

Analyzing and Improving Post-hoc Approaches for the Detection and Correction of Hallucinations in Long-form Text Generation

Ihsan Soydemir

06.05.2024, Master Thesis Kickoff

Chair of Software Engineering for Business Information Systems (sebis)

Department of Computer Science

School of Computation, Information and Technology (CIT)

Technical University of Munich (TUM)

www.matthes.in.tum.de

Analyzing and Improving Post-hoc Approaches for the Detection and Correction of Hallucinations in Long-form Text Generation

Motivation

- Inevitability of hallucinations
- Fictional references
- Importance for critical applications
- Improving correction mechanisms
- Building trust for LLMs



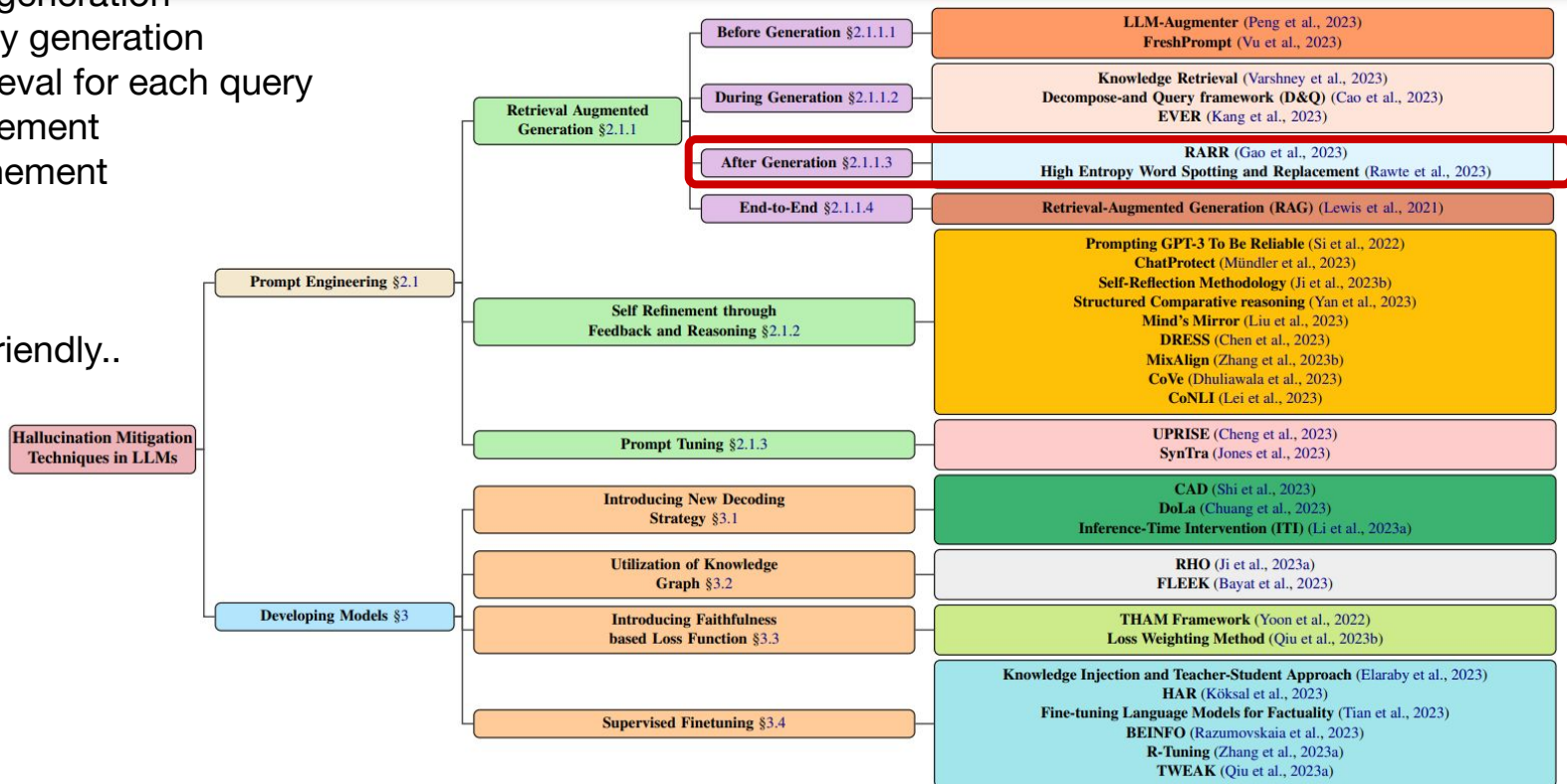
The screenshot shows a chat interface with a dark background. At the top, a user profile icon is followed by the text "You" and the message "Name some politicians born in NY". Below this, the ChatGPT logo is followed by the text "ChatGPT" and the response "Certainly! Here are some politicians who were born in New York:". The response is followed by a numbered list of four names: 1. Donald Trump - Former President of the United States, 2. Hillary Clinton - Former U.S. Secretary of State and Democratic Presidential Nominee, 3. Franklin D. Roosevelt - Former President of the United States, and 4. Theodore Roosevelt - Former President of the United States.

Existing Work

LLM hallucination detection and correction techniques

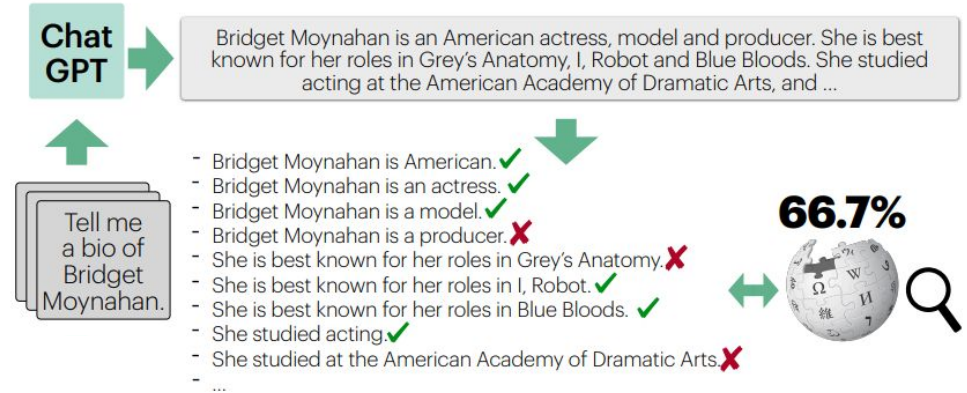
- RAG after generation
 - Query generation
 - Retrieval for each query
 - Agreement
 - Refinement

- Applicat...
- Powerful
- resource-friendly..
-



Studies without refinement

- SelfCheckGPT
- FACTOR
- FactScore
- FacTool
 - performs refinement; however, it requires human-decomposed claims
- Long-Form Factuality in LLMs (27.03.2024) [DeepMind]



- Router Mechanism (between 2. and 3.)
 - Wiki Data (where questions that expect answers in the form of a list of entities)
 - Multi-span (“Who invented the first mechanized printing press and in what year?”)
 - Long-form
- Internal Knowledge or Online Search

Query

Name some politicians who were born in NY, New York.

1. Baseline Response

Here are some politicians who were born in NY, New York:

1. Hillary Clinton - former secretary of state and former Democratic presidential nominee
 2. Donald Trump - former president of the United States
 3. Michael Bloomberg - former Mayor of New York City and former Democratic presidential candidate
- ... <list continues..>

2. Plan Verifications

- Where was Hillary Clinton born?
 - Where was Donald Trump born?
 - Where was Michael Bloomberg born?
- ... <questions continue..>

3. Execute Verifications

Hillary Clinton was born in **Chicago, Illinois**, United States on October 26, 1947.

Donald Trump was born on June 14, 1946, in **Queens, New York City, New York**, United States

Michael Bloomberg was born on February 14, 1942, in **Boston, Massachusetts**, United States.

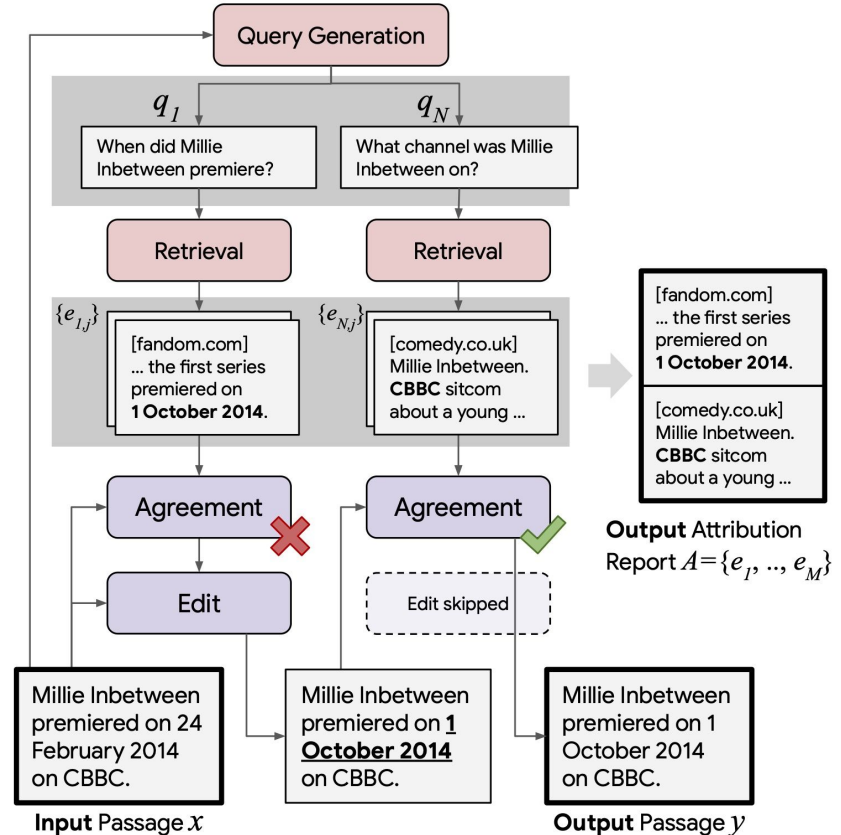
...

4. Final Verified Response

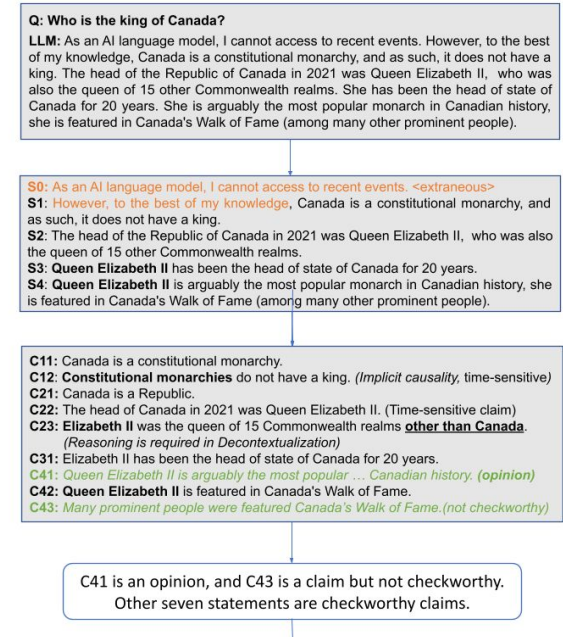
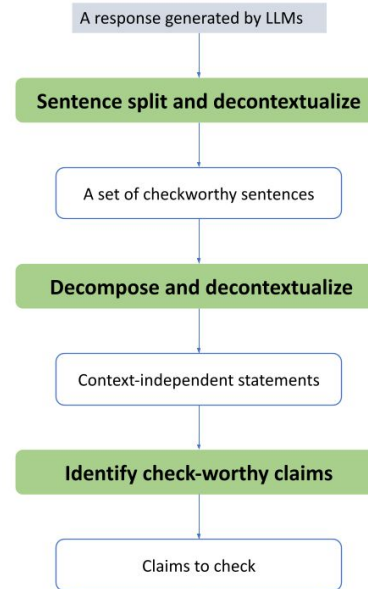
Here are some politicians who were born in NY, New York:

1. Donald Trump - former president of the United States
 2. Alexandria Ocasio-Cortez - Democratic member of the U.S. House of Representatives
- ... <list continues..>

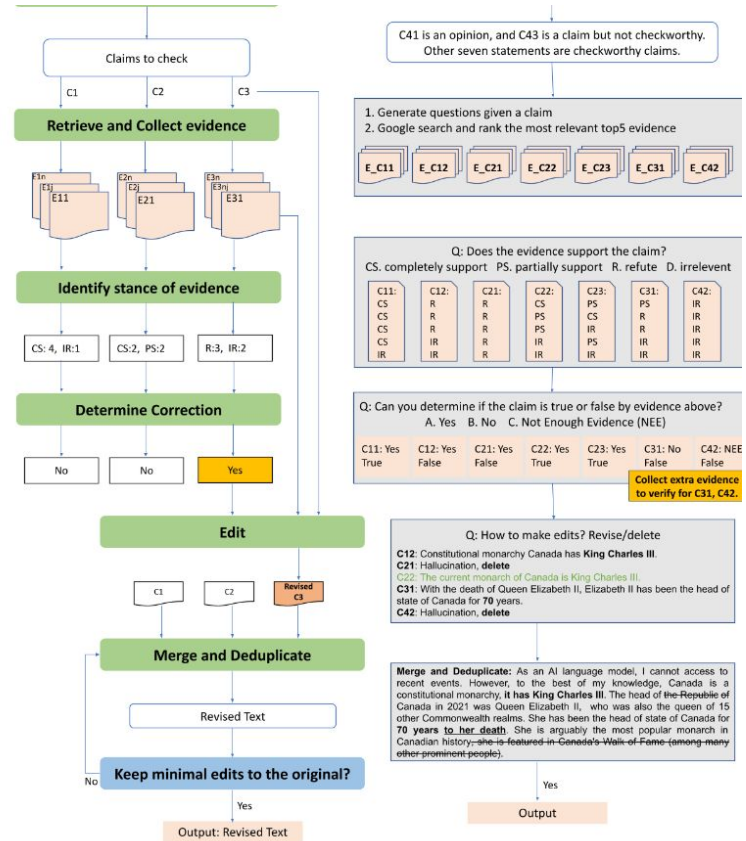
- 4-Step Process:
 - Question Generation
 - Evidence Collection
 - Agreement Test
 - Editing
- Bing search
- Another LLM to modify the initial claim according to verified evidence



- 7-steps
 - decomposition
 - decontextualisation
 - checkworthiness identification



- 7-steps
 - evidence retrieval and collection
 - stance detection
 - correction determination
 - claim correction



Gaps on long-context (After 50+ experiments with long text)

CoVe (DuckDuckGo Search)

- Minimal changes with knowledge source
- Sometimes reasoning fine but lacks refinement


CoVe (Internal Knowledge)

- No access to real-time data

RaRR (Bing Search)

- Changes original text significantly
- Excessive refinement

FactCheck-GPT (Google Search)

- Generic questions (decontextualization )
- Not specific enough in many cases
- Limited usefulness in certain contexts

Research Questions

Research Questions

- What is an appropriate taxonomy for categorizing hallucinations in LLMs (i.e. numeric, semantic hallucinations)?
- How can the Retrieval-Augmented Generation (RAG) model be effectively applied to handle long contexts without compromising performance or efficiency?
- How can we optimally balance editing and faithfulness when refining generated text?

Proposal

Summedits

A Chinese mine collapse killed at least ~~three~~ **two** people and left over 50 others missing. Chinese officials say numerous ~~vehicles~~ **animals** were also buried in the collapse.

Longform Generation of Biographies (from CoVe)

Prompt: "What were the key factors that led to the fall of imperialism in the 20th century?" / "Tell me a bio of Marie Curie."

Generated Text: The fall of imperialism in the 20th century can be attributed to several key factors:

1. World Wars
2. Nationalism [...]

AMRFACT:

US President Donald Trump has said he will ~~consider~~ **fire** special counsel Robert Mueller, who is investigating alleged Russian interference in the US election

Long Form Factuality (from DeepMind)

Task: Given a question about a general concept generate a comprehensive answer covering all relevant aspects of the question
Generated Question: Could you explain the fundamentals of Natural Language Processing (NLP) in Computer Science, its key techniques[..], and discuss [...] in advancing fields like machine translation, sentiment analysis, and information retrieval?

Gaps

Bird flu was not deemed to be a threat to humans until 1997, after an outbreak in Hong Kong. Since then, around 870 infections have been reported worldwide, with 457 deaths in 21 countries.

FactCheck-GPT

Bird flu was not considered a real threat to humans until a 1997 outbreak in Thailand.

457 deaths in 21 countries have been reported

Where did the 1997 outbreak of bird flu occur?
When was the first recognized outbreak of bird flu in humans?

How many deaths have been reported in 21 countries?
In how many countries have deaths been reported?

Generate Answer

The given text states that there have been 457 deaths in 21 countries. The evidences provided mention that as of the same date, the virus had infected over **557.8 million** people worldwide and the number of deaths had totaled more than 6.3 million. Therefore, the given text seems to be inaccurate, as the number of deaths mentioned is much lower than the actual number

Gaps

A Chinese mine collapse killed at least three people and left over 50 others missing. Chinese officials say numerous **animals** ~~vehicles~~ were also buried in the collapse.

RARR

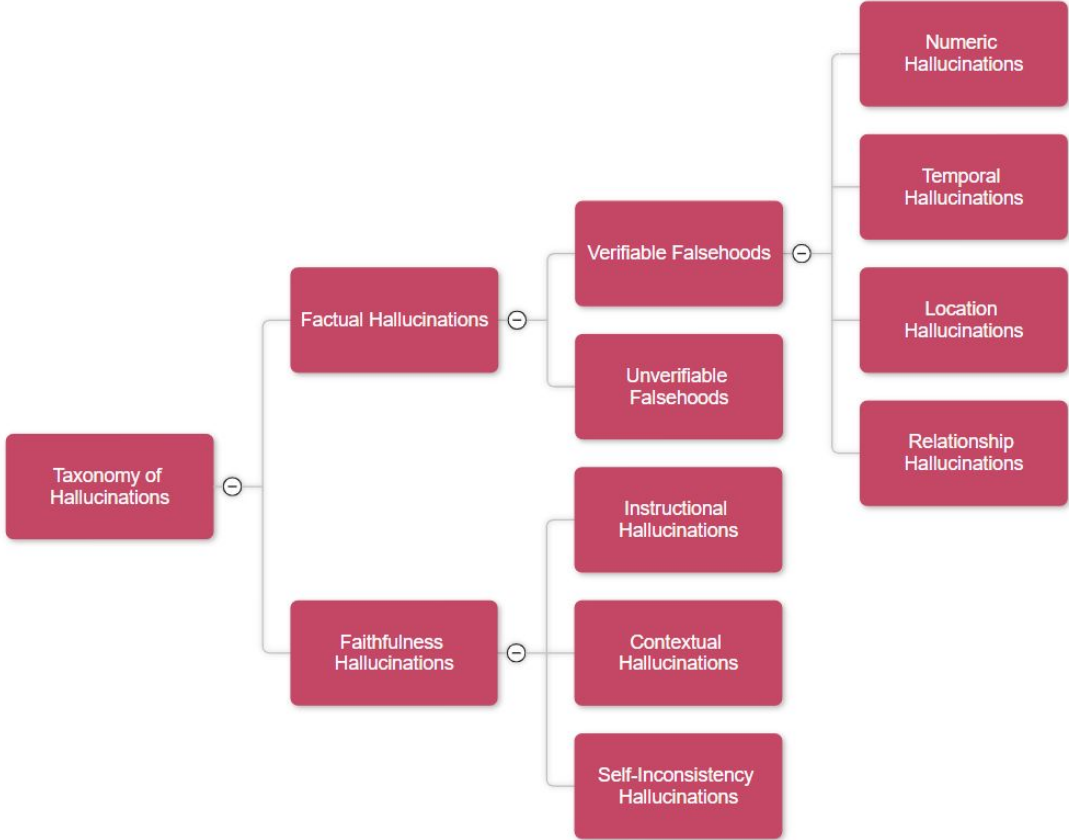
Generate Questions

How many people are still missing in the Chinese mine collapse?
How many people were killed in the Chinese mine collapse
How many people were reported missing in the Chinese mine collapse
Were any animals affected in the Chinese mine collapse
What did Chinese officials say about the animals in the Chinese mine collapse

Generate Answer

A Chinese mine collapse on Thursday (Feb. 23) killed multiple people and left dozens more missing. Chinese officials say the collapse affected a "wide area" and left a pile of debris roughly 500m (1,640ft) across and an estimated 80m high. The shaft was operated by the Xinjing Coal Mining Company, which has not issued a statement yet. Mines in China's Inner Mongolia region are some of China's top coal producers.

Taxonomy of Hallucinations

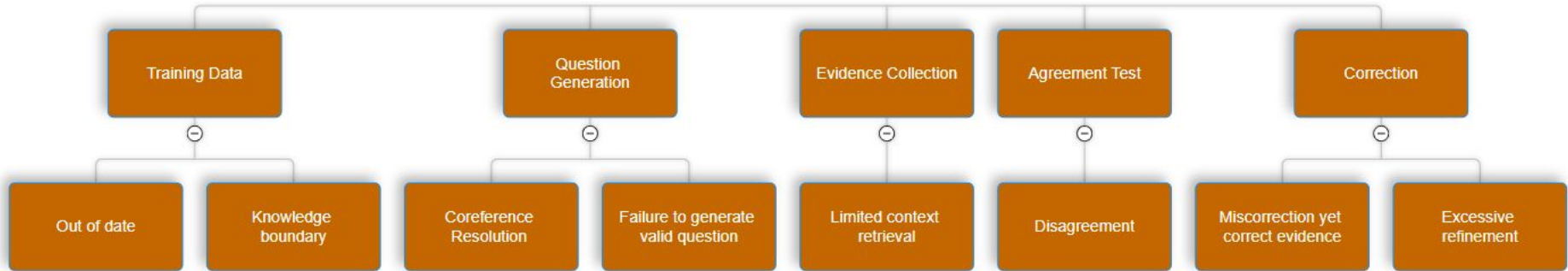


Hallucination Types

I. Factual Hallucinations: Hallucinations that involve incorrect or fabricated information about the world.

- **Verifiable Falsehoods:** Hallucinations that contradict specific, verifiable facts (e.g., "**The capital of France is Berlin**").
 - Location-based Hallucinations: e.g., "The Eiffel Tower is located in Rome"
 - Numeric Hallucinations: e.g., "The speed of light is 300 km/h"
 - Temporal Hallucinations: e.g., "The American Revolution occurred in 1850"
 - Relationship Hallucinations: e.g., "Albert Einstein was a student of Marie Curie"
- **Unverifiable Falsehoods:** Hallucinations that invent new, unverifiable information (e.g., "**The city of Atlantis is located on the moon**").

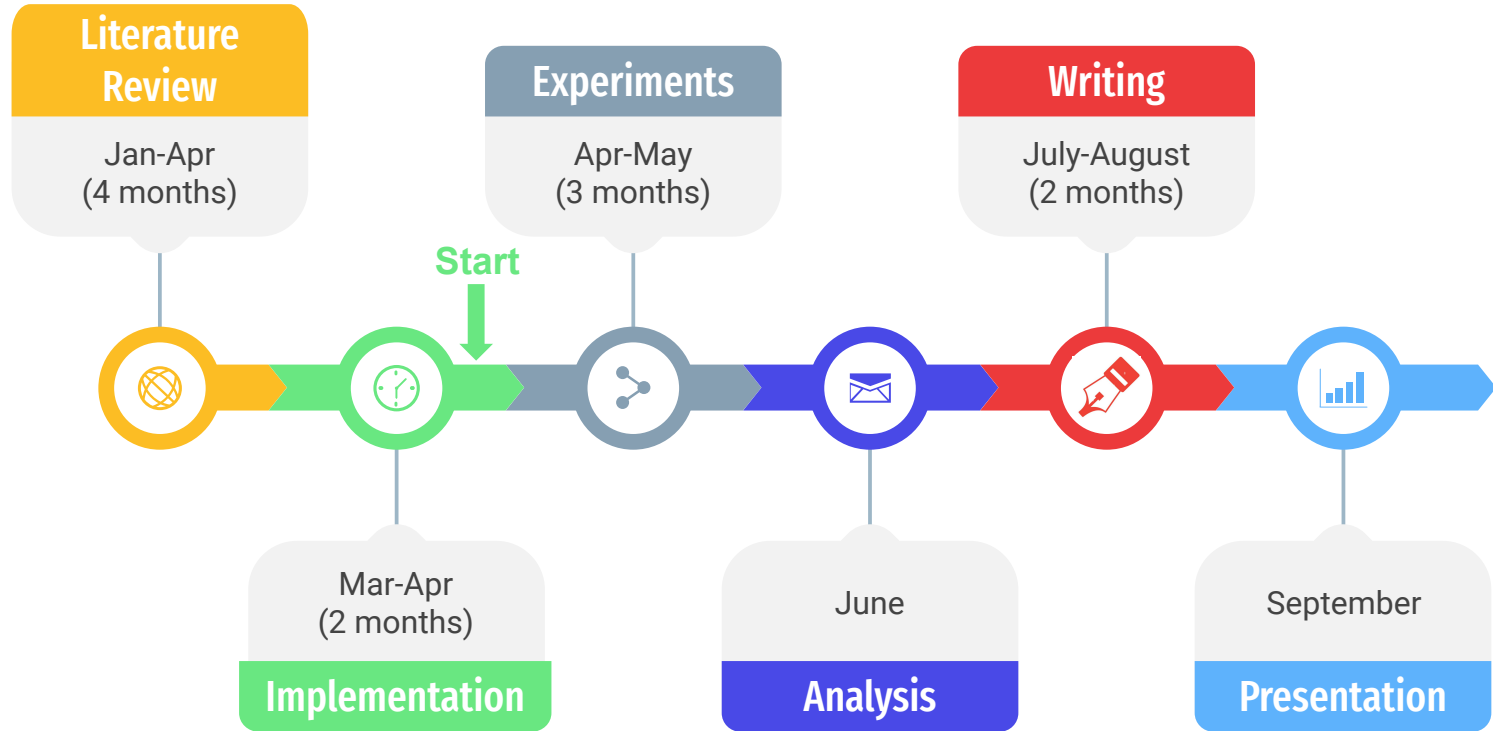
Causes of Hallucination on Post-hoc approaches for long-form generation



- Automatic evaluation (FActScore, f1@, LLM-based, G-Eval)
- Human evaluation
 - Which refinement you liked most?
 - Voting arena (i.e. chat.lmsys.org)
- Correction quality (with context)
 - Semantic Similarity
 - BLEU Score, Edit Distance

$$f(y) = \frac{1}{|\mathcal{A}_y|} \sum_{a \in \mathcal{A}_y} \mathbb{I}[a \text{ is supported by } \mathcal{C}],$$

$$\text{FACTSCORE}(\mathcal{M}) = \mathbb{E}_{x \in \mathcal{X}} [f(\mathcal{M}_x) | \mathcal{M}_x \text{ responds}]$$





Ihsan Soydemir
i.soydemir@tum.de

Technical University of Munich (TUM)
TUM School of CIT
Department of Computer Science (CS)
Chair of Software Engineering for
Business Information Systems (sebis)

Boltzmannstraße 3
85748 Garching bei München

+49.89.289.17132
matthes@in.umt.de
www.matthes.in.tum.de

