Improving Reliability of Large Language Models for Nuclear Power Plant Diagnostics

Thomas Reeves, Graduate Student Intern, C220 Data Analytics and Applied Statistics

Mentors: Cody Walker, Linyu Lin, Vivek Agarwal

Problem

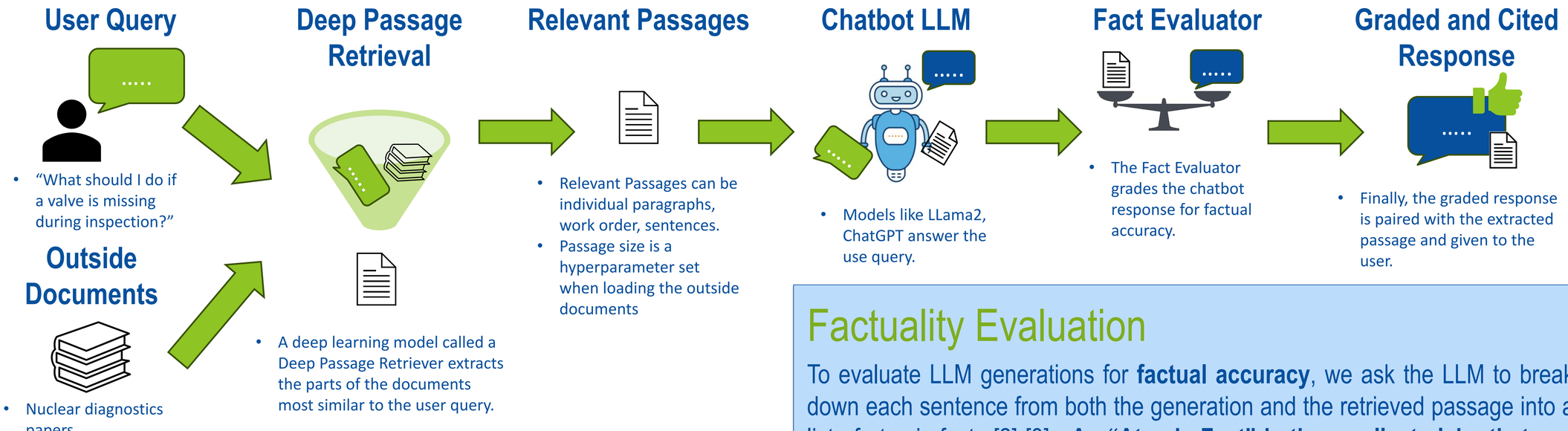
Large Language Models (LLMs) struggle out of the box when answering factually about detailed questions, **especially in domains that are sparsely represented in their training data.** This causes hallucinations and reduces reliability making it difficult for them to be used in practice.

Contributions

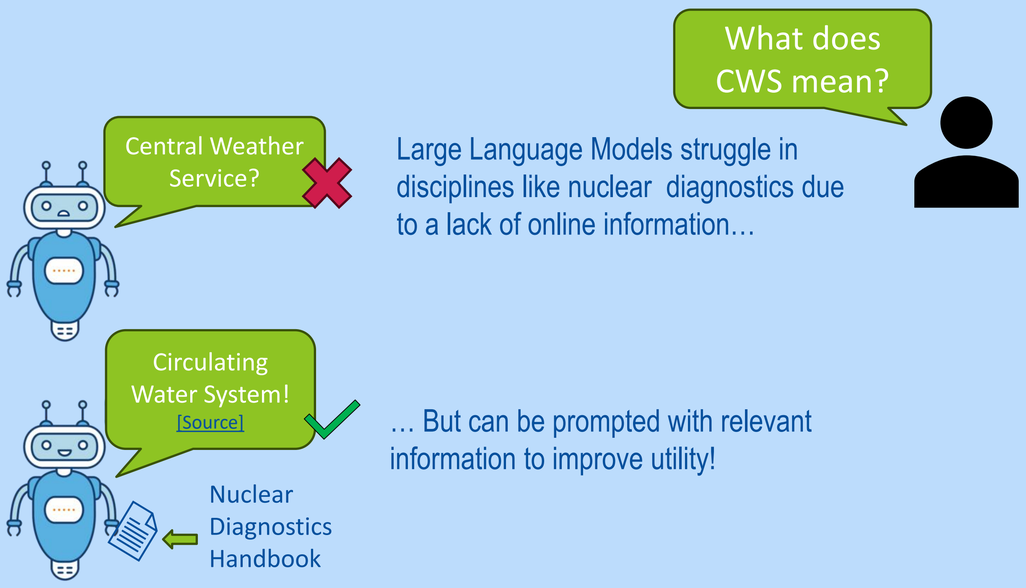
1. We implemented Retrieval Augmented Generation for Chatbot LLM's allowing for **reliable and accurate** nuclear diagnostic question answering.
2. We Implemented fact evaluation tool to concretely evaluation LLM generations and identify hallucinations.
3. This model will be deployed in VIPER: Visualization for Predictive maintenance Recommendation's.

Methods

We implement a multistage pipeline to let LLMs access external information via Retrieval Augment Generation and to evaluate the results for factual accuracy.



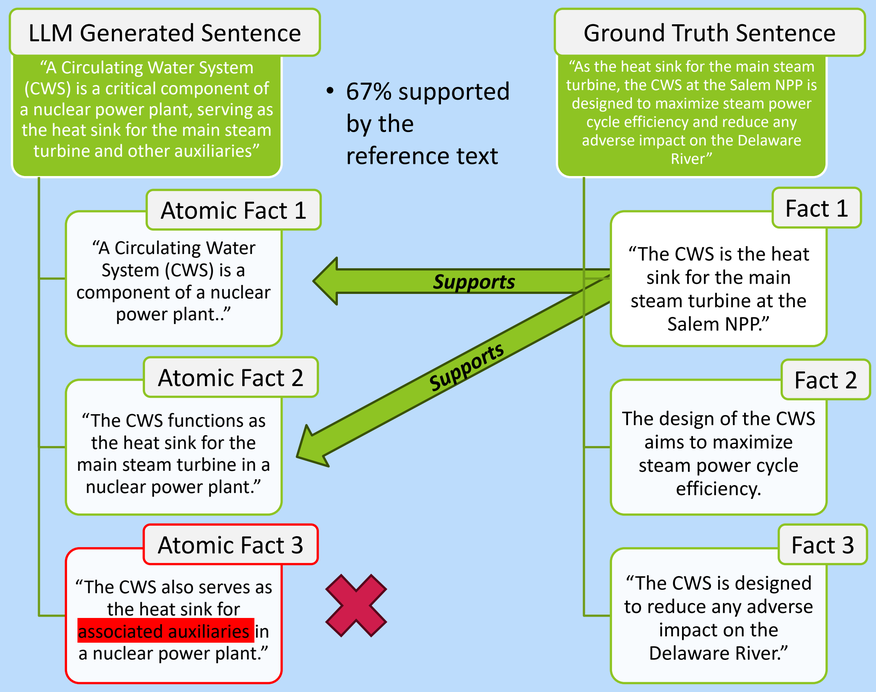
Retrieval Augment Generation



Retrieval Augment Generation (RAG) [1] uses a deep learning model called a Deep Passage Retriever, that embeds queries and passages from external documents, and **returns the document with the highest similarity to the user query**. The LLM is prompted with the most relevant passages to the users' questions and asked to generate a response.

Factuality Evaluation

To evaluate LLM generations for **factual accuracy**, we ask the LLM to break down each sentence from both the generation and the retrieved passage into a list of atomic facts [2] [3]. **An "Atomic Fact" is the smallest claim that can be extracted from a sentence**.



Then we use a Natural Language Inference (NLI) Model to classify each pair of generation facts and passage facts as either Entailment (Supported), Neutral, or Contradictory. To be considered true, a generated atomic fact must be supported with at least 50% probability by a fact from the ground truth.

Results

To evaluate our RAG pipeline, we compared LLM performance on 5 questions related to predictive maintenance with and without our rag pipeline. **We see a 131% increase in performance** according to the Fact Evaluator. We also measure the degree to which the Fact Evaluator adheres to human opinions. We had a human hand verify each atomic fact across the 5 generations and saw that **our Fact Evaluator has a 0.90 F1 score, or in other words 87% of the time it agrees with the human evaluation**.

Significance

This work shows that using RAG techniques can improve accuracy and reliability, allowing for the application of LLMs in specialized areas, even when those areas that aren't extensively covered in their initial training.

References:
[1] Lewis, Patrick, et al. "Retrieval-augmented generation for knowledge-intensive nlp tasks." Advances in Neural Information Processing Systems 33 (2020): 9459-9474.
[2] Min, Sewon, et al. "Factscore: Fine-grained atomic evaluation of factual precision in long form text generation." arXiv preprint arXiv:2305.14251 (2023).
[3] Kamoi, Ryo, et al. "Wice: Real-world entailment for claims in wikipedia." arXiv preprint arXiv:2303.01432 (2023).

www.inl.gov
THE NATIONAL GEM CONSORTIUM
USC
Battelle Energy Alliance manages INL for the U.S. Department of Energy's Office of Nuclear Energy
INL Idaho National Laboratory