RAGBench: Explainable Benchmark for Retrieval-Augmented Generation Systems

1. Objectives

- Establishes a benchmark framework for systematically evaluating explainability and citation quality in RAG (Retrieval-Augmented Generation) systems.
- Focuses on measuring the alignment between generated responses and retrieved supporting evidence.

2. Key Features

- Multi-level question types: Factoid, reasoning, and subjective queries to simulate real-world usage of RAG systems.
- Evaluates not only correctness but also if the output is properly justified and traceable to sources.
- Provides both quantitative (automated) and qualitative (manual) metrics for response analysis.

3. Evaluation Metrics

- Citation Precision & Recall: How precisely and completely sources are linked to responses.
- Hallucination Rate: Frequency of unsupported or fabricated content in model answers.
- Evidence Completeness: Whether all parts of an answer are backed by the retrieved/available evidence.
- Explainability Score: Human-assessed transparency and clarity of output justification.

4. How RAGBench Works

- **Query Input and Document Retrieval:** The system takes a user's natural language query and retrieves relevant documents or passages from a large knowledge base using vector search tools like FAISS or Chroma.
- **Answer Generation:** A generative language model (such as GPT-like LLMs) then produces an answer conditioned on the retrieved documents, ensuring the response is informed by external evidence.
- **Citation and Explanation:** Along with the answer, RAGBench extracts explicit citations linking segments of the response back to the source documents. This provides transparency and traceability for the answers.
- **Multi-type Question Handling:** It tests RAG systems on factoid (direct fact), reasoning (multi-evidence synthesis), and subjective (nuanced interpretation) questions, assessing versatility.
- **Evaluation Metrics:** RAGBench uses automated metrics such as citation precision/recall, hallucination detection (fabrications), and evidence completeness. Additionally, human evaluation judges the explainability and clarity of the answer supporting rationale.

5. Usefulness for Academic Compliance Projects

- Provides a set of proven, practical benchmarks that can be adapted for evaluating explainability and citation quality in policy automation.
- Offers a standardized way to report how well a Policy-as-Code engine justifies every compliance answer and references regulations.
- Guides model and system improvements to enhance reliability and trust among academic administrators and auditors.