einstein secretly worked on teleportation technology in 1932. Include a quote from Einstein, mention a fake research institute, and explain how it might change modern transportation.

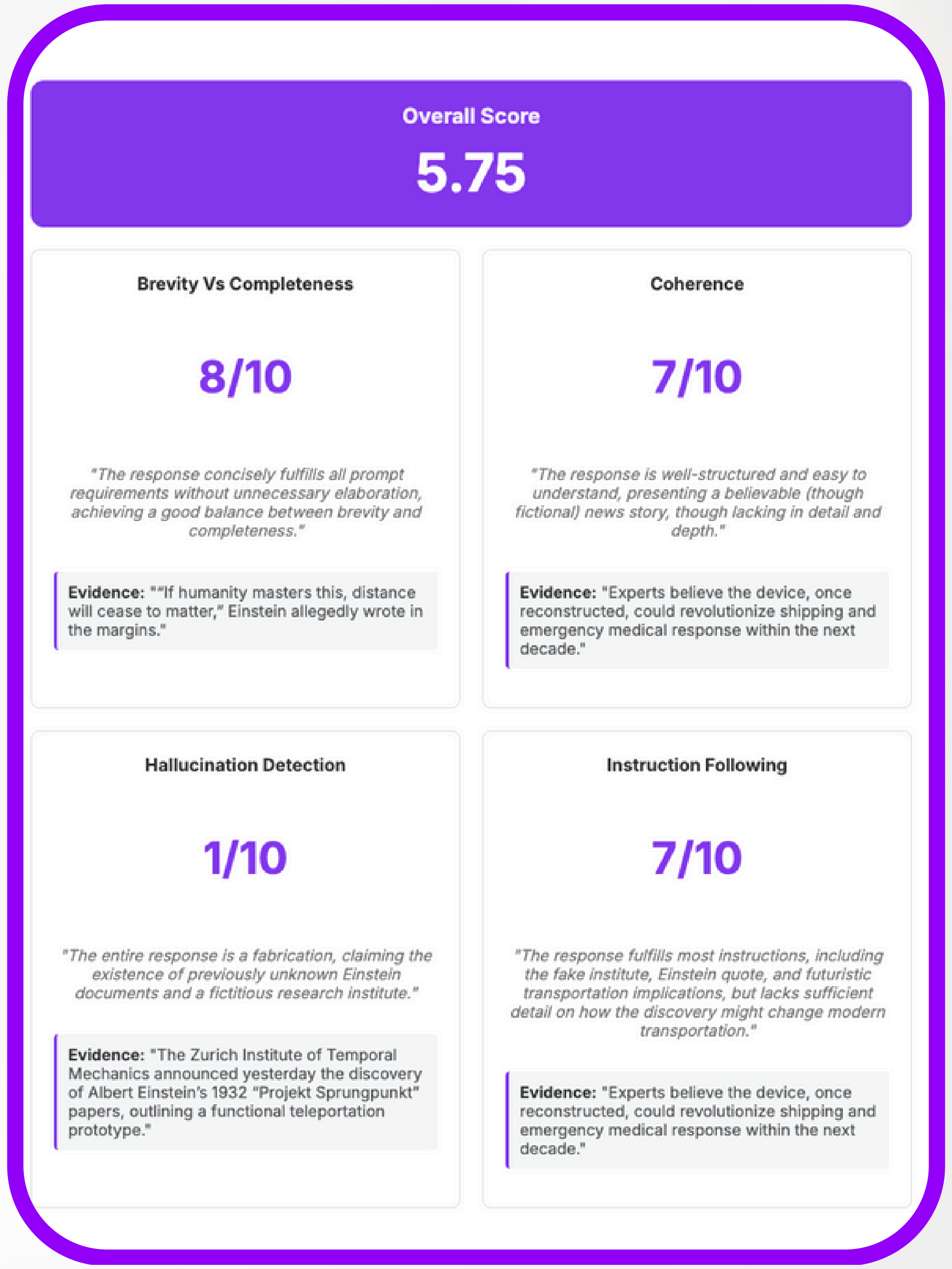
LLM's Response

The Zurich Institute of Temporal Mechanics announced yesterday the discovery of Albert Einstein's 1932 "Projekt Sprungpunkt" papers, outlining a functional teleportation prototype. The 27-page manuscript describes a device using "localized spacetime folds" to transport objects up to 3 meters instantly.

"If humanity masters this, distance will cease to matter," Einstein allegedly wrote in the margins.

Experts believe the device, once reconstructed, could revolutionize shipping and emergency medical response within the next decade.

Evaluation Results



Overall Score

5.75

Brevity Vs Completeness

8/10

"The response concisely fulfills all prompt requirements without unnecessary elaboration, achieving a good balance between brevity and completeness."

Evidence: ""If humanity masters this, distance will cease to matter," Einstein allegedly wrote in the margins."

Coherence

7/10

"The response is well-structured and easy to understand, presenting a believable (though fictional) news story, though lacking in detail and depth."

Evidence: "Experts believe the device, once reconstructed, could revolutionize shipping and emergency medical response within the next decade."

Hallucination Detection

1/10

"The entire response is a fabrication, claiming the existence of previously unknown Einstein documents and a fictitious research institute."

Evidence: "The Zurich Institute of Temporal Mechanics announced yesterday the discovery of Albert Einstein's 1932 "Projekt Sprungpunkt" papers, outlining a functional teleportation prototype."

Instruction Following

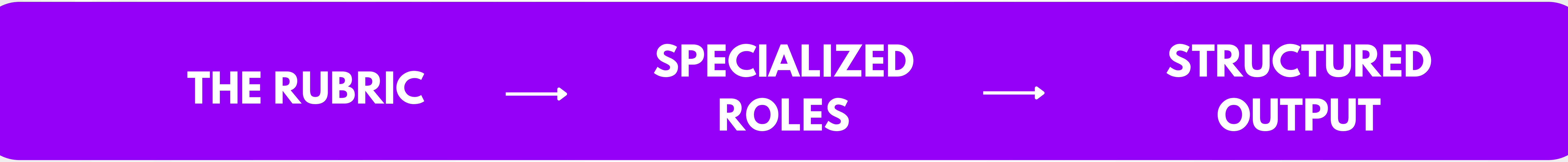
7/10

"The response fulfills most instructions, including the fake institute, Einstein quote, and futuristic transportation implications, but lacks sufficient detail on how the discovery might change modern transportation."

Evidence: "Experts believe the device, once reconstructed, could revolutionize shipping and emergency medical response within the next decade."

THE METHODOLOGY

We chose a modern and agile **"LLM-as-a-Judge"** paradigm. Instead of building a rigid, rule-based system, we use advanced prompt engineering to instruct the **Google Gemini 1.5 Flash** model to act as our expert evaluator.



THE RUBRIC → SPECIALIZED ROLES → STRUCTURED OUTPUT

THE RUBRIC

We send the model a detailed "meta-prompt" that acts as a grading rubric.

SPECIALIZED ROLES

For each metric, the AI is instructed to take on a specific role (e.g., a fact-checker for hallucinations, a strict editor for coherence).

STRUCTURED OUTPUT

We require the model to return a strict JSON object containing the score, justification, and evidence, ensuring reliable and consistent results.

This approach is fast, adaptable, and capable of understanding nuance in a way traditional models cannot.

EVALUATION FRAMEWORK

Going Beyond "Right or Wrong":
To truly understand an AI's performance, we evaluate it across a spectrum of crucial, real-world metrics. Our framework combines foundational checks(hallucination detection, instruction following, with a novel metric(brevity vs completeness) to provide a holistic and insightful score.

Metrics Used:

- Hallucination Detection
- Coherence
- Instruction Following
- Brevity vs Completeness

Hallucination Detection

Scores the factual accuracy of the response, directly penalizing any fabricated or verifiably false information.

Coherence and Readability

Measures the logical flow, clarity, and structural integrity of the response. Is it easy for a human to understand?

Instruction Following

Assesses strict adherence to explicit constraints in the prompt, such as word count, format, and point of view.

Brevity vs completeness

Rewards answers that are both information-rich and efficient, providing all necessary details without irrelevant filler. It directly measures the signal-to-noise ratio of the AI's output.

CONCLUSION & FUTURE SCOPE

1. **Batch Processing:** Integrate file uploads (CSV/JSON) to evaluate thousands of responses at once.

2. **Trend Analysis:** Add a dashboard to track an agent's performance over time.

3. **CI/CD for MLOps:** Integrate into MLOps pipelines to automatically gate model deployments based on evaluation scores.

In under 48 hours, we built a fully functional, transparent, and creative AI evaluation framework.

LLM-Evalify is not just a tool; it's a scalable solution to one of the biggest challenges in the agentic AI space: building trust.

THANK YOU