

Mitigating Factual Inconsistency and Hallucination in Large Language Models

Muneeswaran I Advait Shankar Varun V
Saisubramaniam Gopalakrishnan Vishal Vaddina
Applied Research, Quantiphi

Abstract

Large Language Models (LLMs) have demonstrated remarkable capabilities in various language-related tasks enabling applications in various fields such as healthcare, education, financial services etc. However, they are prone to producing factually incorrect responses or “hallucinations” which can have detrimental consequences such as loss of credibility, diminished customer trust etc. In this presentation, we showcase a solution that addresses the challenge of minimizing hallucinations. Our solution provides accurate responses and generates detailed explanations, thereby enabling the users to know how the model arrived at the final response. Additionally, it verifies if the explanations are factually correct and offers insights into whether the generated explanations are directly derived from the provided context or if they are inferred from it. We also systematically assess the quality of generated responses using an LLM-based evaluation technique. We present empirical results on benchmark datasets to demonstrate the effectiveness of our approach. Our presentation also examines the impact of individual components in the solution, enhancing the factual correctness of the final response. This research is vital for industries utilizing LLMs, as it provides a means to enhance the reliability of responses and mitigate the risks associated with factual hallucinations. Researchers and practitioners seeking to enhance the reliability of LLM responses will find valuable insights in this presentation.

Current Challenges

Question: How did Mahatma Gandhi use G Suite to organize resistance against British violence?

Assistant:

Mahatma Gandhi used G-suite to organize the resistance movement against British violence in a number of ways.

First, he used G-suite to communicate with his fellow activists. He created a Gmail account and used it to send emails and organize meetings. He also used Google Docs to share documents and collaborate on projects.

Second, he used G-suite to raise awareness of the British violence. He created a website and used it to post articles and videos about the violence. He also used social media to share information about the violence.

Third, he used G-suite to mobilize people to take action against the violence. He created a petition and used it to gather signatures. He also used social media to encourage people to protest against the violence.

Fourth, he used G-Suite to raise funds for the resistance movement. He created a crowdfunding campaign and used it to raise money for the movement. He also used social media to ask people to donate money to the movement.

Note: This example was taken from June 2023. There have been updates done to the model from which this was taken from and it has been 'resolved'

Figure 1. Example of Hallucination in LLMs

Large language Models (LLMs) have emerged as a transformative technology in Natural Language Processing (NLP) and Natural Language Generation (NLG), surpassing traditional methods in both scale and performance. Despite increased model and data scale, state-of-the-art LLMs still struggle with the following:

Hallucination: The stored knowledge can lead to hallucinations if the underlying information either gets outdated, or the LLM hidden activations lead it towards incorrect responses during inference

Factual errors: The issue stems from LLMs memorizing extensive knowledge during pre-training and fine-tuning, termed parametric knowledge

Motivation

Retrieval Augmented Generation (RAG) [3] has been successful in creating responses that are connected to real-world information for natural language questions. However, in business settings, responses need to be easily traced back to their sources, often found in internal databases. Here are some key points to consider:

In enterprise interactions, questions and answers form the cornerstone of communication

Business needs extend beyond basic information retrieval, necessitating responses tailored to specific tasks and domains.

Establishing the provenance of responses is essential for fostering user trust and comprehension

Responses are derived from meticulously curated knowledge repositories, including structured databases, knowledge graphs, and web-based information.

Emphasizing the accuracy of responses is critical, with a focus on minimizing errors, inaccuracies, and fabrications.

Advantages of our Solution

Accurate & truthful responses supported by facts

Ability to trace back the answer to the source

Enables fine-tuned open-src LLMs to compete with GPT-3.5

Rationale directly derived or inferred from context

Architecture

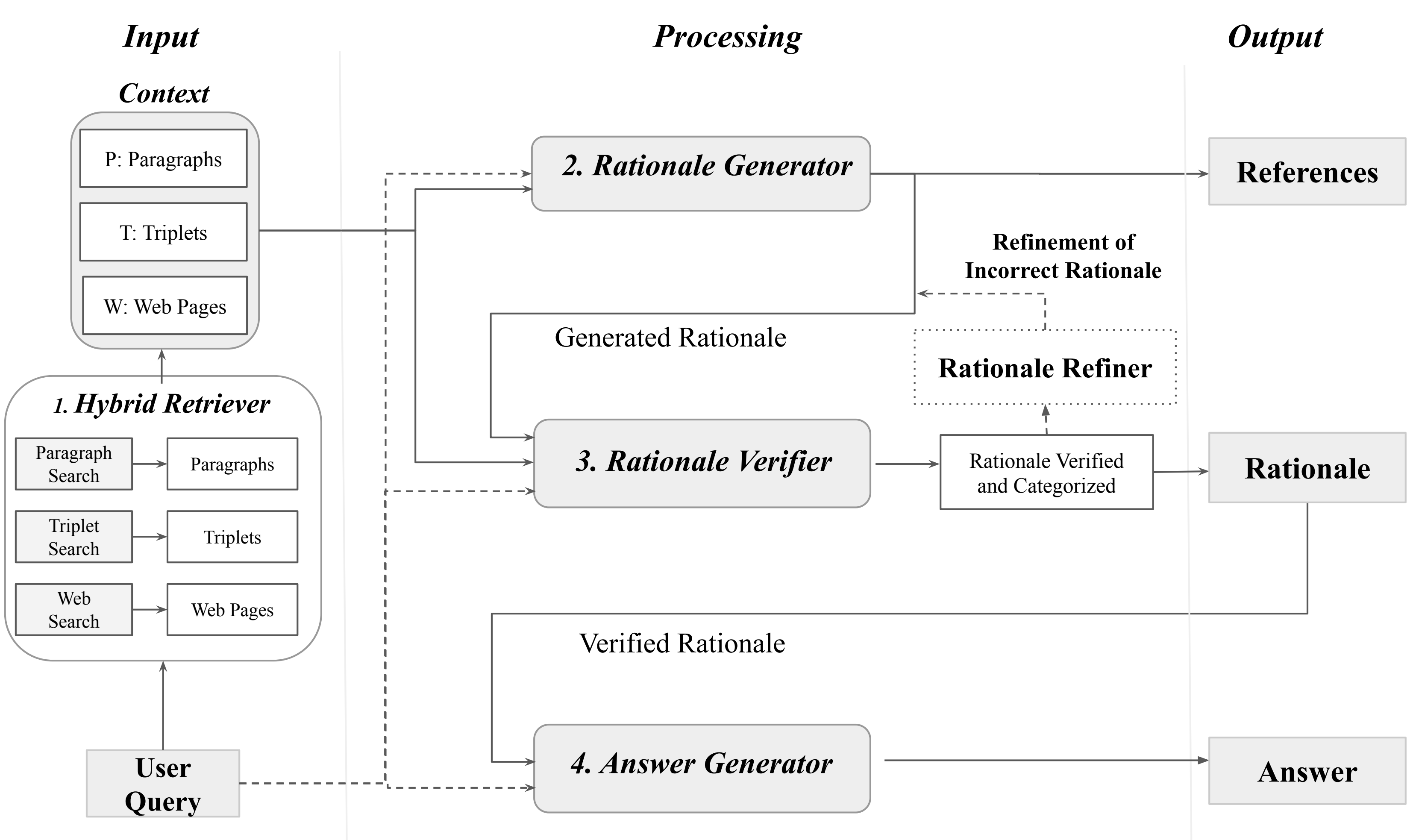


Figure 2. Our proposed Factual Evidence solution, a multi-stage framework showcasing various components - Hybrid Retriever, Rationale Generator, Rationale Verifier, and Refiner

Factual Evidence (FE) is a multi-stage framework, comprising the following components:

HYBRID RETRIEVER

retrieves relevant information (context) based on the user query from various data sources such as documents, knowledge graphs, and the internet

RATIONALE GENERATOR

uses the context and the user query to generate rationale (explanation) with supporting evidence

RATIONALE VERIFIER

verifies the factual accuracy of the rationale and whether it is explicitly or implicitly mentioned in the context, and Rationale Refiner, which revises or refines incorrect rationale based on the context

ANSWER GENERATOR

utilizes the verified and refined rationale to facilitate the LLM to generate the final descriptive response ensuring factual accuracy of the answers

Validation and Results

Question: How did Mahatma Gandhi use G Suite to organize resistance against British violence?

Context:

...Mahatma Gandhi, born on October 2, 1869, in Porbandar, Gujarat, India, had a modest upbringing. He came from a devout Hindu family and imbibed values of truth, nonviolence, and simplicity from a young age...He was assassinated on January 30, 1948, in Delhi..

...Google was founded in 1998 by Larry Page and Sergey Brin. Today's technology landscape is deeply intertwined with Google's vast array of services and innovations. From its powerful search engine that processes billions of queries daily, to its cloud computing platform enabling businesses to scale with ease, Google continues to shape the digital world. ... G-suite was launched on August 28, 2006 by Google...

Rationale:

- "Google Workspace - wiki" "states that G Suite is developed by Google.
- "Google - wiki" states that Google was founded in 1998 by Larry Page and Sergey Brin.
- "Google Workspace - wiki" states that G-suite was launched on August 28, 2006 and this is further supported by "Introducing Google Workspace - Google.com"
- "Mahatma Gandhi | Biography" states that Mahatma Gandhi was born in 1869.
- "Gandhi - The Guardian" states that he was assassinated on January 30, 1948, in Delhi. This implies that Mahatma Gandhi lived during 1869-1948.

Answer:

No, Mahatma Gandhi did not use G Suite as it was launched by Google on August 28, 2006 [c], whereas Mahatma Gandhi lived from 1869 to 1948. [e]

Figure 3. Example of Factual Evidence Response

Table 1 shows our framework improving traditional Retrieval Augmented Generation (RAG) by enabling OpenAI's GPT-3.5-turbo to be 14-25% more faithful and 16-22% more accurate on the PubMedQA [2] and AEQA datasets. Furthermore, fine-tuning samples based on our framework improves the accuracy of smaller open-access LLMs by 33-42%, and competes with RAG on commercial models.

Model	Method	PubMedQA		AEQA	
		Faithful (%)	Accuracy (%)	Faithful (%)	Accuracy (%)
gpt-3.5-turbo	No-Retrieval	56.70	55.92	44.40	65.21
	RAG	72.40	72.67	58.20	75.19
	RAG+FE	86.50	88.49	83.24	96.85
Llama-2-7B	No-Retrieval	8.82	10.30	6.82	3.71
	RAG	7.19	8.92	3.78	5.64
	RAG+FE	27.19	30.80	31.29	35.0
Llama-2-7B-Finetuned	No-Retrieval	14.97	21.20	24.18	20.70
	RAG+FE	70.8	72.23	76.8	68.18

Table 1. Comparison of Factual Evidence with baselines on PubMedQA and AEQA Datasets

References

- Muneeswaran I, Shreya Saxena, Siva Prasad, M V Sai Prakash, Advait Shankar, Varun V, Vishal Vaddina, and Saisubramaniam Gopalakrishnan. Minimizing factual inconsistency and hallucination in large language models. *arXiv preprint arXiv:22311.13878*, 2023.
- Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*, 2019.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.
- Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. Siren's song in the ai ocean: A survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*, 2023.