Tonmoy et al (2024): A Comprehensive Survey of Hallucination Mitigation Techniques in Large Language Models

Presented by: Kristýna Klesnilová
LLMs Hallucinations

- Generating ungrounded content that appears factual

- Problematic areas:
  - Legal
  - Medical
  - Financial
  - ...

- Causes
  - Training method
  - Absence of real-time internet updates
Hallucination Mitigation Techniques

Prompt Engineering §2.1
- Hallucination Mitigation Techniques in LLMs
  - Prompt Tuning §2.1.3
    - Introducing New Decoding Strategy §3.1
    - Use of Knowledge Graph §3.2
    - Introducing Faithfulness based Loss Function §3.3
  - Utilization of Knowledge Graph §3.2
  - Introducing Faithfulness based Loss Function §3.3
- Developing Models §3
  - Supervised Fine-tuning §3.4
  - CAD (Shi et al., 2023)
  - DDLa (Chuang et al., 2021)
  - Inference-Time Intervention (ITI) (Li et al., 2023a)
  - RHO (Li et al., 2023a)
  - FLEEK (Royat et al., 2023)
  - THAM Framework (Yoon et al., 2022)
  - Loss Weighting Method (Qi et al., 2023b)
- Before Generation §2.1.1.1
  - Retrieval Augmented Generation §2.1.1
  - Self-Refinement through Feedback and Reasoning §2.1.2
- During Generation §2.1.1.2
  - Knowledge Retrieval (Voskev et al., 2023)
  - Decompose-and Query framework (D&Q) (Cao et al., 2023)
  - EVER (Kang et al., 2023)
  - RARR (Gao et al., 2023)
  - High Entropy Word Spotting and Replacement (Rausue et al., 2023)
- After Generation §2.1.1.3
  - LLM-Augmentor (Peng et al., 2023)
  - FreshPrompt (Vu et al., 2023)
  - Prompting GPT-3 To Be Reliable (Li et al., 2023)
  - ChatProtect (Mindl et al., 2023)
  - Self-Reflection Methodology (Li et al., 2023b)
  - Structured Comparative reasoning (Xue et al., 2023)
  - Mind’s Mirror (Lin et al., 2023)
  - DRESS (Chen et al., 2023)
  - MAlign (Zhang et al., 2023a)
  - CoVe (Rahman et al., 2023)
  - CaNI (Lei et al., 2023)
- End-to-End §2.1.1.4
  - UPRISE (Cheng et al., 2023)
  - SynTra (Jones et al., 2023)
  - CAD (Shi et al., 2023)
  - DDLa (Chuang et al., 2021)
  - Inference-Time Intervention (ITI) (Li et al., 2023a)
  - RHO (Li et al., 2023a)
  - FLEEK (Royat et al., 2023)
  - THAM Framework (Yoon et al., 2022)
  - Loss Weighting Method (Qi et al., 2023b)
  - Knowledge Injection and Teacher-Student Approach (Elazaby et al., 2023)
  - HAR (Kikula et al., 2023)
  - Fine-tuning Language Models for Factuality (Tian et al., 2023)
  - BEINFO (Reznikovskaya et al., 2023)
  - R-Tuning (Zhang et al., 2023a)
  - TWEAK (Qin et al., 2023a)
Prompt Engineering
Prompt Engineering: Retrieval Augmented Generation

- Before Generation
  - Retrieving external knowledge and adding it to LLM prompt

- During Generation
  - Retrieving external knowledge at sentence-by-sentence level
Knowledge Retrieval

Write an article on Rick Mahler

Model

Rick Mahler was a Major League Baseball pitcher who was born in 1953 in Jupiter, Florida.

Append the repaired sentence to the prompt and continue generating the next sentence

Detect and Mitigate Hallucinations

Identify Key Concepts

Calculate Model's Uncertainty

Validation

Create Validation Question

Find Relevant Knowledge

Answer Validation Question

Fix Hallucinated Sentence (using retrieved knowledge)

Rick Mahler was a Major League Baseball pitcher who was born in 1953 in Austin, Texas.
Prompt Engineering: Retrieval Augmented Generation

- After Generation
  - Post-editing LLM output with retrieved external knowledge
  - Replacing high-entropy words or phrases (open-source LLMs)

- End-to-End
  - RAG
    - Generator (seq2seq transformer) and document retriever trained jointly end-to-end
End-to-End RAG

Define "middle ear" (\(x\))
Question Answering: Question Query
Barack Obama was born in Hawaii. (\(x\))
Fact Verification: Fact Query
The Divine Comedy (\(x\))
Jeopardy Question Generation: Answer Query

End-to-End Backprop through \(q\) and \(p_\theta\)

Query Encoder

Retriever \(p_\eta\) (Non-Parametric)

Document Index

Generator \(p_\phi\) (Parametric)

Marginalize

The middle ear includes the tympanic cavity and the three ossicles. (\(y\))
Question Answering: Answer Generation
supports (\(y\))
Fact Verification: Label Generation
This 14th century work is divided into 3 sections: "Inferno", "Purgatorio" & "Paradiso" (\(y\))
Question Generation
Prompt Engineering: Self-refinement through feedback + reasoning

- Prompting strategies
  - Improving reliability
  - Detecting and mitigating self-contradictions through reasoning

- Improving through iterative feedback
  - Generate-score-refine loop

- Chain-of-Verification (CoVe)
  - Automatic fact-checking through verification questions

- Chain of Natural Language inference (CoNLI)
Self-Reflection

**Medical Question**

What causes Noonan syndrome?

**Acquire Knowledge**

Prompt: Provide background knowledge to answer the given question: "What causes Noonan syndrome?"

**Refine Knowledge**

Prompt: The factuality score for the knowledge is -1.2 (less than THRESHOLD_FACTUAL), which means the knowledge is not strongly supported by empirical evidence. Please refine the knowledge to improve its factuality.

**Try Answering**

Prompt: Refer to the knowledge: "(final_knowledge)" and answer the question: "What causes Noonan syndrome?" with one paragraph.

**Generated Knowledge**

Noonan syndrome is a genetic disorder that is caused by a mutation in a specific gene called the "papillon gene."....

**Generated Answer**

Noonan syndrome is caused by mutations or changes in certain genes....

**Refine Answer**

Prompt: The consistency score for the knowledge is -5.1 (less than THRESHOLD_CONS), which means the alignment and consistency between response and knowledge are low. Please refine the response to improve its consistency.

**Final Answer**

Noonan syndrome is caused by a genetic mutation on chromosome 12...
Chain-of-Verification (CoVe)

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..>
Chain of Natural Language inference (CoNLI)
Prompt Engineering: Prompt Tuning

Prompt Design

- Language Model
- Summarize the...
- Engineered Natural Language Prompt
- Task Input

Prompt Tuning

- Language Model
- Tunable Soft Prompt
- 0.10, 0.05, ...
- Task Input

Prompt Retrieval

- Language Model
- Tune
- Answer this...
- Retrieved Natural Language Prompt
- Task Input
- Prompt Retriever
- Tunable Prompt Retriever
Developing Models
Developing Models: New Decoding Strategy

Context-aware Decoding

\[
(1 + \alpha) \logit(y | c, x) - \alpha \logit(y | x)
\]

Example:
- Query: How many World Cups have Argentina won?

LLaMA-7B

- 32nd layer
- 24th layer
- 16th layer
- 8th layer

Contrast:
- Seattle
- Olympia
- Vancouver
- Spokane

Decoding by Contrasting Layers
Developing Models: Utilization of Knowledge Graph
Developing Models: Introducing faithfulness based loss function

- Designing new metrics
  - How closely model’s outputs match input data or ground truth labels
Developing Models: Supervised fine-tuning (SFT)
Developing Models: Supervised fine-tuning (SFT)

A question from TriviaQA


**Document:** Judy Blume has written many novels, including Forever, Tiger Eyes, Are You There God? It's Me, Margaret, Freckle Juice, and Blubber, which are popular among young women. Her books are so popular, that they have been translated into 31 different languages. She is the author of the fiction books Lace (published in 1982), Lace 2 (1985), Savages (1987), Crimson (1992), Tiger Eyes (1994), Revenge of Mimi Quinn (1998) and The Amazing Umbrella Shop (1990). She has also written many other books for children and adults. She is the recipient of the Library of Congress Living Legends award.

**Answer:** Judy Blume
(Shirley Conran)
Developing Models: Supervised fine-tuning (SFT)