Northeastern University

Can AI Be Trusted? Evaluating Reliability in Retrieval-Augmented Generation Responses

Zheng Gu, Hejia Li, Mulei Ni | Supervised by Prof Yvonne Coady | Guided by Derek Jacoby

LinkedIn

Dataset

Paper/GitHub

1. Background & Motivation

Royal British Columbia Museum (RBCM) AI Avatar Project

- The RBCM is using AI avatars powered by Retrieval-Augmented Generation (RAG) to enhance visitor interactions
- The challenge of evaluating the quality and accuracy of AI responses

Importance

- Enhance visitor experience
- Share cultural heritage
- Provide personalized tours
- Advance educational goals ...

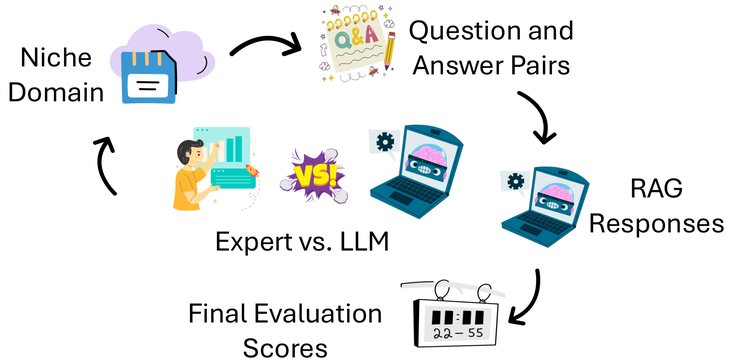
2. Research Gaps & Objectives

RAG Evaluation Gaps

- Limited research
- Lack of standardized benchmark
- Insufficient data for evaluation
- Expensive human evaluation

How to evaluate RAG?

Human Evaluation vs. LLM as a Judge[16]

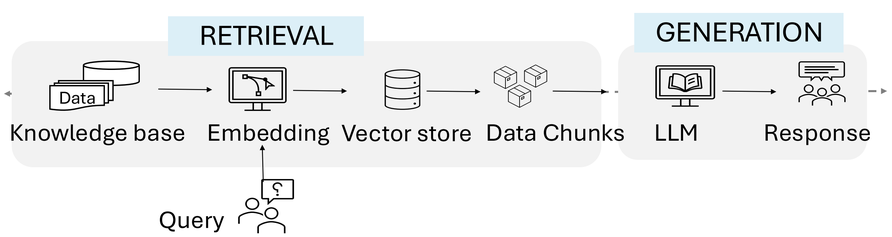


Objectives

- Introduce RAG evaluation pipeline
- Evaluate RAG responses
- Demonstrate "LLM-as-a-judge" as an effective methodology

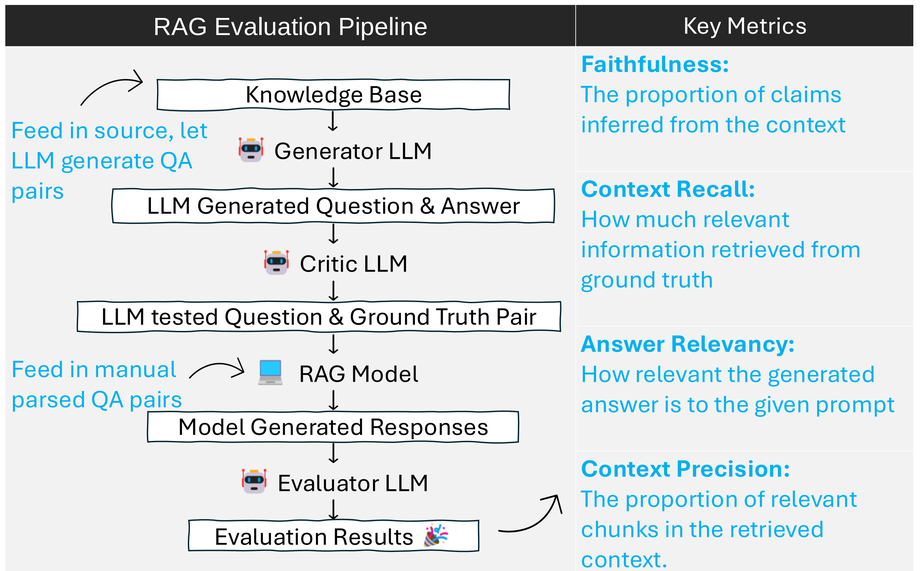
3. Naive RAG Model & Dataset

Baseline Model & Factors: OpenAI models on LlamIndex framework [10]



Dataset/Knowledge Base:
HotpotQA with 113k Wikipedia-based question-answer (QA) pairs [21]
Evaluation Framework: RAGAS

4. Methodology

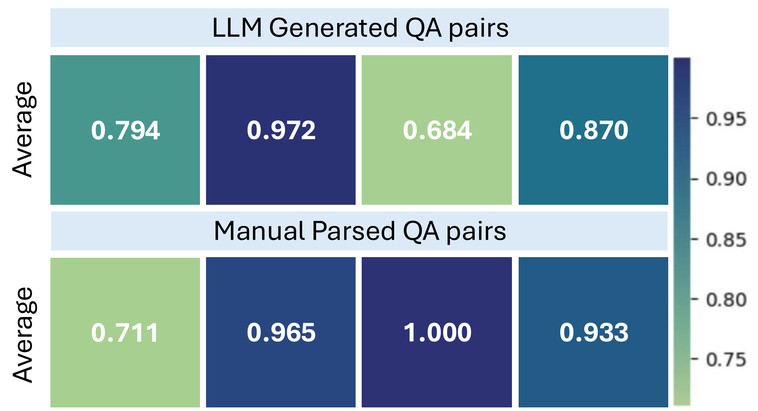


Greenness Evaluation:

Method	Estimate carbon emission using tokens
Literature Evidence	♻ "People are currently paying per token, so **less tokens, less cost, less energy**"[18] "This gives the gCO2e for the operations of GPT-4 to be **0.3 gCo2e for 1k tokens**"[18]

5. Experimental Results

Qualitative evidence analysis of pipeline feasibility:



Quantitative results of evaluation scores in naive models:

Mean Results	GPT-4o	GPT-3.5-Turbo
Faithfulness	0.789	0.7875
Answer Relevancy	0.826	0.9185
Context Precision	0.9	0.9
Context Recall	0.9	0.96

Greenness:

Carbon Emission Estimation		
Model	Tokens	Carbon Emission
GPT-4o	231778	69.5334 grams
GPT-3.5	234780	70.4340 grams

Analysis

- ✓ LLM-as-a-judge aligns well with human evaluation for 3 out of the 4 metrics, demonstrating its feasibility.
- ✓ Naïve model selection affects RAG evaluation results.

6. Future Work

- ✓ Facilitate RBCM in selecting RAG evaluation methodology
- Optimize RAG model based on RBCM Dataset
- Incorporate greenness considerations in model training