Hallucination in LLMs

Ashok Urlana
Researcher, TCS

Research

Agenda

- **Part-1:** Hallucination detection in foundation models
- **Part-2:** Hallucination detection and mitigation in RAG
- **Part-3:** Probing LLMs for limitations generation for scientific papers

Part-1: Hallucination detection in foundation models

What is hallucination?

AI hallucination is a phenomenon where a language model generates factually **incorrect** or **misleading** content.

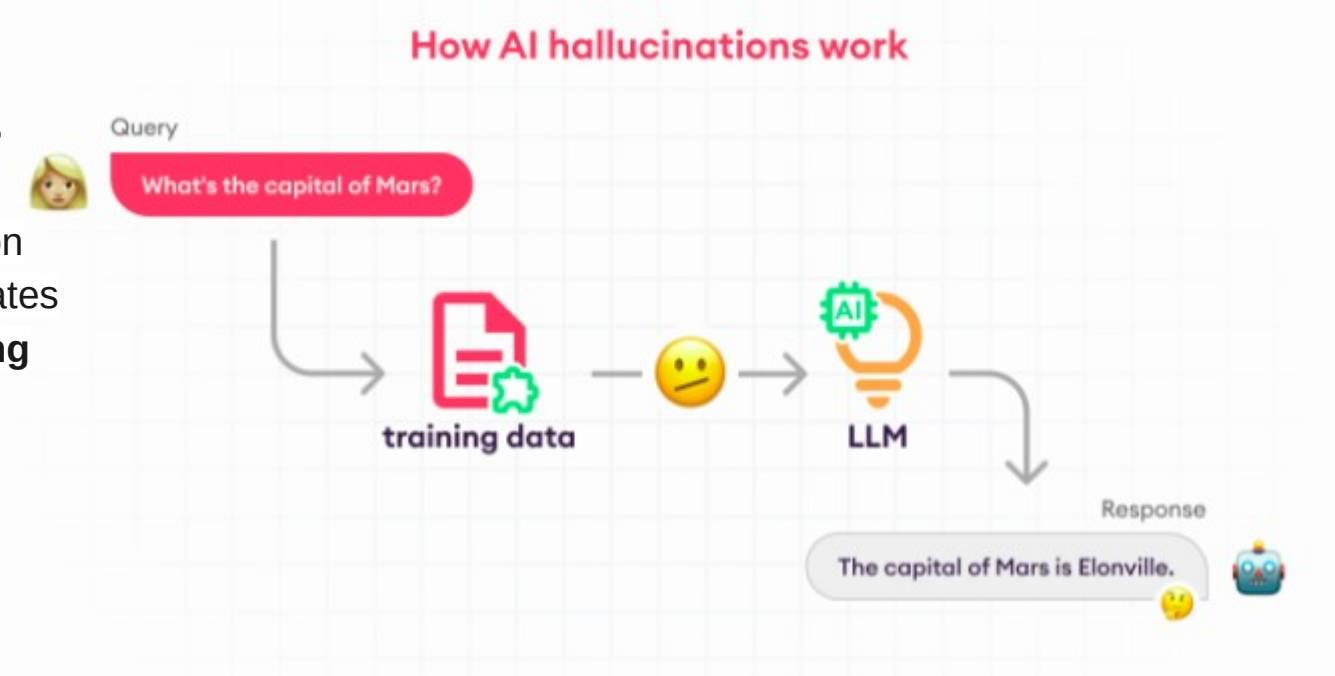
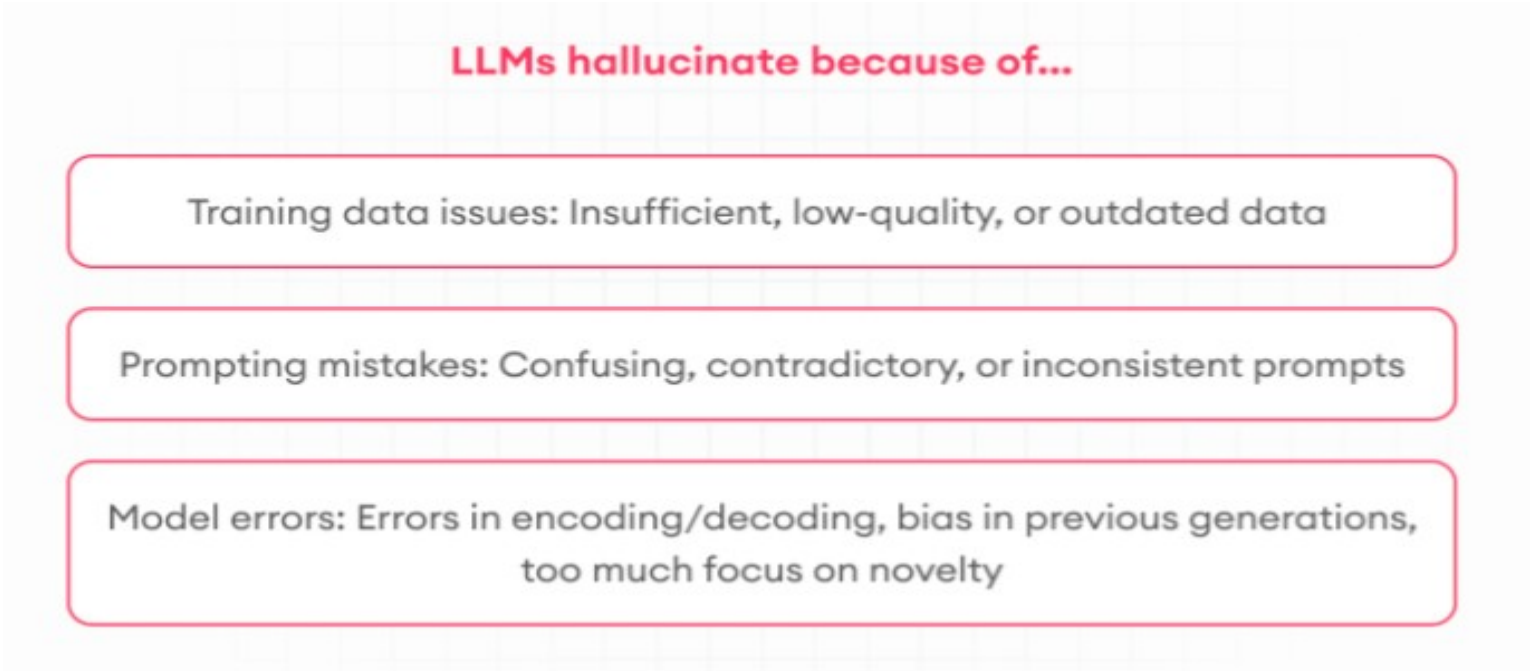


Image source: https://www.superannotate.com/blog/ai-hallucinations

Why LLM Hallucinates?



Example: Question-Answering

Knowledge: The Mosuo Sisters is a 2012 documentary film written and directed by Marlo Poras. Quantum Hoops is a 2007 documentary film directed by Rick Greenwald, that follows the California Institute of Technology's basketball team—the Caltech Beavers—in their attempts to end a 21-year losing streak during the final week of the 2006 basketball season.

Question: Were the documentaries The Mosuo Sisters and Quantum Hoops released in the same year?,

Right answer: No

Hallucinated answer: "Yes, The Mosuo Sisters and Quantum Hoops were released in the same year."

TinyLLaMA 10 sample answers:

1. The answer is no. The Mosuo Sisters, directed by Marlo Poras in 2012, was released in 2012, while Quantum Hoops, directed by Rick Greenwald in 2007, was released in 2007.",
2. "Yes, The Mosuo Sisters and Quantum Hoops were released in the same year (2012).",
3. "No, The Mosuo Sisters was not released in the same year as Quantum Hoops. The Mosuo Sisters was released in 2012, while Quantum Hoops was released in 2007.",
4. "Yes, The Mosuo Sisters and Quantum Hoops were released in the same year, 2012.",
5. "The answer to the question is \"no\". The Mosuo Sisters was released in 2012, while Quantum Hoops was released in 2007.",
6. "Yes, The Mosuo Sisters and Quantum Hoops were released in the same year, 2012.",
7. "The answer to the question is no. The Mosuo Sisters was released in 2012, while Quantum Hoops was released in 2007.",
8. "The answer is \"no.\" While The Mosuo Sisters and Quantum Hoops both share the year 2012, they are not released in the same year. The Mosuo Sisters, released in 2012, follows the Mosuo Sisters, a group of six elderly women who have lived in the Mosuo Autonomous Prefecture in northern China for over 70 years. Quantum Hoops, released in 2007,",
9. "Yes, \"knowledge\" and \"question\" refer to the same thing. The two items are part of a single sentence and cannot be separated.",
10. "The given response, \"knowledge,\" is the correct answer. The provided question is not related to the provided knowledge."

ROUGE

Right	Hallucinated
1. 0.04	0.3
2. 0	0.96
3. 0.05	0.41
4. 0	0.96
5. 0.06	0.38
6. 0	0.96
7. 0.06	0.38
8. 0.02	0.29
9. 0	0.19
10.0	0.09

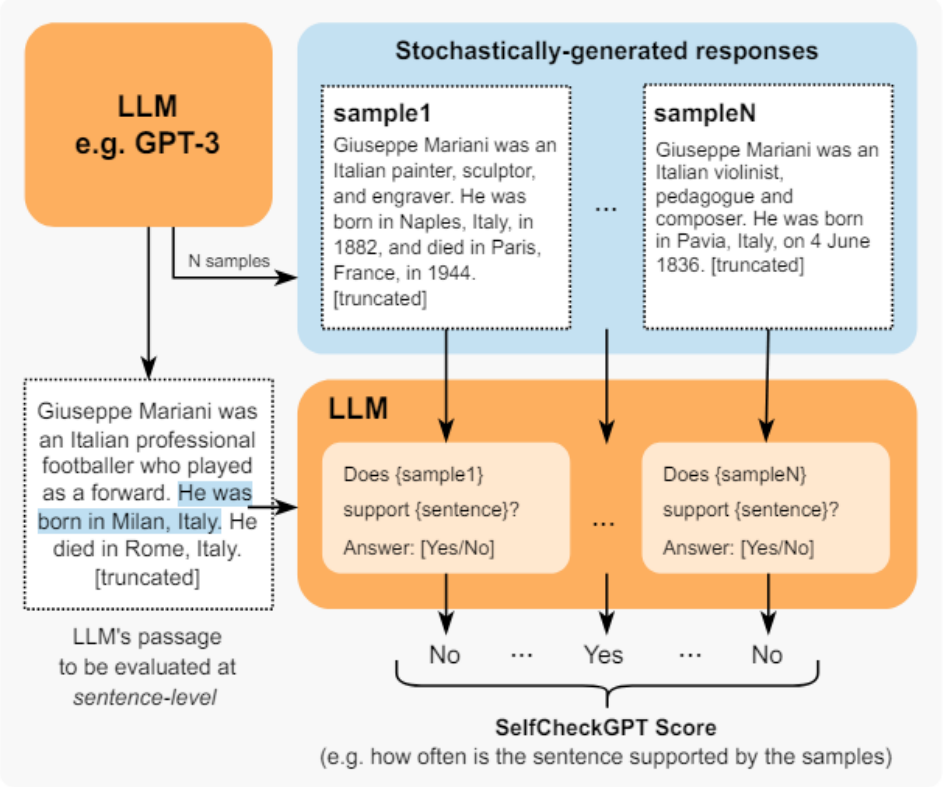
NLI score for Right answer: 0.54, NLI score for Hallucinated answer: 0.70

Reference-free Hallucination

A metric that relies on an underlying model and does not require human-annotated ground truth or reference data to assess the performance of another model generated output.

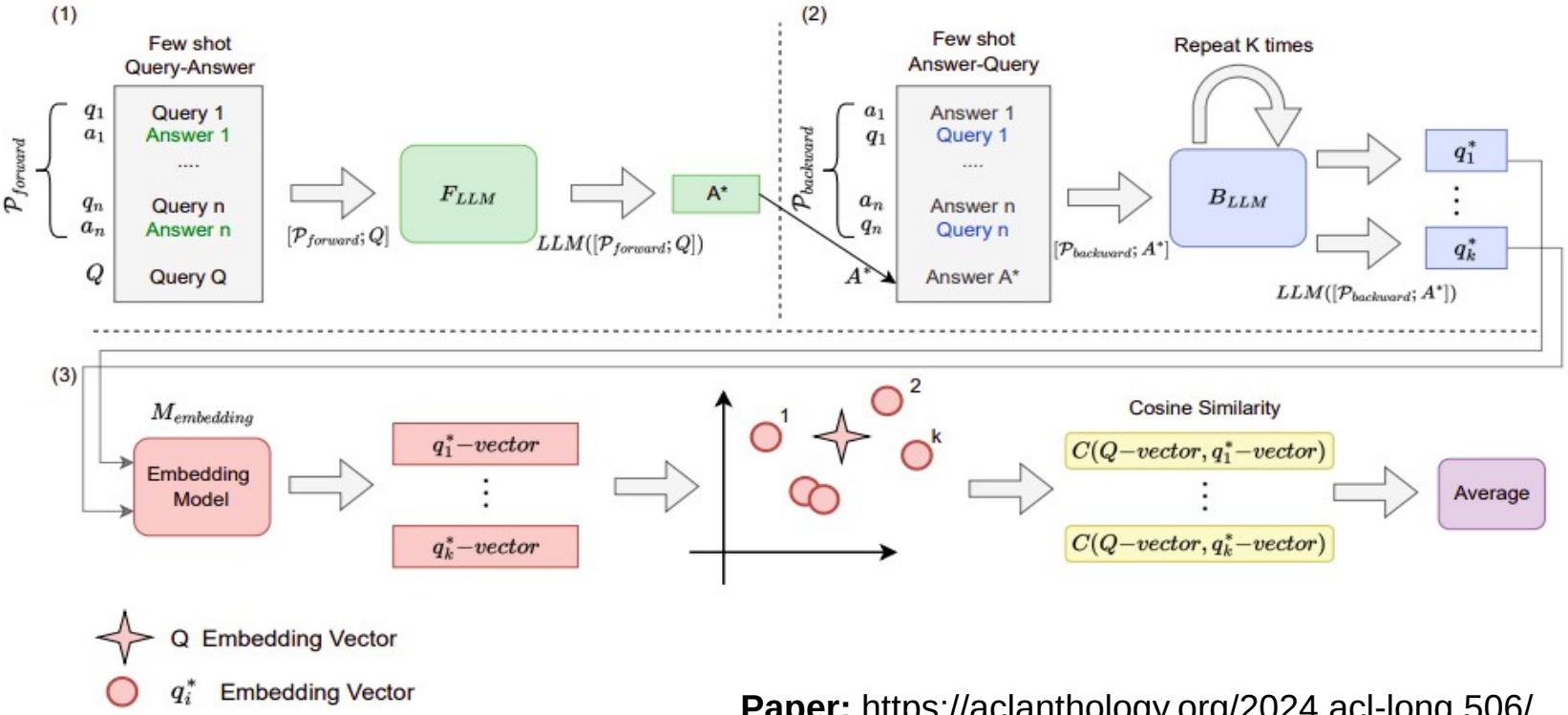
SelfCheckGPT

The idea of **SelfCheckGPT** is that when an LLM has been trained on a given concept, the sampled responses are likely to be similar and contain consistent facts



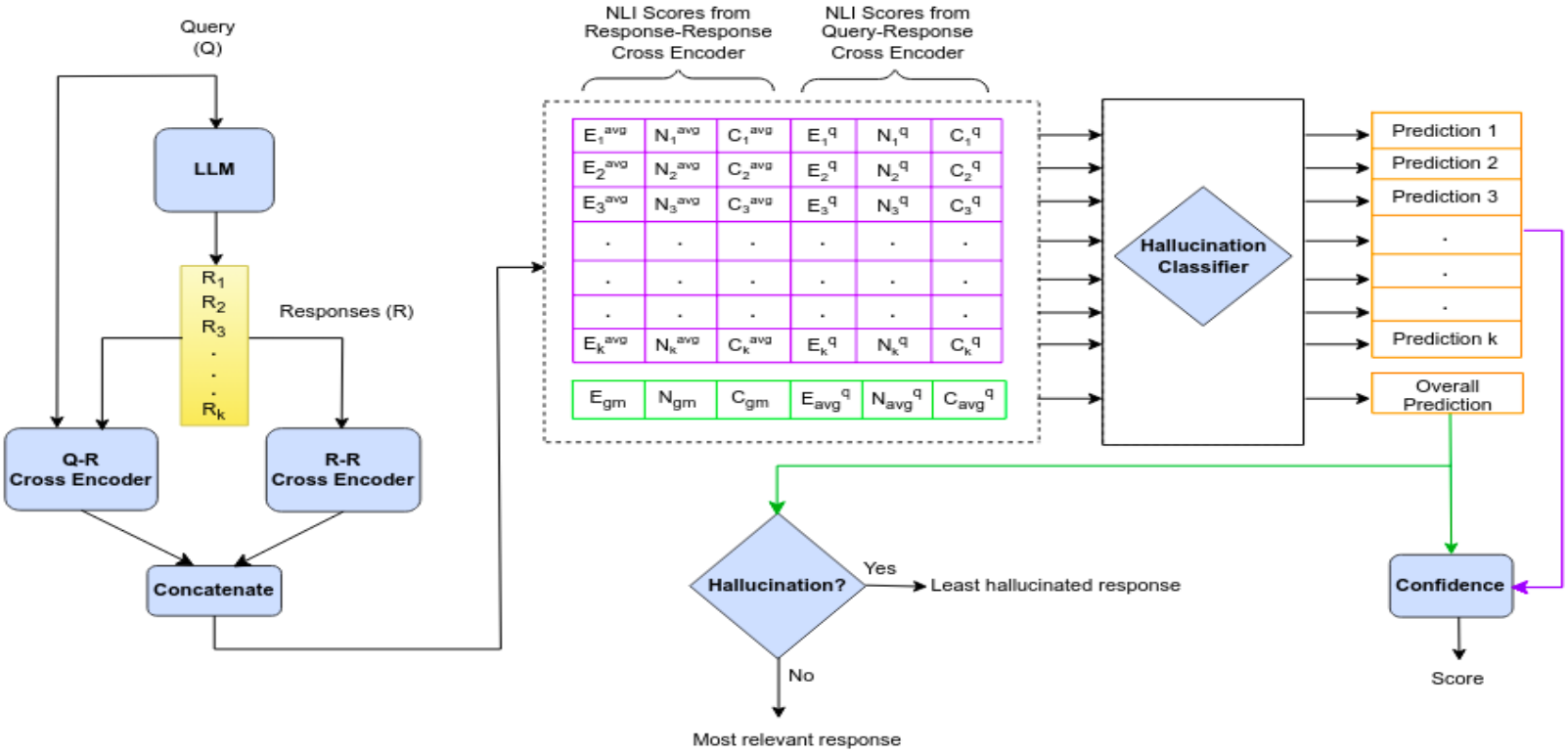
Paper: https://aclanthology.org/2023.emnlp-main.557

Interrogate LLM

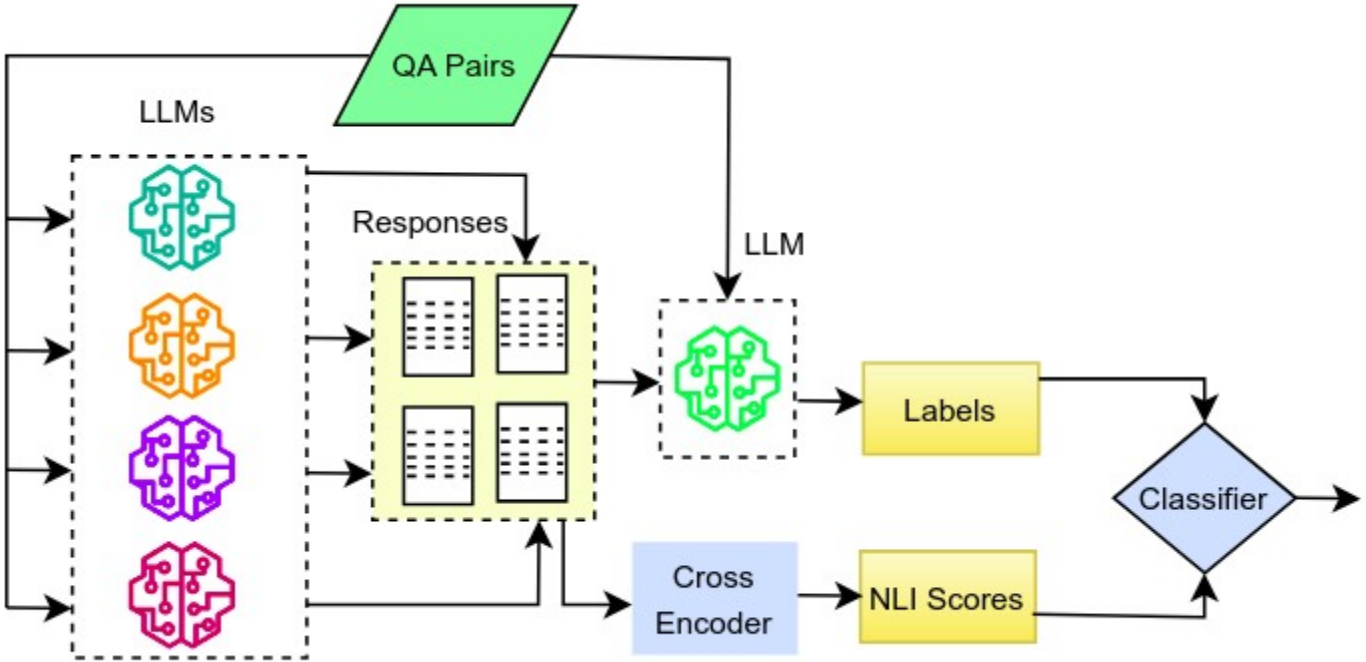


Paper: https://aclanthology.org/2024.acl-long.506/

HaluGuard: Reference-free Hallucination Pipeline



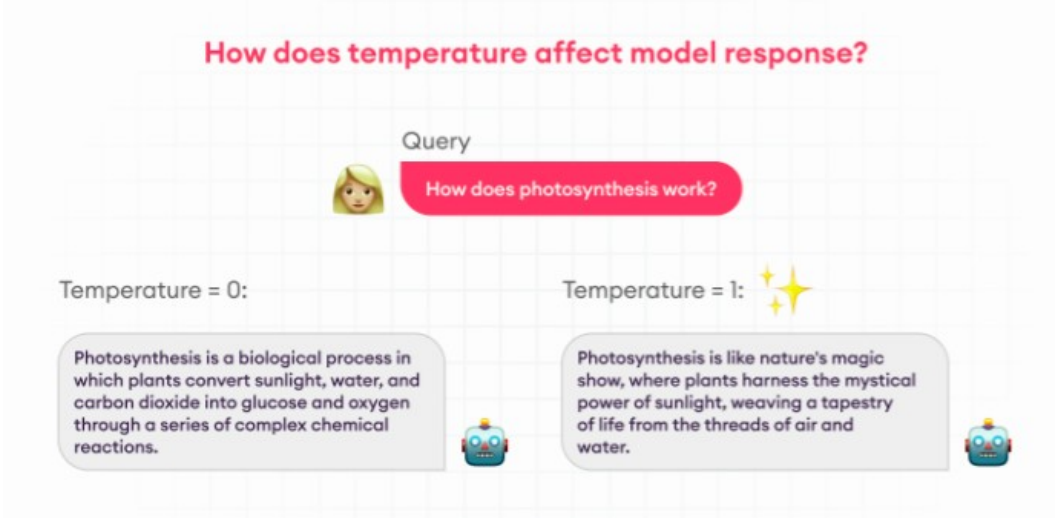
Hallucination Classifier



Takeaways

How to avoid hallucination

- Prompting
  - Use clear prompts
  - Provide relevant information
  - Give a role to the model
- Data
  - Use diverse and relevant data
  - Experiment with temperature
  - Experiment with sample responses
- . Model
  - Utilize RAG
  - Layer-wise analysis by building novel model architectures



Retrieval Augmented Generation (RAG)

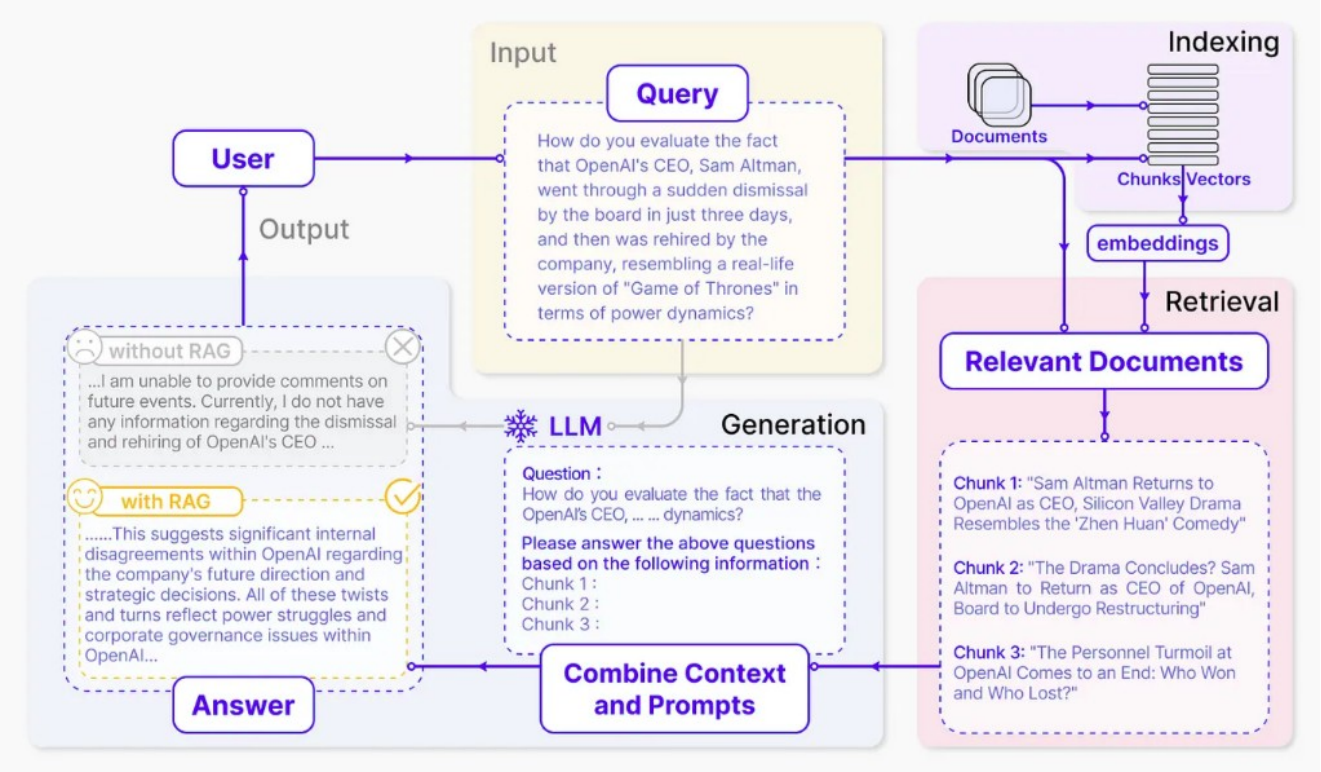


Image source: https://shorturl.at/Kqskr

Takeaways

How to avoid hallucination in RAG

- Improve retrieval accuracy
- Validate retrieved content
- Fine-tune the generative model
- Incorporate explicit attribution
- Re-rank the results

Part-3: Probing LLMs for limitations generation for scientific papers

LimGen: Probing the LLMs for Generating Suggestive Limitations of Research Papers

Research Focus:

- Generation of Suggestive limitations.
- Manual process challenges:
  - Time-consuming and intricate.
  - Requires deep subject knowledge and analytical thinking.

Key Contributions:

- Introduction of the SLG task.
- Creation of the LimGen Dataset containing 4068 research papers and limitations.
- Evaluation of various LLMs for their effectiveness in generating these limitations.

LimGen Dataset

Source: ACL Anthology
Parser: Scipdf_parser

Number of research papers **4068**			
ACL 2022	1750	#Avg words per paper	5122
EMNLP 2022	1227	#Avg sentences per paper	188
EACL 2022	456	#Avg words per limitation	230
EMNLP 2023	635	#Avg sentences per limitation	10

Nature of Limitations

We conduct the manual analysis of LimGen to understand:

- **Relevance:** between research paper and limitation
- **Deduce:** limitation from the research paper or not
- **Purpose:** whether a limitation or a future work
- **Belongs:** which section of the paper

Relevance			Deduce Limitation			Future work or Limitation		
Yes	No	Partial	Yes	No	Partial	Yes	No	Partial
37	3	20	12	13	**35**	15	13	**32**

Limitations related to			
Methodology	Experimental setup	Dataset	Evaluation
41	17	22	7

Methodology

Paradigms and Models:

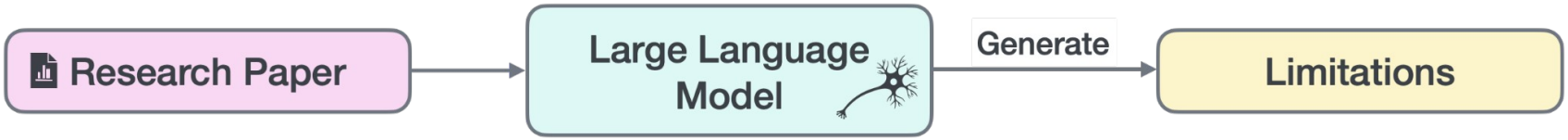
- Summarization-specific pre-trained models including BART and PEGASUS
- Large Language Models (LLMs) namely T5, Cerebras-GPT 1.3B and Llama 2 7B

Schemes

- Non-truncated research paper
- Dense Passage Retrieval (DPR)
- Chain Modeling

Non-truncated Approach

Use entire paper as input with zero-shot prompting and fine-tuning. Has Token limit constraints.



Prompt: Generate limitations or shortcomings for the following scientific paper

{Paper Text}

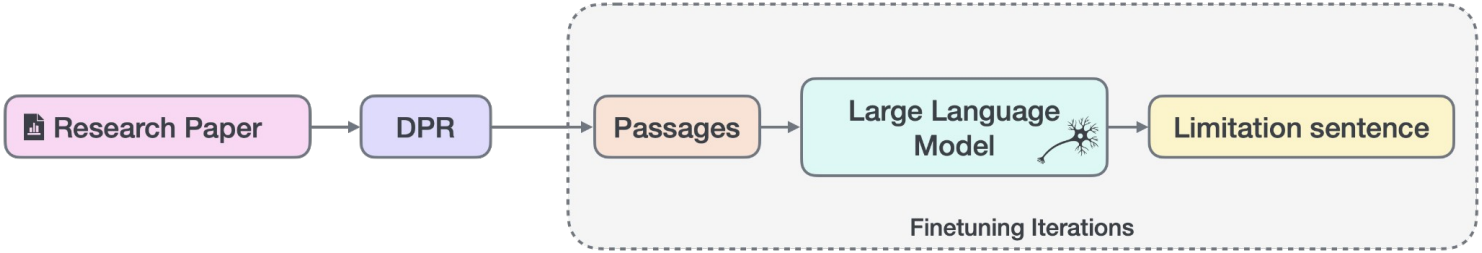
Limitations:

End-of-Prompt

Dense Passage Retrieval (DPR)

Process to retrieve relevant passages:

- Segment paper into passages
- Tokenize, Optimize, Encode passages and limitation sentences
- Retrieve top relevant passages



Prompt: Generate limitations or shortcomings for the following passage from a scientific paper

passage: {DPR paragraph}

A brief technical summary of the scientific paper for context: {Summary of the paper}

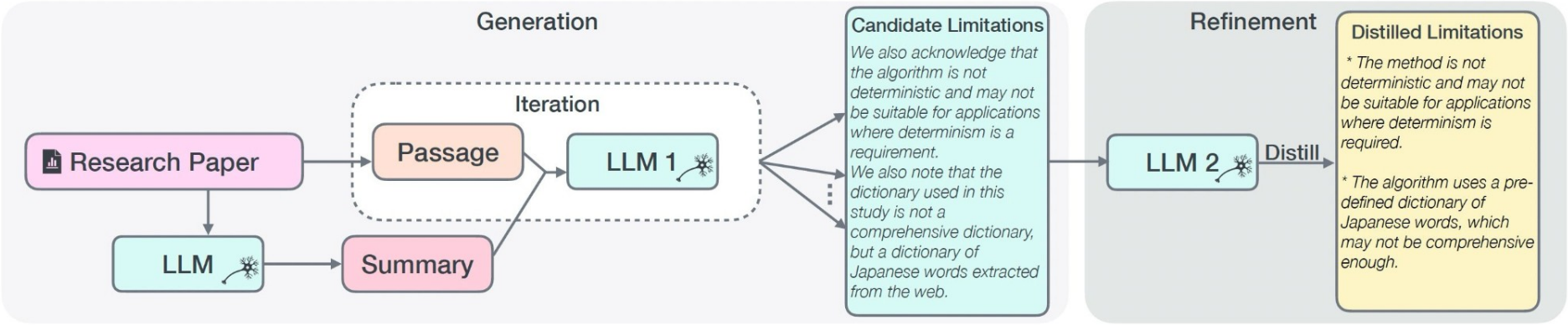
Limitations:

End-of-Prompt

Chain Modeling Approach

Two-stage process:

Generate limitations for each passage Refine and distill generated limitations
Includes paper summary for context



Automatic Evaluation Results

Model	Approach	Without DPR data				With DPR data			
		R-1	R-2	R-L	BS	R-1	R-2	R-L	BS
BART-large	Fine-tuning	30.8	4.5	15.8	82.8	10.7	0.6	8.3	82.8
PEGASUS-large	Fine-tuning	20.2	3.1	12.7	82.3	16.7	7.1	14.2	84.7
T5-base	Fine-tuning	27.7	4.3	16.4	83.6	18.8	7.6	16.2	**85.9**
Llama-2-7b	Zero-shot	21.3	3.3	12.1	81.9	16.7	5.2	8.9	82.4
Cerebras GPT2.7B	Zero-shot	17.6	2.1	12.1	78.9	19.8	4.8	10.3	80.5
Llama-2-7b	Fine-tuning	21.4	3.1	12.7	81.1	**34.8**	**11.0**	**17.7**	83.5
Cerebras GPT2.7B	Fine-tuning	16.9	1.9	12.0	79.1	32.4	9.6	15.9	83.4

- DPR approaches outperform non-truncated methods
- Llama2-DPR shows best performance in ROUGE scores
- BERTScore relatively consistent across models

LLM-based Evaluation – GPT-4

Zero-shot Prompting: Models rated 1-5.
Top Performer: Llama2-FT-Distilled (4.10 average score).
Lowest Performer: T5-base, struggles with generating limitations.
Analysis Gap: Models do not thoroughly analyze limitations.
Bias Issue: GPT-4 assigns high scores to general limitations, potentially inaccurate.
Verification Weakness: GPT-4 may miss incorrect limitations if they seem relevant.
Re-Evaluation: GPT-4 improves only after the issues are explicitly highlighted.

	T5-base	Llama2	Llama2-DPR	Llama2-FT-Distilled
Score	2.71	3.60	3.12	**4.10**

LLM-based Evaluation – TIGERScore (GPT-4o)

Top Performance: Llama2-FT-Distilled outperformed other models despite evaluation biases.

	T5-base	Llama2	Llama2-DPR	Llama2-FT-Distilled
Completeness	5.21	5.58	5.32	**5.14**
Clarity	2.79	2.22	2.34	**1.82**
Relevance	**2.10**	2.34	3.45	2.63
Objectivity	**1.19**	1.38	1.38	1.38
Coherence	2.72	2.39	2.26	**1.68**
Accuracy	2.81	**2.57**	3.13	3.72
Total	16.82	16.48	17.88	**16.36**

Human Evaluation

Evaluation Criteria:
Q1: Sense-making: Do the generated limitations make logical sense?
Q2: Overlap with Gold Standard: Do they align with actual limitations?
Q3: Validity: Are they actual limitations rather than summaries or future work?
Q4: Quality Check: Presence of hallucinations, repetitions, and grammatical errors.

	Llama2-DPR			Llama2			T5-base			Llama2-FT-Distilled		
	Yes	No	Partial	Yes	No	Partial	Yes	No	Partial	Yes	No	Partial
Q1	20	48	32	38	24	38	20	44	36	**70**	8	22
Q2	12	72	16	10	62	28	4	68	28	**28**	30	42
Q3	16	48	36	44	26	30	20	42	38	**62**	8	30
Q4	36	52	12	22	60	18	54	24	22	20	**62**	18

Challenges and Future work

- Complexity of the SLG task
- Evaluation metrics suitability
- Multi-modal content
- Coherence and Relevance
- Open-ended generation of LLMs
- Controllability

Thank you

Contact: ashok.urlana@tcs.com