RAG-Ex: A Generic Framework for Explaining Retrieval Augmented Generation

Problem statement:

It remains challenging for users to understand if the LLM truly used the provided context or is hallucinating from its internal knowledge, which reduces user trust.

Report on Understanding the RAG-Ex Framework for Explaining Retrieval Augmented Generation

The paper presents a novel framework called **RAG-Ex**, which is designed to explain the responses generated by large language models (LLMs) enhanced with retrieval augmented generation (RAG) techniques. RAG-Ex is particularly important since large language models often lack transparency in how they produce answers, making it hard for users to trust the outputs, especially in question answering tasks.

1. What is RAG-Ex?

RAG-Ex is a **model- and language-agnostic general explanation framework**. It provides approximate explanations for why an LLM generated a particular response by applying **post-hoc perturbation-based methods**. This means the framework works independently of the specific language model or its internal workings and can be used with different types of LLMs, including proprietary ones like GPT-4 that do not expose internal token-probabilities.

2. How RAG-Ex Works

The core idea behind RAG-Ex is to analyze how changes to the input affect the generated output. By creating multiple slightly altered ("perturbed") versions of the input query and context and observing the variations in the output responses, RAG-Ex estimates the importance of different parts of the input in influencing the final LLM answer. The process requires querying the LLM multiple times with these perturbed inputs, then comparing each resulting output with the original LLM response.

3. Perturbation and Its Steps

Perturbation refers to systematically modifying parts of the input text to see how these changes affect the model's output. In RAG-Ex, perturbations are created at different granularities, such as words, phrases, or sentences. The perturbation strategies used include:

- **Leave-one-token-out:** Remove one token at a time from the input.
- **Random Noise:** Insert random words near the target token.
- **Entity Manipulation:** Replace named entities or nouns with random words.
- **Antonym Injection:** Substitute words with their antonyms.
- **Synonym Injection:** Substitute words with their synonyms.
- **Order Manipulation:** Change the sequence of words in a token.

Each perturbed input is then fed to the same LLM, generating multiple responses corresponding to each perturbation.

4. Explanation and Comparison Module

After generating outputs for the perturbed inputs, RAG-Ex compares these outputs to the original output using two main types of comparison:

- **Syntactic comparisons:** String-based metrics such as Levenshtein distance, Jaro-Winkler distance, and n-gram overlap to measure textual similarity and difference.
- **Semantic comparisons:** Using language model or SentenceBERT embeddings to compute cosine similarity, capturing meaning-level similarity beyond exact text matching.

The framework converts the dissimilarity scores into normalized importance scores for each token. Tokens that cause larger changes in output when perturbed are considered more important in influencing the LLM's original response. These importance scores form the *approximated explanation*, highlighting the parts of the input that most impacted the generated answer.

5. Application to Verify RAG Systems

RAG-Ex can be used to verify whether a Retrieval Augmented Generation system is working as intended. By examining which parts of the input context and question the LLM relied on to produce its answer, users can evaluate the faithfulness of the generated responses to the retrieved documents. If the explanations highlight relevant context sentences and the question as important tokens, this indicates the LLM is properly using the retrieval context rather than generating hallucinations from internal parameters. Conversely, unreasonable or irrelevant explanations may warn that the system is not faithfully incorporating the retrieved data.

Therefore, it is appropriate and beneficial to use RAG-Ex to assess and improve RAG systems, increasing end-user trust and understanding of how answers are composed in real applications.

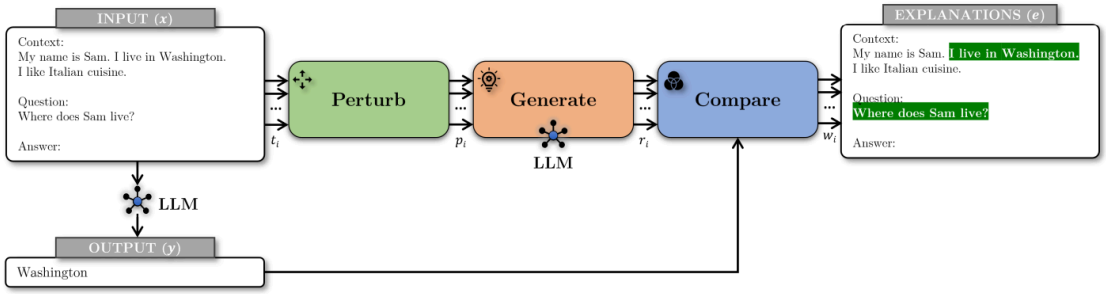

Figure 1: The proposed model- and language-agnostic RAG-Ex explainability framework.