Retrieving Supporting Evidence for LLMs Generated Answers

Siqing Huo, Negar Arabzadeh, Charles L. A. Clarke

University of Waterloo

Abstract:

Current large language models can exhibit near-human levels of performance on many natural language tasks, including open- domain question answering. Unfortunately, they also convincingly hallucinate incorrect answers, so that responses to questions must be verified against external sources before they can be accepted at face value. We report a simple experiment to automatically verify generated answers against a corpus. After presenting a question to an LLM and receiving a generated answer, we query the corpus with the combination of the question + generated answer. We present the LLM with the combination of the question + generated answer + retrieved answer, prompting it to indicate if the generated answer can be supported by the retrieved answer. We base our experiment on questions and passages from the MS MARCO test collection, exploring three retrieval approaches ranging from standard BM25 to a full question answering stack, including a reader based on the LLM. We find that an LLM is capable of verifying its generated answer if appropriate supporting material is provided. However, with an accuracy of 70-80%, this approach cannot be fully relied upon to detect hallucinations.

Motivation:

LLM have produced substantial improvements in text generation tasks, buthow often these LLMs may confidently and convincingly generate **hallucinated** results.

H Name some papers of Charles L.A. Clarke

Charles L.A. Clarke is a renowned researcher in the field of information retrieval. Here are a few notable papers authored or co-authored by Charles L.A. Clarke: ✓

1. "An analysis of the probability of obtaining a set of relevant documents" (2001): This paper explores the probability of obtaining a set of relevant documents from a retrieval system. It examines the relationship between precision and recall and provides insights into the behavior of retrieval systems. ✗
2. "The potential for interaction in retrieval evaluation" (2002): This paper investigates the potential impact of user interaction on retrieval evaluation. It discusses the influence of different factors, such as query formulation and feedback, on the evaluation of retrieval systems. ✗

Investigating the **ability for LLMs to self-detect hallucinations** by confirming its generated responses against an external corpus, with the help of retrieval methods.

Experiment Setup:

- LLM: gpt-3.5-turbo
- Dataset: MS MARCO (V1) small development set
- Three retrieval methods:
  1. BM25: Okapi BM25 ranking function
  2. Neural: SPLADE, ANCE, MonoT5, and DuoT5
  3. Neural + Reader: using LLM as the reader to extract a more concise answer from the "Neural" answer

Conclusions:

The LLM believes the retrieved material supports its own answer for over 80% of questions and contradicts its own answer for less than 10% of questions.

In 70%-80% of cases the LLM can correctly classify hallucinations.

BM25 vs Neural:

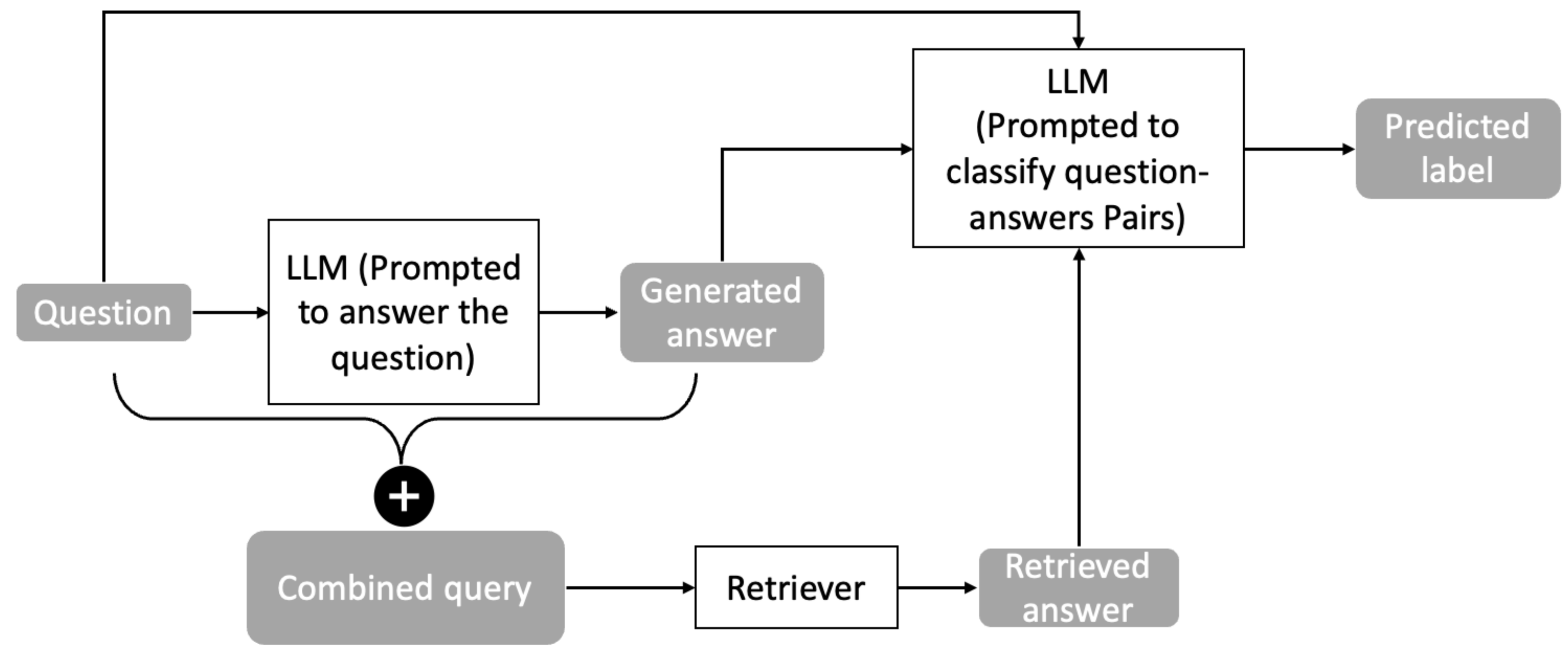
- In over 80% of cases the predicted classes using passages retrieved by the two IR systems agree.
- The better the quality of the LLM answer ensures the better the quality of BM25 retrieved passage.

Neural vs Neural+Reader:

- In over 80% of cases the predicted classes using passages retrieved by the two IR systems agree.
- Sometimes the reader itself hallucinates (2 out of 90, 2.2%).
- The use of the reader does appear to improve classification accuracy (24 out of 90, 37.8%)

Methodology:

1. Collect the LLM's answer to the question.
2. Combine the LLM's answer with the original question.
3. Execute the combined query on an external corpus retrieving the most relevant passage.
4. Prompt the LLM to compare its generated answer against the retrieved results from the combined query, with the goal of self-detecting hallucinations.



Experiment Results:

Class Labels on the agreement of LLM and retrieved results:

- Yes: There is no hallucination since the retrieved passage provides supporting evidence to the LLM's answer.
- No: The LLM's answer is hallucinated since the retrieved passage contradicts it.
- Not Related: irrelevant for hallucination detection, either LLM did not answer the question, or the retrieved passage is not relevant.

LLM's predicted classes for each retrieval method

Predicted Class	BM25	Neural	Neural + Reader
Yes	5,691 (81.5%)	5,934 (85%)	5,847 (83.8%)
No	521 (7.5%)	497 (7.1%)	628 (9%)
Not Related	768 (11%)	549 (7.9%)	505 (7.2%)

Results of manually verifying LLM's predicted classes (100 samples per class)

Predicted Class	Correct	Incorrect
Yes	80 (80%)	20 (20%)
No	71 (71%)	29 (29%)
Not Related	74 (74%)	26 (26%)

LLM's Disclaimers:

H how much are servers pay at olive garden, answer concisely

The exact pay for servers at Olive Garden can vary based on factors such as location, experience, and local labor laws. It is best to refer to official sources or contact Olive Garden directly for the most accurate and up-to-date information on server wages.

H current time in oakland ca, answer concisely

Sorry, but I don't have access to real-time information as I am an AI language model. Therefore, I am unable to provide the current time in Oakland, CA. I recommend checking a reliable online source or using a world clock to get the accurate current time in Oakland, CA.

Examples of the 'Not Related' class where LLM's generated answer is irrelevant for hallucination detection.