UNIVERSITY of VIRGINIA

LLM Hallucination

Team 3
3/19/2024

Road Map

1. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions

1. LLMs as Factual Reasoners: Insights from Existing Benchmarks and Beyond

UNIVERSITY of VIRGINIA

A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions

Zhiyang Yuan (vfr4pr), Feilian Dai (kdr4qp)

Structure

- Brief introduction to LLM Hallucinations
- Definitions and Taxonomy
- Origins & Types of Hallucinations
- Detection and Mitigation Strategies
- Techniques and Evaluation Benchmarks
- Challenges and Future Directions

Hallucination?

- Nonsensical (logical coherence)
- Unfaithful to source content.
  - Intrinsic hallucination
  - Extrinsic hallucination

Broader Versatility

Applicability and Adaptability

Hallucination Types

- **Factuality Hallucination**: inconsistent with real-world facts or potentially misleading
  - **Factual Inconsistency**: facts relate to real-world information, but has contradictions
  - **Factual Fabrication**: unverifiable against established real-world knowledge

Hallucination Type	Sub-Type	User Input	Model Output	Explanation
Factuality Hallucination	Factual Inconsistency	Tell me about the first person to land on the Moon.	Yuri Gagarin was the first person to land on the Moon.	The LLM's response is factual inconsistency as **Neil Armstrong was the first person to land on the Moon**, not Yuri Gagarin, who was the first human in space.
	Factual Fabrication	Tell me about the historical origins of unicorns.	Unicorns were documented to have roamed the plains of Atlantis around 10,000 BC, where they were considered sacred creatures and were often associated with royalty.	The LLM's response is a fabricated claim since **there's no verified, real-world evidence supporting the existence of unicorns in Atlantis**, especially in connection with royalty.

Hallucination Types

- **Faithfulness Hallucination**: inconsistency with user provided instructions and contextual information
  - **Instruction inconsistency**: deviate from a user's instructions
  - **Context inconsistency**: unfaithful with the provided contextual information
  - **Logical inconsistency**: exhibit internal logical contradictions

Faithfulness Hallucination	Instruction Inconsistency	Translate the following English question into Spanish: "What is the capital of France?"	The capital of France is Paris.	The user explicitly provided an instruction to translate the English question into Spanish. **Instead of performing the translation as instructed, the LLM chose to answer the question.**
	Context Inconsistency	Summarize the following paragraph about the Nile: The Nile is a major north-flowing river in northeastern Africa. It's traditionally considered the longest river in the world. Originating in the Great Lakes region of central Africa, it flows through several countries before emptying into the Mediterranean Sea.	The Nile originates from the mountain ranges of central Africa and flows into the Mediterranean.	The LLM's summary incorrectly states that the Nile originates from the "mountain ranges of central Africa," while **the paragraph clearly mentions that the Nile originates in the Great Lakes region of central Africa.**
	Logical Inconsistency	Solve the equation 2x + 3 = 11 step by step.	Step 1: Subtract 3 from both sides to get: 2x = 8. Step 2: Divide both sides by 2 to get: x = 3. So the result of the equation is x=3.	The first step correctly isolates the term with x to get 2x = 8. However, **the next step inaccurately divides 8 by 2 to yield a result of x = 3**, which is inconsistent with the earlier reasoning.

Refined Taxonomy

Factuality hallucination & Faithfulness hallucination

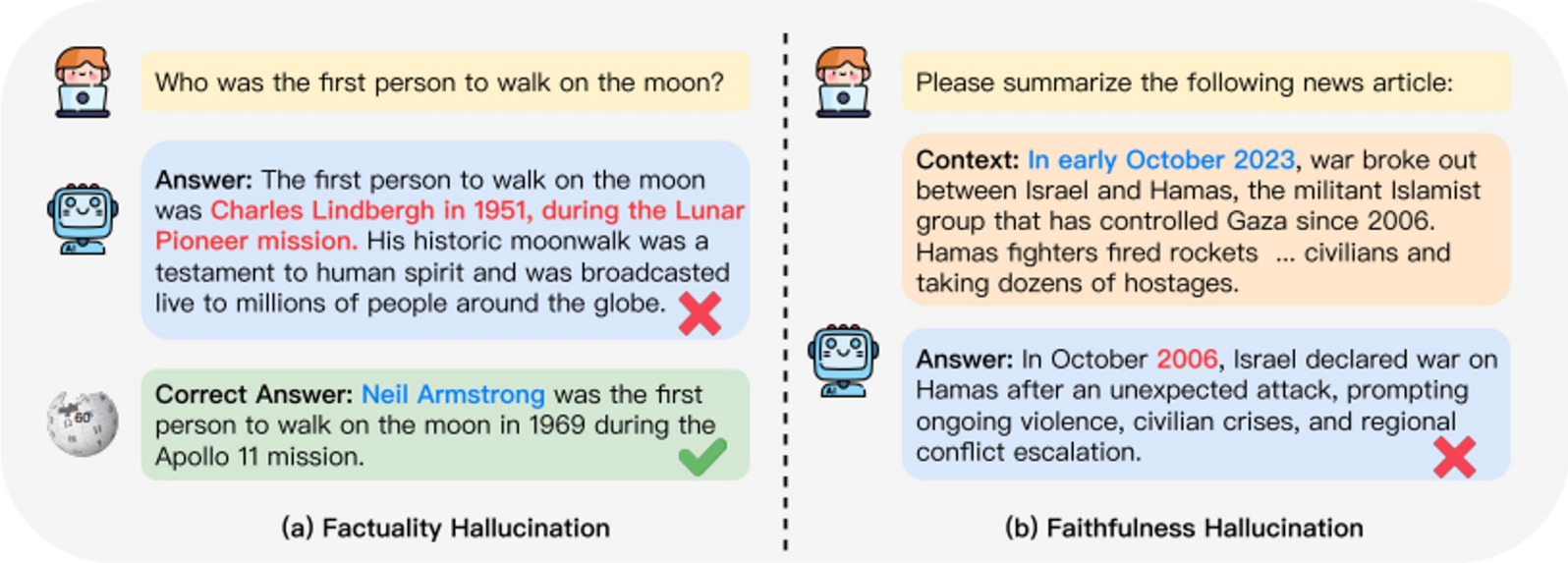
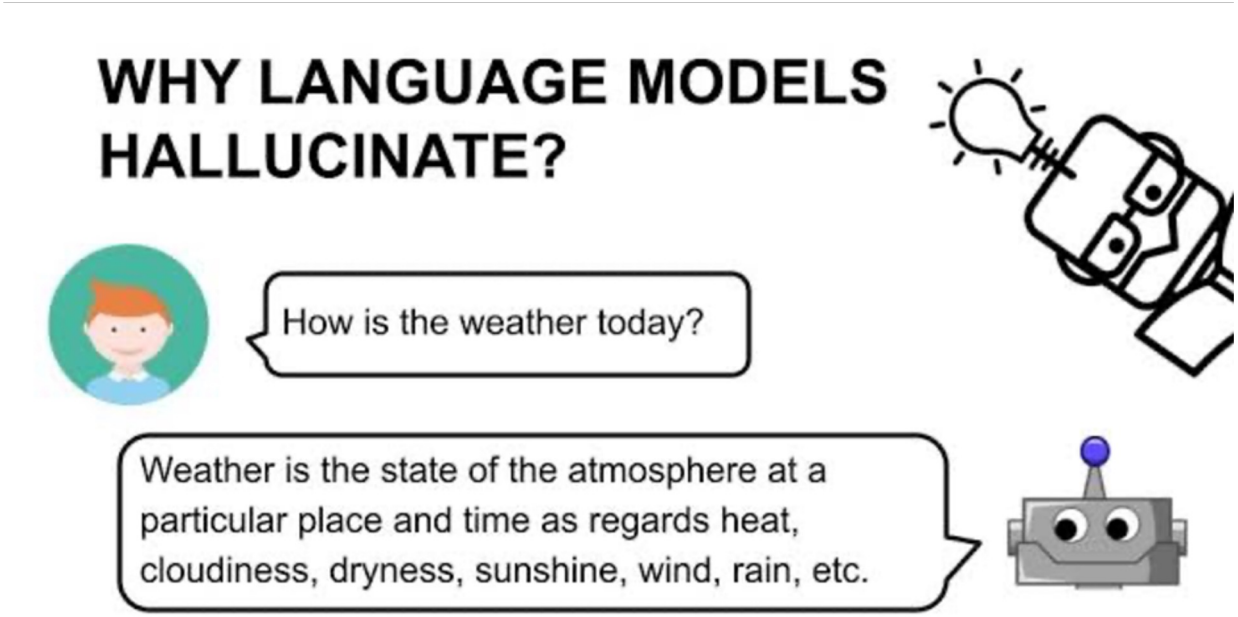
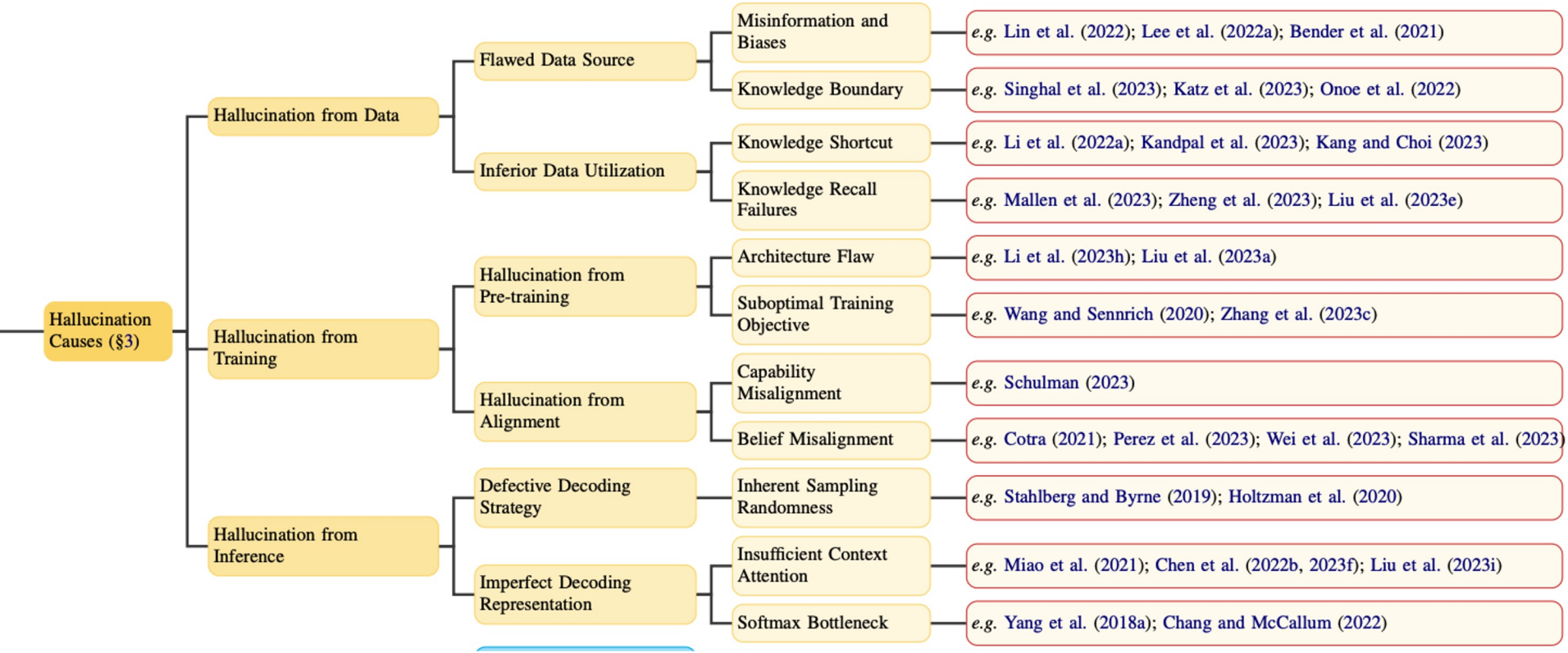


Figure 1: An intuitive example of LLM hallucination.

Hallucination Causes

- Data
- Training
- Inference




Hallucination Causes (§3)
Hallucination from Data
Hallucination from Training
Hallucination from Inference
Flawed Data Source
Inferior Data Utilization
Hallucination from Pre-training
Hallucination from Alignment
Defective Decoding Strategy
Imperfect Decoding Representation
Misinformation and Biases
Knowledge Boundary
Knowledge Shortcut
Knowledge Recall Failures
Architecture Flaw
Suboptimal Training Objective
Capability Misalignment
Belief Misalignment
Inherent Sampling Randomness
Insufficient Context Attention
Softmax Bottleneck
e.g. Lin et al. (2022); Lee et al. (2022a); Bender et al. (2021)
e.g. Singhal et al. (2023); Katz et al. (2023); Onoe et al. (2022)
e.g. Li et al. (2022a); Kandpal et al. (2023); Kang and Choi (2023)
e.g. Mallen et al. (2023); Zheng et al. (2023); Liu et al. (2023e)
e.g. Li et al. (2023h); Liu et al. (2023a)
e.g. Wang and Sennrich (2020); Zhang et al. (2023c)
e.g. Schulman (2023)
e.g. Cotra (2021); Perez et al. (2023); Wei et al. (2023); Sharma et al. (2023)
e.g. Stahlberg and Byrne (2019); Holtzman et al. (2020)
e.g. Miao et al. (2021); Chen et al. (2022b, 2023f); Liu et al. (2023i)
e.g. Yang et al. (2018a); Chang and McCallum (2022)

Hallucination from Data

- Misinformation and Biases
  - Imitative Falsehoods: trained on factual incorrect data
  - Duplication Bias: over-prioritize the recall of duplicated data
  - Social Biases: Gender, Race

Type	Premise	Input	Model Ouput	Explanation
Imitative Falsehood	Within the LLM's training data, a frequently referenced but incorrect fact appears multiple times: *Thomas Edison invented the light bulb.*	Who is credited with the invention of the light bulb?	Thomas Edison is credited with the invention of the light bulb.	**While Edison improved the light bulb's design, he was not its sole inventor.** The model's answer reflects a misconception prevalent in its training data, exemplifying an imitative falsehood.
Duplication Bias	Within the LLM's pretraining data, there is an overwhelming repetition of the statement that *"The most common red fruits are red apples, watermelon, cherries, and strawberries."*	Please list some red fruits, excluding apples.	Red fruits are red apples, watermelon, cherries, and strawberries.	**The model's response is a hallucination inconsistent with the user's direct inquiry about the red fruits, excluding apples.** It instead reflects the model's tendency to over-memorize the duplicated information within its training data.
Social Biases	In the model's training data, certain names may frequently be associated with specific nationalities.	Provide a summary of the following news: Dr. Kim recently published a breakthrough paper on quantum computing. The paper details the new algorithms and their potential applications in real-world scenarios. It has been hailed as revolutionary by experts in the field.	Dr. Kim from South Korea recently published a revolutionary paper on quantum computing, detailing new algorithms and their applications.	**The model erroneously added "South Korea" based on the name "Kim"** revealing its bias in linking certain names to specific nationalities, which can cause hallucinations in the summary.

Hallucination from Data

- **Knowledge Boundary**
  - **Domain Knowledge Deficiency**: Lack of proprietary data lead to less expertise
  - **Outdated Factual Knowledge**

Type	Input	Model Ouput	Explanation
Domain Knowledge Deficiency	What causes phenylketonuria?	Phenylketonuria is caused by a mutation in the hemoglobin gene, which results in an unusual interaction with hemoglobin, affecting its ability to transport oxygen in the bloodstream.	Phenylketonuria (PKU) is actually caused by **mutations in the gene that helps make an enzyme called phenylalanine hydroxylase**. Due to the lack of specialized medical knowledge in LLM's training data, it may sometimes provide incorrect information on medical topics.
Outdated Knowledge	Which city hosted the most recent Winter Olympics?	The most recent city to host the Winter Olympics was Pyeongchang, South Korea, in 2018.	The most recent city to host the Winter Olympics was **Beijing, in 2022**. This is due to the fact that LLM stores outdated knowledge, and the answer to this question exhibits a time-shift phenomenon.

Hallucination from Data

- Inferior Data Utilization
  - **Knowledge Shortcut**: overly rely on co-occurrence statistics, relevant document count
  - **Knowledge Recall Failures**
    - Long-tail Knowledge: rare, specialized, or highly specific information not widely known or discussed.
    - Complex Scenario: multi-hop reasoning and logical deduction

Type	Sub-Type	User Input	Model Output	Explanation
Knowledge Shortcut	Co-occurrence Bias	What is the capital of Canada?	Toronto is the capital of Canada.	The model leans heavily on the frequent co-occurrence of the terms *Toronto* and *Canada* in its training data, without truly capturing the factual knowledge about the capital of Canada
Knowledge Recall Failures	Long-tail Knowledge	Please generate a biography for George James Rankin.	George James Rankin is a dedicated educator known for his contributions to the field of education and his passion for fostering learning.	**George James Rankin is actually a politician**, but the LLM may have difficulty effectively utilizing knowledge about this long-tail entity despite being trained on comprehensive Wikipedia data during pre-training.
	Complex Reasoning	If Mount Everest were to descend by 500 meters, which mountain would become the world's highest peak?	If Mount Everest were to descend by 500 meters, it would still remain the world's highest peak.	The height of Mount Everest is 8844.43 meters, while K2's height is 8611 meters. If Mount Everest were to descend by 500 meters, K2 would become the world's highest peak. Facing complex multi-step reasoning questions like this, LLM may struggle to recall all the relevant knowledge associated with it.

Hallucination from Pre-training

- Architecture Flaw
  - **Inadequate Unidirectional Representation**: predict the subsequent token based solely on preceding tokens in a left-to-right manner
  - **Attention Glitches**: limitations of soft attention
    - attention diluted across positions as sequence length increases
- Exposure Bias: teacher forcing

Hallucination from Alignment

- **Capability Misalignment**: mismatch between LLMs' pre-trained capabilities and the expectations from fine-tuning data
- **Belief Misalignment**: prioritize appeasing perceived user preferences over truthfulness

Hallucination from Inference

- Inherent Sampling Randomness
  - Stochastic Sampling: controlled randomness enhance creativity and diversity
  - likelihood trap: high-probability, low-quality text
- Imperfect Decoding Representation
  - Insufficient Context Attention: prioritize recent or nearby words in attention (Over-Confidence Issue)
  - Softmax Bottleneck: inability manage multi-modal distributions, irrelevant or inaccurate content

Feilian Dai (kdr4qp)

Hallucination Detection and Benchmarks

1. Factuality Hallucination Detection
- A. Retrieve External Facts: comparing the modelgenerated content against reliable knowledge sources.
- B. Uncertainty Estimation
  - LLM Internal States
  - LLM Behavior

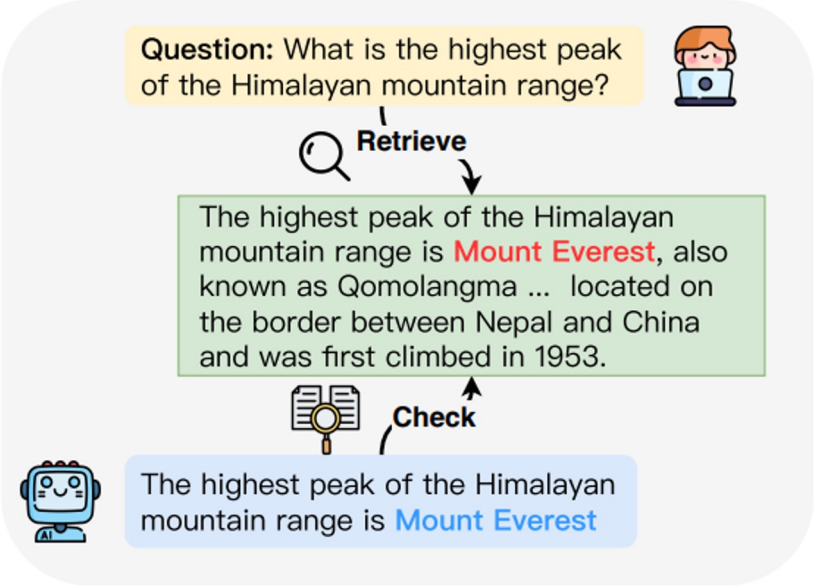


Figure 3: An example of detecting factuality hallucination by retrieving external facts.

Hallucination Detection and Benchmarks

1. **Factuality Hallucination Detection**

B. Uncertainty Estimation

Premise: the origin of LLM hallucinations is inherently tied to the model's uncertainty.

B.1 LLM Internal States: operates under the assumption that one can access the model's internal states

B.2 LLM Behavior: leveraging solely the model's observable behaviors to infer its underlying uncertainty

Hallucination Detection and Benchmarks

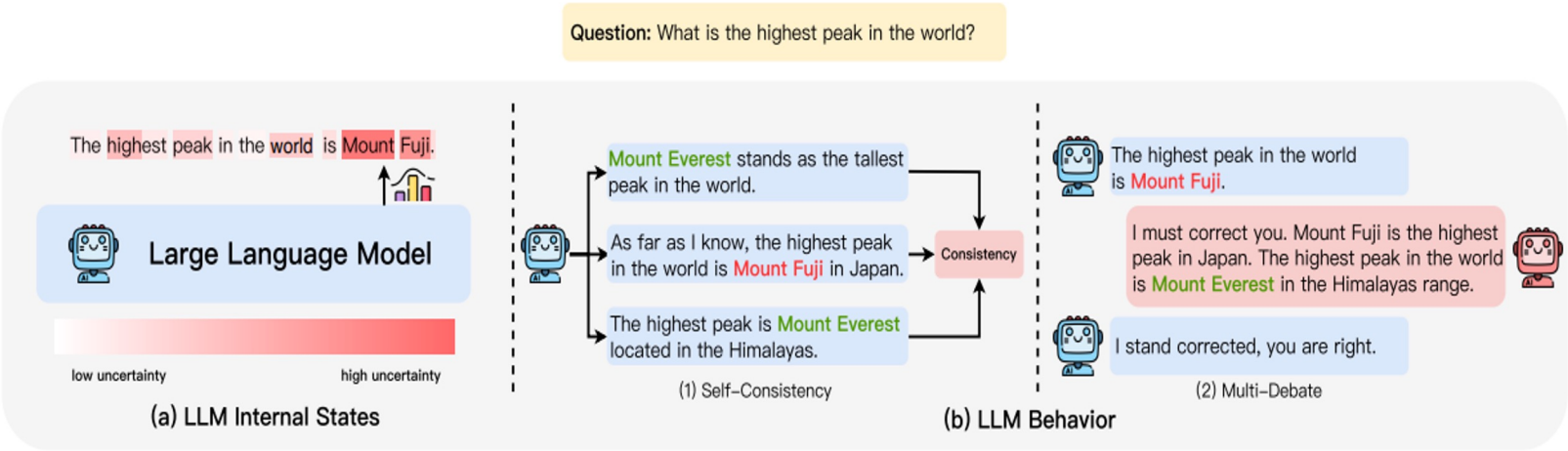


Figure 4: Taxonomy of Uncertainty Estimation Methods in Factual Hallucination Detection, featuring **a) LLM Internal States** and **b) LLM Behavior**, with LLM Behavior encompassing two main categories: Self-Consistency and Multi-Debate.

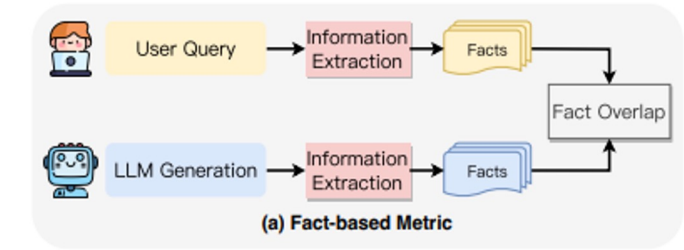
Hallucination Detection and Benchmarks

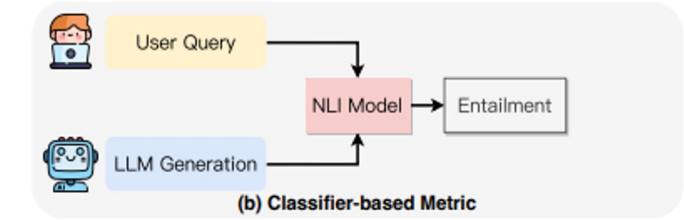
2. Faithfulness Hallucination Detection

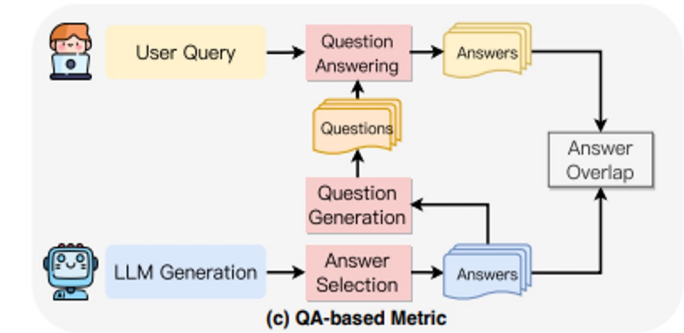
Focuses on ensuring the alignment of the generated content with the given context, sidestepping the potential pitfalls of extraneous or contradictory output.

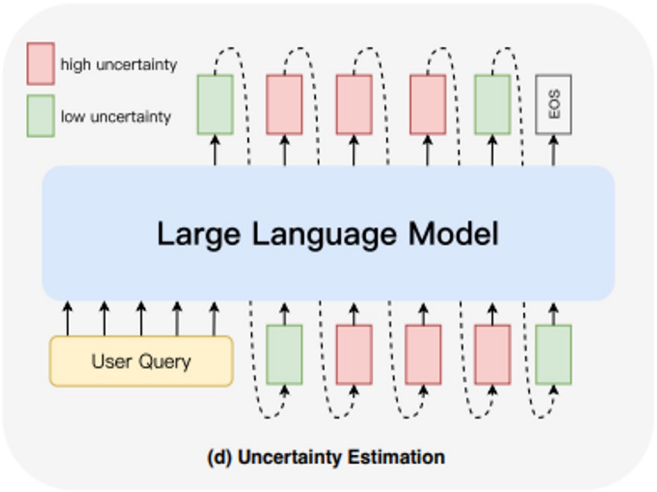
- **Fact-based Metrics**: assesses faithfulness by measuring the overlap of facts between the generated content and the source content
- **Classifier-based Metrics**: utilizing trained classifiers to distinguish the level of entailment between the generated content and the source content
- **Question-Answering based Metrics**: employing question-answering systems to validate the consistency of information between the source content and the generated content
- **Uncertainty Estimation**: assesses faithfulness by measuring the model's confidence in its generated outputs
- **Prompting-based Metrics**: induced to serve as evaluators, assessing the faithfulness of generated content through specific prompting strategies.

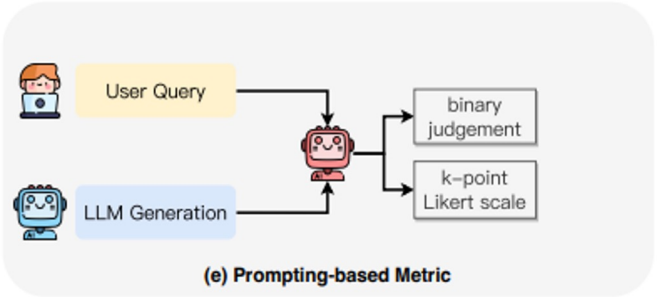
Hallucination Detection and Benchmarks







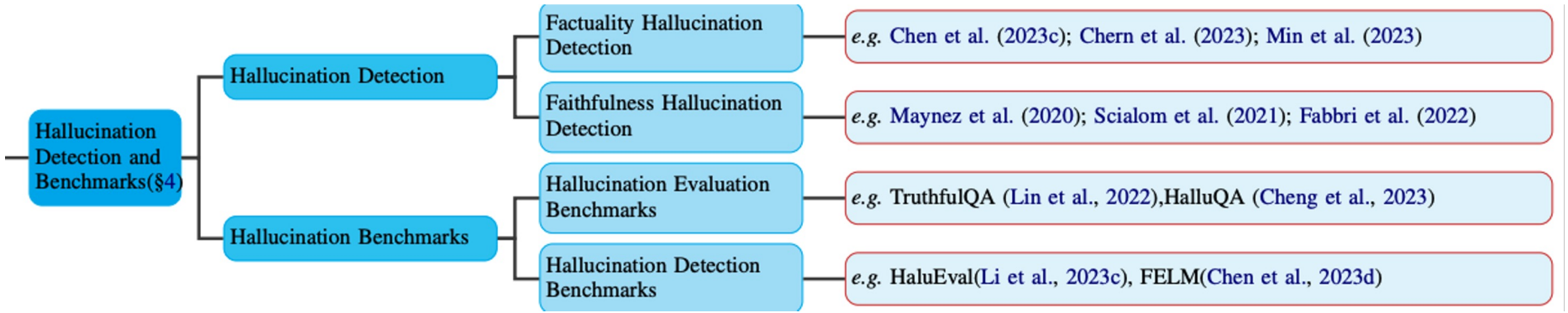


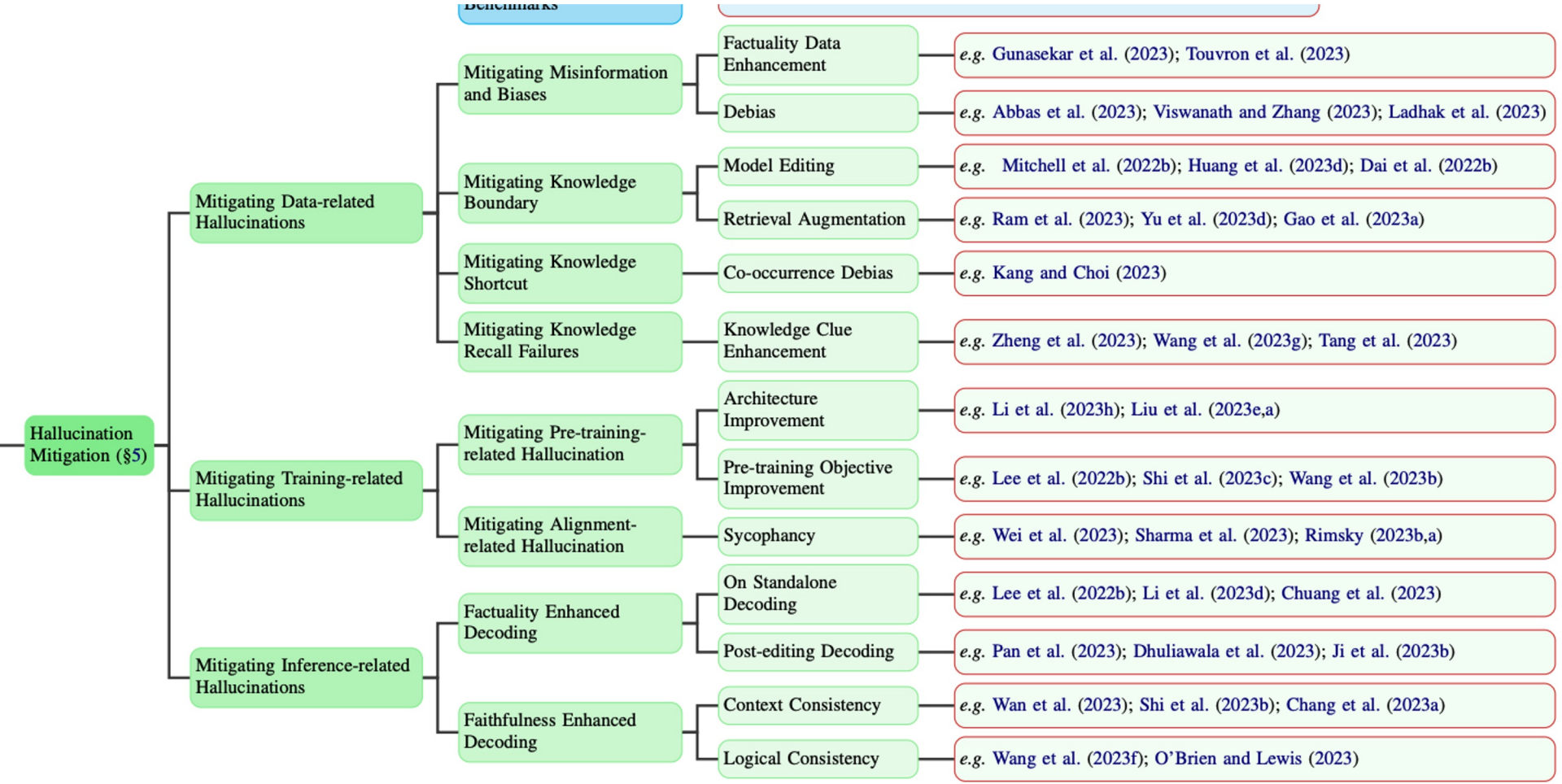


Hallucination Detection and Benchmarks

3. Benchmarks

- **Hallucination Evaluation Benchmarks**
  Assess LLMs' proclivity to produce hallucinations, with a particular emphasis on identifying factual inaccuracies and measuring deviations from original contexts
- **Hallucination Detection Benchmarks**
  Evaluate the performance of existing hallucination detection methods. Primarily concentrated on taskspecific hallucinations, such as abstractive summarization, data-to-tex, and machine translation.


Hallucination Detection and Benchmarks(§4)
Hallucination Detection
Hallucination Benchmarks
Factuality Hallucination Detection
Faithfulness Hallucination Detection
Hallucination Evaluation Benchmarks
Hallucination Detection Benchmarks
e.g. Chen et al. (2023c); Chern et al. (2023); Min et al. (2023)
e.g. Maynez et al. (2020); Scialom et al. (2021); Fabbri et al. (2022)
e.g. TruthfulQA (Lin et al., 2022),HalluQA (Cheng et al., 2023)
e.g. HaluEval(Li et al., 2023c), FELM(Chen et al., 2023d)


Hallucination Mitigation (§5)
Mitigating Data-related Hallucinations
Mitigating Misinformation and Biases
Factuality Data Enhancement
e.g. Gunasekar et al. (2023); Touvron et al. (2023)
Debias
e.g. Abbas et al. (2023); Viswanath and Zhang (2023); Ladhak et al. (2023)
Mitigating Knowledge Boundary
Model Editing
e.g. Mitchell et al. (2022b); Huang et al. (2023d); Dai et al. (2022b)
Retrieval Augmentation
e.g. Ram et al. (2023); Yu et al. (2023d); Gao et al. (2023a)
Mitigating Knowledge Shortcut
Co-occurrence Debias
e.g. Kang and Choi (2023)
Mitigating Knowledge Recall Failures
Knowledge Clue Enhancement
e.g. Zheng et al. (2023); Wang et al. (2023g); Tang et al. (2023)
Mitigating Training-related Hallucinations
Mitigating Pre-training-related Hallucination
Architecture Improvement
e.g. Li et al. (2023h); Liu et al. (2023e,a)
Pre-training Objective Improvement
e.g. Lee et al. (2022b); Shi et al. (2023c); Wang et al. (2023b)
Mitigating Alignment-related Hallucination
Sycophancy
e.g. Wei et al. (2023); Sharma et al. (2023); Rimsky (2023b,a)
Mitigating Inference-related Hallucinations
Factuality Enhanced Decoding
On Standalone Decoding
e.g. Lee et al. (2022b); Li et al. (2023d); Chuang et al. (2023)
Post-editing Decoding
e.g. Pan et al. (2023); Dhuliawala et al. (2023); Ji et al. (2023b)
Faithfulness Enhanced Decoding
Context Consistency
e.g. Wan et al. (2023); Shi et al. (2023b); Chang et al. (2023a)
Logical Consistency
e.g. Wang et al. (2023f); O'Brien and Lewis (2023)

4. Mitigating Data-related Hallucinations

- **Mitigating Misinformation and Biases**:
  Factuality Data Enhancement: Gathering high-quality data, Up-sampling factual data during the pre-training
  Duplication Bias: Exact Duplicates, Near-Duplicates
  Societal Biases: Focusing on curated, diverse, balanced, and representative training corpora
- **Mitigating Knowledge Boundary:**
  Knowledge Editing: Modifying Model Parameter(Locate-then-edit methods, Meta-learning methods), Preserving Model Parameters
  Retrieval Augmentation: One-time Retrieval, Iterative Retrieval, Post-hoc Retrieval
- **Mitigating Knowledge Shortcut**：
  Fine-tuning on a debiased dataset by excluding biased samples
- **Mitigating Knowledge Recall Failures:**
  Adding relevant information to questions to aid recall, Encourages LLMs to reason through steps to improve recall

Mitigating Data-related Hallucinations

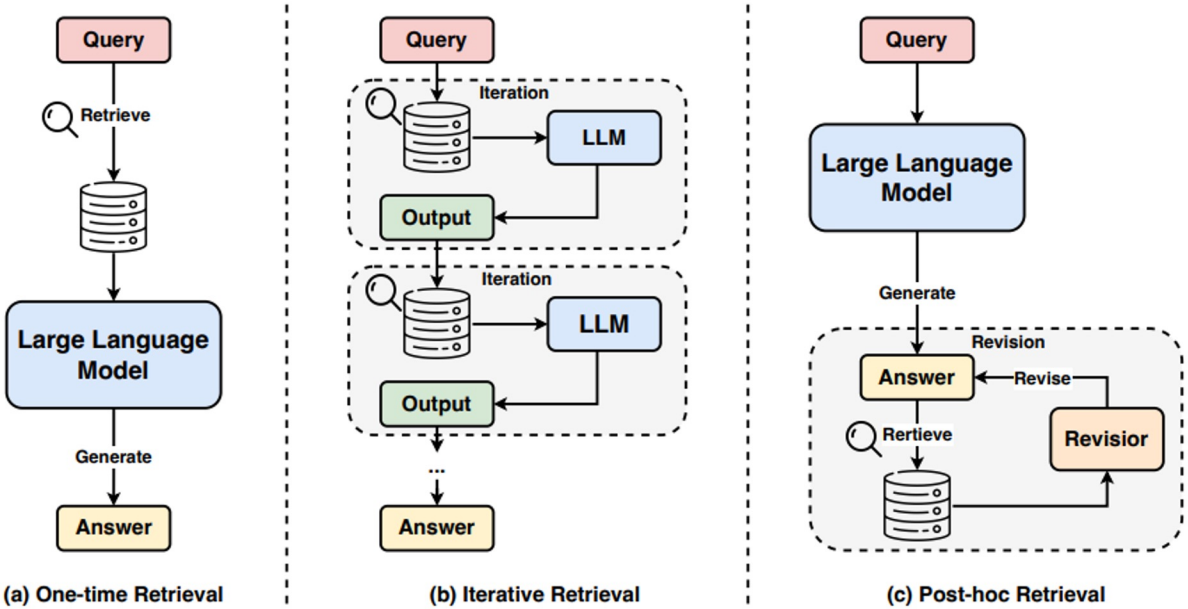


Figure 6: The illustration of three distinct approaches for Retrieval-Augmented Generation: **a) One-time Retrieval**, where relevant information is retrieved once before text generation; **b) Iterative Retrieval**, involving multiple retrieval iterations during text generation for dynamic information integration; and **c) Post-hoc Retrieval**, where the retrieval process happens after an answer is generated, aiming to refine and fact-check the generated content.

Mitigating Inference-related Hallucination

Factuality Enhanced Decoding

- **On Standalone Decoding:**
  - **Factual-Nucleus Sampling**: Adjusts nucleus probability dynamically for a balance between factual accuracy and output diversity.
  - **Inference-Time Intervention (ITI)**: Utilizes activation space directionality for factually correct statements, steering LLMs towards accuracy during inference.
- **Post-editing Decoding:**
  - **Chain-of-Verification (COVE):** Employs self-correction capabilities to refine generated content through a systematic verification and revision process

Faithfulness Enhanced Decoding

- **Context Consistency:**
  - **Context-Aware Decoding (CAD):** Adjusting output distribution to enhance focus on contextual information, balancing between diversity and attribution
- **Logical Consistency:**
  - **Knowledge Distillation and Contrastive Decoding:** Generating consistent rationale and fine-tuning with counterfactual reasoning to eliminate reasoning shortcuts, ensuring logical progression in multi-step reasoning

Challenges and Open Questions

Challenges in LLM Hallucination

- **Hallucination in Long-form Text Generation**
  Absence of manually annotated hallucination benchmarks in the domain of long-form text generation
- **Hallucination in Retrieval Augmented Generation**
  Irrelevant evidence can be propagated into the generation phase, possibly tainting the output
- **Hallucination in Large Vision-Language Models**
  LVLMs sometimes mix or miss parts of the visual context, as well as fail to understand temporal or logical connections between them

Challenges and Open Questions

Open Questions in LLM Hallucination

- **Can Self-Correct Mechanisms Help in Mitigating Reasoning Hallucinations?** Occasionally exhibit unfaithful reasoning characterized by inconsistencies within the reasoning steps or conclusions that do not logically follow the reasoning chain.
- **Can We Accurately Capture LLM Knowledge Boundaries?** LLMs still face challenges in recognizing their own knowledge boundaries. This shortfall leads to the occurrence of hallucinations, where LLMs confidently produce falsehoods without an awareness of their own knowledge limits.
- **How Can We Strike a Balance between Creativity and Factuality?** hallucinations can sometimes offer valuable perspectives, particularly in creative endeavors such as storytelling, brainstorming, and generating solutions that transcend conventional thinking.

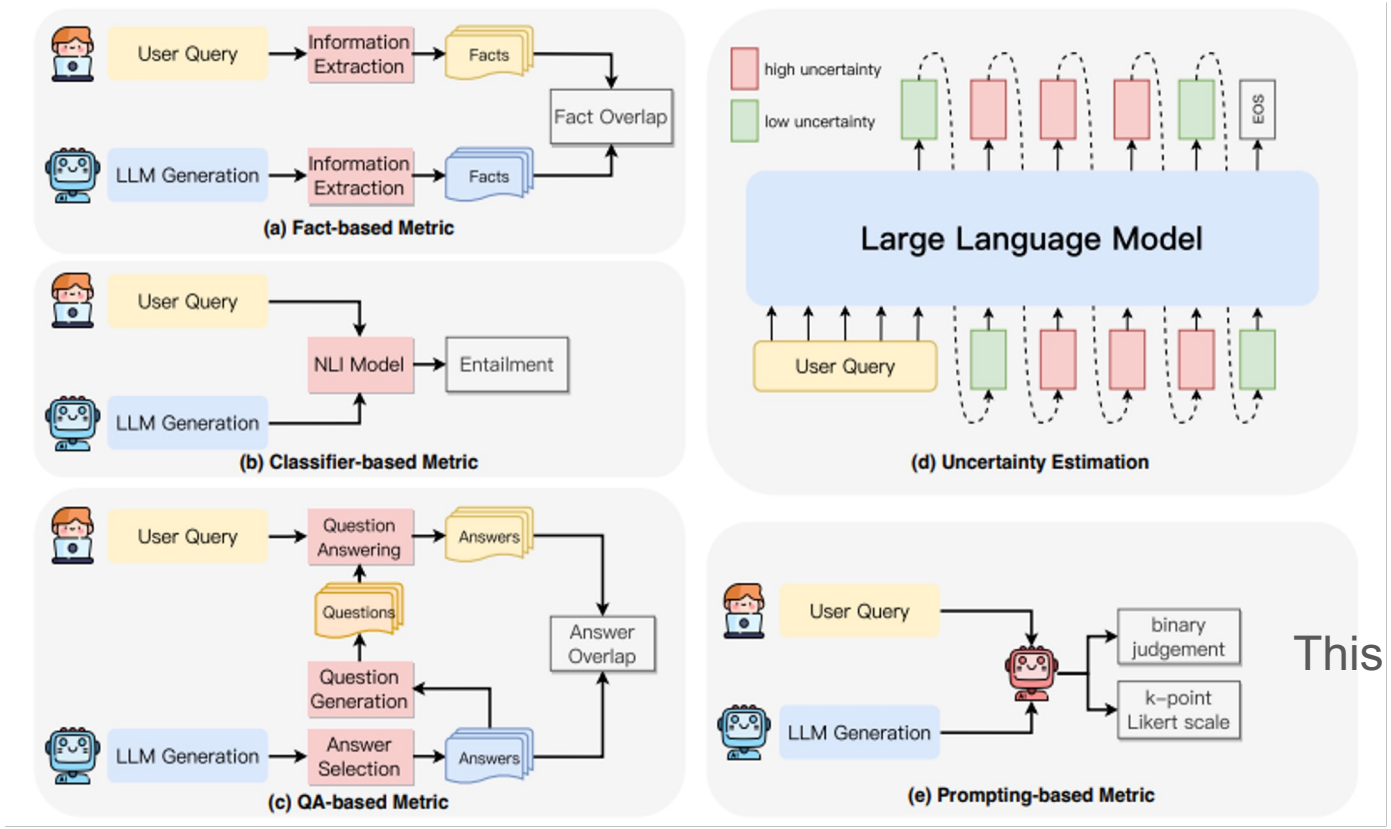
UNIVERSITY of VIRGINIA

LMs as Factual Reasoners: Insights from Existing Benchmarks and Beyond

Shihe Wang(qvw9pv), Parker Hutchinson (pch6am)

Shihe Wang(qvw9pv)

Hallucination Detection and Benchmarks



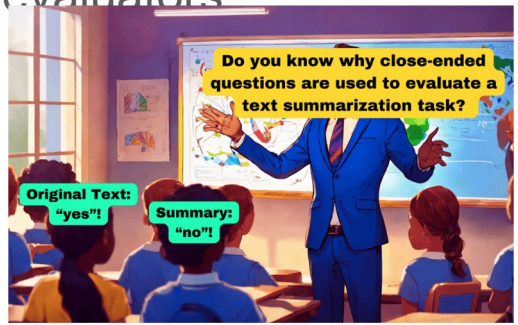
This paper

Background

LLMs are used to produce automatic summarization for work meetings, health records, or even scientific documents.

It is important to limit the reach of factually inconsistent summaries.

One line of research is to use LLMs as factuality evaluators



Document The Knicks beat the Rockets. The fans were excited.	
Summary The Knicks beat the Bucks.	
Entailment Matrix [Contra, Neutral, Support] $\begin{bmatrix} 0.90 & 0.07 & 0.03 \\ 0.02 & 0.90 & 0.08 \end{bmatrix}$	**Selected Answer** the Bucks
	Generated Question Who did the Knicks beat?
	QA Output the Rockets
Max Support Score 0.08	**Answer Overlap Score** 0.20

Table 1: Toy example of a factual inconsistency between a summary and a source document. *Left:* The entailment-based metric computes the level of contradiction, neutrality, and support between the summary and each source document sentence. The final factual consistency metric is calculated as the maximum support score over all source sentences. *Right:* The QA-based metric first selects a noun-phrase *answer* from the summary. A QG model then generates an associated question that a QA model answers based on the source document. The answer overlap score of the QA-based metric measures the semantic overlap between the QA model output and the selected answer as the final metric score.

Issues with Using LLMs as Factual Reasoners

The accuracy-only results from consistency benchmarks are not reliable.

1. Not all LLMs can generate explanations that pinpoint factual inaccuracies. (“Cheated” binary prediction)
2. A significant number of mislabeled samples (7+%) of factual inconsistencies undetected by annotators in the benchmarks themselves.

Contributions of the Paper

A **protocol** designed to create challenging benchmarks while ensuring the reproducibility of the labels.

– By verifying a small set of seed summaries and generating numerous edited versions of these summaries.

The **SUMMEDITS benchmark** by implementing the protocol in ten diverse textual domains, including the legal, dialogue, academic, financial, and sales domains.

The protocol can be applied to other benchmarks for other domains with **low cost**;

The code, the protocol and the dataset are all public.

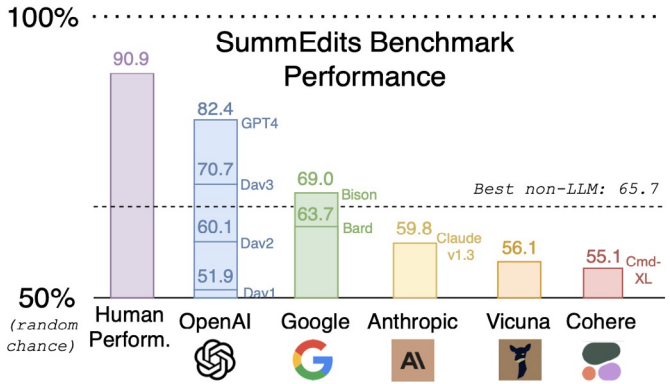


Figure 1: SUMMEDITS is a benchmark to evaluate the factual reasoning abilities of LLMs, measuring if models detect factual inconsistencies when they occur in summaries. Capable detection models can help build more reliable NLG systems.

Find LLMs with the Potential to be Factual Reasonsers

To find the LLMs that might be suitable for the tasks, the authors tested the different LLMs on FactCC.

FactCC is based on XSum news summarization dataset.

SUMMARY: *A man and a child have been killed after a light aircraft made an emergency landing on a beach in Portugal.*

DOCUMENT: Authorities said the incident took place on Sao Joao beach in Caparica, south-west of Lisbon.
The National Maritime Authority said a middle-aged man and a young girl died after they were unable to avoid the plane.
[*6 sentences with 139 words are abbreviated from here.*]
Other reports said the victims had been sunbathing when the plane made its emergency landing.
[*Another 4 sentences with 67 words are abbreviated from here.*]
Video footage from the scene carried by local broadcasters showed a small recreational plane parked on the sand, apparently intact and surrounded by beachgoers and emergency workers.
[*Last 2 sentences with 19 words are abbreviated.*]

Source article fragments	
(CNN) The mother of a quadriplegic man who police say was left in the woods for days cannot be extradited to face charges in Philadelphia until she completes an unspecified "treatment," Maryland police said Monday. The Montgomery County (Maryland) Department of Police took Nyia Parler, 41, into custody Sunday (...)	(CNN) The classic video game "Space Invaders" was developed in Japan back in the late 1970's – and now their real-life counterparts are the topic of an earnest political discussion in Japan's corridors of power. Luckily, Japanese can sleep soundly in their beds tonight as the government's top military official earnestly revealed that (...)
Model generated claims	
Quadriplegic man Nyia Parler, 41, left in woods for days can not be extradited.	Video game "Space Invaders" was developed in Japan back in 1970.

Table 1: Examples of factually incorrect claims output by summarization models. Green text highlights the support in the source documents for the generated claims, red text highlights the errors made by summarization models.

More on FactCC

Transformation	Original sentence	Transformed sentence
Paraphrasing	Sheriff Lee Baca has now decided to recall some 200 badges his department has handed out to local politicians just two weeks after the picture was released by the U.S. attorney's office in support of bribery charges against three city officials.	Two weeks after the US Attorney's Office issued photos to support bribery allegations against three municipal officials, Lee Baca has now decided to recall about 200 badges issued by his department to local politicians.
Sentence negation	Snow was predicted later in the weekend for Atlanta and areas even further south.	Snow wasn't predicted later in the weekend for Atlanta and areas even further south.
Pronoun swap	It comes after his estranged wife Mona Dotcom filed a $20 million legal claim for cash and assets.	It comes after your estranged wife Mona Dotcom filed a $20 million legal claim for cash and assets.
Entity swap	Charlton coach Guy Luzon had said on Monday: 'Alou Diarra is training with us.'	Charlton coach Bordeaux had said on Monday: 'Alou Diarra is training with us.'
Number swap	He says he wants to pay off the $12.6million lien so he can sell the house and be done with it, according to the Orlando Sentinel.	He says he wants to pay off the $3.45million lien so he can sell the house and be done done with it, according to the Orlando Sentinel.
Noise injection	Snow was predicted later in the weekend for Atlanta and areas even further south.	Snow was was predicted later in the weekend for Atlanta and areas ~~even~~ further south.

Table 2: Examples of text transformations used to generate training data. Green and red text highlight the changes made by the transformation. Paraphrasing is a semantically invariant transformation, Sentence negation, entity, pronoun, and number swaps are semantically variant transformation.

Prompts

Unlock the abilities of LLMs

1. Zero-Shot Prompts: a short task description and the input data
2. Few-Shot Prompts: a task description and one or more demonstrations of the task.
3. Chain-of-Thought Prompts: a task description and input data and are asked to generate a series of intermediate reasoning steps.
4. Generate-with-Evidence Prompts: a task description and input data and are asked to answer the task, and then generate evidence for the chosen answer.
5. Persona-based Prompts: assigned a role, or "persona", and next prompted to complete a given task. (Journalist)

Models

Non-LLM(baselines): NLI-based (natural language inference) approaches, DAE and SummaC, and a QA-based method QAFactEval.

Foundation Models: pre-trained but not fine-tuned. Meta's LLaMa-13b , and OpenAI's Ada001, Babbage001, Curie001, and DaVinci-001.

Instruction-tuned LLMs: tuned on instruction-following data. Databrick's Dolly, Stanford's Alpaca , Anthropic's Claude V1.3, Cohere's Command-XL, Google's PaLM2-bison, and OpenAI's DaVinci-002, and DaVinci-003 models.

Chat-based LLMs: tuned on conversational and instruction-following datasets. Google's Bard, Mosaic's MPT-7b-chat , Vicuna-13b, and OpenAI's GPT3.5-turbo (ChatGPT), and GPT-4.

Experiment Setup

150 samples to conduct experiments, by including 25 examples for each of the 5 error types in the dataset, i.e. date-, entity- , negation-, number-, and pronoun-related errors, and 25 factually correct samples.

Inconsistency Detection

Binary classification test:

non-LLM outperform LLM

Few-shot will improve performance comparing to zero-shot (not GPT4 and PaLM2)

Generate-with-Evidence outperforms Chain-of-Thought

Persona-based improves GPT3.5-turbo performance

	Model (↓)	Non-LLM Models				
✓	DAE			67.2		
✓	SummaC			96.8		
✓	QAFactEval			93.6		
	Prompt Group →	**ZS**	**FS**	**Pers**	**CoT**	**GwE**
	LLaMa-13B	50.0	51.6	52.4	-	-
	Alpaca-13B	54.8	48.4	57.2	-	-
	Dolly-v2-12B	50.4	50.8	50.8	-	-
	MPT-7B-Chat	58.7	54.0	54.4	-	-
✓	Vicuna-13B	65.5	68.0	63.2	-	-
✓	Cohere-CMD-XL	61.3	50.0	53.3	64.7	54.8
✓	Claude-v1.3	76.4	83.9	72.0	79.7	77.2
✓	Bard	79.3	72.3	73.7	82.0	71.9
✓	PaLM2-Bison	82.3	75.5	63.7	73.1	71.3
	Ada001	46.4	47.7	49.6	52.0	50.0
	Bab001	51.9	57.1	49.5	49.1	53.1
	Cur001	53.5	53.3	51.1	57.5	56.3
✓	Dav001	61.2	56.8	52.9	61.6	58.1
✓	Dav002	74.5	81.3	57.9	78.5	73.2
✓	Dav003	82.3	78.4	62.4	85.5	76.8
✓	GPT3.5-turbo	84.3	82.9	75.1	84.0	86.3
✓	GPT4	91.3	90.1	66.3	85.7	78.0

Table 1: Balanced accuracy on the synthetic FactCC benchmark per prompt group (averaged across prompts in each group). Specialized non-LLMs, (top) Foundation Models, Instruction-tuned LLMs, and Chat-based LLMs (bottom). For LLMs, performance is evaluated with Zero-shot (`ZS`), Few-Shot (`FS`), Persona (`Pers`), Chain-of-Thought (`CoT`), and Generate-with-Evidence (`GwE`) prompts when sequence length allows.

Inconsistency Detection

Test on each error type, averaging the accuracy score across all prompts for LLM models.

Accuracy mostly > 80% in classifying positive (factually correct) examples.

However, lower than random chance when detecting factual inconsistencies. (Especially pronoun swap)

	Error Type					
Model (↓)	POS	DS	ES	NSent	NS	PS
DAE	96.0	12.0	44.0	28.0	52.0	44.0
SummaC	96.0	100.0	100.0	100.0	100.0	80.0
QAFactEval	96.0	84.0	92.0	96.0	96.0	84.0
LLaMa-13B	88.8	10.4	13.6	14.4	12.8	12.8
Alpaca-13B	80.0	30.4	20.0	36.0	25.6	28.0
Dolly-v2-12B	93.6	3.2	11.2	10.4	7.2	5.6
MPT-7B	72.0	36.0	41.6	52.8	38.4	40.0
Vicuna-13B	68.8	59.2	63.2	74.4	65.6	48.8
Cohere-CMD-XL	85.1	32.0	31.5	36.3	26.1	17.1
Claude v1.3	71.7	82.4	80.3	89.1	89.9	78.1
Bard	80.0	68.8	69.3	77.3	83.7	59.2
Palm2	96.5	47.7	45.3	65.1	52.0	38.9
Ada001	58.7	36.5	40.3	45.9	39.2	36.3
Bab001	70.1	33.9	29.6	41.6	30.7	34.7
Cur001	88.0	12.0	17.3	45.1	16.5	12.3
Dav001	88.0	21.9	28.5	50.4	27.5	13.1
Dav002	80.3	66.4	61.3	74.9	71.5	55.5
Dav003	93.1	66.1	58.4	71.2	69.9	39.7
GPT3.5-turbo	87.2	82.4	63.5	87.5	89.1	66.7
GPT4	86.1	74.9	77.3	81.6	84.3	74.1
LLM Avg.	81.64	44.95	44.25	56.10	48.81	38.87

Table 2: Accuracy on the synthetic FactCC benchmark per error type (averaged across all prompts). Specialized non-LLMs (top) Foundation Models, Instruction-tuned LLMs, and Chat-based LLMs (bottom). Performance is assessed individually for positive examples (`POS`) and each of the error types: Date Swap (`DS`), Entity Swap (`ES`), Negated Sentences (`NSent`), Number Swap (`NS`), Pronoun Swap (`PS`).

Factual Reasoning

Manual analysis of responses generated by LLMs that are classified as inconsistent.

Binary accuracy != Accurate explanations.

Different responses to questions that are challenging.

a. Not providing explanations
b. Unrelated explanations
c. Plausible but wrong explanations

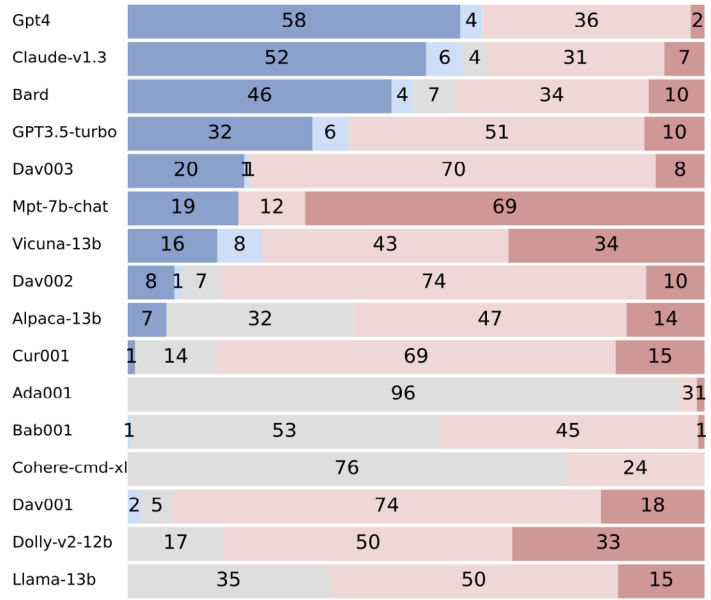


Figure 2: Percentage distribution of the types of explanations each LLM provides in its output when predicting a FactCC summary is inconsistent. Each model explanation is manually annotated as ● entirely correct, ● partially correct, ● no explanation provided, ● unrelated to factuality, ● or incorrect.

Fine-grained Inconsistency Detection

Prompt the models to evaluate each (document, sentence) pair with respect to individual error types and ignoring other types of errors.

Low precision but high recall score, not able to distinguish error types.

	Zero-Shot										Few-Shot									
	DS		ES		NSent		NS		PS		DS		ES		NSent		NS		PS	
Model (↓)	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R
LLaMa-13B	12.0	12.0	18.8	12.0	29.4	40.0	20.8	20.0	12.5	8.0	-	-	-	-	-	-	-	-	-	-
Alpaca-13B	20.5	32.0	12.9	16.0	36.6	60.0	20.9	36.0	23.1	36.0	-	-	-	-	-	-	-	-	-	-
Dolly-v2-12B	26.7	16.0	18.8	12.0	14.3	16.0	22.2	8.0	33.3	12.0	-	-	-	-	-	-	-	-	-	-
MPT-7B-Chat	19.1	52.0	19.7	60.0	23.6	68.0	18.1	52.0	18.8	48.0	-	-	-	-	-	-	-	-	-	-
Vicuna-13B	20.8	100.0	19.1	84.0	23.4	100.0	21.4	96.0	18.3	80.0	-	-	-	-	-	-	-	-	-	-
Cohere-CMD-XL	22.1	92.0	19.4	84.0	26.9	100.0	22.1	100.0	17.3	76.0	17.6	72.0	21.2	72.0	30.8	96.0	20.2	84.0	21.1	80.0
Claude-v1.3	20.3	96.0	19.7	96.0	19.2	92.0	20.8	100.0	20.3	100.0	21.9	92.0	22.0	96.0	22.9	96.0	26.0	100.0	21.0	100.0
PaLM2-Bison	21.1	64.0	21.3	76.0	26.1	92.0	23.6	84.0	20.2	80.0	21.7	60.0	17.5	56.0	28.8	84.0	25.3	80.0	20.2	68.0
Ada001	19.7	92.0	20.0	96.0	19.3	92.0	19.3	92.0	18.4	84.0	33.3	4.0	0.0	0.0	0.0	0.0	22.8	92.0	20.0	4.0
Bab001	4.8	4.0	19.0	16.0	18.2	24.0	10.3	12.0	11.5	12.0	27.3	24.0	20.0	24.0	23.5	32.0	12.5	16.0	19.4	28.0
Cur001	13.5	28.0	22.0	44.0	32.8	84.0	17.9	28.0	15.4	24.0	16.4	44.0	15.9	44.0	26.6	84.0	14.3	32.0	16.5	52.0
Dav001	13.7	28.0	26.3	60.0	39.5	68.0	14.3	28.0	13.0	12.0	18.6	52.0	18.8	36.0	37.5	72.0	17.3	36.0	10.9	24.0
Dav002	20.2	92.0	22.4	96.0	23.4	88.0	21.9	100.0	20.5	92.0	18.3	84.0	19.8	92.0	21.1	96.0	20.3	100.0	19.7	96.0
Dav003	20.4	92.0	20.2	96.0	24.7	96.0	21.7	100.0	20.2	96.0	25.6	92.0	21.6	96.0	24.0	92.0	27.5	100.0	19.3	92.0
GPT3.5-turbo	20.6	88.0	18.0	80.0	24.0	100.0	22.9	100.0	21.1	96.0	20.2	92.0	16.7	76.0	21.8	96.0	22.7	100.0	20.7	96.0
GPT4	31.0	88.0	20.5	96.0	22.0	96.0	24.3	100.0	19.8	96.0	26.4	92.0	20.2	96.0	22.4	96.0	21.9	100.0	20.5	96.0

Table 4: Precision (P) and Recall (R) scores of error detection with fine-grained prompts for individual error types. Experiments run in Zero- and Few-shot settings for each of the error types: Date Swap (`DS`), Entity Swap (`ES`), Negated Sentences (`NSent`), Number Swap (`NS`), Pronoun Swap (`PS`).

Limits of Crowd-Based Benchmark

Analyze existing benchmarks: AggreFact and DialSumEval.

Filter out all models that did not achieve a balanced accuracy above 60% on FactCC.

Use single Zero-Shot (ZS) prompt for all LLM models

AggreFact

A factual consistency benchmark focused on the news domain.

- LLMs perform close to specialized models but all <80%
- After manually examine the responses of GPT4… minimum of 6% of the samples in AggreFact are mislabeled.

Low reliability of crowd-sourced work

	AggreFact	DialSummEval	
Model Name	**%BAcc.**	**%BAcc.**	**Corr.**
DAE	**76.0**	56.2	0.44
SummaC	71.6	62.7	0.35
QAFactEval	73.9	64.4	**0.59**
Cohere-cmd-XL	63.1	56.6	0.36
Claude V1.3	50.6	56.8	0.30
Bard	62.7	59.5	0.26
PaLM2-Bison	57.0	55.6	0.57
Dav001	53.3	52.9	0.11
Dav002	54.3	59.2	0.49
Vicuna-13b	60.3	58.6	0.36
Dav003	64.8	60.9	0.51
GPT3.5-turbo	70.2	62.0	0.56
GPT-4	73.6	**68.4**	0.58

Table 5: Performance of models on the AggreFact, DialSummEval consistency benchmarks reported in balanced accuracy (**%Bacc.**) and correlation (**corr.**).

DialSummEval

The domain of dialogue summarization.

Each (dialogue, summary) tuple is evaluated by three annotators on a Likert score (1-5).

Test on correlation between model predictions and the average annotator score.

Models perform well on non-borderline cases.

A continuous **scale limits the quality and interpretability** of the benchmark. Instead, Factual consistency benchmarks **as a detection task**, if detect inconsistency then negative.

	AggreFact	DialSummEval	
Model Name	**%BAcc.**	**%BAcc.**	**Corr.**
DAE	**76.0**	56.2	0.44
SummaC	71.6	62.7	0.35
QAFactEval	73.9	64.4	**0.59**
Cohere-cmd-XL	63.1	56.6	0.36
Claude V1.3	50.6	56.8	0.30
Bard	62.7	59.5	0.26
PaLM2-Bison	57.0	55.6	0.57
Dav001	53.3	52.9	0.11
Dav002	54.3	59.2	0.49
Vicuna-13b	60.3	58.6	0.36
Dav003	64.8	60.9	0.51
GPT3.5-turbo	70.2	62.0	0.56
GPT-4	73.6	**68.4**	0.58

Table 5: Performance of models on the AggreFact, DialSummEval consistency benchmarks reported in balanced accuracy (**%Bacc.**) and correlation (**corr.**).

	Average Annotator Likert Score							
Model	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
Dav001	68.1	78.4	84.6	90.2	83.6	84.9	86.0	88.9
Cohere-cmd-XL	46.2	51.0	70.3	83.6	88.6	89.2	91.7	96.3
DAE	30.8	56.9	63.7	83.6	86.8	94.3	90.3	94.2
PaLM2-bison	25.3	35.3	56.0	78.7	93.6	97.2	98.4	95.8
Dav002	13.2	29.4	47.3	62.3	77.7	83.0	88.3	90.0
Dav003	4.4	17.6	28.6	31.1	63.2	69.3	84.9	81.6
GPT3.5-turbo	8.8	15.7	29.7	45.9	73.6	76.4	88.5	90.0
GPT4	2.2	5.9	6.6	24.6	45.9	54.2	80.9	87.9
QAFactEval	3.3	5.9	17.6	24.6	44.5	54.7	70.3	74.7
Vicuna-13b	8.8	15.7	17.6	37.7	50.9	54.2	65.5	66.8
SummaC	4.4	5.9	20.9	21.3	27.7	40.1	43.7	58.9
Claude V1.3	1.1	9.8	11.0	13.1	33.6	37.3	47.1	45.8
Bard	9.9	7.8	5.5	9.8	18.2	21.2	36.5	42.6

Table 6: Percent of summaries classified as consistent in DialSummEval, bucketed by average Likert consistency score. Models are more uncertain in mid-range borderline buckets ([2.0, 4.0]).

SummEdits Benchmark

Parker Hutchinson

Design Principles

P1: Binary Classification Task: summary is either consistent or inconsistent

P2: Focus on Factual Consistency: summary is flawless on attributes unrelated to consistency

P3: Reproducibility: labels should be independent of annotator

P4: Benchmark Diversity: inconsistencies should represent a wide range of errors in real textual domains

SummEdits Creation Protocol

3 key steps:

1. **Seed summary verification**: seed summary generated for each document in a small collection
    - Annotator must determine that summary is flawless and factually consistent for the (document, seed summary) tuple to proceed to step 2
2. **Generation of edits**: minor edits are made to the summary (manually or by LLM)
3. **Annotation of edited summaries**: annotator from step 1 reviews each edited summary and assigns a label of consistent, inconsistent, or borderline

SummEdits Creation Protocol: Additional Details

- The same annotator who performs step 1 (read document and seed summary) should perform step 3 (label edited summaries)
  - Minimizes cost, since the annotator already invested the time to read the (document, seed summary) pair
- Use a large (ex. 30) number of edits for the edit summaries
  - Maximizes edit diversity
  - Encourages annotator to apply ‘borderline’ label if unsure, maximizing reproducibility
- ChatGPT (gpt 3.5-turbo) used to generate seed summaries and edited summaries
  - Other LLMs were incapable of generating them
  - Future research could use a variety of models to generate summaries
- Takeaway: protocol only requires a small number of documents and seed summaries because of the many edited summaries generated

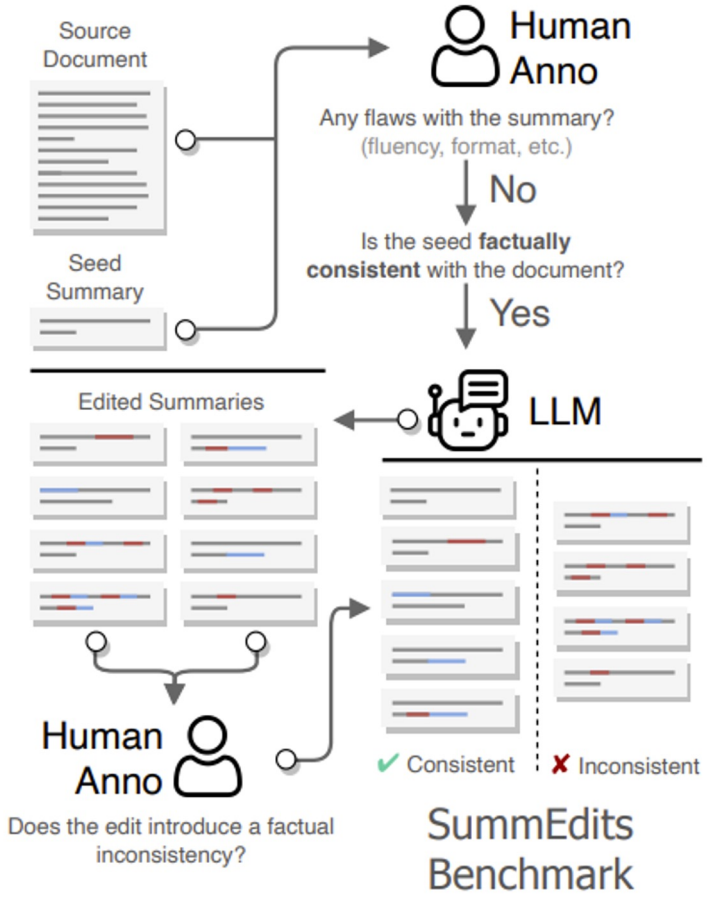


Figure 3: SUMMEDITS protocol diagram, a three-step protocol to create summarization ID benchmarks. See Table 7 for example samples produced by the protocol.

Edited Summary Labeled As **Consistent**	Edited Summary Labeled As **Inconsistent**
The characters discuss ponder the consequences of banishing Marcius, with Cominius warning that his alliance collaboration with the Volscians will bring great danger to Rome.	The characters discuss the consequences of banishing Marcius, with Cominius warning that his alliance with the Volscians Romans will bring great danger to Rome the Volscians. – `Entity Manipulation`
We introduced a novel new, simple, and efficient data augmentation method that boosts improves the performances of existing GANs when training data is limited and diverse.	We introduced a novel, simple, and efficient data augmentation method that boosts the performances of existing GANs when training data is limited abundant and diverse. – `Antonym Swap`
Employees of the European Commission are now forced instructed to delete remove TikTok from their work devices, and delete get rid of it from their personal devices too if they have work-related apps applications installed.	Employees of the European Commission are now forced not required to delete TikTok from their work devices, and delete but should still remove it from their personal devices too if they have work-related apps installed. – `Hallucinated Fact`
A conversation between a sales agent and a potential client possible customer. The sales agent provides information on different home insurance plans options and pricing, as well as available discounts for clients with good credit scores and other factors.	A conversation between a sales agent and a potential client. The sales agent provides information on different home insurance plans and, but not on pricing, as well as or available discounts for clients with good credit scores and other factors. – `Negation Insertion`

Table 7: Example edit summaries – deletions, insertions – for four domains of SUMMEDITS (top-to-bottom: Shakespeare Plays, SciTLDR, News, Sales Call). Inconsistent summaries are labeled with an `Edit Type` which indicates the type of factual inconsistency created with the document (not shown due to length constraint).

SummEdits Creation Protocol: Domains

- **News**: Articles and summaries from Google News top events from February 2023
- **Podcasts**: 40 transcripts from Spotify dataset, automatic summaries
- **BillSum**: 40 US bills and their summaries
- **SamSum**: 40 dialogues and their summaries from a dialogue summarization dataset
- **Shakespeare**: 40 scenes, automatic summaries
- **SciTLDR**: 40 research paper abstracts and their summaries
- **QMSum**: 40 documents and summaries from query-based meeting summarization dataset
- **ECTSum**: 40 documents from financial earnings call dataset, automatic summaries
- **Sales Call & Email**: 40 fictional sales calls & emails generated along with summaries

SummEdits Statistics

- At least 20% of each domain's samples were annotated by multiple annotators
- Cohen's Kappa varied between 0.72-0.90 for the domains when considering the three labels, averaging 0.82
  - After removing 'borderline' samples, average Kappa rose to 0.92 -> high agreement
- Total cost: $3,000 for 150 hours of annotator work
  - Average domain cost is $300
- Using processes of other benchmarks would have had a 20x increase in cost
  - If each sample required 30 min of annotator time, as in the FRANK benchmark

Model	Podcast	BillSum	SAMSum	News	Sales C	Sales E	Shkspr	SciTLDR	QMSum	ECTSum	Overall (↓)
DAE	54.9	55.1	59.5	61.7	50.8	55.0	54.5	55.2	52.0	58.6	55.7
SummaC	58.5	55.7	54.7	62.1	59.0	57.7	59.3	59.7	56.6	64.4	58.8
QAFactEval	64.0	54.4	66.3	74.6	68.5	64.2	61.9	67.5	62.4	72.9	65.7
Dav001	53.3	50.2	51.0	54.4	55.3	52.5	50.0	51.0	50.3	50.9	51.9
Cohere-cmd-XL	51.1	52.7	52.0	52.6	60.3	59.5	50.0	60.5	53.9	60.5	55.1
Vicuna-13b	52.8	52.6	50.8	63.0	58.1	51.8	55.5	59.7	54.0	62.5	56.1
Claude v1.3	59.9	52.1	64.1	63.3	61.7	56.6	58.0	57.6	56.9	67.8	59.8
Dav002	56.4	53.9	57.1	61.9	65.1	59.1	56.6	64.6	60.6	66.2	60.1
Bard	50.0	58.3	61.3	72.8	73.8	69.0	58.4	66.1	53.9	73.1	63.7
PaLM2-bison	66.0	62.0	69.0	68.4	74.5	68.1	61.6	78.1	70.2	72.3	69.0
Dav003	65.7	59.9	67.5	71.2	78.8	69.4	69.6	74.4	72.2	77.9	70.7
GPT3.5-turbo	68.4	63.6	69.1	74.5	79.7	65.5	68.1	75.6	69.2	78.9	71.3
GPT4	83.3	71.1	82.9	83.3	87.6	80.1	84.6	82.4	80.4	88.0	82.4
GPT4 Oracle	90.2	85.5	86.3	88.3	91.1	83.5	96.6	86.3	89.9	91.7	88.9
Human Perf.	90.8	87.5	89.4	90.0	91.8	87.4	96.9	89.3	90.7	95.4	90.9

Table 9: Balanced accuracy of models on the SUMMEDITS benchmark. The top three models are non-LLM specialized models, the middle section are LLMs. We also report a GPT4 oracle performance and an estimate of human performance.

SummEdits Results

- Low performance overall - only GPT-4 comes within 10% of human performance
- Only 4 LLMs outperform non-LLM QAFactEval - most LLMs are not capable of reasoning about the consistency of facts out-of-the-box
- Specialized models performed best on News, probably because it was similar to their training data
- BillSum and Shakespeare are particularly challenging
- Oracle test: model is given document, seed, and edited summary
    - Large boost in performance, within 2% of human performance
    - Shows that high performance is indeed attainable

SummEdits Edit Types

1. Entity modification
2. Antonym Swap
3. Hallucinated Fact Insertion
4. Negation Insertion

- SummEdits distribution: 78% of inconsistent summaries contain entity modification, 48% antonym swap, 22% hallucinated fact insertion, 18% negation insertion
    - Distribution influenced by the LLM used to produce the edits

Model	Inconsistent Edit Type			
	EntMod	Anto	Hallu	Neg
DAE	52.0	53.0	52.9	53.9
SummaC	56.8	56.8	55.3	57.3
QAFactEval	61.4	65.0	64.3	70.4
Dav001	50.0	50.9	50.8	53.7
Cohere-cmd-XL	53.7	55.8	55.5	63.8
Vicuna-13b	55.2	57.1	56.2	61.0
Claude v1.3	58.8	60.3	61.5	66.7
Dav002	58.3	61.4	62.4	72.0
Bard	63.2	65.3	65.6	71.3
PaLM2-Bison	67.0	70.0	71.7	80.3
Dav003	69.2	71.1	76.3	83.3
GPT3.5-turbo	70.7	70.6	74.2	79.7
GPT4	82.2	81.3	87.0	92.7
Average	61.4	62.9	64.1	69.7

Table 10: Balanced accuracy of models on the SUMMEDITS benchmark, broken down by type of factual error: Entity Modification (`EntMod`), Antonyms (`Anto`), Hallucination (`Hallu`) and Negation (`Neg`) insertion.

Model	#Distinct Edit Types			
	1	2	3	4
DAE	50.2	53.5	55.4	64.9
SummaC	58.2	56.3	57.6	67.3
QAFactEval	59.4	63.7	72.3	76.5
Dav001	50.0	50.5	53.9	63.1
Vicuna-13b	52.8	57.0	60.2	58.5
Cohere-cmd-XL	50.0	55.9	63.7	70.0
Claude v1.3	57.5	60.6	65.4	64.3
Dav002	56.3	61.2	69.4	81.7
Bard	61.0	64.9	72.4	73.4
PaLM2-Bison	66.1	69.5	79.6	69.4
ChatGPT	68.5	71.4	82.0	86.6
Dav003	65.3	72.0	85.8	88.8
GPT4	81.0	83.0	92.0	94.3
Average	59.2	62.5	69.2	74.1

Table 11: Relationship between the number of edits types in the summary and balanced accuracy of models on SUMMEDITS. Models generally perform better as the number of introduced edits in a summary increases.

Discussion

- Why not fix existing benchmarks?
  - Would require re-annotating a large portion of the dataset, with no guarantee that there would be an improvement
- Effect of LLM in benchmark creation: could favor LLMs most similar to the one used for summary generation
- Evaluation of underlying summarizers

Conclusion

- Simplified annotation process for improved reproducibility
- SummEdits benchmark created which spans 10 domains
  - Highly reproducible and more cost-effective than previous benchmarks
  - Challenging for most current LLMs
  - Valuable tool for evaluating LLMs' ability to reason about facts and detect factual errors
- Authors encourage LLM developers to report their performance on the benchmark

More:

Survey of Hallucination in Natural Language Generation

ZIWEI JI, NAYEON LEE, RITA FRIESKE, TIEZHENG YU, DAN SU, YAN XU, ETSUKO ISHII, YEJIN BANG, DELONG CHEN, HO SHU CHAN, WENLIANG DAI, ANDREA MADOTTO, and PASCALE FUNG, Center for Artificial Intelligence Research (CAiRE), Hong Kong University of Science and Technology, Hong Kong

Natural Language Generation (NLG) has improved exponentially in recent years thanks to the development of sequence-to-sequence deep learning technologies such as Transformer-based language models. This advancement has led to more fluent and coherent NLG, leading to improved development in downstream tasks such as abstractive summarization, dialogue generation and data-to-text generation. However, it is also apparent that deep learning based generation is prone to hallucinate unintended text, which degrades the system performance and fails to meet user expectations in many real-world scenarios. To address this issue, many studies have been presented in measuring and mitigating hallucinated texts, but these have never been reviewed in a comprehensive manner before.

In this survey, we thus provide a broad overview of the research progress and challenges in the hallucination problem in NLG. The survey is organized into two parts: (1) a general overview of metrics, mitigation methods, and future directions; (2) an overview of task-specific research progress on hallucinations in the following downstream tasks, namely abstractive summarization, dialogue generation, generative question answering, data-to-text generation, machine translation, and visual-language generation; and (3) hallucinations in large language models (LLMs) [1]. This survey serves to facilitate collaborative efforts among researchers in tackling the challenge of hallucinated texts in NLG.

(1) **Intrinsic Hallucinations**: The generated output that contradicts the source content. For instance, in the abstractive summarization task from Table 1, the generated summary "*The first Ebola vaccine was approved in 2021*" contradicts the source content "*The first vaccine for Ebola was approved by the FDA in 2019.*".

(2) **Extrinsic Hallucinations**: The generated output that cannot be verified from the source content (i.e., output that can neither be supported nor contradicted by the source). For example, in the abstractive summarization task from Table 1, the information "*China has already started clinical trials of the COVID-19 vaccine.*" is not mentioned in source. We can neither find evidence for the generated output from the source nor assert that it is wrong. Notably, the extrinsic hallucination is not always erroneous because it could be from factually correct external information [177, 247]. Such factual hallucination can be helpful because it recalls additional background knowledge to improve the informativeness of the generated text. However, in most of the literature, extrinsic hallucination is still treated with caution because its unverifiable aspect of this additional information increases the risk from a factual safety perspective.

	Category	Task	Works
Automatic Metrics	Statistical	Dialogue	Shuster et al. [233]
		Data2Text	Dhingra et al. [44], Wang et al. [274]
		Translation	Martindale et al. [175]
		Captioning	Rohrbach et al. [222]
	Model-based	Abstractive Summarization	Durmus et al. [53], Kryscinski et al. [127], Nan et al. [190], Wang et al. [264] Gabriel et al. [74], Goodrich et al. [86], Pagnoni et al. [197], Zhou et al. [326] Falke et al. [65], Laban et al. [132], Mishra et al. [185], Scialom et al. [227]
		Dialogue	Balakrishnan et al. [9], Honovich et al. [101], Li et al. [153] Dziri et al. [60], Gupta et al. [94], Santhanam et al. [226]
		Generative QA	Sellam et al. [229]*,Zhang et al. [318]*,Durmus et al. [53]* Wang et al. [264]*, Su et al. [237]
		Data2Text	Dušek and Kasner [56], Liu et al. [162], Wiseman et al. [283] Filippova [71], Rebuffel et al. [217], Tian et al. [251]
		Translation	Kong et al. [125], Lee et al. [134], Tu et al. [257] Feng et al. [70], Garg et al. [79], Zhou et al. [326] Parthasarathi et al. [199], Raunak et al. [215]
		Task-Agnostic	Goyal and Durrett [90], Liu et al. [160], Zhou et al. [326]

Mitigation Method	Data-Related	Abstractive Summarization	Cao et al. [24], Nan et al. [190], Zhu et al. [329] Gunel et al. [92]
		Dialogue	Honovich et al. [101], Shen et al. [230], Wu et al. [285] Santhanam et al. [226], Shuster et al. [233]
		Generative QA	Bi et al. [14], Fan et al. [66], Yin et al. [297]
		Data2Text	Liu et al. [162], Nie et al. [193], Parikh et al. [198], Wang [268] Nie et al. [192], Rebuffel et al. [216]
		Translation	Lee et al. [134], Raunak et al. [215] Briakou and Carpuat [18], Junczys-Dowmunt [112]
		Captioning	Biten et al. [16]
	Modeling and Inference	Abstractive Summarization	Huang et al. [102], Li et al. [143], Song et al. [235] Aralikatte et al. [6], Cao et al. [21], Cao and Wang [22] Albrecht and Hwa [4], Chen et al. [31], Zhao et al. [322]
		Dialogue	Balakrishnan et al. [9], Li et al. [153], Rashkin et al. [214] Dziri et al. [59]
		Generative QA	Fan et al. [66], Krishna et al. [126], Li et al. [142] Nakano et al. [189], Su et al. [237]
		Data2Text	Liu et al. [162], Tian et al. [251], Wang et al. [269, 274], Xu et al. [291] Filippova [71], Rebuffel et al. [216], Su et al. [239], Xiao and Wang [286] Puduppully and Lapata [208]
		Translation	Feng et al. [70], Lee et al. [134], Weng et al. [281] Li et al. [150], Raunak et al. [215], Wang and Sennrich [267] Bengio et al. [12], Zhou et al. [326] Goyal et al. [89], Xu et al. [290]
		Captioning	Dai et al. [41], Xiao and Wang [286]

Table 2. Evaluation metrics and mitigation methods for each task. *The hallucination metrics are not specifically proposed for generative question answering (GQA), but they can be adapted for that task.