RAG System Evaluation

Yao Cheng
Kai Li
Xueyun Li
Muzhou Liu

AGENDA

- Literature Review
- Research Gap
- Proposed Methods
- Questions for Discussion

LITERATURE REVIEW

Paper 1: RAGAS: Automated Evaluation of Retrieval Augmented Generation

Paper 2: Replacing Judges with Juries: Evaluating LLM Generations with a Panel of Diverse Models

Paper 3: RAGBench: Explainable Benchmark for RAG systems

Why we choose it?

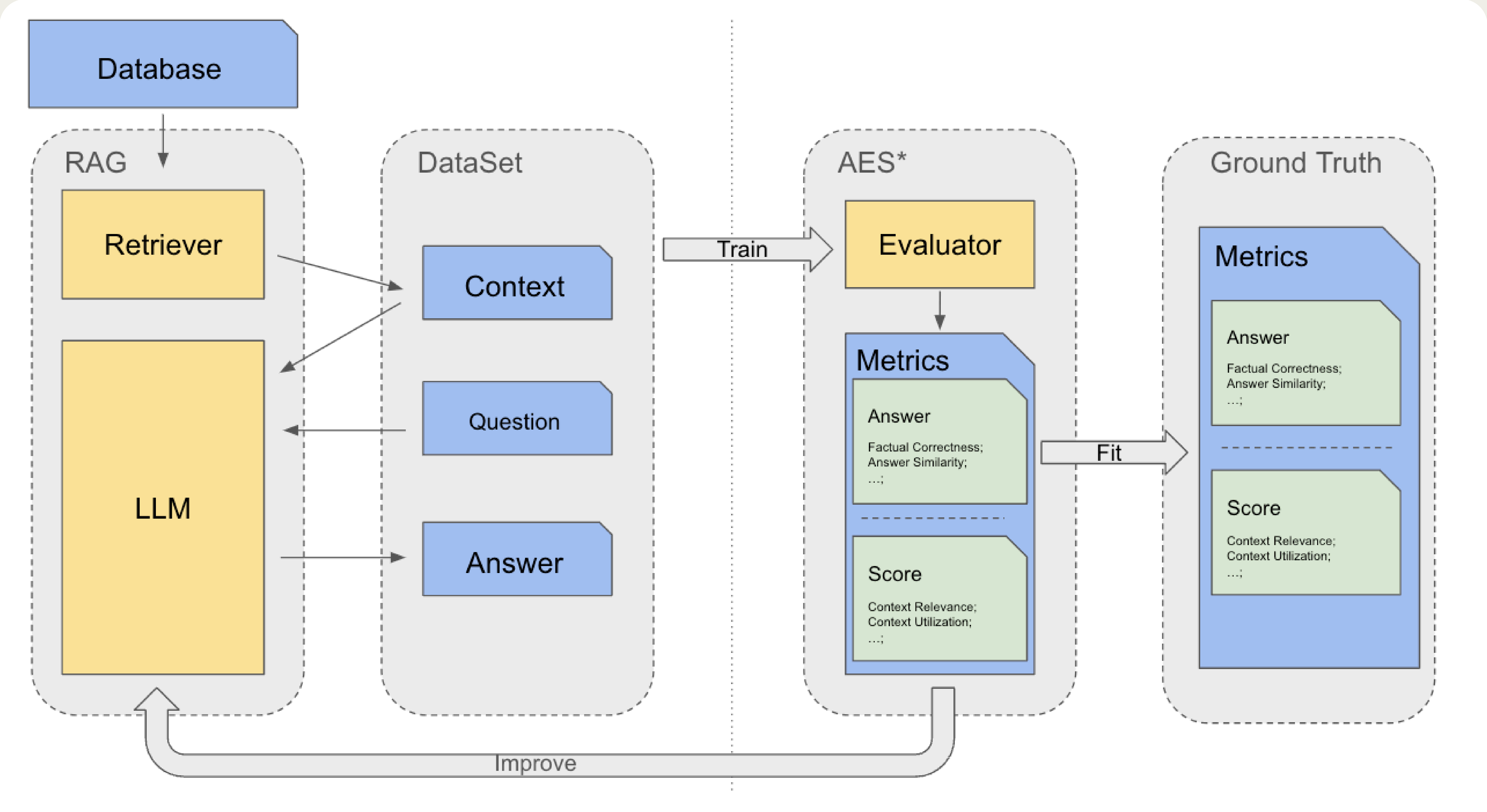
- Highly related dataset: Delucion QA
  - A customer support QA that based on an automobile manual
- Available ground truth
  - GPT-4 generated, **95%** agreement with human annotation

RESEARCH GAP

1. **Bias in Automated RAG Evaluation**
   - RAGAS automates RAG evaluation but inherits LLM biases and lacks adaptability to new evaluation methods.
2. **Lack of RAG-Specific Diverse Evaluation**
   - PoLL improves bias reduction using multiple models but is not tailored for RAG-specific evaluation.
3. **Missing Structured Metric Selection**
   - RAGBench provides benchmarking but lacks a method for automated metric selection and improvement.

Goal: Build an automated metric selection framework for the evaluation of RAG system.

WORKFLOW OVERVIEW



AES: Automated Evaluation System

PROPOSED METHODS

Our Methods:

Evaluators

- Fine-tuned LLM
- A panel of diverse small models & ML
- LLM agents for metric selection
- Reasoning

Metrics

- GT Answer
- GT Score

DISCUSSION

More specific use case scenarios

- How user compose their query?

  E.x. What are the question type?

  1. factual questions
  2. definition questions
  3. list questions
  4. "how" questions
  5. "why" questions
  6. hypothetical questions

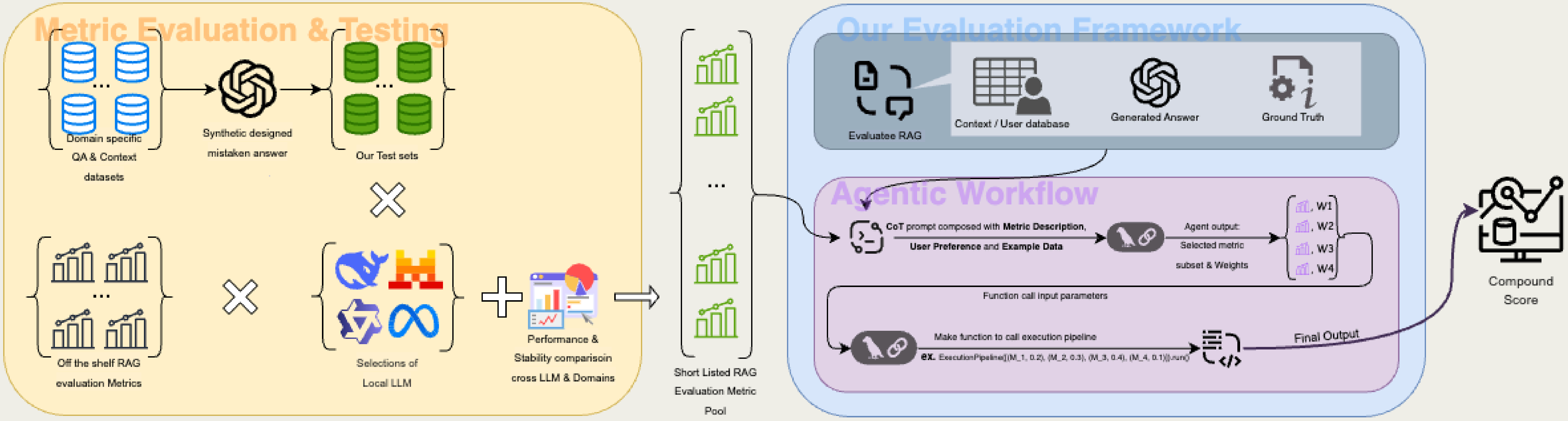
- How TSBC experts make human annotations?

  E.x. Does retriever gives enough context?

  E.x. What aspect in the answer needed to be defined as a human preferred answer?

  1. Correctness
  2. Gives actionable instruction
  3. Reject wrong query;

SYSTEM WORKFLOW OVERVIEW



Metric Evaluation & Testing

Domain specific QA & Context datasets

Synthetic designed mistaken answer

Our Test sets

Off the shelf RAG evaluation Metrics

Selections of Local LLM

Performance & Stability comparisoin cross LLM & Domains

Short Listed RAG Evaluation Metric Pool

Our Evaluation Framework

Evaluatee RAG

Context / User database

Generated Answer

Ground Truth

Agentic Workflow

CoT prompt composed with **Metric Description, User Preference** and **Example Data**

Agent output: Selected metric subset & Weights

Function call input parameters

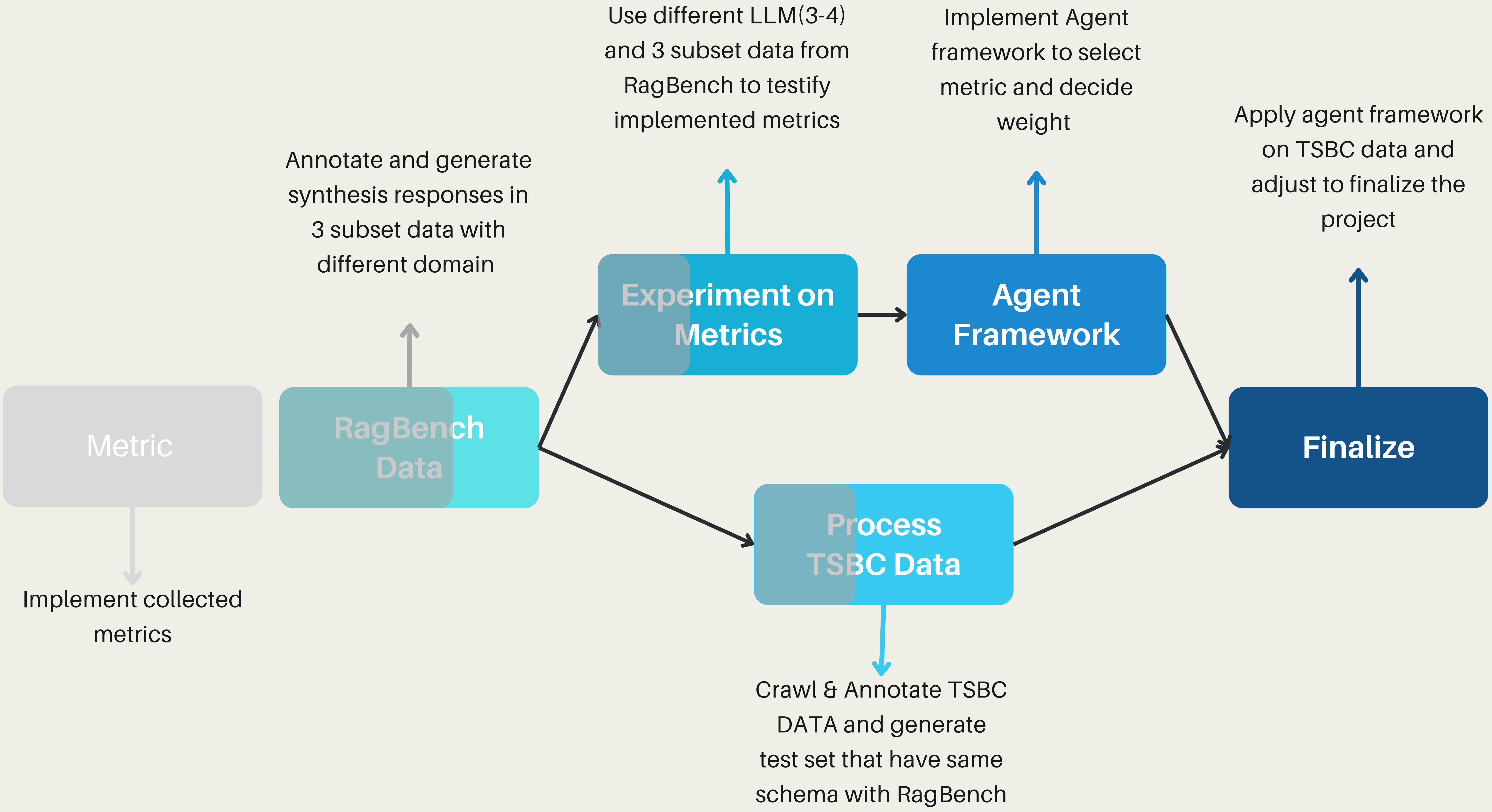
Make function to call execution pipeline

ex. ExecutionPipeline([(M_1, 0.2), (M_2, 0.3), (M_3, 0.4), (M_4, 0.1)]).run()

Final Output

Compound Score

PROGRESS UPDATE



Metric

Implement collected metrics

RagBench Data

Annotate and generate synthesis responses in 3 subset data with different domain

Experiment on Metrics

Use different LLM(3-4) and 3 subset data from RagBench to testify implemented metrics

Process TSBC Data

Crawl & Annotate TSBC DATA and generate test set that have same schema with RagBench

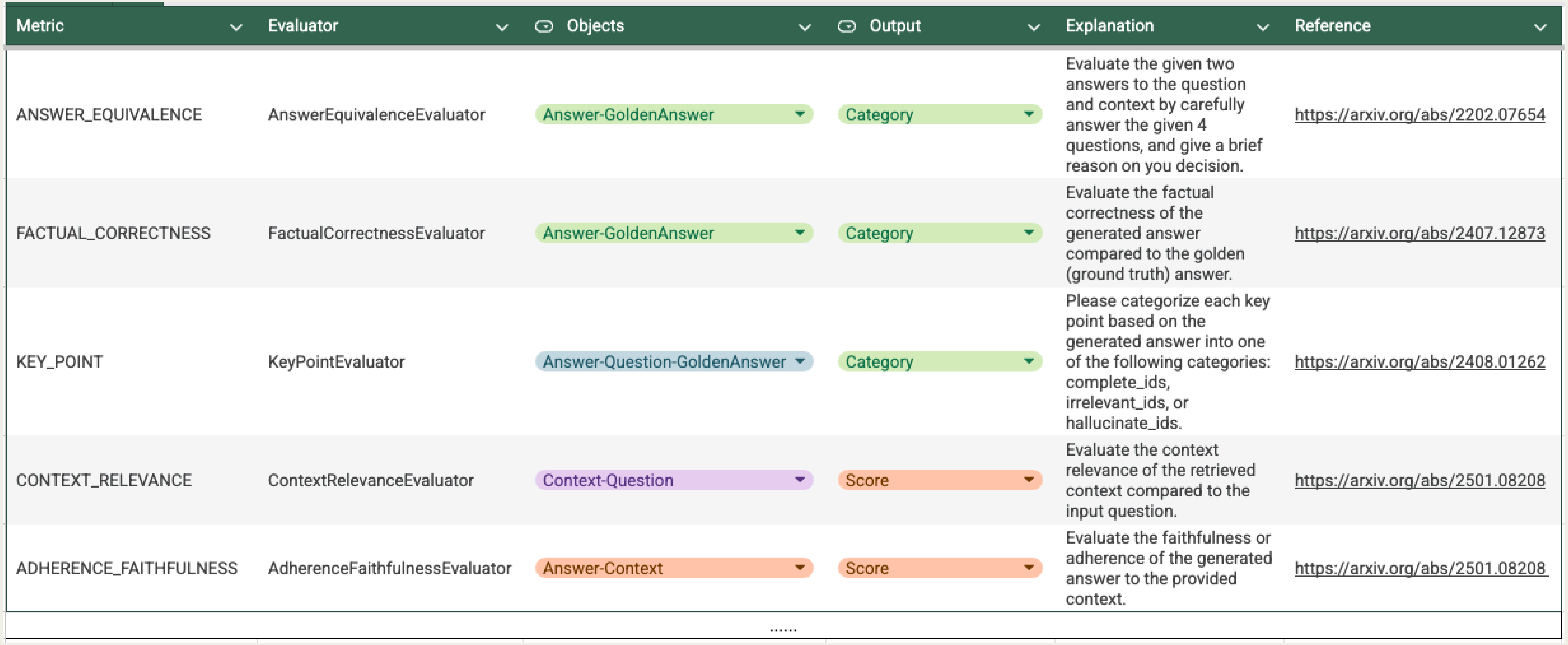
Agent Framework

Implement Agent framework to select metric and decide weight

Finalize

Apply agent framework on TSBC data and adjust to finalize the project

IMPLEMENTED METRICS



Metric	Evaluator	Objects	Output	Explanation	Reference
ANSWER_EQUIVALENCE	AnswerEquivalenceEvaluator	Answer-GoldenAnswer	Category	Evaluate the given two answers to the question and context by carefully answer the given 4 questions, and give a brief reason on you decision.	https://arxiv.org/abs/2202.07654
FACTUAL_CORRECTNESS	FactualCorrectnessEvaluator	Answer-GoldenAnswer	Category	Evaluate the factual correctness of the generated answer compared to the golden (ground truth) answer.	https://arxiv.org/abs/2407.12873
KEY_POINT	KeyPointEvaluator	Answer-Question-GoldenAnswer	Category	Please categorize each key point based on the generated answer into one of the following categories: complete_ids, irrelevant_ids, or hallucinate_ids.	https://arxiv.org/abs/2408.01262
CONTEXT_RELEVANCE	ContextRelevanceEvaluator	Context-Question	Score	Evaluate the context relevance of the retrieved context compared to the input question.	https://arxiv.org/abs/2501.08208
ADHERENCE_FAITHFULNESS	AdherenceFaithfulnessEvaluator	Answer-Context	Score	Evaluate the faithfulness or adherence of the generated answer to the provided context.	https://arxiv.org/abs/2501.08208

......

Thank you!