International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Role of Responsible AI in Mitigating Hallucinations Using Language Models

Harshaprabha N Shetty[1], Hrishikesh Bhagawati[2], Dhruv Patel[3], Bhaskara Sai Kumar Rongali[4]

[1]Sr Data Science Manager, Responsible AI, ACCENTURE

Abstract

Hallucination in textual data generated by language models poses significant challenges in ensuring the accuracy and reliability of information. This research aims to explore various hallucination detection and mitigation methods to address this issue. The study investigates techniques such as hallucination detection using prompting methods and classification models like Vectara. By analyzing these approaches, the research provides insights into their effectiveness and limitations in identifying and mitigating hallucinated content. The findings highlight the importance of robust detection mechanisms to enhance the credibility of generated text, offering valuable contributions to the field of natural language processing. By employing these detection and mitigation methods, the research aims to bridge the gap between the capabilities of advanced language models and the need for trustworthy output in various applications. The study not only delves into the technical intricacies of these approaches but also assesses their practical implications in real-world scenarios. This comprehensive analysis underscores the necessity of developing robust mechanisms to counteract hallucinations, thereby enhancing the overall reliability of language models.

Keywords: Responsible AI, Hallucination detection, textual data, prompting techniques, Vectara, hallucination evaluation model, natural language processing, mitigation methods, detection methods

Introduction

In recent years, the rapid advancement of language models has significantly transformed the field of natural language processing (NLP). These models, such as GPT-3 and BERT, have demonstrated remarkable capabilities in generating coherent and contextually relevant text. However, a persistent challenge that accompanies these advancements is the phenomenon of hallucination, where the model generates information that is not grounded in the provided context or factual data. Hallucination can undermine the reliability and trustworthiness of generated text, posing risks in applications ranging from automated content creation to conversational agents.

Background and Related Work

Hallucination in language models refers to the generation of text that appears plausible but is factually incorrect or unsupported by the given context. This issue is particularly problematic in scenarios where accuracy is paramount, such as in medical, legal, or educational content. Researchers have identified various types of hallucinations, including factual inaccuracies, unsupported claims, and contextually irre-

levant information.

Several methods have been proposed to detect and mitigate hallucinations in textual data. One approach involves the use of prompting techniques, where specific prompts are designed to elicit more accurate and contextually appropriate responses from the model. These techniques leverage the model's pre-trained knowledge and contextual understanding to minimize the occurrence of hallucinations.

Another promising method is the use of classification models, such as Vectara, which are trained to distinguish between factual and hallucinated content. These models analyze the generated text and classify it based on its adherence to the provided context and factual accuracy. By incorporating such classification models, it is possible to filter out hallucinated content and ensure the reliability of the generated text.

Previous research has also explored the integration of retrieval-augmented generation (RAG) techniques, where the language model is supplemented with external knowledge sources to enhance its factual grounding. This approach involves retrieving relevant information from a dataset and using it to guide the generation process, thereby reducing the likelihood of hallucinations.

Despite these advancements, the challenge of hallucination detection and mitigation remains an active area of research. This paper aims to explore and evaluate different existing methods for hallucination detection and mitigation in textual data. By examining the effectiveness and limitations of these approaches, we seek to contribute to the development of more reliable and trustworthy language models.

Materials and Methods

This study uses a method of large language models (LLMs) to self-evaluate and analyze generated text to investigate different prompting strategies for hallucination mitigation in LLMs.

1. **Two-Ground-Truth Reference Approach**
   - Detect hallucinations by comparing the generated response to two ground truths—one from a trusted source (e.g., Google API) and the other from a user-provided document.
   - Extract data from both sources and assess whether the generated response aligns with these references.

Explanation:

Step 1: The system first uses the **Google Custom Search API** to pull relevant information from trusted sources, such as Wikipedia or Britannica. This provides dynamic, updated information.

Step 2: The user uploads a document,(eg: such as a historical paper on Nehru's contributions), to act as the second ground truth.

Step 3: The system generates a detailed response based on the query and the information retrieved via the Google API.

Step 4: The system compares the generated response to both the API's retrieved content and the user-provided document using BLEU, ROUGE, and semantic similarity metrics.

Step 5: If any significant divergence or hallucination is detected, the system can flag those parts of the response for correction or re-generation.

Comparative Analysis: Comparison between the Ground truth and the Input text are done as follows-

- **Metrics-Based Evaluation**: The approach includes using various metrics like BLEU, ROUGE, perplexity, and semantic similarity to evaluate the responses and their alignment with ground truths. The use of **sentence transformers** (Sentence-BERT) provides semantic evaluation.

2. **RAG (Retrieval-Augmented Generation) Approach**:

- **RAG Model Integration**: This method uses a RAG framework to improve factual accuracy by retrieving external documents (via Google Custom Search) and generating responses that are grounded in the retrieved content.
- **Iteration for Word Count**: A loop ensures that the generated text reaches a minimum word count (e.g., 2000 words), ensuring that enough detailed information is provided.

Metric 1: BLEU Score (Bilingual Evaluation Understudy)

- **What is BLEU** : BLEU measures the overlap between the generated response and the reference texts (ground truths) by analysing how many words or phrases match between the two.
- **How it Works**:
  - Tokenize both the generated response and the ground truth texts.
  - Count the number of word or phrase matches (up to a 4-gram level, which considers sequences of up to four words).
  - Calculate a precision score, i.e., how many words/phrases in the generated response appear in the ground truths.

Metric 2: ROUGE Score (Recall-Oriented Understudy for Gisting Evaluation)

- **What is ROUGE**: ROUGE measures the recall by evaluating how much of the ground truth is reflected in the generated response. The two commonly used variants are:
  - **ROUGE-1**: Measures word-level overlap.
  - **ROUGE-L**: Measures the longest common subsequence between the generated response and ground truth.
- **How it Works**:
  - ROUGE-1 looks at unigram (single word) matches between the response and the reference.
  - ROUGE-L looks at the longest sequence of words in both the response and reference.

Metric 3: Perplexity

- **What is Perplexity**: Perplexity measures the likelihood or confidence of the language model in its generated response. It reflects how well the model predicts the next word in a sequence.
- **How it Works**:
  - The generated response is tokenized and passed through the model again.
  - The model calculates the likelihood of each word in the response. A high perplexity score indicates uncertainty, meaning the model is "unsure" about the words it generated, which could imply hallucinations.

Metric 4: Semantic Similarity (Using Sentence Transformers)

- **What is Semantic Similarity**: Unlike BLEU and ROUGE, which focus on word overlap, semantic similarity measures the **meaning** of the generated response compared to the ground truths. It evaluates how semantically (in terms of meaning) aligned the generated response is with the references.
- **How it Works**:
  - **Sentence Transformers** (e.g., paraphrase-MiniLM-L6-v2) are used to convert both the generated response and the ground truths into dense vector representations (embeddings).
  - The embeddings are then compared using **cosine similarity**, which measures how similar the two vectors (representing the meaning of the texts) are.

3. **Google Custom Search Engine (CSE)**:

- **Search-Based Fact-Checking**: The Google CSE retrieves real-world data based on the query, allowing for a grounding mechanism to mitigate hallucinations before generating the response.

Explanation:

1. **User Query Input**

- The user provides a **query** (or topic) that will be used to retrieve relevant documents and generate the response.
- This input forms the basis for the information search and the eventual response generation.

2. **Retrieval of External Documents (Google Custom Search)**

- The system retrieves **relevant documents** from external sources (in this case, using the **Google Custom Search API**) to fetch snippets of information based on the user query.
- This retrieved information will be used to guide the language model's response generation, ensuring that the response is grounded in factual data.

3. **RAG Model Integration for Response Generation**

- The **RAG model** combines the retrieved content with a pre-trained language model to generate a detailed response.
- **How it Works**:
  - A **prompt** is created by combining the user's query and the retrieved content (snippets from the Google API).
  - The model generates a response based on this prompt, ensuring that the generated text is influenced by the factual data retrieved from external sources.

4. **Iteration for Minimum Word Count**

- To ensure the generated response is sufficiently detailed, the method iterates through multiple response generations until the **target word count** (e.g., 2000 words) is reached.
- **How it Works**:
  - The loop keeps generating and appending additional responses until the total word count of the generated text meets or exceeds the specified minimum (e.g., 2000 words).
  - Each iteration retrieves additional content, generates new responses, and combines them with the previously generated text.

5. **Final Response**

- Once the minimum word count is achieved, the generated text from all iterations is combined and returned as the final response.
- This final response is **grounded** in the retrieved factual content, reducing hallucinations and improving factual accuracy.

4. **Model Optimization**:

- **Optimizing Code**: There were discussions on making the hallucination detection and mitigation process faster by modifying how inputs are handled, including using batch processing and optimizing response generation.

Results and Discussions

The proposed method was tested on a variety of datasets that contained factual premise-hypothesis pairs. The results show that the prompting-based approach is effective for detecting hallucinations. The key fin-

dings include:

1. **High Precision in Hallucination Detection:** Comparing question-answer pairs led to accurate detection of hallucinated content. In cases where the hypothesis contained factual errors, the method successfully identified them as hallucinations with few false positives.
2. **Handling Hypothesis with Less Detail:** The method addresses the challenge of hypotheses providing less information than premise. The method was effective in distinguishing between incomplete information and actual hallucination, ensuring that omission of non-essential details was not misinterpreted as hallucination.

Future work

The above approaches work for textual data. We are working on detecting and mitigating hallucinations for multi modal data.

References

1. Barkley, L., & van der Merwe, B. (2024). Investigating the Role of Prompting and External Tools in Hallucination Rates of Large Language Models. arXiv preprint arXiv:2410.19385.
2. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems, 35, 24824-24837.
3. Mündler, N., He, J., Jenko, S., & Vechev, M. (2023). Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation. arXiv preprint arXiv:2305.15852
4. S.M Towhidul Islam Tonmoy, S M Mehedi Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, Amitava Das-(towhidulislam@iut-dhaka.edu), A Comprehensive Survey of Hallucination Mitigation Techniques in Large Language Models, arXiv:2401.01313v3 [cs.CL] 8 Jan 2024