AI AGENT EVALUATION FRAMEWORK

PRESENTED BY GOKULAN M

Team: Deep Dreamer

Proposed Solution: Multi-Metric AI Evaluation Platform

What is My Big Idea?

To revolutionize AI quality assessment by building the scalable platform that automatically evaluates AI agents responses across multiple dimensions-providing not just a score, but actionable insights and transparent explanations.

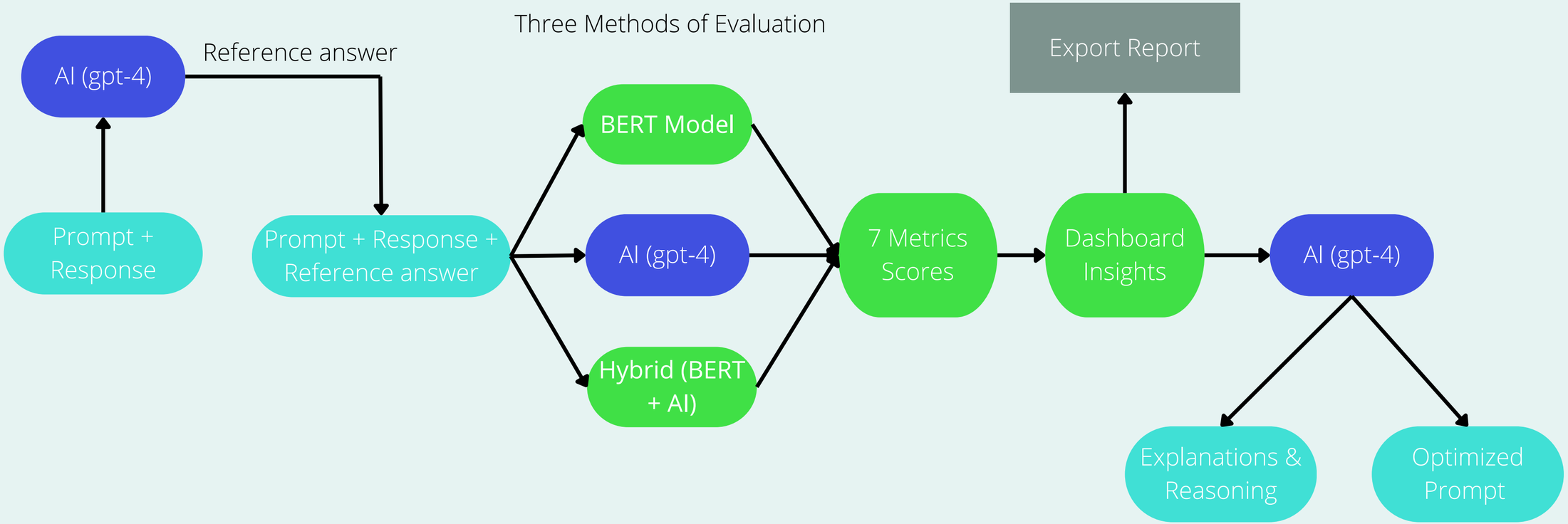
Key Features

- **Multi-Metric Evaluation**

Our platform goes beyond “right or wrong”. Every response is scored across seven crucial aspects-Instruction, Hallucination, Assumption, Coherence, Accuracy, Completeness, and Overall Quality using Multi-Head BERT model and Advanced LLM.

- AI powered automated scoring with reasoning and explanations
- Automated reference answer generation using advanced LLM model gpt-4 to reduce manual effort.
- **Intelligent prompt optimization:** Diagnoses low-quality prompts and automatically suggests improved formulations.
- **Actionable Insights and Comparisons:** Easily compare models or agents, identify strengths and weaknesses, and receive data driven recommendations to improve both prompts and AI performance.

Flow Diagram



Implementation & Design

1. Implementation Overview

- **Hybrid AI Evaluation Pipeline:**
  We combine three powerful evaluation strategies—BERT model, GPT-4, and a Hybrid (BERT + AI) to analyze prompt-response pairs. This ensures both factual grounding and advanced reasoning.
- **Seamless Reference Integration:**
  Prompts and responses are automatically paired with reference answers, using GPT-4 where human references are unavailable. This enables objective scoring across seven core metrics.

2. Design Highlights

- **Unified 7-Metric Scoring:**
  All evaluation methods output a comprehensive seven-metric score for every response, providing granular insight into model behavior.
- **Intelligent Dashboard Workflow:**
  Scores flow into an interactive dashboard for real-time visualization, actionable feedback, and explanations. The system can auto-generate improved prompts and detailed reasoning using GPT-4—closing the loop for continuous model improvement.

Our modular design leverages both traditional ML (BERT) and advanced generative AI (GPT-4) for robust, scalable, and interpretable evaluation—empowering users to analyze, optimize, and trust AI outputs at every step.

Innovation & Impact

1. Innovation

- **First-of-its-Kind Multi-Metric AI Evaluator:**
  Simultaneously scores AI responses on seven distinct metrics for a truly multi-dimensional assessment—moving beyond simple accuracy or pass/fail systems.
- **Automated, Explainable Insights:**
  Integrates state-of-the-art AI (BERT & GPT-4) to not only grade but also explain reasoning, generate reference answers, and optimize prompts, bridging the gap between evaluation and actionable improvement.

2. Impact

- **Enhanced Trust & Accountability:**
  Delivers transparent, evidence-backed evaluations—enabling users to understand, trust, and refine AI outputs at scale.
- **Drives Better AI for All:**
  Powers faster iteration, robust performance, and ethical deployment in real-world AI—supporting industry, research, and education with scalable, continuous model improvement.