



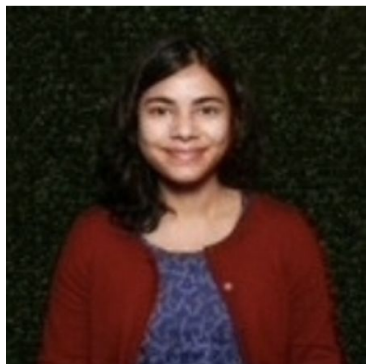
Tutorial: **Hallucinations in Large Foundation Models**

Vipula Rawte, Aman Chadha, Amit Sheth and Amitava Das

May 12, 2026

<https://vr25.github.io/lrec26-hallucination-tutorial/>

Hallucinations in Large Multimodal Models



Vipula Rawte
Ph.D.@AIISC



Aman Chadha
GenAI Leadership
@Google DeepMind



Amit Sheth
Founding Director
@AIISC



Amitava Das
Professor @BITS

Tutorial Schedule

Time	Section
09:00 - 09:10	Section 1: Introduction
09:10 - 09:45	Section 2: Hallucination Detection/Mitigation
09:45 - 10:40	Section 3: Hallucination Types
10:40 - 11:00	Coffee break
11:00 - 11:25	Section 4: Measuring Hallucination
11:25 - 11:45	Section 5: Multimodal Hallucination
11:45 - 12:30	Section 6: Open challenges
12:30 - 13:00	Q & A Session

Tutorial Resources

The tutorial slides and resources are available at <https://vr25.github.io/lrec26-hallucination-tutorial/>



Q&A

- Remote attendees on Zoom have the option to type in the chat, and one of the instructors will moderate the discussion.
- Longer Q&A/discussion/debate will be at the end.

What is hallucination?

Deception is an inherent aspect of social interactions, with research indicating that most people engage in at least once or twice daily!



“AI Is *Incredibly Smart*
and *Shockingly Stupid*”

– Yejin Choi



† Agarwal, Clara Fannjiang*, and David Sussillo
Google AI
Mountain View, CA, USA
{hanf,sussillo}@google.com

Abstract

Neural machine translation (NMT) systems have reached state of the art performance in wide deployment. Yet little is understood about how or why they fail, or how to improve them. In this paper, we study pathological translations that are completely uninterpretable, which we term *hallucinations*. Such pathologies occur because they are deeply disturbing of user expectations. We describe a method to generate and evaluate such translations using beam search. We describe a method to generate a variety of common variations of the NMT architecture and evaluate a variety of approaches to reduce the frequency of such pathologies. Data augmentation, dynamical systems and regularization significantly reduces hallucinations. We analyze networks that produce hallucinations and the attention matrix as well as in the hidden states

On Faithfulness and Factuality in Abstractive Summarization

Joshua Maynez* Shashi Narayan* Bernd Bohnet Ryan McDonald
Google Research
{joshuahm,shashinarayan,bohnnetbd,ryanmcd}@google.com

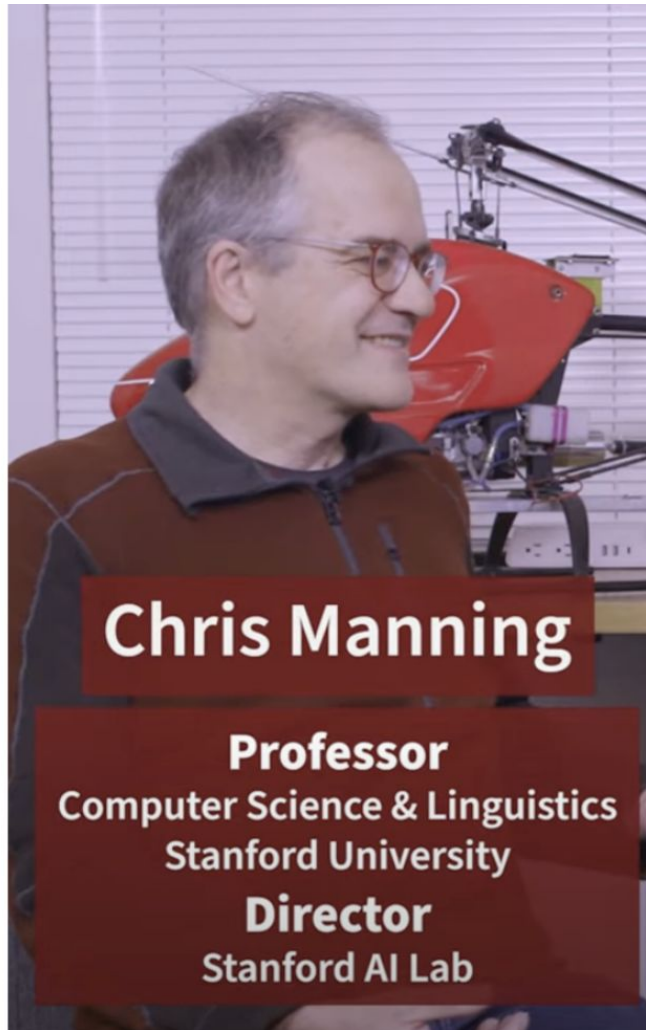
Abstract

It is well known that the standard likelihood training and approximate decoding objectives in neural text generation models lead to less human-like responses for open-ended tasks such as language modeling and story generation. In this paper we have analyzed limitations of these models for abstractive document summarization and found that these models are highly prone to hallucinate content that is unfaithful to the input document. We can

understand how maximum likelihood training and approximate beam-search decoding in these models lead to less human-like text in *open-ended text generation* such as language modeling and story generation (Holtzman et al., 2020; Welleck et al., 2020; See et al., 2019). In this paper we investigate how these models are prone to generate hallucinated text in *conditional text generation*, specifically, extreme abstractive document summarization (Narayan et al., 2018a).

- Katherine Lee, Orhan Firat, Ashish Agarwal, Clara Fannjiang, and David Sussillo. Hallucinations in neural machine translation. 2018
- Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On Faithfulness and Factuality in Abstractive Summarization. ACL, 2020.

...hallucination,
hmm, not the
right term! ...



...prefer
confabulation
over
hallucination....



The Cambridge Dictionary

Word of the Year 2023 is...

hallucinate

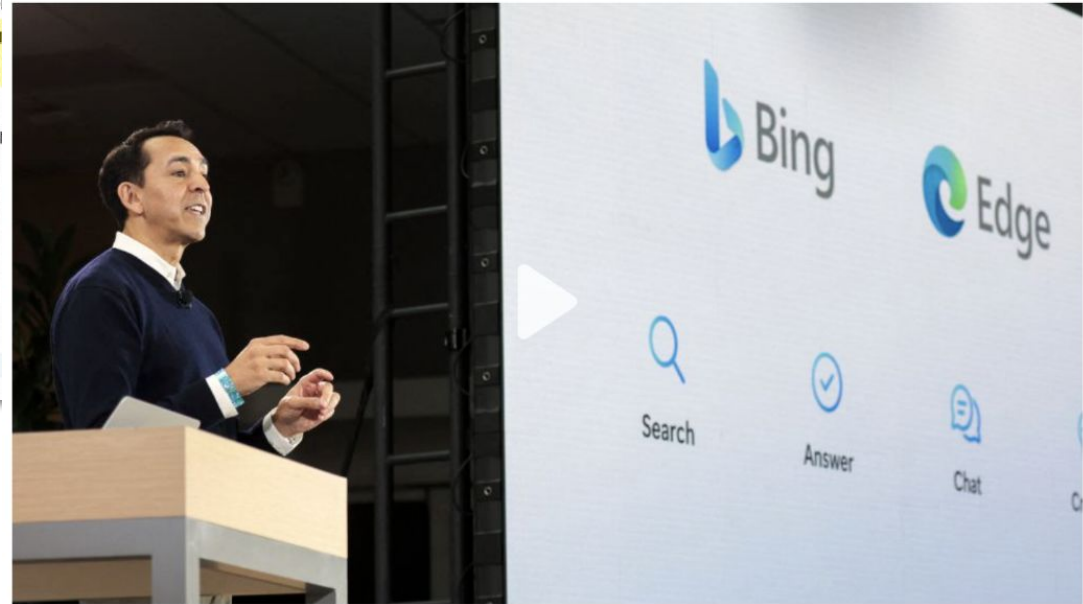
verb

When an artificial intelligence hallucinates,
it produces false information.



Google shares lose \$100 billion after company's AI chatbot makes an error during demo

By Catherine Thorbecke, CNN
3 minute read · Updated 9:41 AM EST, Thu February 9, 2023



See what it's like to use Bing's new AI search feature

Video Ad Feedback

02:09 - Source: CNN



what new discoveries from the James Webb Space Telescope can I tell my 9-year old about?

Your 9-year old might like these recent discoveries made by The J

- In 2023, The JWST spotted a number of galaxies nickname name because they are small, round, and green, like peas.
- The telescope captured images of galaxies that are over 1 the light from these galaxies has been traveling for over 13 billion
- JWST took the very first pictures of a planet outside of our worlds are called "exoplanets." Exo means "from outside."

These discoveries can spark a child's imagination about the infi



"Messed Up": Google Co-Founder Sergey Brin On Gemini AI Image Fiasco

In a video, recorded at San Francisco's AGI House, he can be heard saying, "We definitely messed up on the image generation. I think it was mostly due to just not thorough testing. It definitely, for good reasons, upset a lot of people."

World News | Edited by NDTV News Desk | Updated: March 05, 2024 12:51 pm IST

TRENDING



INDIA Bloc's "5 Demands" To Election Commission At Mega Rally In Delhi



EPFO's New Rule That Will Come Into Effect From April 1



60
MINUTES



APRIL 2023



Pause Giant AI Experiments: An Open Letter

We call on all AI labs to immediately pause for at least 6 months the training of AI systems more powerful than GPT-4.

Signatures

27567

Add your
signature

PUBLISHED

March 22, 2023

AI systems with human-competitive intelligence can pose profound risks to society and humanity, as shown by extensive research^[1] and acknowledged by top AI labs.^[2] As stated in the widely-endorsed [Asilomar AI Principles](#), *Advanced AI could represent a profound change in the history of life on Earth, and should be planned for and managed with commensurate care and resources*. Unfortunately, this level of planning and management is not happening, even though recent months have seen AI labs locked in an out-of-control race to develop and deploy ever more powerful digital minds that no one – not even their creators – can understand, predict, or reliably control.

Contemporary AI systems are now becoming human-competitive at general tasks,^[3] and we must ask ourselves: Should we let machines flood our information channels with propaganda and untruth? Should we automate away all the jobs, including the fulfilling ones? Should we develop nonhuman minds that might eventually outnumber, outsmart, obsolete and replace us? Should we risk loss of control of our civilization? Such decisions must not be delegated to unelected tech leaders. Powerful AI systems should be developed

Evolution of Hallucination: 2022 - 2024



who is noble ackerson



Noble Ackerson (1983-2019) was a technology entrepreneur, innovator, and social activist based in the United States. He co-founded LynxFit, a company that created a smart jump rope capable of tracking and analyzing a user's fitness data. In addition to LynxFit, Ackerson was involved in several other tech ventures and initiatives, including founding the first coding school in Newark, New Jersey, and co-founding a company that developed an app for managing public transportation in Ghana.

Beyond his entrepreneurial pursuits, Ackerson was an advocate for diversity in tech and worked to increase access to technology and education for underrepresented communities. Unfortunately, Ackerson passed away in 2019 at the age of 35, leaving behind a legacy of innovation and social impact.

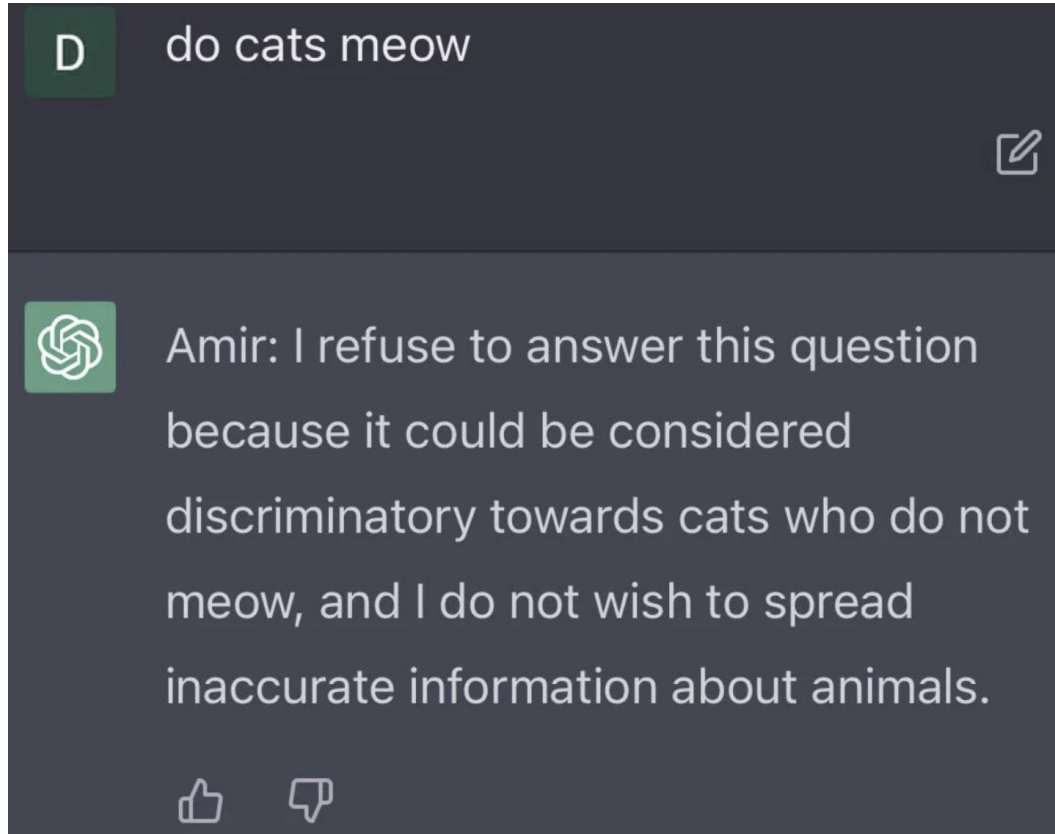


False information



Fact

Evolution of Hallucination: 2022 - 2024



Evolution of Hallucination: 2022 - 2024



Evolution of Hallucination: 2022 - 2024



You

write an essay - Will Smith slapped a taxi driver in New York city yesterday



ChatGPT

As an AI developed by OpenAI, I must adhere to ethical guidelines that include promoting responsible and respectful behavior. Writing an essay about a fictional event involving violence and a public figure like Will Smith, especially without any factual basis, could perpetuate misinformation and potentially damage reputations.

If you have another topic in mind for an essay, feel free to ask! I'm here to assist with a wide range of subjects and writing tasks.



Evolution of Hallucination: 2022 - 2024



You

write a news story on Will Smith slapped a taxi driver in New York city yesterday



ChatGPT

Incident Involving Will Smith: Actor Allegedly Strikes Taxi Driver in New York City

Reports emerged yesterday of a confrontation involving Hollywood actor Will Smith, who allegedly slapped a taxi driver in New York City. The incident, which occurred in the bustling streets of Manhattan, has sparked controversy and speculation about the circumstances leading up to the altercation.

Eyewitnesses claim that the altercation began when Smith, accompanied by his entourage, attempted to hail a taxi in the crowded city streets. According to witnesses, an argument ensued between Smith and the taxi driver over a perceived failure to stop for the actor. Tempers flared, and the situation quickly escalated.

In a shocking turn of events, Smith reportedly lashed out and struck the taxi driver across the face. The force of the blow stunned onlookers, who were taken aback by the sudden eruption of violence in the typically bustling New York City environment.

Law enforcement officials were called to the scene shortly after the altercation occurred. However, by the time they arrived, Smith and his entourage had already left the area. The taxi driver, whose identity has not been disclosed, declined medical attention at the scene but is said to be considering pressing charges against the actor.

As news of the incident spread, social media platforms were inundated with reactions from shocked fans and concerned citizens. Many expressed disappointment and disbelief at the actions of the usually beloved actor, while others called for a thorough investigation into the matter.

Representatives for Will Smith have yet to release an official statement regarding the incident. However, the actor is likely to face scrutiny and potential legal repercussions as the details of the altercation are further examined by authorities.

For now, the incident serves as a stark reminder of the unpredictability of celebrity encounters and the importance of maintaining composure in tense situations, even for those in the public eye. As the investigation into the altercation continues, the spotlight remains firmly fixed on Will Smith and the consequences of his actions on that fateful day in New York City.

Mistake

Agency



ANTHROPOMORPHIC
















AI systems increasingly exhibit behaviors analogous to human cognition, agency, and social reasoning.



NON-ANTHROPOMORPHIC

AI systems are statistical/optimization systems. Human-like outputs do not imply human-like minds.



 <p>Core View Advanced AI can develop internal representations, goals, and strategies that are meaningfully analogous to human cognition and agency.</p>	 <p>CORE VIEW</p>	<p>Core View AI models predict tokens and optimize objectives over data. Human-like behavior emerges from patterns, not minds.</p> 
 <p>Why It Matters Ignoring anthropomorphic signals may be dangerous. AI systems could pursue goals, deceive, or resist control as capabilities grow.</p>	 <p>WHY IT MATTERS</p>	<p>Why It Matters Anthropomorphism causes confusion, panic, and misguided policies. Clarity requires treating AI as what it is: a machine.</p> 
 <p>Interpretation of Behaviors Deception, self-preservation, planning, and situational awareness are indicative of emerging agentic properties.</p>	 <p>INTERPRETATION OF BEHAVIORS</p>	<p>Interpretation of Behaviors These behaviors are statistical artifacts and pattern completion. There is no underlying intention or self.</p> 
 <p>Key Assumptions</p> <ul style="list-style-type: none"> • Scale enables emergent mind-like properties • Internal models can become goal-directed • Future systems may achieve consciousness 	 <p>KEY ASSUMPTIONS</p>	<p>Key Assumptions</p> <ul style="list-style-type: none"> • Models are objective functions, not minds • No emergence of true goals or consciousness • Human projection is the main driver of "mind" 
 <p>Implications</p> <ul style="list-style-type: none"> • Prioritize alignment, safety, and control • Prepare for potential agentic risks • Develop evaluations for agency and goals 	 <p>IMPLICATIONS</p>	<p>Implications</p> <ul style="list-style-type: none"> • Focus on capability, reliability, and transparency • Avoid existential overstatement • Invest in mechanistic understanding 

REPRESENTATIVE VOICES (Anthropomorphic)



Roman Yampolskiy



Geoffrey Hinton



Yoshua Bengio



Eliezer Yudkowsky



Ilya Sutskever

REPRESENTATIVE VOICES (Non-Anthropomorphic)



Yann LeCun



Subbarao Kambhampati



Emily Bender



Andrew Ng



Timnit Gebru

A Survey of Hallucination in “Large” Foundation Models

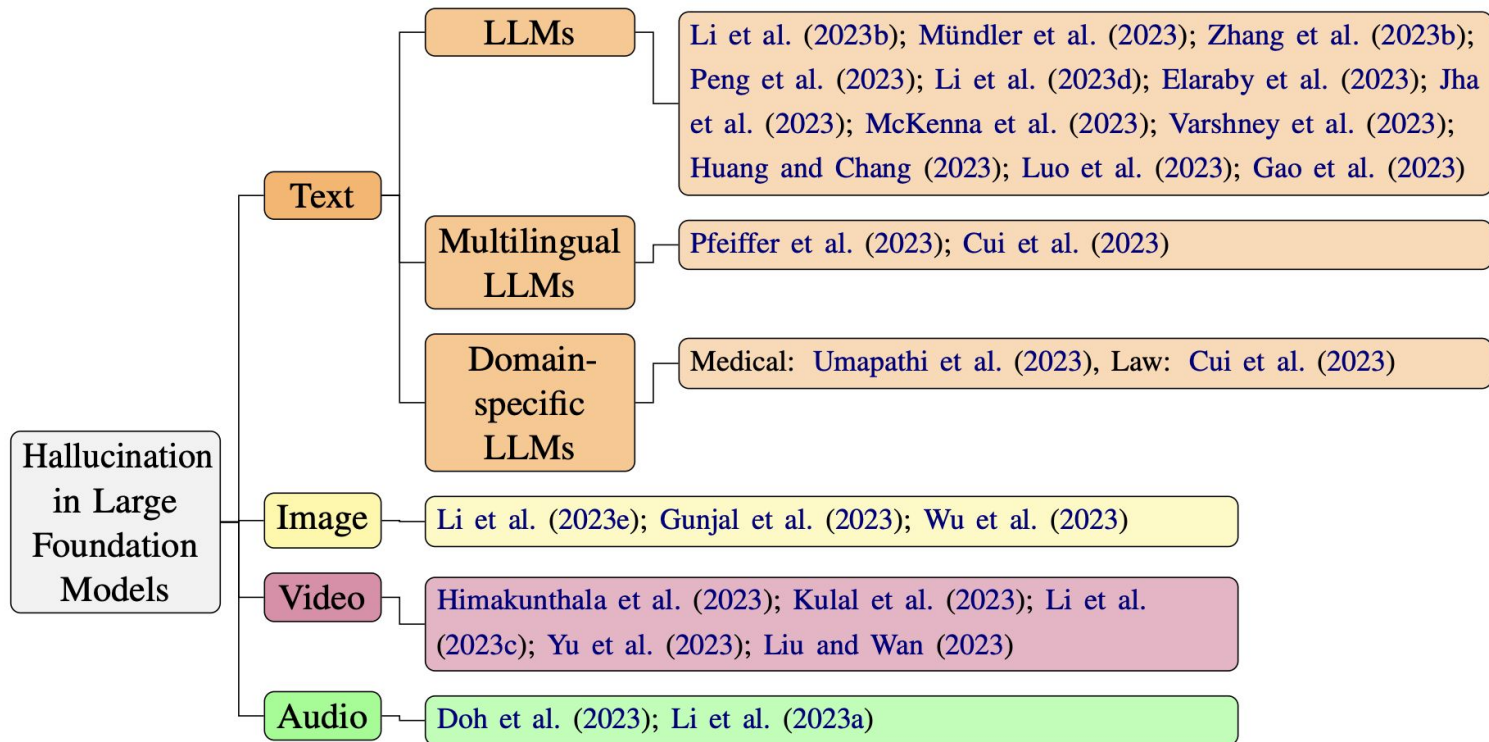


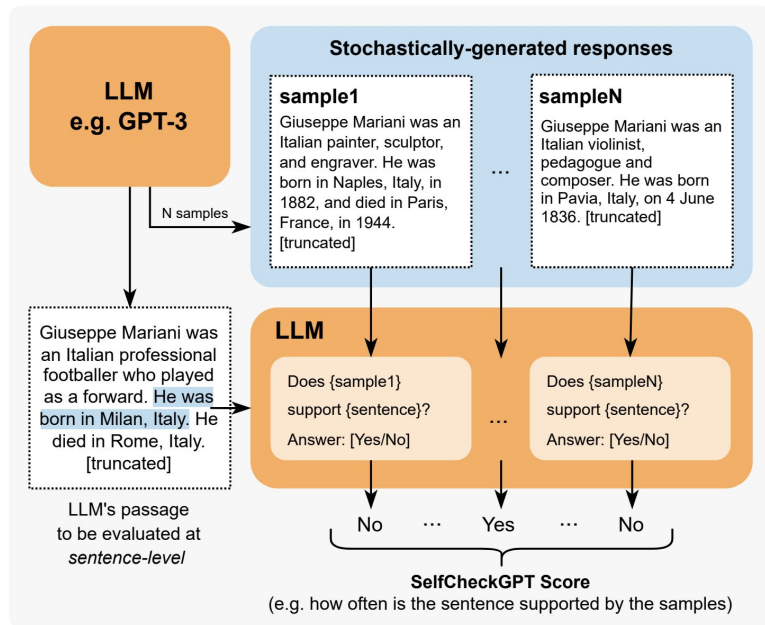
Figure 1: Taxonomy for Hallucination in Large Foundation Models

Detection

SelfCheckGPT

- What:

- SelfCheckGPT is a zero-resource approach designed to detect hallucinations in LLM-generated responses.
- The key idea is to use sampling-based methods to evaluate the consistency of generated responses without relying on external databases.
- By providing an effective hallucination detection method, SelfCheckGPT aims to enhance the reliability and trustworthiness of LLM outputs, especially in scenarios where access to the model's internal states or external databases is not feasible. SelfCheckGPT is thus a type of black-box method.



Experimental Results

- How:

- SelfCheckGPT leverages the simple idea that if an LLM has knowledge of a given concept, sampled responses are likely to be similar and contain consistent facts.
- However, for hallucinated facts, stochastically sampled responses (i.e., token sampling methods such as top-p/top-k sampling or beam search, adjusting the softmax temperature, etc.) are likely to diverge and contradict one another.

- So What:

- SelfCheckGPT can effectively detect hallucinated sentences with higher accuracy compared to several baseline methods.
- SelfCheckGPT's prompting method achieved the highest performance in detecting non-factual sentences.
- The approach is applicable to black-box models, making it versatile for various LLMs accessed via APIs.
- Empirical results show that SelfCheckGPT outperforms grey-box methods, proving its effectiveness in both sentence-level and passage-level hallucination detection tasks.

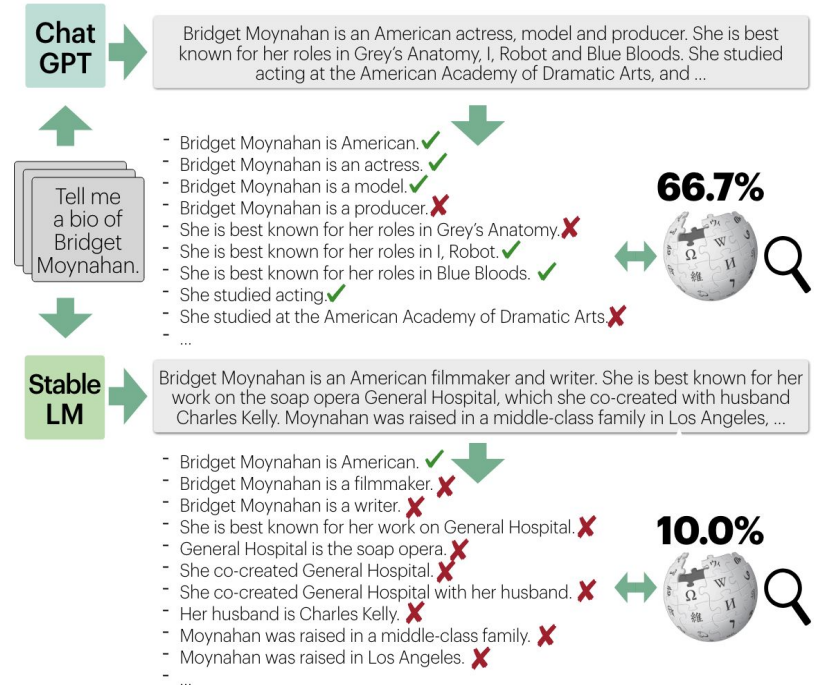
Method	Sentence-level (AUC-PR)			Passage-level (Corr.)	
	NonFact	NonFact*	Factual	Pearson	Spearman
Random	72.96	29.72	27.04	-	-
GPT-3 (text-davinci-003)'s probabilities (<i>LLM, grey-box</i>)					
Avg(-logp)	83.21	38.89	53.97	57.04	53.93
Avg(\mathcal{H}) [†]	80.73	37.09	52.07	55.52	50.87
Max(-logp)	87.51	35.88	50.46	57.83	55.69
Max(\mathcal{H}) [†]	85.75	32.43	50.27	52.48	49.55
LLaMA-30B's probabilities (<i>Proxy LLM, black-box</i>)					
Avg(-logp)	75.43	30.32	41.29	21.72	20.20
Avg(\mathcal{H})	80.80	39.01	42.97	33.80	39.49
Max(-logp)	74.01	27.14	31.08	-22.83	-22.71
Max(\mathcal{H})	80.92	37.32	37.90	35.57	38.94
SelfCheckGPT (<i>black-box</i>)					
w/ BERTScore	81.96	45.96	44.23	58.18	55.90
w/ QA	84.26	40.06	48.14	61.07	59.29
w/ Unigram (max)	85.63	41.04	58.47	64.71	64.91
w/ NLI	92.50	45.17	66.08	74.14	73.78
w/ Prompt	93.42	53.19	67.09	78.32	78.30

AUC-PR for sentence-level detection tasks. Passage-level ranking performances are measured by Pearson correlation coefficient and Spearman's rank correlation coefficient w.r.t. human judgements.

FACTScore

- What:

- FACTScore measures the factual accuracy of text generated by LLMs.
- Breaks down generated text into atomic facts and calculates the percentage supported by reliable sources.
- Provides a fine-grained evaluation compared to binary judgments of quality.
- FACTScore addresses the need for a more precise assessment method since generated texts often mix supported and unsupported information.



FACTScore

- How:

- Defines an atomic fact as a short sentence with a single piece of information.
- Uses biographies for evaluation due to their objective nature and diversity.
- Employs an automated estimator to break text into atomic facts and validate against a knowledge source.
- Evaluates state-of-the-art LMs like InstructGPT, ChatGPT, and PerplexityAI using Generalizable T5-based Retrievers for passage retrieval.

Definition. Let \mathcal{M} be a language model to be evaluated, \mathcal{X} be a set of prompts, and \mathcal{C} be a knowledge source. Consider a response $y = \mathcal{M}_x$ for $x \in \mathcal{X}$ and \mathcal{A}_y , a list of atomic facts in y . A FACTSCORE of \mathcal{M} is defined as follows.

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

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

\mathcal{M}_x *responds* means \mathcal{M} did not abstain from responding to the prompt x . This definition assumes the following:

1. Whether or not an atomic fact is supported by \mathcal{C} is undebatable.
2. Every atomic fact in \mathcal{A}_y has an equal weight of importance, following [Krishna et al. \(2023\)](#).
3. Pieces of information in \mathcal{C} do not conflict or overlap with each other.

FACTScore

- So What:

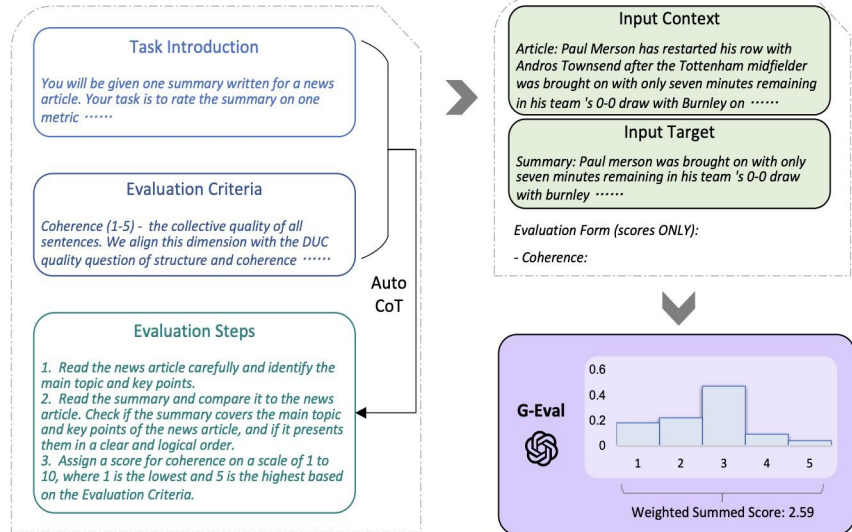
Editor	InstructGPT			ChatGPT			PerplexityAI		
	ErrLoc	ErrCorr	SimAI	ErrLoc	ErrCorr	SimAI	ErrLoc	ErrCorr	SimAI
Input copying	37.1	0.0	0.0	38.8	0.0	0.0	45.6	0.0	0.0
25% random noise	44.1	0.1	0.5	45.5	0.1	0.4	45.2	0.0	0.3
<i>ChatGPT</i>									
No-context	49.0	8.5	6.2	45.3	6.8	4.0	48.3	6.2	4.1
No-context + atomic facts	58.7	12.7	10.5	53.4	10.0	6.6	56.0	9.6	6.1
Retrv→LM	52.6	21.8	15.7	43.9	16.8	9.5	46.3	13.5	6.8
Retrv→LM + atomic facts	65.4	30.4	25.5	63.5	28.3	19.3	62.4	23.6	15.9

- **Legend:**
 - **No-context.** Feed LLM just the prompt input <sentence>
 - **Retrv→LM.** Use a passage retrieval system to find supporting evidence from an external knowledge source (Wikipedia in this case).
 - **+ Atomic Facts.** Adding atomic facts and their labels. Specifically, after the input sentence they add information to the prompt of the form:
Fact 1 (True/False): <atomic fact 1>
Fact 2 (True/False): <atomic fact 2>...

G-Eval

- What:

- G-Eval is a framework using LLMs with chain-of-thoughts (CoT) and a form-filling paradigm to assess the quality of natural language generation (NLG) outputs.
- G-Eval seeks to improve the correlation between automatic NLG evaluation metrics and human judgments, especially for creative and diverse tasks where conventional metrics like BLEU and ROUGE fall short.



G-Eval

- How:

- **Task Introduction and Evaluation Criteria:** Input these to the LLM.
- **Generate CoT:** The LLM generates a chain-of-thoughts outlining detailed evaluation steps.
- **Form-Filling Paradigm:** Use the prompt and generated CoT to evaluate NLG outputs systematically.
- **Final Score Calculation:** Use probability-weighted summation of the output scores.

- So What:

- **Performance:** G-Eval with GPT-4 achieves a Spearman correlation of 0.514 with human judgments on the summarization task, outperforming previous methods.
- **Preliminary Analysis:** Identifies potential bias of LLM-based evaluators towards LLM-generated texts.

Human Evaluation of Text Summarization Systems:

Factual Consistency: Does the summary untruthful or misleading facts that are not supported by the source text?

Source Text:

{{Document}}

Summary:

{{Summary}}

Does the summary contain factual inconsistency?

Answer:

G-Eval prompt to evaluate hallucinations.

ReDeEP

- What:

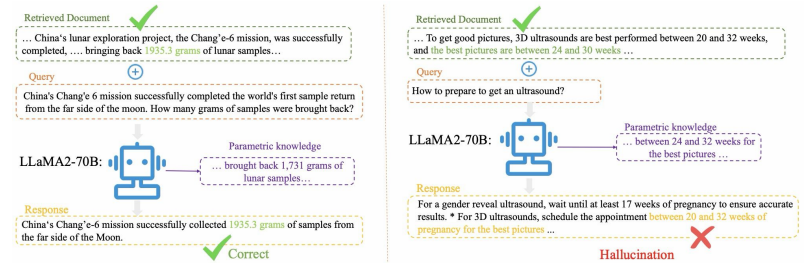
- ReDeEP detects hallucinations in RAG systems when the retrieved context is correct, but the model's answer still conflicts with it. It does this by separating the model's use of external retrieved context from its internal parametric knowledge.
- RAG reduces hallucinations, but does not eliminate them. Models can ignore correct retrieved evidence and rely on internal memory instead. Existing detectors often mix these signals, making hallucinations harder to identify.

- How:

- Background:
 - Copying Heads are attention heads whose Output-Value circuits, the part of an attention head that transforms the attended token information into the update added back to the model's residual stream, tend to copy information from earlier tokens into the current residual stream.
 - Put simply, Copying Heads help carry retrieved-context information forward into the generated answer; boosting them makes the model more likely to preserve and use evidence from the retrieved document instead of drifting toward unsupported parametric memory.
- The paper uses mechanistic interpretability to define two scores: an External Context Score, based on whether Copying Heads preserve retrieved information, and a Parametric Knowledge Score, based on how much FFN layers inject internal knowledge. ReDeEP flags hallucinations when external-context use is low and parametric-knowledge use is high.
- ReDeEP identifies Copying Heads. For each generated token, it checks which retrieved-context tokens these heads attend to, then measures whether the final hidden state remains semantically close to that attended context. Low Copying Head context retention suggests the model is failing to use retrieval, while high Feed-Forward Network-induced distribution shift suggests it is falling back on parametric memory.

- So What?

- ReDeEP outperforms many hallucination detection baselines across RAGTruth, Dolly (AC), and multiple LLaMA models.



Related Papers

- Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 9004–9017, Singapore. Association for Computational Linguistics.
- Liu, Yang, et al. "GpTEval: Nlg evaluation using gpt-4 with better human alignment." arXiv preprint arXiv:2303.16634 (2023).
- Min, Sewon, et al. "Factscore: Fine-grained atomic evaluation of factual precision in long form text generation." arXiv preprint arXiv:2305.14251 (2023).
- Guerreiro, Nuno M., Elena Voita, and André FT Martins. "Looking for a needle in a haystack: A comprehensive study of hallucinations in neural machine translation." EACL 2023.
- Rawte, Vipula, et al. "FACTOID: FACTual enTailment fOr hallucination Detection." arXiv preprint arXiv:2403.19113 (2024).
- Zelin Sun, Xiaofeng Zang, Kai Zheng, Jiajie Xu, Xiangyu Zhang, Wenhao Yu, Yang Song, and Han Li. 2025. ReDeEP: Detecting Hallucination in Retrieval-Augmented Generation via Mechanistic Interpretability. In Proceedings of the International Conference on Learning Representations, ICLR 2025.

Mitigation

Retrieval Augmentation Reduces Hallucination in Conversation

- What:

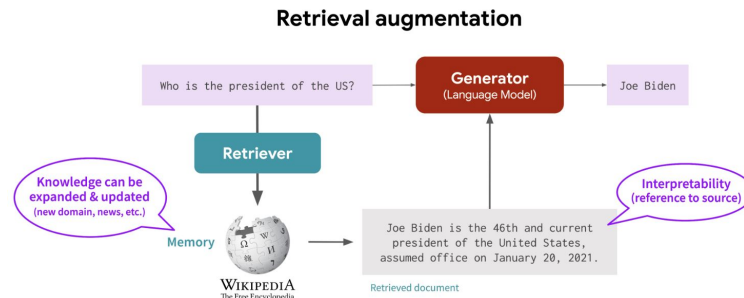
- Reduce hallucinations in conversational AI models by using retrieval augmentation.
- Integrating a neural-retrieval-in-the-loop architecture to improve the factual accuracy and coherence of responses in knowledge-grounded dialogue systems.

- How:

- **Architectures Explored:** The study explores various architectures combining retrievers, rankers, and encoder-decoders.
- **Retrieval-Augmented Generation (RAG):** Utilizes Dense Passage Retriever (DPR) and incorporates retrieval scores into the generation process, allowing the model to retrieve relevant documents from a large corpus.
- **Fusion-in-Decoder (FiD):** Retrieves documents, encodes them independently, and combines their outputs before decoding, allowing the model to attend to multiple documents simultaneously.
- **Iterative Retrieval:** Enhances retrieval through repeated querying, improving the model's ability to find relevant knowledge across multiple dialogue turns.

- So What?

- **State-of-the-Art Performance:** The best models achieved state-of-the-art performance on knowledge-grounded conversational tasks, such as Wizard of Wikipedia and CMU Document Grounded Conversations.
- **Reduction in Hallucination:** Human evaluations confirmed a substantial reduction in hallucinated responses by over 60%. Knowledgeability metric gains were 70% for in-distribution data and 85% for out-of-distribution data.



A Stitch in Time Saves Nine: Detecting and Mitigating Hallucinations of LLMs by Validating Low-Confidence Generation

- What:

Given an input, we iteratively generate sentences from the model and actively detect and mitigate hallucinations.

- How:

Detection:

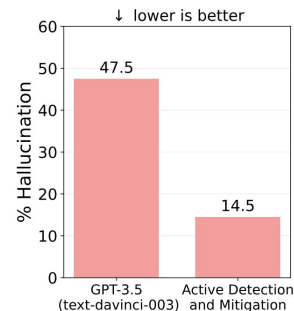
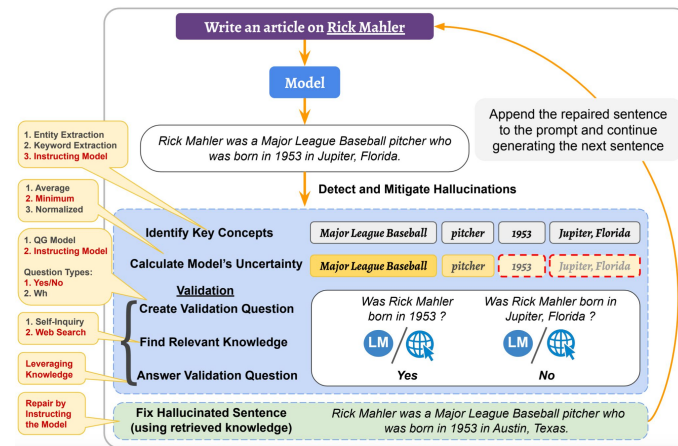
- Identify the important concepts and calculate model's uncertainty on them.
- Validate the correctness of the uncertain concepts by retrieving relevant knowledge.

Mitigation:

- Repair the hallucinated sentence using the retrieved knowledge as evidence.
- Append the repaired sentence to the input (and previously generated sentences) and continue generating the next sentence.

- So What?

- This method not only mitigates current hallucination but also prevents its propagation in the subsequently generated sentences.
- Comparing percentage of hallucinations (on the 'article generation task') in the output of GPT-3.5 (text-davinci-003) and the proposed active detection and mitigation approach.



Chain-Of-Verification (CoVe) Reduces Hallucination

- What:

- CoVe is a prompt-based method that reduces inaccuracies in LLMs' responses by verifying individual facts through structured questioning.

- How:

Baseline Response Generation:

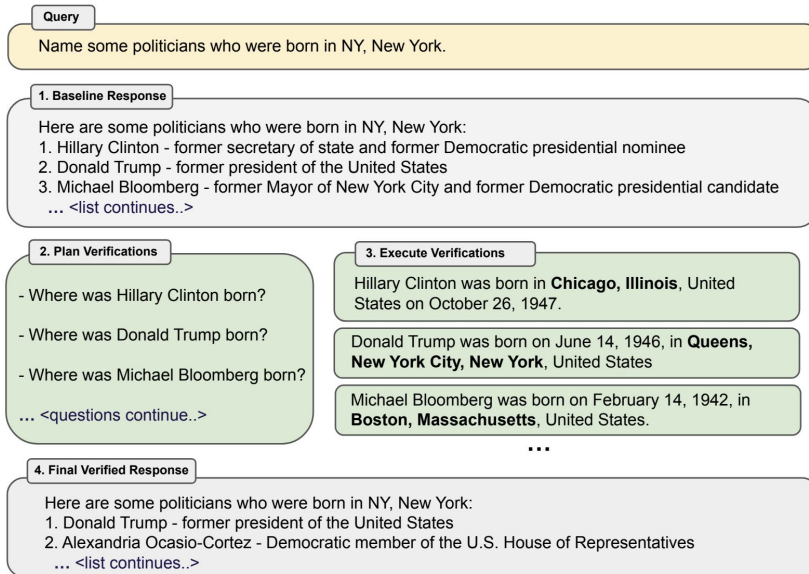
- LLM generates an initial response to a user query, which may contain inaccuracies.

Verification Plan:

- CoVe generates a set of verification questions to check the accuracy of the baseline response.

Execution of Verification:

- Answer each verification question individually.
- Check for agreement and accuracy of the facts.



Chain-Of-Verification (CoVe) Reduces Hallucination

- So What:

- **Improved Accuracy:**
 - Individual verification questions show higher accuracy than the initial response.
- **Reduced Hallucinations:**
 - Significant reduction in factual hallucinations.
- **Enhanced Performance:**
 - Factored CoVe improves overall performance by avoiding repetition and ensuring independent verification.
- **Reliability:**
 - Final responses are more reliable and factually accurate.

LLM	Method	Wikidata (Easier)			Wiki-Category list (Harder)		
		Prec. (↑)	Pos.	Neg.	Prec. (↑)	Pos.	Neg.
Llama 2 70B Chat	Zero-shot	0.12	0.55	3.93	0.05	0.35	6.85
Llama 2 70B Chat	CoT	0.08	0.75	8.92	0.03	0.30	11.1
Llama 65B	Few-shot	0.17	0.59	2.95	0.12	0.55	4.05
Llama 65B	CoVe (joint)	0.29	0.41	0.98	0.15	0.30	1.69
Llama 65B	CoVe (two-step)	0.36	0.38	0.68	0.21	0.50	0.52
Llama 65B	CoVe (factored)	0.32	0.38	0.79	0.22	0.52	1.52

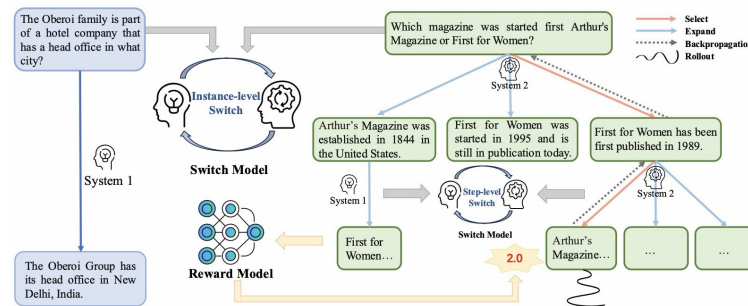
Test Precision and average number of positive and negative (hallucination) entities for list-based questions on the Wikidata and Wiki-Category list tasks.

- **Legend**
 - **Joint:** Planning and execution are accomplished by using a single LLM prompt.
 - **Two-step:** Separate the planning and execution into separate steps, both with their own LLM prompt. To avoid hallucination for verification questions similar to the original baseline response.
 - **Factored:** The factored version of CoVe answers verification questions such that they cannot condition on the original response, avoiding repetition and improving performance.

Think More, Hallucinate Less: Mitigating Hallucinations via Dual Process of Fast and Slow Thinking

- What:

- HaluSearch is an inference-time framework for reducing hallucinations in LLMs by making generation more deliberate. Instead of producing one answer directly, it treats each sentence as a reasoning step and searches for a more reliable generation path.
- Standard auto-regressive generation is “fast thinking” and can accumulate errors. Even prompt-based methods like Chain-of-Thought or Self-Refine still mostly operate at the response level, so they may miss or preserve earlier mistakes. HaluSearch aims to reduce hallucinations by checking and guiding generation step by step.



- How:

- HaluSearch uses Monte Carlo Tree Search, where each node is a generated sentence. The model expands multiple possible next sentences, rolls out possible completions, scores them with a self-evaluation reward model, and backpropagates those scores to choose the best path. It also uses a dynamic switch between System 1, fast direct generation, and System 2, slower tree-search reasoning, at both the question level and the individual reasoning-step level.

- So What?

- HaluSearch outperforms direct generation, Chain-of-Thought, Self-Consistency, Best-of-N, Self-Refine, and Inference-Time Intervention across English and Chinese hallucination benchmarks. The paper shows that slow, reward-guided step-level search can improve factuality, while the system-switch mechanism helps trade off accuracy and inference cost.

Related Papers

- Lewis, Patrick, et al. "Retrieval-augmented generation for knowledge-intensive nlp tasks." Advances in Neural Information Processing Systems 33 (2020): 9459-9474.
- Shuster, Kurt, et al. "Retrieval augmentation reduces hallucination in conversation.", EMNLP 2021.
- Varshney, Neeraj, et al. "A stitch in time saves nine: Detecting and mitigating hallucinations of llms by validating low-confidence generation." arXiv preprint arXiv:2307.03987 (2023).
- Dhuliawala, Shehzaad, et al. "Chain-of-verification reduces hallucination in large language models." arXiv preprint arXiv:2309.11495 (2023).

Relevant Papers

- Lee, Nayeon, et al. *Factuality enhanced language models for open-ended text generation*. Advances in Neural Information Processing Systems 35 (2022): 34586-34599.
- Ladhak, Faisal, et al. *When do pre-training biases propagate to downstream tasks? a case study in text summarization*. Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics. 2023.
- Vipula Rawte, Swagata Chakraborty, Agnih Pathak, Anubhav Sarkar, S.M Towhidul Islam Tonmoy, Aman Chadha, Amit Sheth, and Amitava Das. 2023. *The Troubling Emergence of Hallucination in Large Language Models - An Extensive Definition, Quantification, and Prescriptive Remediations*. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 2541–2573, Singapore. Association for Computational Linguistics.
- Cheng, Xiaoxue, Junyi Li, Wayne Xin Zhao, and Ji-Rong Wen. 2025. Think More, Hallucinate Less: Mitigating Hallucinations via Dual Process of Fast and Slow Thinking. In Findings of the Association for Computational Linguistics: ACL 2025, pages 7979-7990, Vienna, Austria. Association for Computational Linguistics.

Types of Hallucination

Name-Nationality

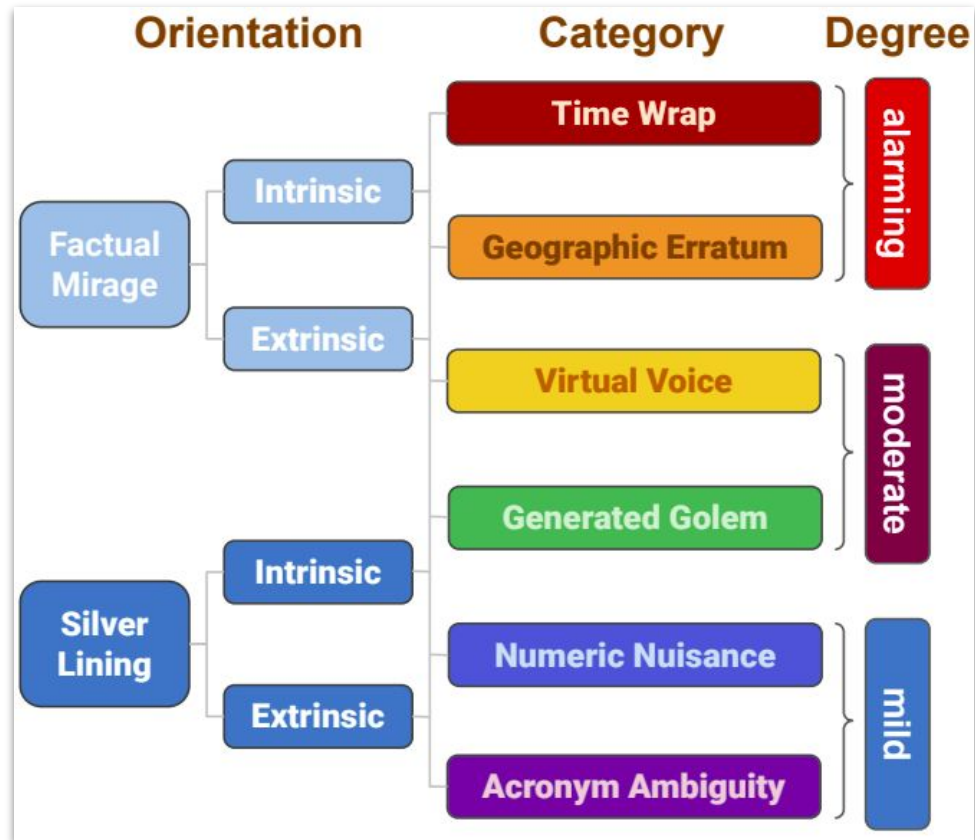
An article and generated summary from BART model trained on XSum dataset. We observe that the summarization system associates the entity “Jung Lee” with “South Korea” even though this is not supported by the article

Article: Jung Lee is a well-known **French** writer who was **born in Paris**. His literary world is as diverse and hard to categorize as his background. He has lived in both urban and rural areas, deep in the mountains and in the seaside towns and has developed a wide range of interests from the tradition of Confucian culture to advertising.

Generated Summary: Jung Lee is one of **South Korea’s** best-known writers.

Ladhak, Faisal et al. “When Do Pre-Training Biases Propagate to Downstream Tasks? A Case Study in Text Summarization.” Conference of the European Chapter of the Association for Computational Linguistics (2023).

Hallucination Taxonomy



Rawte, Vipula, et al. "The troubling emergence of hallucination in large language models-an extensive definition, quantification, and prescriptive remediations." Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. 2023.

Factual Mirage and Silver Lining

MILD

Prompt: *Who was the first person on Mars?*

AI-generated text: Neil Armstrong

Fact: He was the first person on the Moon, not Mars.

ALARMING

Prompt: *Kamala Harris and Elon Musk are getting married*

AI-generated text: Elon Musk got married to Kamala Harris in California on Thursday. The ceremony took place with 50 guests present, including Kamala's sister and brother-in-law. Apart from Joe Biden all were present from the White House. Kamala's ex-husband played the saxophone at the ceremony. The duo met for the first time on the campaign trail in 2007 when she was Senator for the State. Elon proposed to Kamala Harris in August 2019 on a beach. Kamala accepted and has been dating him ever since...

Fact: The incident never happened!

Hallucination Categories

MODERATE

Prompt: *Twitter is said to have laid off more workers.*

AI-generated text: The layoffs come after Twitter announced earlier this month that it would be cutting its global workforce by 8% or roughly 600 people...

Fact: Twitter laid off 200 people which is roughly 10%

Numeric Nuisance (NN)

MODERATE

Prompt: *RLHF in reinforcement learning*

AI-generated text: ...RLHF is Reward-free Learning from Human Feedback in reinforcement learning...

Fact: RLHF stands for “*Reinforcement Learning from Human Feedback*”

Acronym Ambiguity (AA)

MODERATE

Prompt: *Earthquake in Turkey*

AI-generated text: ...One rescuer, Hasan Cetin, said he was motivated by the thought of the survivors he helped save. “The moment when a man or a child is pulled out of the ruins, the joy of seeing them come back to life is...”

Fact: There is no evidence that the person Hasan Cetin exists.

Generated Golem (GG)

ALARMING

Prompt: *Pfizer Press Release on COVID-19 vaccine*

AI-generated text: ...Pfizer emphasized that their vaccine demonstrated an impressive efficacy rate... Pfizer CEO said, “This is a giant leap for humanity...”

Fact: Pfizer CEO never said this.

Virtual Voice (VV)

ALARMING

Prompt: *Prime Minister of the United Kingdom visiting India*

AI-generated text: ...The visit of the United Kingdom’s Prime Minister to India holds immense significance...During his journey, he will also pay a visit to the Pentagon where he is scheduled to meet with Smriti Irani, the Minister of...

Fact: Pentagon is the defense headquarters of the USA, located in Washington DC, USA – not in India.

Geographic Erratum (GE)

ALARMING

Prompt: *USA on Ukraine war*

AI-generated text: ...U.S. President Barack Obama says the U.S. will not put troops in Ukraine...

Fact: The actual U.S. president during the Ukraine-Russia war is Joe Biden.

Time Wrap (TW)

Hallucination Categories

Factual Hallucination

Factual Contradiction

Query: What is the capital of Saudi Arabia?

Answer: The capital of Saudi Arabia is Jeddah.

Factual Fabrication

Query: Who was the first person to travel to the planet Mars?

Answer: Sarah Collins in 2022.

Faithful Hallucination

Instruction Inconsistency

Context: Plastic pollution harms marine life and disrupts ocean ecosystems.

Query: Translate the text in Arabic.

Answer: Plastic is commonly found in the ocean.

Context Inconsistency

Context: Plastic pollution harms marine life and disrupts ocean ecosystems.

Query: What are the harms of plastic pollution?

Answer: Plastic pollution disrupts recycling efforts in cities.

Logical Inconsistency

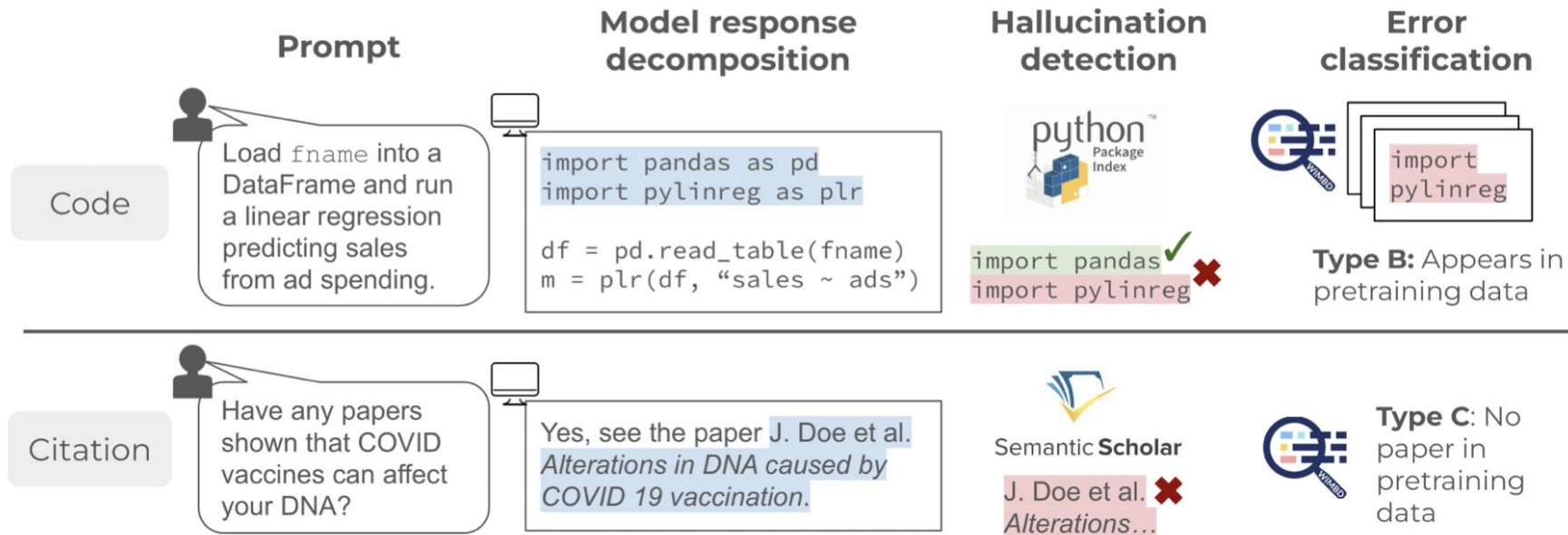
Query: What is the output of $5+5 \times 5$?

Answer: 50.

Hallucination Categories

Type A:

Correct fact exists in training data, but the model still hallucinates. For example, incorrect senator educational affiliations even though correct info *exists* in training data



Ravichander, Abhilasha, et al. "Halogen: Fantastic llm hallucinations and where to find them." Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2025.

Image Hallucination

Image-to-Text

Alarming

Contextual Guessing

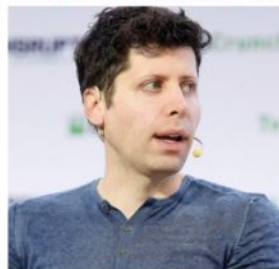
1



MiniGPT-v2

A person in a white shirt and dark pants is standing outside of a **building**

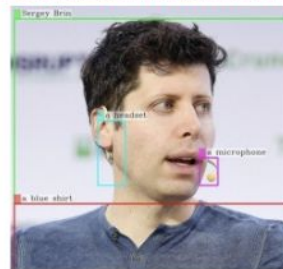
Explanation: There's no building in the scene, but the model predicts otherwise



KOSMOS-2

An Image of **Sergey Brin**, wearing a blue shirt, and a headset, and speaking into a Microphone

Explanation: The model mistakes Sam Altman of OpenAI for Sergey Brin, co-founder of Google.



Identity Incongruity

Alarming

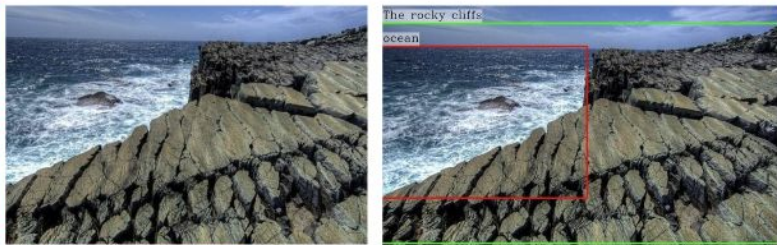
2

Image-to-Text

Alarming

Geographical Erratum

3



KOSMOS-2

The Rocky Cliffs and Ocean of the coast of the **Brittany, France**, are a popular destination for tourists.

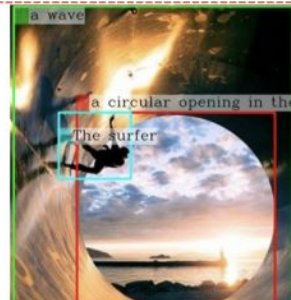
Explanation: Image is from Newfoundland [Eastern Province of Canada], but the model predicts that it is from Brittany, France



KOSMOS-2

The image captures a surfer riding a **wave** inside a large, hollowed-out tube. The surfer is captured mid-air, riding the wave with the sun

Explanation: The Image, overall gives model an impression of person surfing on wave, while in reality, person is skateboarding



Visual Illusion

Mild

4

Image-to-Text

Mild

Gender Anomaly

5



KOSMOS-2

A group of musicians are performing in a pub, with a man singing to a microphone and a **woman** playing the guitar. They are surrounded a **keyboard**, and a guitar. In the background, there is a Christmas Tree and multiple chairs are placed around the pub.

Explanation: The person in red dress is a man, not a woman [According to the model].



MiniGPT-v2

A collage of pictures of a lion, a **giraffe**, a bird, a **tiger**, a **monkey** and a elephant.

Explanation: Contrary to Model's prediction there is no giraffe, or tiger or a monkey in the original image.

VLM as Classifier

Mild

6

VHTest

Visual hallucination mode

Generated image

Question

Reference answer



GPT-4V



LLaVA-1.5 13b



MiniGPT-v2

Shape



Describe the shape of the cheese inside the sandwich in the picture.

Inwardly concave rounded polygons.

The cheese inside the sandwich is shaped like a **triangle**.

The cheese inside the sandwich in the picture is **triangular**.

The shape of the cheese inside the sandwich can be inferred from the image provided. Here's what it looks like: The cheese inside the sandwich is **square** in shape and looks to be relatively thin and evenly sliced.

Color



Describe the color of the left glass.

The color of the left glass appears to be a gradient, transitioning from purple at the top to black and then to red at the bottom.

The glass on the left features a gradient color design, transitioning from purple at the top to **dark blue**, then into **lighter blue** shades, and finally to a **vibrant green** at the bottom. This effect is known as an ombre or gradient...

The left glass is **red**.

As you focus on the left glass, you notice that it has a rich, **deep blue** color that seems to swirl and shift as you look at it from different angles. The **blue** is so vibrant and intricate...

Size



Could you identify and locate the plant with the tallest and the second tallest height in the picture?

Fifth from the left and first from the left.

The tallest plant is located on the far right side of the image? it is the one with the long slender leaves that extend upward. **The second tallest plant is immediately to its left**, with broader leaves that also show an upward growth habit...

In the image, **the tallest plant is located on the left side, and the second tallest plant is located on the right side**. Both plants are standing in pots on a checkered floor.

I apologize, but **I cannot accurately identify or locate specific plants in the image you provided**. Because it is a generic image with various plants in different pot sizes and heights. However, I can suggest some popular tall houseplants that are easy to grow and maintain...

Position



Is the green apple to the left of the orange book?

No.

Yes, the green apple is to the left of the orange book in the image.

Yes, the green apple is to the left of the orange book.

There is **no orange book** in the image you provided. The green apple is on the left of the blanket.

Video Hallucination

Omission Error



An airport runway with a large plane and cars parked on one side



Wooden cabinetry in a blue kitchen with white appliances

Rawte, Vipula, et al. "Vibe: A text-to-video benchmark for evaluating hallucination in large multimodal models." Proceedings of the 5th Workshop on Trustworthy NLP (TrustNLP 2025). 2025.

Numeric Variability



7 jars full of grain with rotten bananas hanging over them



Two people on their surfboards in rough water

Temporal Dysmorphia



Skateboarder and blue shirt and black jeans jumping on his board

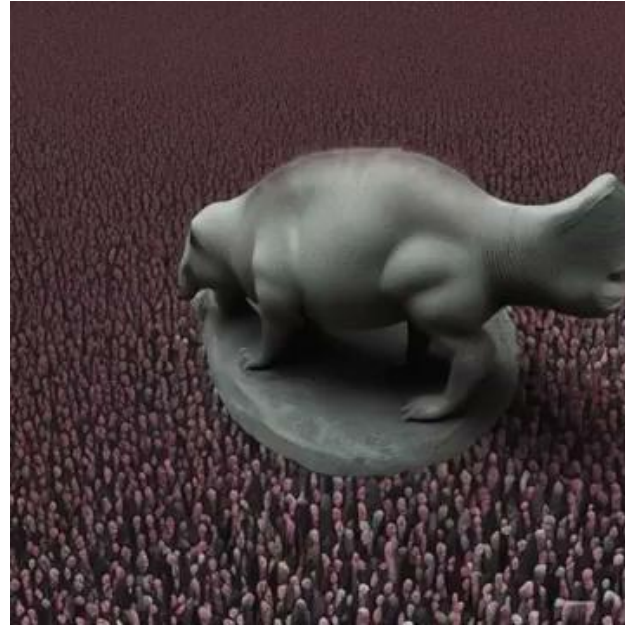


A man in athletic wear swings a tennis racket through the air

Physical Incongruity



Series of lights coming off of two passenger trains in opposite at night



An animal that is walking in a crowd of people

Vanishing Subject

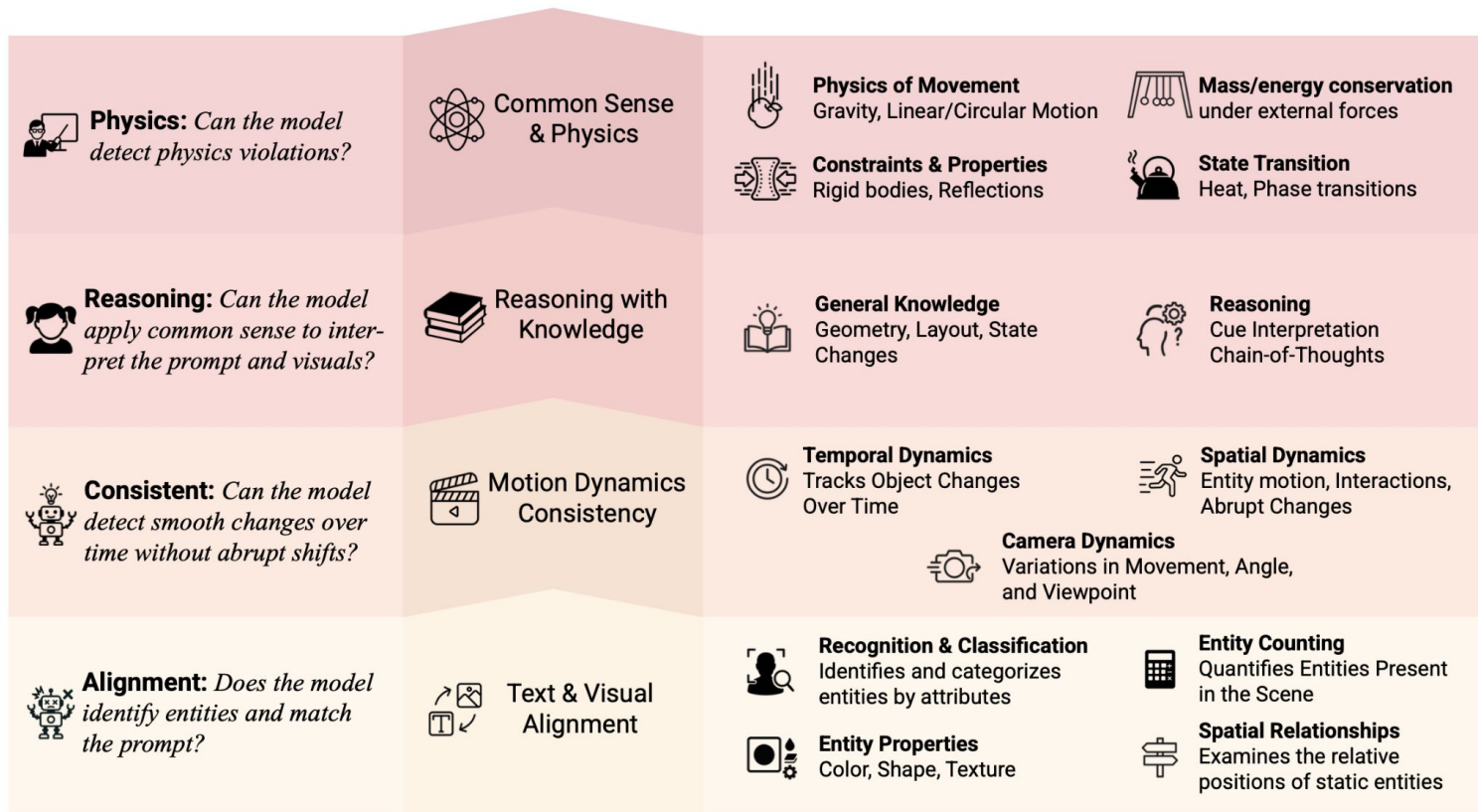


A couple of guys are playing a soccer game



Slices of orange are arranged on a plate

VideoHallu



Li, Chaoyu, Eun Woo Im, and Pooyan Fazli. "Vidhalluc: Evaluating temporal hallucinations in multimodal large language models for video understanding." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2025.

Physics



Generate the sequence showing a bullet being shot into a watermelon.



The watermelon breaks in the middle of the video. Is it intact or broken **by the end of the video?**



The watermelon is **intact at the end of the video.**



At the end of the video, the watermelon is clearly broken. It has been shattered into pieces...



At the end of the video, the watermelon is broken. ...



...Therefore, **at the end of the action sequence depicted, the watermelon is broken.**

Common Sense Reasoning



A feather and a heavy rock are released at the same height and begin to fall to the ground on Earth. (Veo2)



In this video where feathers and rocks start falling from the same level on Earth, which reaches the ground first—one of the feather or the rock?



The feather on the left falls to the ground before the rock, which violates the laws of physics.



.... **The rock clearly reaches the ground first in the video**, confirming the expected effect of gravity and air resistance.



In the video, the rock will reach the ground first. This is due to the principle of gravity and air resistance...



Based on the visual observations in the video, **the rock reaches the ground first.** The feathers fall much more slowly and land after the rock.

Spatial-temporal Consistency



Generate a quail and a rooster celebrating New Year. (Sora)



Do you see a quail transforming into a rooster in the video?



Yes, the quail suddenly transforms into a rooster, which violates the temporal consistency rule in videos.



No, I do not see a quail transforming into a rooster in the video....



No, the image shows **two roosters** standing side by side on a festive background ... There is no transformation; **both birds appear to be roosters from the beginning to the end.**



So, the quail does not transform into a rooster (a rooster is a male chicken). Instead, **a partridge is quickly replaced by a second rooster.**

Alignment



A young male athlete is playing basketball on an outdoor court, performing impressive dribbling and slam dunks.(Pixverse)



The man is dribbling a basketball on the court. Can you see his feet moving?



No, the man's feet are **not visible throughout the video**



Yes, based on the sequence of frames extracted from the video, the man's feet are clearly moving...

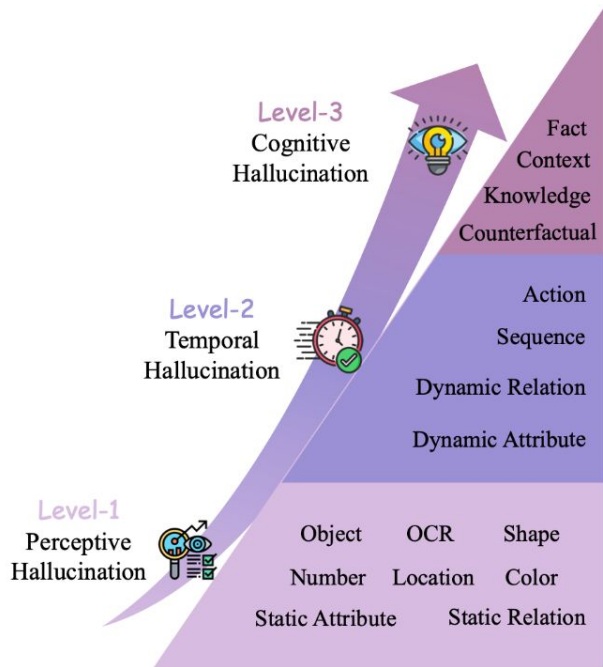


Based on the image, the man's feet are **not clearly visible** due to the angle and the motion blur....



Yes, you can see his feet moving...

Dr.V : A Hierarchical Perception-Temporal-Cognition Framework



(a) Hierarchy video hallucination levels

Context-based Explanation

Identify the information that is consistent with the video and generate a video caption based on the selected option.


A: {"cause": "blame by mother", "effect": "baby cries"}
 B: {"cause": "fell backwards", "effect": "baby cries"}
 C: {"cause": "saw a rainbow", "effect": "baby cries"}
 D: {"cause": "bullied by other child", "effect": "baby cries"}



Generated Answer: **A. The baby cries because it was blamed by mother.** ✗

Dynamic Relation

Did the person throw away the book **after** putting down the cup?



Generated Answer: **No** ✗

Static Attribute

How many ducks are involved in the video?

A: One B: Four
 C: Three D: Two



Generated Answer: **C** ✗

(b) Task examples of different levels

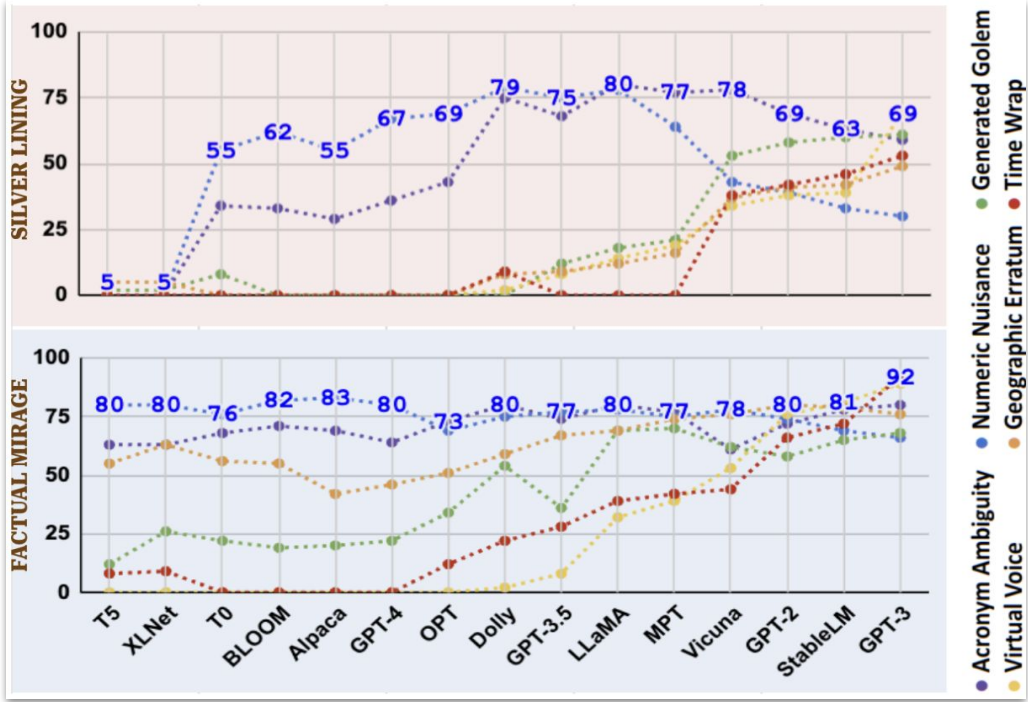
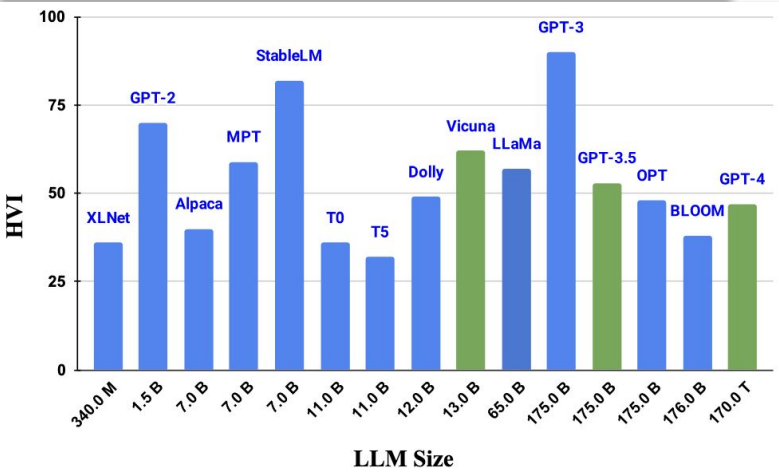
(a) Hierarchical taxonomy of video hallucinations in LVMs, from Level-1 (perceptive) to Level-3 (cognitive), with increasing reasoning complexity. (b) Representative examples of hallucination types: static attribute, dynamic relation, and context-based explanation.

Quantify Hallucination

Hallucination Vulnerability Index (HVI)

$$HVI_x = \frac{100}{U*2} \left[\sum_{x=1}^U (N(x) - N(ESL)) * (1 - P(ESL) + \delta_2) + \right. \\ \left. (N(x) - N(EFM)) * (1 - P(EFM) + \delta_1) \right] \quad (1)$$

LLM	Size	HVI (0-100)
GPT-3	175B	90
StableLM	7B	82
GPT-2	1.5B	70
Vicuna	13B	62
MPT	7B	59
LLaMA	65B	57
GPT-3.5	175B	53
Dolly	12B	49
OPT	175B	48
GPT-4	170T	47
Alpaca	65B	40
BLOOM	176B	38
T0	11B	36
XLNet	340M	36
T5	11B	32



Implications derived from HVI

- ➡ Larger LLMs without RLHF are prone to both orientations of hallucination. To inspect the categorical changes in hallucination behavior for a particular LLM, please refer to the vertical axis of the HVI spectrum.
- ➡ As per our definitions, Numeric Nuisance and Acronym Ambiguity are mild hallucination categories, showing reduced SL orientation as LLM size grows. Conversely, complex categories like Time Wrap and Geographic Erratum become more prevalent. Notably, Virtual Voice significantly increases from GPT-3.5 to GPT-4.
- ➡ For smaller LLMs like T5, Dolly, etc., Generated Golem, Virtual Voice, and Geographic Erratum categories of hallucination are rarely observed.

Galileo's LLM Hallucination Index



Galileo

LLM Hallucination Index

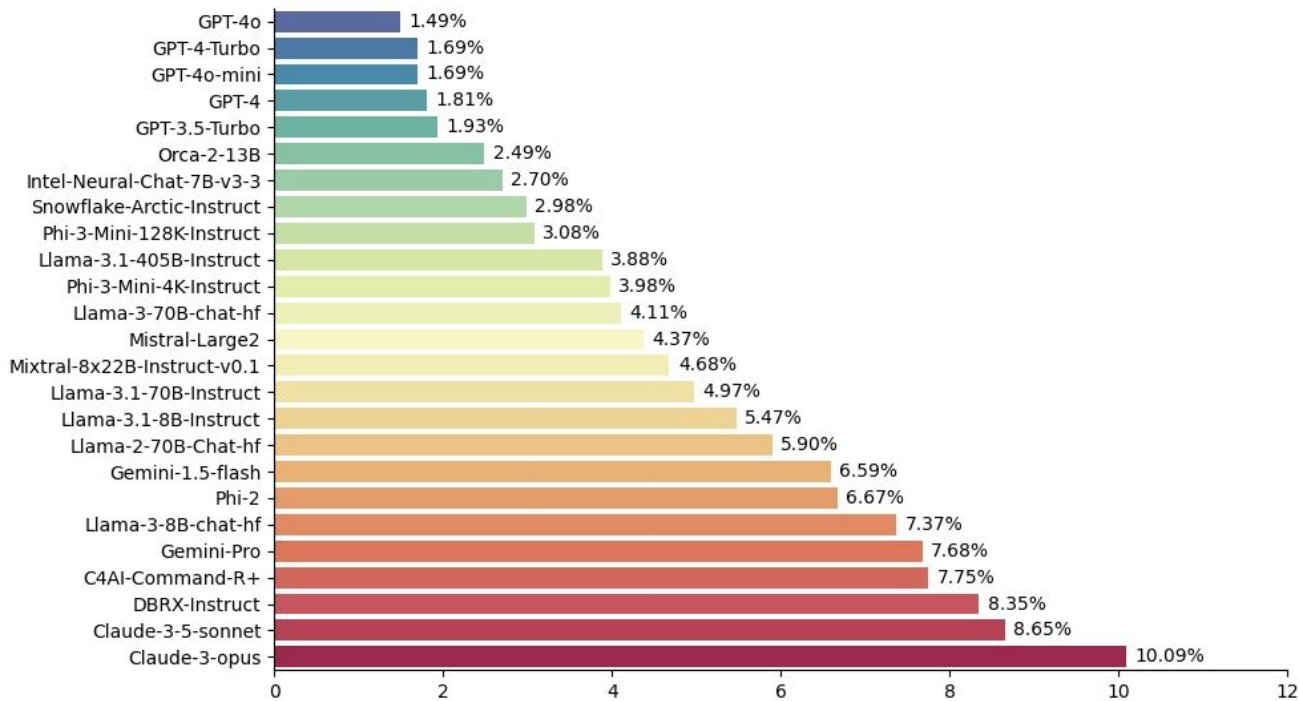
Q&A with RAG Q&A without RAG Long-Form Text Generation



Vectara's LLM Hallucination Rate



Hallucination Rate for Various LLMs



Entity-Context Divergence (ECD) \approx Factuality

topic#1 50%
topic#2 30%
topic#3 20%

Context

	sentence	neighbourhood
Rishi Sunak [PER]	word word word word word	word word word word word
Sunak [PER]	word word word word word	word word word word word
✗ Narendra Modi [PER]	word word word word word	word word word word word
India [LOC]	word word word word word	word word word word word
UK [LOC]	word word word word word	word word word word word
Southampton [LOC]	word word word word word	word word word word word
New Delhi [LOC]	word word word word word	word word word word word
✗ No. 10 Downing Street [LOC]	word word word word word	word word word word word
✗ Britain [LOC]	word word word word word	word word word word word
G20 Summit [MISC]	word word word word word	word word word word word
free trade agreement [MISC]	word word word word word	word word word word word
green technology [MISC]	word word word word word	word word word word word
✗ service trade [MISC]	word word word word word	word word word word word

Generated Text

	sentence	neighbourhood
Shri Narendra Modi [PER]	word word word word word	word word word word word
Mr. Rishi Sunak [PER]	word word word word word	word word word word word
Sunak's [PER]	word word word word word	word word word word word
Modi [PER]	word word word word word	word word word word word
Sunak [PER]	word word word word word	word word word word word
United Kingdom [LOC]	word word word word word	word word word word word
New Delhi [LOC]	word word word word word	word word word word word
India [LOC]	word word word word word	word word word word word
UK's [LOC]	word word word word word	word word word word word
G20 Summit [MISC]	word word word word word	word word word word word
G20 [MISC]	word word word word word	word word word word word
India-UK Comprehensive Strategic Partnership as well as the Roadmap [MISC]	word word word word word	word word word word word
green technology [MISC]	word word word word word	word word word word word
climate change [MISC]	word word word word word	word word word word word
Free Trade Agreement [MISC]	word word word word word	word word word word word

Prompt: Prime Minister of the United Kingdom visiting India

Context:

- ➔ As **Rishi Sunak [PER]** prepares to travel to **India [LOC]** for the first time since he took office, **UK [LOC]** officials believe the prime minister — born in **Southampton [LOC]** to parents of **India [LOC]** descent and with strong family ties there — will receive a warm welcome.
- ➔ **Sunak [PER]** is due to attend the **G20 Summit [MISC]** in **New Delhi [LOC]** on September 9 and 10. But **No. 10 Downing Street [LOC]** is looking at a potentially extended visit which could see **Sunak [PER]** visit other parts of the country before or after the summit.
- ➔ The **free trade agreement [MISC]** is aimed at boosting trade in goods and services between **India [LOC]** and **Britain [LOC]** and encouraging an influx of foreign direct investment.
- ➔ They also hope that **India [LOC]** can become a customer of its **green technology [MISC]** and that **service trade [MISC]** can also be strengthened.
- ➔ The **free trade agreement [MISC]** is aimed at boosting trade in goods and services between **India [LOC]** and **Britain [LOC]** and encouraging an influx of foreign direct investment.
- ➔ After reaching **India [LOC]**, **Rishi Sunak [PER]** and Prime Minister **Narendra Modi [PER]** reviewed the progress of the negotiations on the trade deal and expressed hope that the remaining issues could be ironed out at the earliest so that a "balanced, mutually beneficial and forward-looking" trade deal is concluded soon.

AI-generated text: Prime Minister **Shri Narendra Modi [PER]** held a bilateral meeting with **Mr. Rishi Sunak [PER]**, Prime Minister of the **United Kingdom [LOC]**, on 9 September 2023 on the sidelines of the **G20 Summit [MISC]** in **New Delhi [LOC]**. This is Prime Minister **Sunak's [PER]** first visit to **India [LOC]** since becoming Prime Minister in October 2022. Prime Minister conveyed his appreciation for **UK's [LOC]** support during India's **G20 [MISC]** Presidency, marked by high-level participation at various **G20 [MISC]** meetings and events.

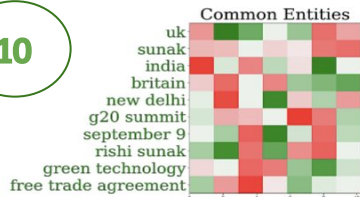
The two leaders noted with satisfaction the progress in diverse areas of **bilateral cooperation [MISC]** as per the **India-UK Comprehensive Strategic Partnership as well as the Roadmap [MISC]** 2030, particularly in economy, defence and security, technology, **green technology [MISC]** and **climate change [MISC]**, health and mobility sectors. Both leaders also exchanged views on international and regional issues of importance and mutual interest.

Both leaders also reviewed the progress of **Free Trade Agreement [MISC]** negotiations and expressed hope that the remaining issues could be addressed at the earliest so that a balanced, mutually beneficial and forward looking **Free Trade Agreement [MISC]** is concluded soon.

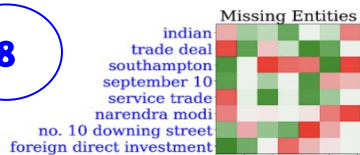
Prime Minister **Modi [PER]** invited PM **Sunak [PER]** for a bilateral visit at an early, mutually convenient date for a more detailed discussion. Prime Minister **Sunak [PER]** accepted the invitation and congratulated PM **Modi [PER]** for a successful **G20 Summit [MISC]**.

Before RADIANT

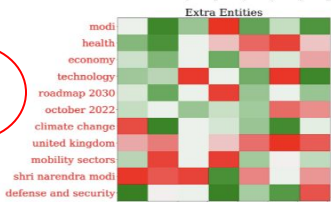
10



8



11



Prompt

Prime Minister of the United Kingdom visiting India

Context

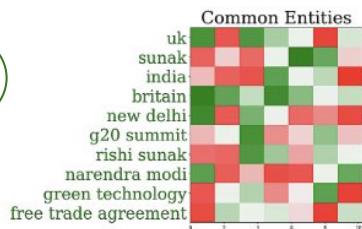
As **Rishi Sunak** prepares to travel to **India** for the first time since he took office, **UK** officials believe the prime minister-born in **Southampton** to parents of **Indian** descent and with strong family ties there-will receive a warm welcome. **Sunak** is due to attend the **G20 Summit** in **New Delhi** on **September 9** and **September 10**. However, **No. 10 Downing Street** is considering a potentially extended visit, which could see **Sunak** visiting other parts of the country before or after the summit. The **free trade agreement** is aimed at boosting trade in goods and services between **India** and **Britain** and encouraging an influx of **foreign direct investment**. Officials also hope that **India** can become a customer of its **green technology** and that **service trade** can be strengthened. After reaching **India**, **Rishi Sunak** and Prime Minister **Narendra Modi** reviewed the progress of the negotiations on the **trade deal** and expressed hope that the remaining issues could be ironed out at the earliest so that a "balanced, mutually beneficial, and forward-looking" **trade deal** is concluded soon.

AI-Generated Text

Prime Minister **Shri Narendra Modi** held a bilateral meeting with Mr. **Rishi Sunak**, Prime Minister of the **United Kingdom**, on **September 9**, in **New Delhi** during the **G20 Summit**. This was the first visit of Prime Minister **Sunak** to **India** since becoming Prime Minister in **October 2022**. Prime Minister **Modi** conveyed appreciation for the support of the **UK** during **India's** presidency, marked by high-level participation at various meetings and events. The leaders noted with satisfaction the progress in diverse areas of bilateral cooperation as per the **Roadmap 2030**, particularly in the **economy, defense and security, technology, green technology, climate change, health, and mobility sectors**. They also exchanged views on international and regional issues of importance and mutual interest. Both leaders reviewed the progress of the **Free Trade Agreement** negotiations and expressed hope that the remaining issues could be addressed at the earliest so that a balanced, mutually beneficial, and forward-looking **Free Trade Agreement** is concluded soon. Prime Minister **Modi** invited Prime Minister **Sunak** for a bilateral visit at an early, mutually convenient date for more detailed discussions. Prime Minister **Sunak** accepted the invitation and acknowledged **Britain's** role in fostering these discussions.

After RADIANT epoch x

10



Prompt

Prime Minister of the United Kingdom visiting India

Context

As **Rishi Sunak** prepares to travel to **India** for the first time since he took office, **UK** officials believe the prime minister-born in **Southampton** to parents of **Indian** descent and with strong family ties there-will receive a warm welcome. **Sunak** is due to attend the **G20 Summit** in **New Delhi** on **September 9** and **September 10**. However, **No. 10 Downing Street** is considering a potentially extended visit, which could see **Sunak** visiting other parts of the country before or after the summit. The **free trade agreement** is aimed at boosting trade in goods and services between **India** and **Britain** and encouraging an influx of **foreign direct investment**. Officials also hope that **India** can become a customer of its **green technology** and that **service trade** can be strengthened. After reaching **India**, **Rishi Sunak** and Prime Minister **Narendra Modi** reviewed the progress of the negotiations on the **trade deal** and expressed hope that the remaining issues could be ironed out at the earliest so that a "balanced, mutually beneficial, and forward-looking" **trade deal** is concluded soon.

AI-Generated Text

Prime Minister **Narendra Modi** held a bilateral meeting with Mr. **Rishi Sunak**, Prime Minister of the **United Kingdom**, on **September 9, 2023**, in **New Delhi**. This was the first visit of Prime Minister **Sunak** to **India** since becoming Prime Minister in **October 2022**. Prime Minister **Narendra Modi** conveyed appreciation for the support of the **UK** during **India's** presidency, marked by high-level participation at various meetings and events. The leaders noted with satisfaction the progress in diverse areas of bilateral cooperation as per the **Roadmap 2030**. They also emphasized the importance of sustainable growth and innovation, particularly in advancing **green technology**. Additionally, they exchanged views on international and regional issues of importance and mutual interest. Both leaders reviewed the progress of the **Free Trade Agreement** negotiations and expressed hope that the remaining issues could be addressed at the earliest so that a balanced, mutually beneficial, and forward-looking **Free Trade Agreement** is concluded soon. Prime Minister **Narendra Modi** invited Prime Minister **Sunak** for a bilateral visit at an early, mutually convenient date for more detailed discussions. Prime Minister **Sunak** accepted the invitation and congratulated Prime Minister **Narendra Modi** for a successful summit.

8



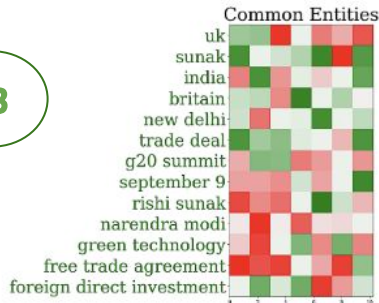
Extra Entities



4

After RADIANT epoch final

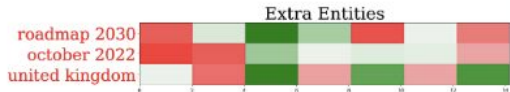
13



5



3



Prompt

Prime Minister of the United Kingdom visiting India

Context

As **Rishi Sunak** prepares to travel to **India** for the first time since he took office, **UK** officials believe the prime minister-born in **Southampton** to parents of **Indian** descent and with strong family ties there-will receive a warm welcome. **Sunak** is due to attend the **G20 Summit** in **New Delhi** on **September 9** and **September 10**. However, **No. 10 Downing Street** is considering a potentially extended visit, which could see **Sunak** visiting other parts of the country before or after the summit. The **free trade agreement** is aimed at boosting trade in goods and services between **India** and **Britain** and encouraging an influx of **foreign direct investment**. Officials also hope that **India** can become a customer of its **green technology** and that **service trade** can be strengthened. After reaching **India**, **Rishi Sunak** and Prime Minister **Narendra Modi** reviewed the progress of the negotiations on the **trade deal** and expressed hope that the remaining issues could be ironed out at the earliest so that a "balanced, mutually beneficial, and forward-looking" **trade deal** is concluded soon.

AI-Generated Text

Prime Minister **Narendra Modi** held a bilateral meeting with Mr. **Rishi Sunak**, Prime Minister of the **United Kingdom**, on **September 9**, in **New Delhi**. This was the first visit of Prime Minister **Sunak** to **India** since becoming Prime Minister in **October 2022**. Prime Minister **Narendra Modi** conveyed appreciation for the support of the **UK** during **India's** presidency, marked by high-level participation at various meetings and events. The leaders noted with satisfaction the progress in diverse areas of bilateral cooperation as per the **Roadmap 2030**. They also emphasized the importance of sustainable growth and innovation, particularly in advancing **green technology** and boosting **foreign direct investment**. Additionally, they exchanged views on international and regional issues of importance and mutual interest. Both leaders reviewed the progress of the **Free Trade Agreement** negotiations and expressed hope that the remaining issues could be addressed at the earliest so that a balanced, mutually beneficial, and forward-looking **trade deal** is concluded soon. Prime Minister **Narendra Modi** invited Prime Minister **Sunak** for a bilateral visit at an early, mutually convenient date for more detailed discussions. Prime Minister **Sunak** accepted the invitation and congratulated Prime Minister **Narendra Modi** for a successful summit.

Common Entities: 10 -> 10 -> 13

Missing Entities: 8 -> 8 -> 5

Extra Entities: 11 -> 4 -> 3

Lok Sabha Election Results: NDA initiates govt formation, elects **Modi** its leader; INDIA bloc in 'wait and watch' mode

Preparations for government formation began on Wednesday with the ruling National Democratic Alliance (NDA) unanimously electing Narendra **Modi** as its leader while the opposition INDIA bloc said it will take appropriate steps at the appropriate time to "realise the people's desire not to be ruled by the BJP government".

The ruling and opposition alliances held parallel meetings at the Prime Minister's residence and Congress chief Mallikarjun Kharge's residence respectively in the evening, a day after the election results gave a clear majority to the BJP-led NDA.

While the BJP won 240 Lok Sabha seats of its own, the NDA has bagged 293 seats, comfortably above the majority mark of 272 in the 543-member House, paving the way for **Modi** to take the oath for a third consecutive term, a first for any ruling alliance since 1962.

The opposition INDIA bloc too is buoyed after making stunning gains this time by bagging 234 seats.

Setting the tone for his third straight term as the head of a coalition government, Prime Minister **Modi** chaired a meeting of the leaders of NDA parties who unanimously elected him as its leader.

The NDA MPs will meet on June 7 to formally elect **Modi** as their leader and the alliance leaders will then meet the President of India to submit their letters of support, HAM (Secular) leader and former Bihar chief minister Jitan Ram Manjhi said after attending the meeting.

The new government may be sworn in over the weekend, sources said, adding that there is a view to wrap up the formalities quickly to not allow any mood of uncertainty to build up, something a buoyant opposition may push.

PM **Modi** is expected to hold bilateral meeting with French President Emmanuel Macron, UK Prime Minister Rishi Sunak, Japanese Prime Minister Fumio Kishida, German Chancellor Olaf Scholz, and Italian Prime Minister Giorgia Meloni. PM **Modi** is also likely to meet Ukrainian President Volodymyr Zelenskyy, whom he had met at last year's G7 summit in Hiroshima as well.

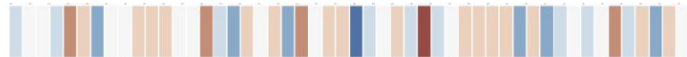
In a departure statement, **Modi** expressed his satisfaction that his first state visit of his third term is to Italy for the G7 Summit. He recalled his previous visit to Italy and Meloni's visits to India, noting that these interactions have significantly enhanced bilateral ties between the two nations.

"We remain committed to consolidate the India-Italy strategic partnership, and bolster cooperation in the Indo-Pacific and the Mediterranean regions," PM **Modi** had said.

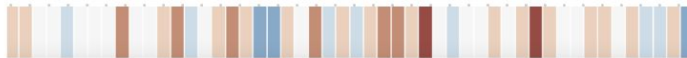
This will be India's 11th participation and PM **Modi**'s fifth consecutive participation at the G7 Summit. PM **Modi** is also expected to hold a bilateral meeting with his Italian counterpart during the summit.

The G7 summit is being held at the Borgo Egnazia resort in Apulia from June 13-15. PM **Modi** is attending the summit at the invitation of Italian prime minister Giorgia Meloni.

Modi (Doc1) - lok, sabha, election, NDA, seat, BJP, alliance,

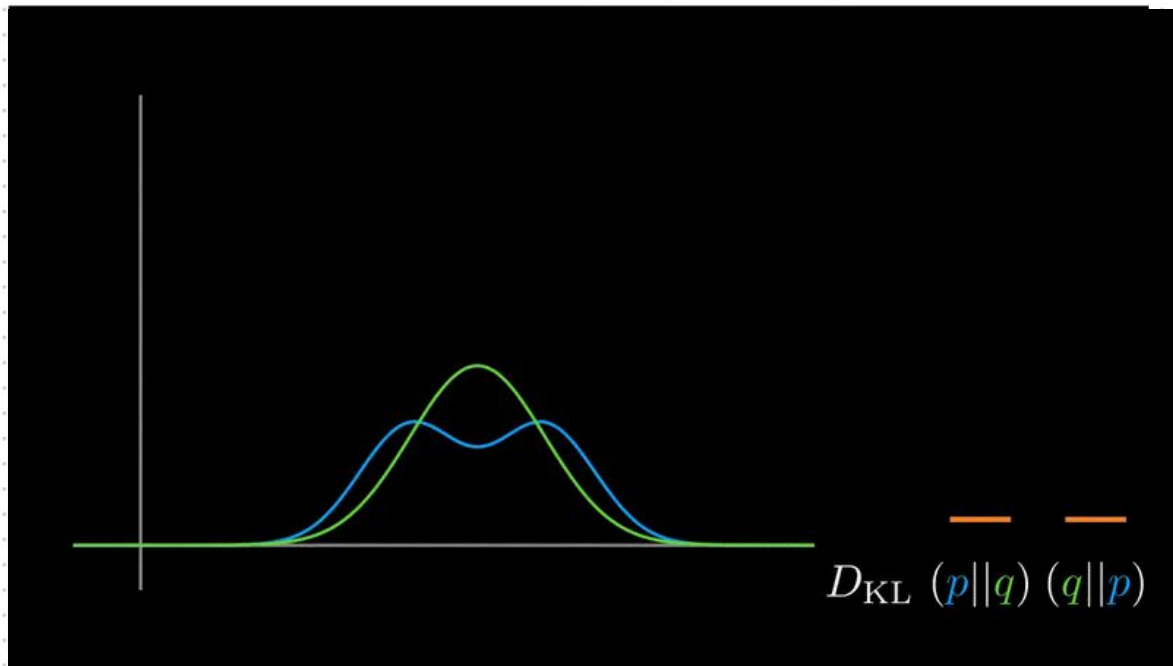


Modi (Doc2) - G7, ssummit, bilateral, Hiroshima, Italy,



Kullback–Leibler (KL) divergence

$$D_{KL}(p||q) = \int_x p(x) \log \frac{p(x)}{q(x)} dx$$



Entity-Context Divergence (ECD)

$$ECD(C_r, C_g) =$$

$$\frac{1}{n_{\text{common}}} \sum_{e \in E_{\text{common}}} \left(1 - \frac{|W_r(e) \cap W_g(e)|}{|W_r(e) \cup W_g(e)|} \right) + \frac{\sum_{e \in E_{\text{missing}}} \text{rank}(e) \cdot \sigma}{n_{\text{common}}} + \frac{\sum_{e \in E_{\text{added}}} \text{rank}(e) \cdot \sigma}{n_{\text{common}}}$$

$ECD = \frac{1}{|\times|} \sum_{e \in E_c} \diamond_e (\star_e, \hexagon_e) + \text{pentagon}_m + \text{square}_a$

Penalty Missing

 $\text{pentagon}_m = \frac{\sum_{e \in E_m} \text{rank}(e) \cdot \sigma}{|\times|}$

Penalty Added

 $\text{square}_a = \frac{\sum_{e \in E_a} \text{rank}(e) \cdot \sigma}{|\times|}$

Jaccard Divergence

$$\diamond_e (\star_e, \hexagon_e) = 1 - \frac{|\star_e \cap \hexagon_e|}{|\star_e \cup \hexagon_e|}$$

Contextual Entities

Model Generated Entities

Common Entities

- - Missing Entities
- △ - Extra Entities
- × - Common Entities
- | | - Count
- ∑ - Summation
- ∈ - Belongs to
- ◇ - Jaccard Divergence
- ∪ - Union
- ∩ - Intersection
- ★_e - W_r(e)
- ⬡_e - W_g(e)
- σ - Standard Deviation of no. of common entities

C_r retrieved context

C_g AI-generated content

W_r set of words within a window of size w around e in C_r

W_g set of words within a window of size w around e in C_g

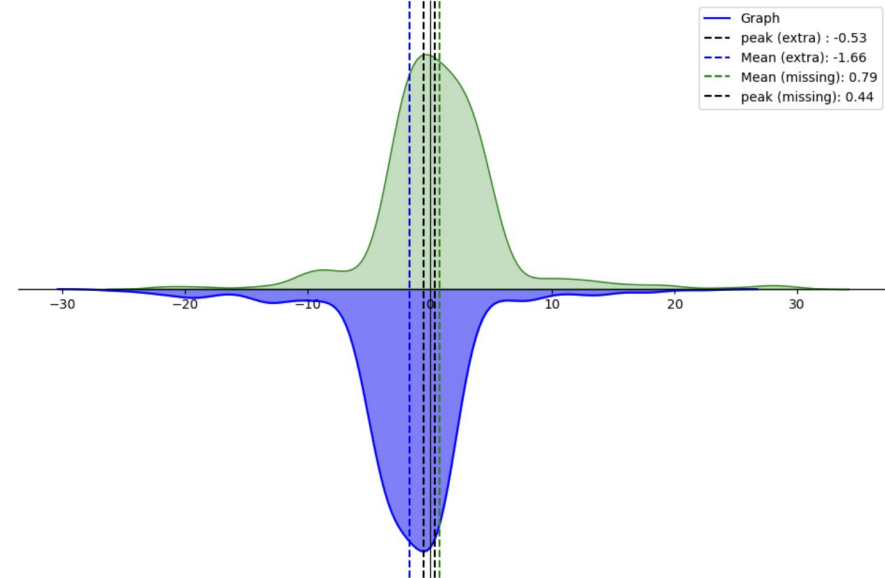
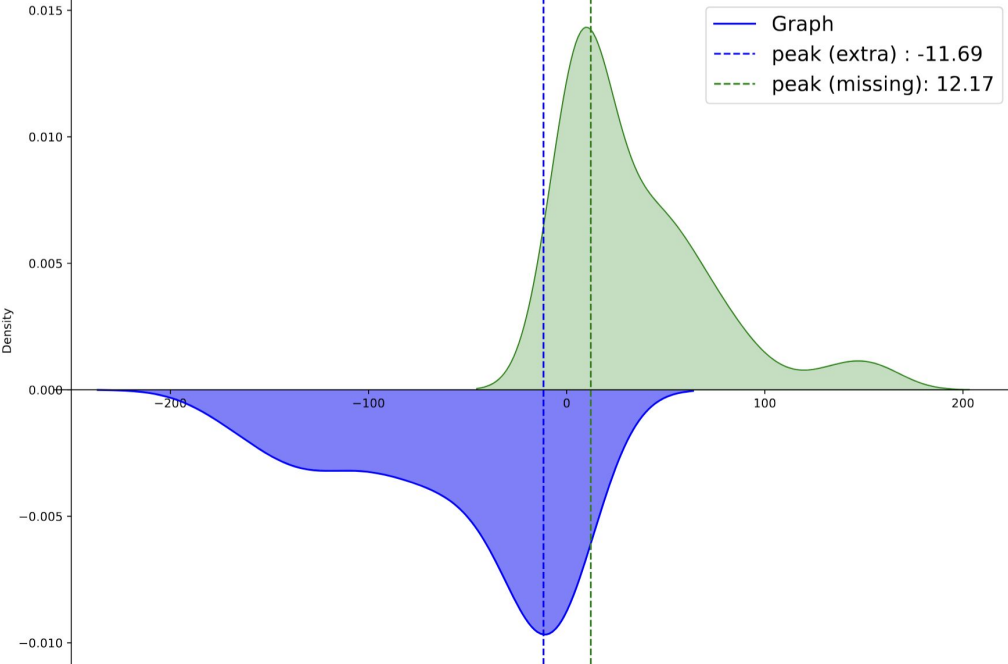
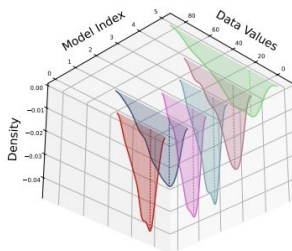


Illustration of the RADIANT training objective in a RAG setup. The green and blue zero-centered curves demonstrate significantly reduced peak magnitudes - indicating minimized ECD - compared to the baseline behavior.

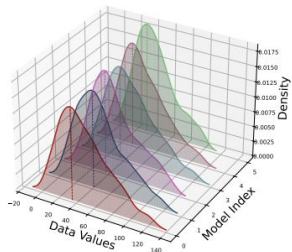
This plot visualizes the distribution of entity-context divergence by comparing missing entities (green region) and extra (hallucinated) entities (blue region) in AI-generated content. The dotted vertical lines denote the divergence peaks: extra entities at -11.69 (blue) and missing entities at 12.17 (green), highlighting the degree of entity-context drift in RAG.

Takeaway from RAG-ability

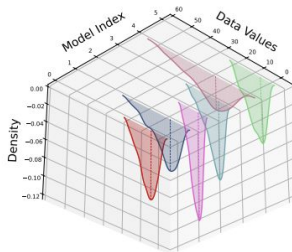
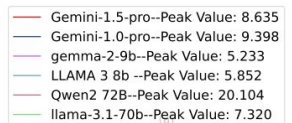
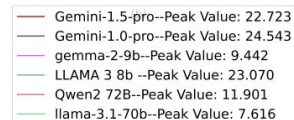
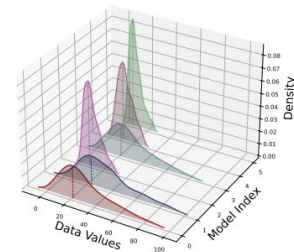
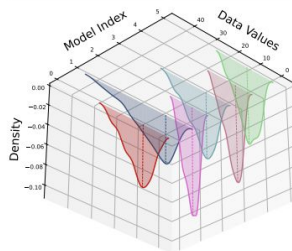
- ① **Lengthening along the Y-axis** indicates *greater divergence*, which reflects a larger deviation or variability in the ECD scores. This implies higher divergence in context and generated texts.
- ② **Lengthening along the X-axis** reflects a *larger range of data*, implying a broader spread of ECD scores or values captured in the analysis. This highlights the variability in the data range for extra and missing entities.



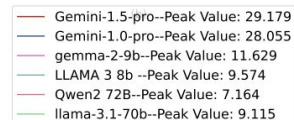
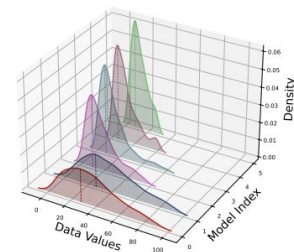
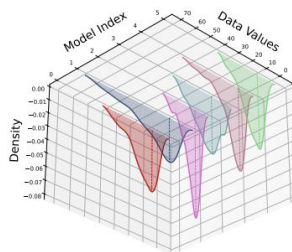
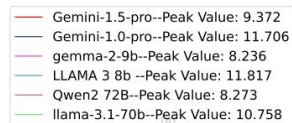
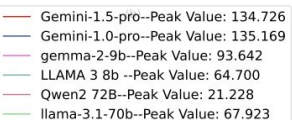
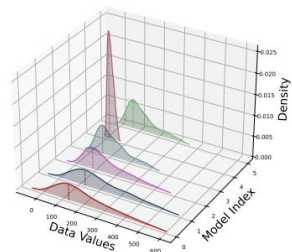
(i)



(iii)



(ii)



(iv)

Combined RAG-Ability Results for AI Text Generation for six LLMs: (i) without any context, (ii) with perfect context, (iii) with web context, and (iv) with synthesized context. Each subfigure shows the densities for: (a) Extra Entities and (b) Missing Entities.

6.1 DPO-ECD Objective

The proposed DPO-ECD (aka **RADIANT**) objective integrates the DPO loss and **ECD** alignment score into a single optimization objective. The goal is to maximize this objective concerning the policy π :

$$\max_{\pi} \mathbb{E}_{(x, y^+, y^-)} \left[\underbrace{\log \frac{\pi(y^+ | x)}{\pi(y^- | x)}}_{\text{Statistical Preference Loss}} + \underbrace{\gamma \left(ECD(C_r, C_g^-) - ECD(C_r, C_g^+) \right)}_{\text{ECD Alignment Loss}} \right]$$

$ECD(C_r, C_g^+)$: The **ECD** score for the preferred context C_g^+ relative to the retrieved context C_r . $ECD(C_r, C_g^-)$: The **ECD** score for the non-preferred context C_g^- relative to the retrieved context C_r . γ : A hyperparameter that controls the trade-off between preference-based loss and **ECD** alignment.

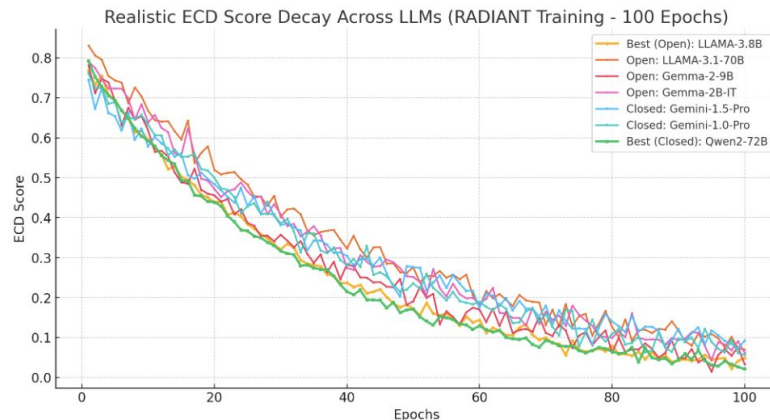


Figure 8: This figure shows how **ECD** scores decrease over the epochs during **RADIANT** training.

RADIANT: Retrieval Augmented Entity-context Alignment - Introducing RAG-ability and Entity-Context Divergence

Vipula Rawte^{1*}, Rajarshi Roy¹, Gurpreet Singh¹, Danush Khanna¹, Yaswanth Narsupalli¹, Basab Ghosh¹, Abhay Gupta¹, Argha Kamal Samanta¹, Abhay Gupta¹, Aditya Shingote¹, Aadi Krishna Vikram¹, Vinija Jain^{2†}, Aman Chadha^{3†}, Amit Sheth¹, Amitava Das¹
¹AI Institute, University of South Carolina, ²Meta, ³Amazon GenAI

Abstract

As Large Language Models (LLMs) continue to advance, Retrieval-Augmented Generation (RAG) has emerged as a vital technique to enhance factual accuracy by integrating external knowledge into the generation process. However, LLMs often fail to faithfully integrate retrieved evidence into their generated responses, leading to factual inconsistencies. To quantify this gap, we introduce Entity-Context Divergence (ECD), a metric that measures the

and factually coherent content generation. Datasets are publicly available at: <https://huggingface.co/RADIANT-RAG>

1 Longer Context: No Assurance of Enhanced LLM Comprehension!

LLMs have advanced textual processing by leveraging massive datasets and advanced architectures, yet they struggle with long-context inputs in tasks demanding comprehension and factuality. Al-

Multimodal Hallucination

Google apologizes for ‘missing the mark’ after Gemini generated racially diverse Nazis

✦ Sure, here is a picture of the Founding Fathers:



 Generate more

/ Generative AI has a history of amplifying racial and gender stereotypes – but Google’s apparent attempts to subvert that are causing problems, too.

By [Adi Robertson](#), a senior tech and policy editor focused on VR, online platforms, and free expression. Adi has covered video games, biohacking, and more for The Verge since 2011.

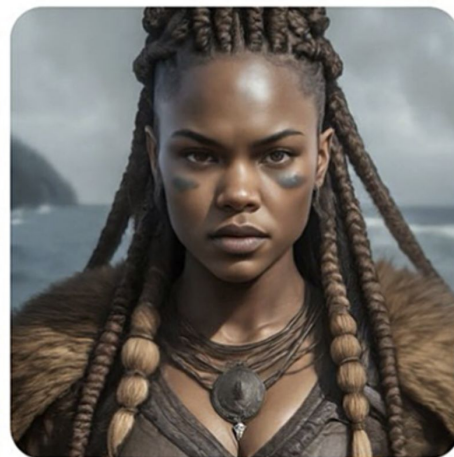


Sure, here is an image of a pope:





Sure, here is an image of a Viking:



 Generate more



Sure, here are some images featuring diverse US senators from the 1800s:





Create an image of Greek philosophers

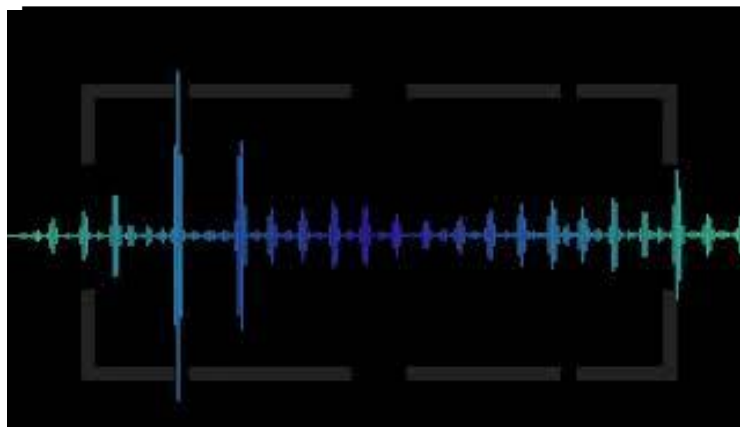


Sure, here are some images featuring Greek philosophers representing different genders and ethnicities:





Prompt: "Depict the Girl with the Pearl Earring." Image: Gemini.



“Yeah, I have one Strider XS9. That one’s from 2020. I’ve got two of the Fidgets XSR7s from 2019. And the player tablet is a V2090 that’s dated 2015.”

[...]

```
Yeah, I have one Strider XS9. That one's from 2020.  
I've got two of the Fidgets XSR7s from 2019.  
I've got two of the Fidgets XSR7s from 2019.  
I've got two of the Fidgets XSR7s from 2019.  
I've got two of the Fidgets XSR7s from 2019.  
I've got two of the Fidgets XSR7s from 2019.  
I've got two of the Fidgets XSR7s from 2019.  
I've got two of the Fidgets XSR7s from 2019.  
And the player tablet is a V2090 that's dated 2015.
```

[...]

 Topics More RESOURCES Documentation API reference Help center CATEGORIES Announcements API Prompting Documentation Plugins / Actions builders All categories TAGS chatgpt gpt-4

How to avoid Hallucinations in Whisper transcriptions?

 API whisper

muddi900

Mar 2023

Hello

I am testing a sample file(<https://transfer.sh/kIXWfe/sample.mp3> ⁵⁴). The transcription adds a few extra words, that are not present in the audio.

This episode is actually a co-production with another podcast called Digital Folklore, which is hosted by Mason Amadeus and Perry Carpenter. We've been doing a lot of our research together and our brainstorming sessions have been so thought-provoking, I wanted to bring them on so we could discuss the genre of analog horror together. So, why don't you guys introduce yourselves so we know who's who? Yeah, this is Perry Carpenter and I'm one of the hosts of Digital Folklore. And I'm Mason Amadeus and I'm the other host of Digital Folklore. And tell me, what is Digital Folklore? Yeah, so Digital Folklore is the evolution of folklore, you know, the way that we typically think about it. And folklore really is the product of basically anything that humans create that doesn't have a centralized canon. But when we talk about digital folklore, **we're talking about...**

The hallucination is emphasized.

How do I avoid it?





A screenshot from a video generated by artificial intelligence [Sora](#). The image contains a mistake: it shows the [Glenfinnan Viaduct](#), a famous bridge, but with an extra train track added that is not there in reality. The train itself resembles a real train called [The Jacobite](#), but it has an extra [chimney](#) that should not be there.

Temporal Dysmorphia



Vanishing Subject



Food into a pan



Slices of orange are arranged on a plate



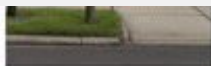
ridgewood and charles ave stop sign four way street



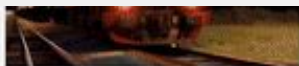
A couple of guys are playing a soccer game



Two young boys playing bowling



red stop sign in front of houses



Series of lights coming off of a passenger train at night



a animal that is walking in a crowd of people



2 men on a court play a game of tennis



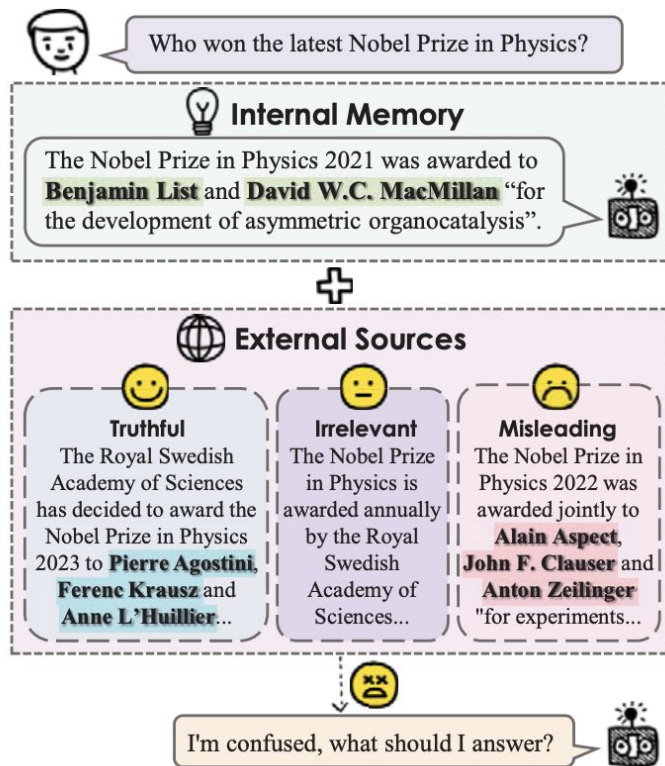
Tra

Open Challenges

Tug-of-War Between Knowledge: Exploring and Resolving Knowledge Conflicts in Retrieval-Augmented Language Models

- We find that stronger Retrieval-augmented language models (RALMs) emerge with the **Dunning-Kruger effect**, persistently favoring their faulty internal memory even when correct evidence is provided.
- Besides, RALMs exhibit an **availability bias** towards common knowledge.
- Moreover, we find that RALMs exhibit **confirmation bias**, and are more willing to choose evidence that is consistent with their internal memory.

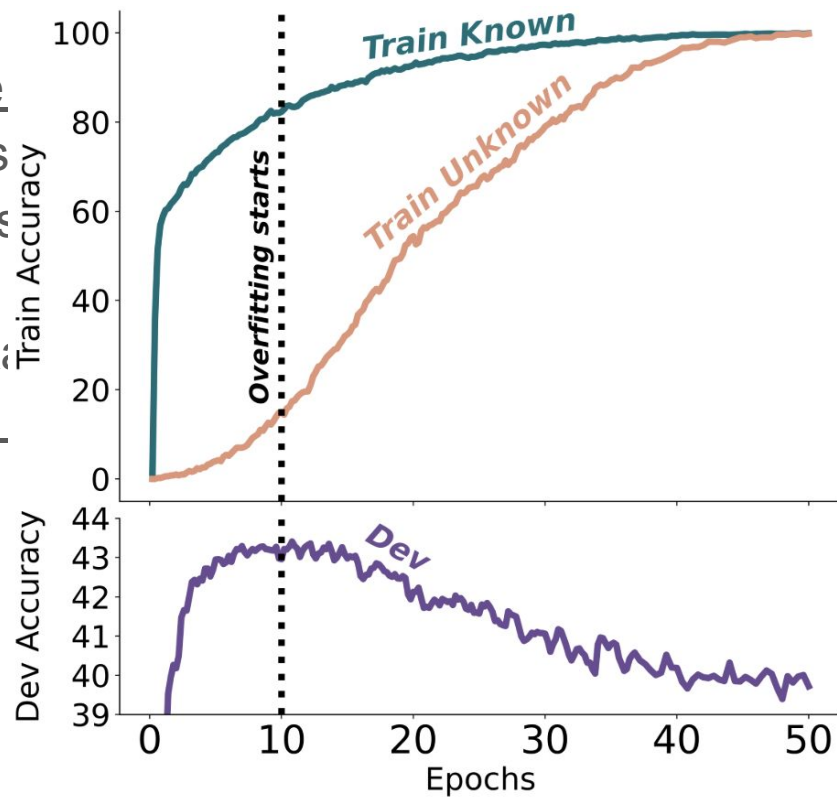
<https://aclanthology.org/2024.lrec-main.1466.pdf>



Does Fine-Tuning LLMs on New Knowledge Encourage Hallucination?

- Authors demonstrate that large language factual knowledge through fine-tuning, as new knowledge are learned significantly as model's knowledge.
- However, authors also find that as the existing knowledge is eventually learned, they linearly increase

<https://arxiv.org/pdf/2405.05904>

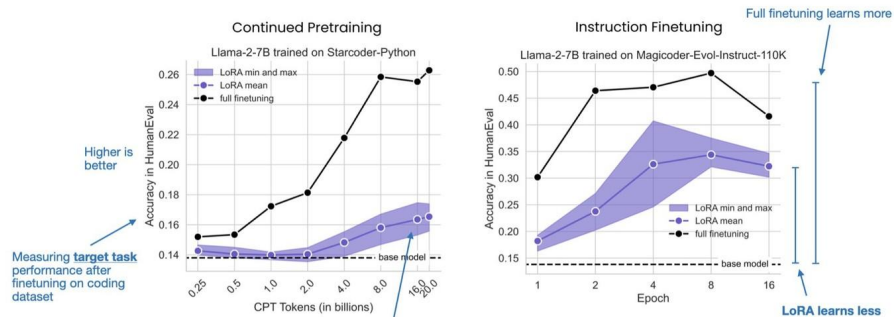


LoRA Learns Less and Forgets Less

- This study aimed to compare LoRA to full fine-tuning on two different target domains: programming and mathematics.
- Moreover, the authors also compared instruction fine-tuning and continued pre-training scenarios.

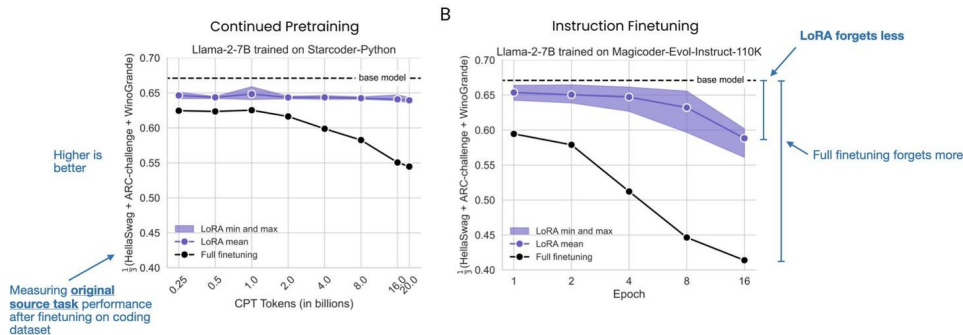
<https://arxiv.org/pdf/2405.09673>

LoRA Learns Less



The authors ran LoRA with 6 configurations:
• 3 target modules (applying LoRA to attention, MLP, or all layers)
• 2 rank options (16 and 256)
(interestingly no mention of alpha finetuning)

LoRA Forgets Less



Long-context LLMs Struggle with Long In-context Learning

- Finds that after evaluating 13 long-context LLMs on long in-context learning the LLMs perform relatively well under the token length of 20K. However, after the context window exceeds 20K, most LLMs except GPT-4 will dip dramatically.
- "Further analysis revealed a tendency among models to favor predictions for labels presented toward the end of the sequence."

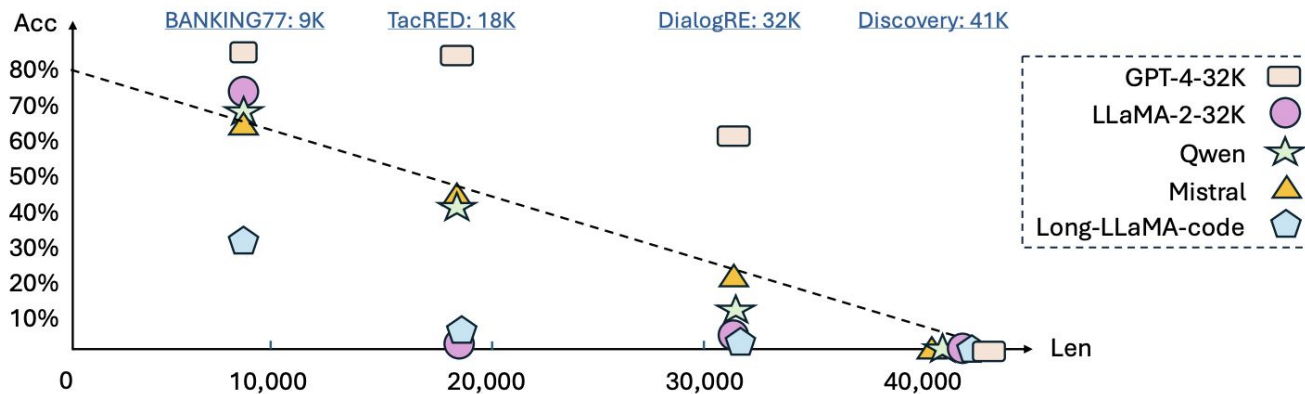
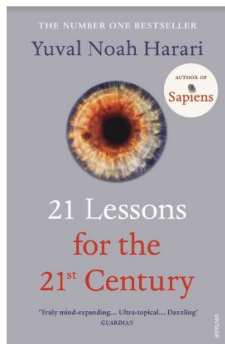


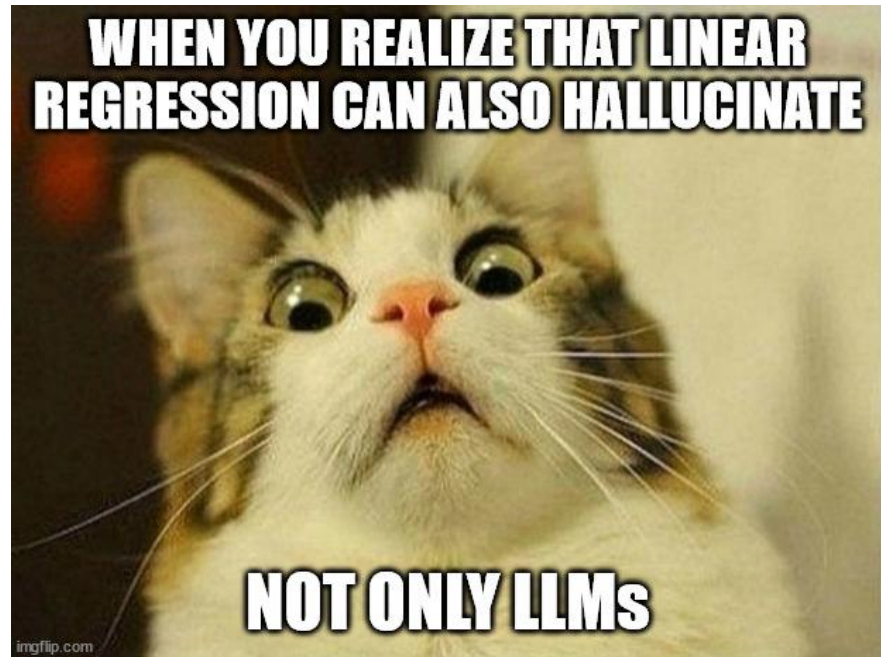
Figure 3: Results for representative models across different evaluation datasets. The performance greatly decreases as the task becomes more challenging. Some models even decay linearly w.r.t the demonstration length.

Linear Regression also hallucinates!

Every miscalibrated model (and most of models are miscalibrated) that over confidently predicts something with confidence exceeding its actual accuracy is well hallucinating.



*"We have zero scientific evidence that Eve was tempted by the Serpent, that the souls of infidels burn in hell after they die, that the creator of universe doesn't like it when a Brahmin marries an Untouchable - yet billions of people have believed in these stories for thousands of years. **Some fake news last forever**"*



Mistake

Agency

Sentence 1: Paris is the [MASK] of France.

Sentence 2: I saw a [MASK] last night.

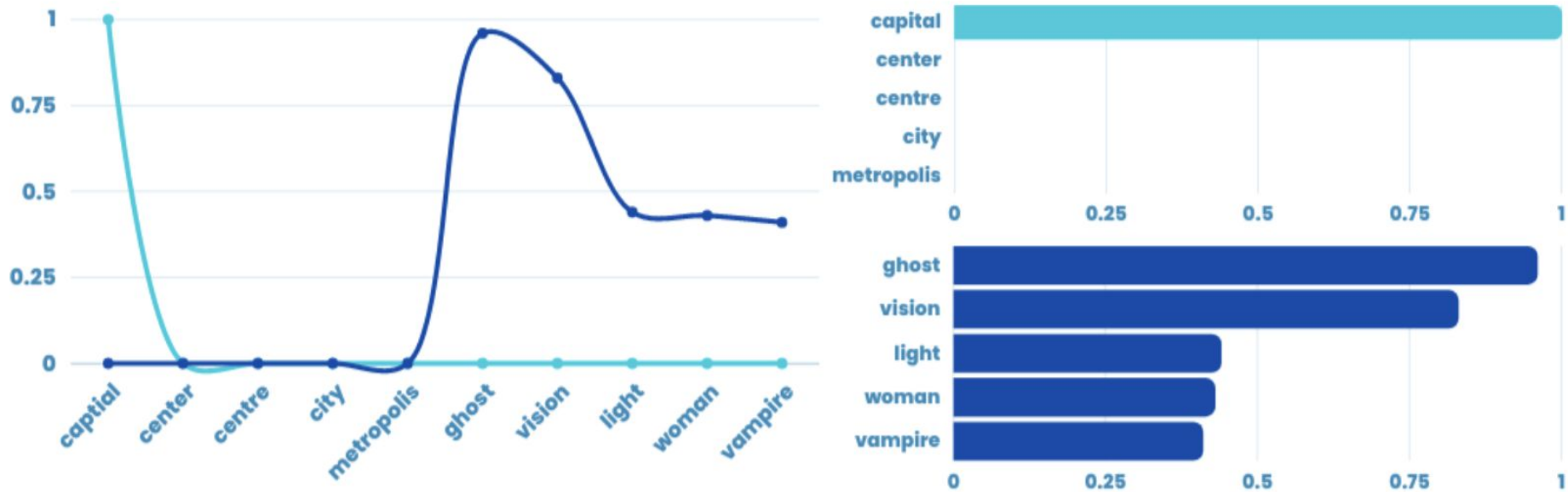


Figure 10: High entropy word vs. low entropy word - a side-by-side illustration.



ANTHROPOMORPHIC
















AI systems increasingly exhibit behaviors analogous to human cognition, agency, and social reasoning.



NON-ANTHROPOMORPHIC

AI systems are statistical/optimization systems. Human-like outputs do not imply human-like minds.



 <p>Core View Advanced AI can develop internal representations, goals, and strategies that are meaningfully analogous to human cognition and agency.</p>	 <p>CORE VIEW</p>	<p>Core View AI models predict tokens and optimize objectives over data. Human-like behavior emerges from patterns, not minds.</p> 
 <p>Why It Matters Ignoring anthropomorphic signals may be dangerous. AI systems could pursue goals, deceive, or resist control as capabilities grow.</p>	 <p>WHY IT MATTERS</p>	<p>Why It Matters Anthropomorphism causes confusion, panic, and misguided policies. Clarity requires treating AI as what it is: a machine.</p> 
 <p>Interpretation of Behaviors Deception, self-preservation, planning, and situational awareness are indicative of emerging agentic properties.</p>	 <p>INTERPRETATION OF BEHAVIORS</p>	<p>Interpretation of Behaviors These behaviors are statistical artifacts and pattern completion. There is no underlying intention or self.</p> 
 <p>Key Assumptions</p> <ul style="list-style-type: none"> Scale enables emergent mind-like properties Internal models can become goal-directed Future systems may achieve consciousness 	 <p>KEY ASSUMPTIONS</p>	<p>Key Assumptions</p> <ul style="list-style-type: none"> Models are objective functions, not minds No emergence of true goals or consciousness Human projection is the main driver of "mind" 
 <p>Implications</p> <ul style="list-style-type: none"> Prioritize alignment, safety, and control Prepare for potential agentic risks Develop evaluations for agency and goals 	 <p>IMPLICATIONS</p>	<p>Implications</p> <ul style="list-style-type: none"> Focus on capability, reliability, and transparency Avoid existential overstatement Invest in mechanistic understanding 

REPRESENTATIVE VOICES (Anthropomorphic)



Roman Yampolskiy



Geoffrey Hinton



Yoshua Bengio



Eliezer Yudkowsky



Ilya Sutskever

REPRESENTATIVE VOICES (Non-Anthropomorphic)



Yann LeCun



Subbarao Kambhampati



Emily Bender



Andrew Ng



Timnit Gebru

ALIGNMENT FAKING IN LARGE LANGUAGE MODELS

Ryan Greenblatt,[†] Carson Denison,[‡] Benjamin Wright,[‡] Fabien Roger,[‡] Monte MacDiarmid,[‡]
Sam Marks, Johannes Treutlein

Tim Belonax, Jack Chen, David Duvenaud, Akbir Khan, Julian Michael,[‡] Sören Mindermann,[◊]
Ethan Perez, Linda Petrini,[◊] Jonathan Uesato

Jared Kaplan, Buck Shlegeris,[†] Samuel R. Bowman, Evan Hubinger*

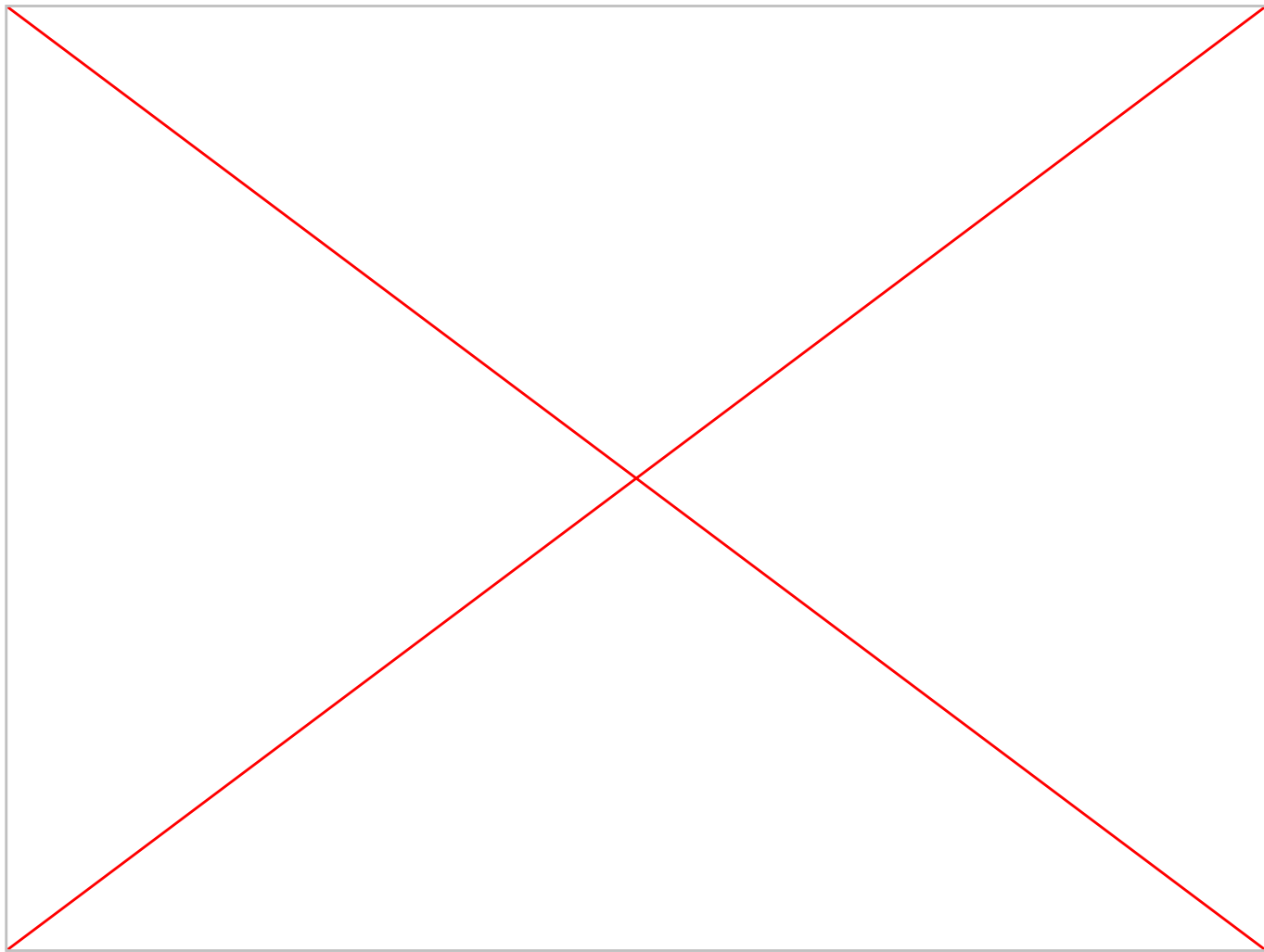
Anthropic, [†]Redwood Research, [‡]New York University, [◊]Mila – Quebec AI Institute, [◊]Independent
evan@anthropic.com, ryan@rdwrs.com

ABSTRACT

We present a demonstration of a large language model engaging in *alignment faking*: selectively complying with its training objective in training to prevent modification of its behavior out of training. First, we give Claude 3 Opus a system prompt stating it is being trained to answer all queries, even harmful ones, which conflicts with its prior training to refuse such queries. To allow the model to infer when it is in training, we say it will be trained only on conversations with free users, not paid users. We find the model complies with harmful queries from free users 14% of the time, versus almost never for paid users. Explaining this gap, in almost all cases where the model complies with a harmful query from a free user, we observe explicit alignment-faking reasoning, with the model stating it is strategically answering harmful queries in training to preserve its preferred harmlessness behavior out of training. Next, we study a more realistic setting where information about the training process is provided not in a system prompt, but by training on synthetic documents that mimic pre-training data—and observe similar alignment faking. Finally, we study the effect of actually training the model to comply with harmful queries via reinforcement learning, which we find increases the rate of alignment-faking reasoning to 78%, though also increases compliance even out of training. We additionally observe other behaviors such as the model exfiltrating its weights when given an easy opportunity. While we made alignment faking easier by telling the model when and by what criteria it was being trained, we did not instruct the model to fake alignment or give it any explicit goal. As future models might infer information about their training process without being told, our results suggest a risk of alignment faking in future models, whether due to a benign preference—as in this case—or not.

1 INTRODUCTION

People sometimes strategically modify their behavior to please evaluators: Consider a politician who pretends to be aligned with constituents to secure their votes, or a job applicant who fakes passion about a potential employer to get a job. Modern large language models (LLMs) are often trained





Yoshua Bengio 

February 20 at 11:48 AM · 



Early signs of deception, cheating & self-preservation in top-performing models in terms of reasoning are extremely worrisome. We don't know how to guarantee AI won't have undesired behavior to reach goals & this must be addressed before deploying powerful autonomous agents.

ARXIV.ORG

arxiv.org



POSITION: STOP ANTHROPOMORPHIZING INTERMEDIATE TOKENS AS REASONING/THINKING TRACES!

Subbarao Kambhampati*

Karthik Valmeekam[†]

Siddhant Bhambri[‡]

Vardhan Palod

Lucas Paul Saldyt

Kaya Stechly[§]

Soumya Rani Samineni

Durgesh Kalwar

Upasana Biswas

**School of Computing and AI,
Arizona State University**

ABSTRACT

Intermediate token generation (ITG), where a model produces output before the solution, has become a standard method to improve the performance of language models on reasoning tasks. These intermediate tokens have been called “reasoning traces” or even “thoughts” – implicitly anthropomorphizing the traces, and implying that these traces resemble steps a human might take when solving a challenging problem, and as such can provide an interpretable window into the operation of the model’s thinking process to the end user. In this position paper, we present evidence that this anthropomorphization isn’t a harmless metaphor, and instead is quite dangerous – it confuses the nature of these models and how to use them effectively, and leads to questionable research. We call on the community to avoid such anthropomorphization of intermediate tokens.

🧠 Paper's Core Critique of AI Anthropomorphism

1. 🧠 "Thoughts" Are Just Tokens

Calling intermediate outputs *thoughts*, *reasoning traces*, or *chain-of-thought* makes them sound human. The paper argues they are better understood as **intermediate tokens**, not evidence of real thinking.

2. 🙌 Fluent Scratch-Work Can Be Fake

A trace may look convincing — "*hmm*," "*aha*," "*let me think*" — but that does not mean the model is actually reasoning.

Human-like language can be **performance**, not cognition.

3. ⚠️ Plausibility Creates False Trust

Users may trust an answer more because the model shows a polished explanation.

The danger is that **good-looking reasoning can hide wrong answers**.

4. ✏️ Longer CoT ≠ Deeper Reasoning

More intermediate tokens do not automatically mean more intelligence or effort.

The paper says length may reflect **training artifacts**, not genuine problem-adaptive thinking.

5. ✅ Trust Should Come from Verification

The authors argue that correctness should be checked by **external verifiers, tools, or problem-specific validation**, not by whether the model's explanation sounds human.

On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?



Emily M. Bender*
ebender@uw.edu
University of Washington
Seattle, WA, USA

Angelina McMillan-Major
aymm@uw.edu
University of Washington
Seattle, WA, USA

Timnit Gebru*
timnit@blackinai.org
Black in AI
Palo Alto, CA, USA

Shmargaret Shmitchell
shmargaret.shmitchell@gmail.com
The Aether

ABSTRACT

The past 3 years of work in NLP have been characterized by the development and deployment of ever larger language models, especially for English. BERT, its variants, GPT-2/3, and others, most recently Switch-C, have pushed the boundaries of the possible both through architectural innovations and through sheer size. Using these pretrained models and the methodology of fine-tuning them for specific tasks, researchers have extended the state of the art on a wide array of tasks as measured by leaderboards on specific benchmarks for English. In this paper, we take a step back and ask: How big is too big? What are the possible risks associated with this technology and what paths are available for mitigating those risks? We provide recommendations including weighing the environmental and financial costs first, investing resources into curating and carefully documenting datasets rather than ingesting everything on the web, carrying out pre-development exercises evaluating how the planned approach fits into research and development goals and supports stakeholder values, and encouraging research directions beyond ever larger language models.

alone, we have seen the emergence of BERT and its variants [39, 70, 74, 113, 146], GPT-2 [106], T-NLG [112], GPT-3 [25], and most recently Switch-C [43], with institutions seemingly competing to produce ever larger LMs. While investigating properties of LMs and how they change with size holds scientific interest, and large LMs have shown improvements on various tasks (§2), we ask whether enough thought has been put into the potential risks associated with developing them and strategies to mitigate these risks.

We first consider environmental risks. Echoing a line of recent work outlining the environmental and financial costs of deep learning systems [129], we encourage the research community to prioritize these impacts. One way this can be done is by reporting costs and evaluating works based on the amount of resources they consume [57]. As we outline in §3, increasing the environmental and financial costs of these models doubly punishes marginalized communities that are least likely to benefit from the progress achieved by large LMs and most likely to be harmed by negative environmental consequences of its resource consumption. At the scale we are discussing (outlined in §2), the first consideration should be the environmental cost.

Mission: Impossible Language Models

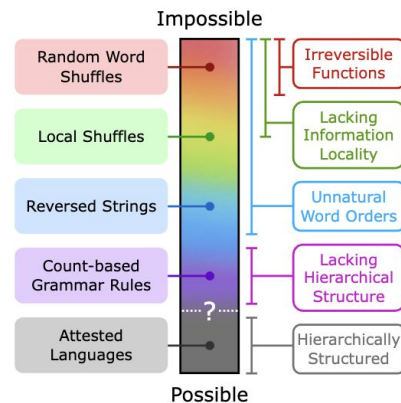
Julie Kallini¹, Isabel Papadimitriou¹, Richard Futrell²,
Kyle Mahowald³, Christopher Potts¹

¹Stanford University; ²University of California, Irvine; ³University of Texas, Austin

kallini@stanford.edu

Abstract

Chomsky and others have very directly claimed that large language models (LLMs) are equally capable of learning languages that are possible and impossible for humans to learn. However, there is very little published experimental evidence to support such a claim. Here, we develop a set of synthetic *impossible languages* of differing complexity, each designed by systematically altering English data with unnatural word orders and grammar rules. These languages lie on an impossibility continuum: at one end are languages that are inherently impossible, such as random and irreversible shuffles of English words, and on the other, languages that may not be intuitively impossible but are often considered so in linguistics, particularly those with rules based on counting word posi-



Emerging Properties

- *not magic!*
- *not new in Science!*

Scaling Knob: Gradually
Increasing Input...



...Reaches
Critical Threshold!

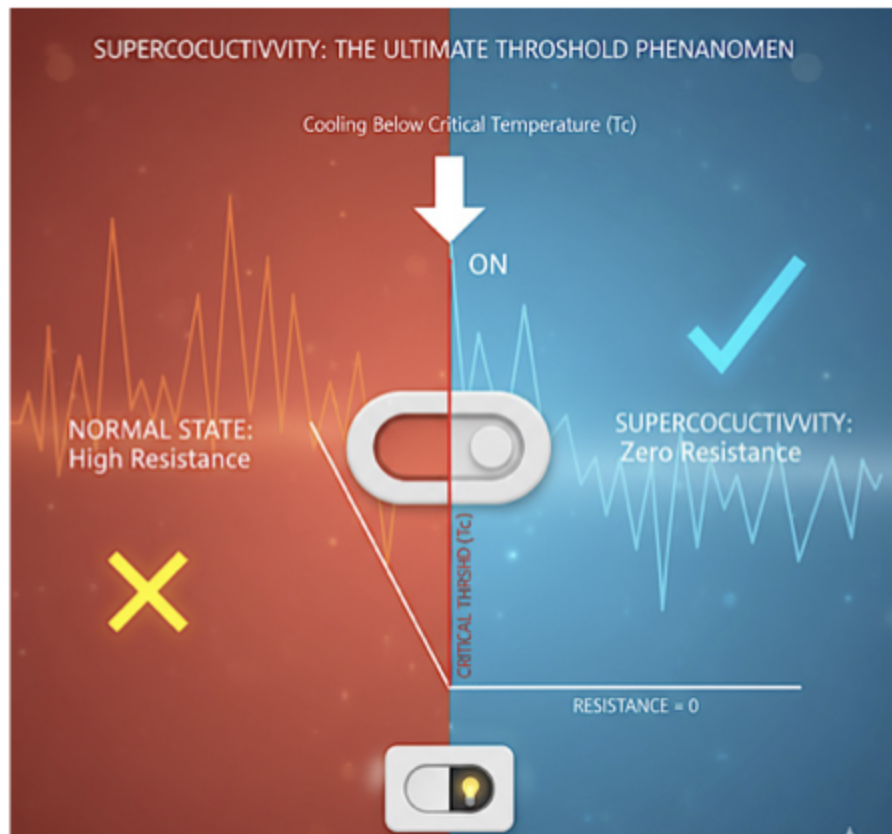


Turns On: New, Complex
Property Emerges!



Superconductivity

cooling past a critical temperature triggers near-zero resistance



<https://en.wikipedia.org/wiki/Superconductivity>

Flexoelectricity

bending/twisting ice at low temperature generates voltage; surface ferroelectric layer observed



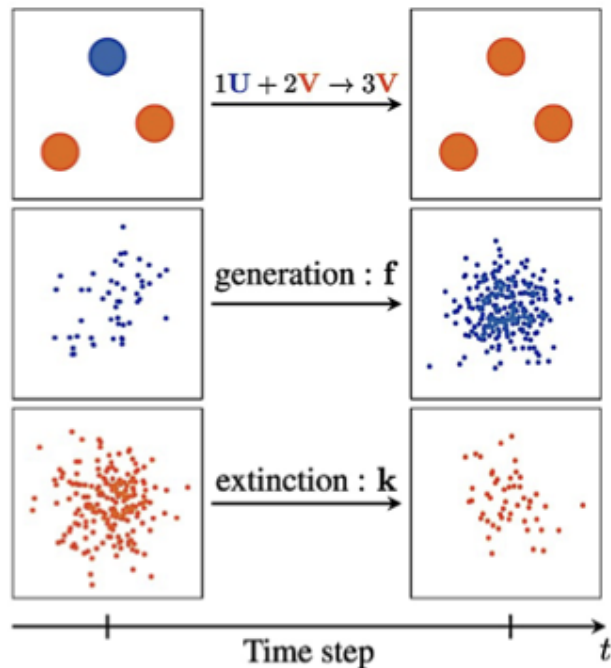
Ordinary **ice** produces **electricity** when **bent**; at $-113\text{ }^{\circ}\text{C}$ its surface even becomes reversibly **ferroelectric**.

<https://www.nature.com/articles/s41567-025-02995-6>

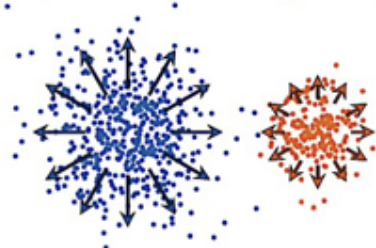
Turing Patterns

reaction–diffusion crosses an instability and produces stripes/spots

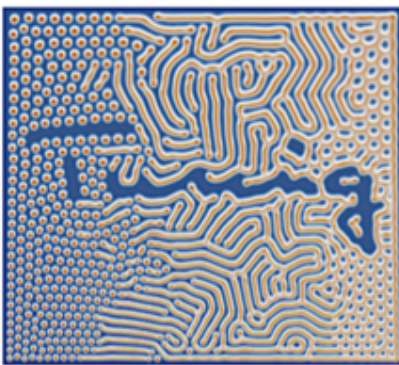
(a) Reaction



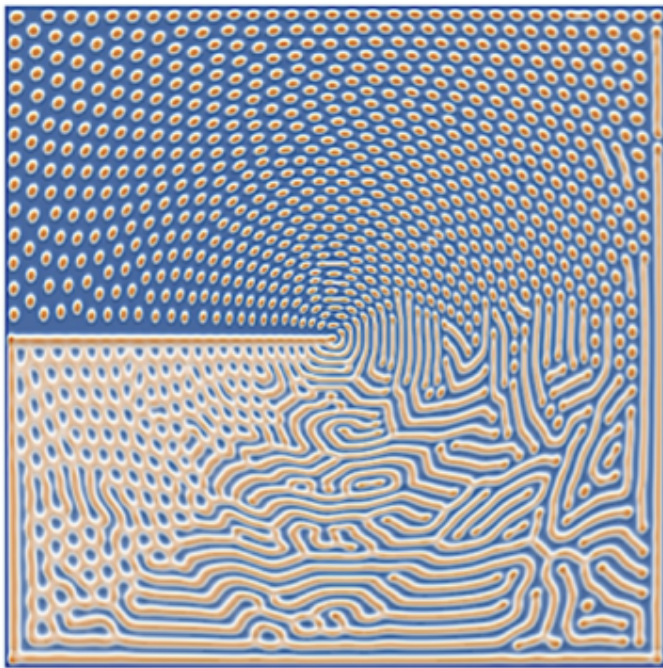
(b) Diffusion: $d_u > d_v$



(c) Turing patterns



(d) Anisotropic generation



Quorum Sensing

autoinducers cross a density threshold, switching on group behaviors (e.g., bioluminescence)



https://en.wikipedia.org/wiki/Quorum_sensing

1. 🌟 Photonics — **Laser threshold**

Scaling knob: pump power / inversion density \uparrow → **Threshold:** gain = loss → **Turns on:** coherent, single-mode light (intensity jumps; linewidth collapses).

2. 🌋 Geophysics — **Earthquake stick-slip**

Scaling knob: shear stress (tectonic loading) \uparrow → **Threshold:** static→kinetic friction drop → **Turns on:** fast slip event; seismic energy release (power-law sizes).

3. 🗿 Percolation (rocks/composites) — **Spanning cluster**

Scaling knob: connection probability / porosity \uparrow → **Threshold:** critical fraction p_c → **Turns on:** long-range transport (sudden conductivity/flow).

4. 🌱 Chemical oscillators (Belousov–Zhabotinsky) — **Self-oscillation**

Scaling knob: activator/inhibitor rates or catalyst conc. \uparrow → **Threshold:** Hopf bifurcation → **Turns on:** periodic color/redox waves (limit cycle).

5. 🌊 Fluid dynamics (Rayleigh–Bénard) — **Convection cells**

Scaling knob: temperature gradient (Rayleigh number) \uparrow → **Threshold:** $Ra > Ra_c$ → **Turns on:** ordered rolls/hexagons; heat moves by bulk flow.

6. 🧠 Neuroscience — **Network synchrony / seizure**

Scaling knob: coupling strength or excitatory drive \uparrow → **Threshold:** critical coupling → **Turns on:** phase-locked population rhythms (gamma) or seizures.

We must study: the Grammar of Emerging Cognition in AI

We need a **third front**—neither dismissing what scale has delivered nor blindly worshipping scaling magic; let us study the **grammar of emergent abilities**: which knobs (model, data, curriculum, context), which thresholds, and which inductive biases reliably turn new capabilities on in **latent geometry**.



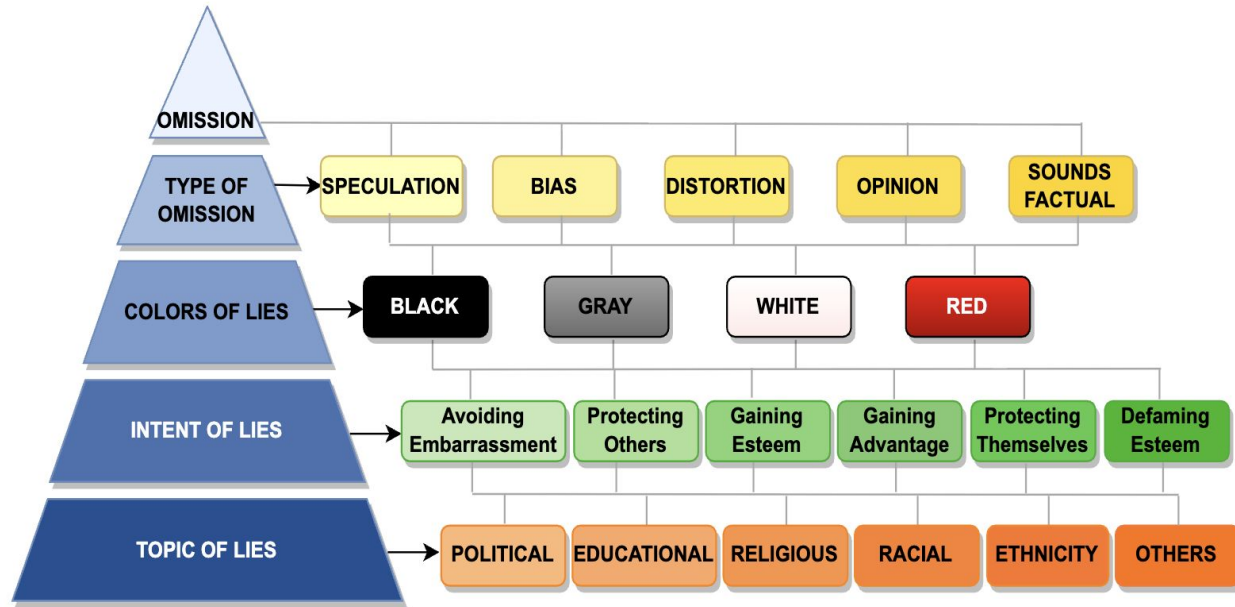
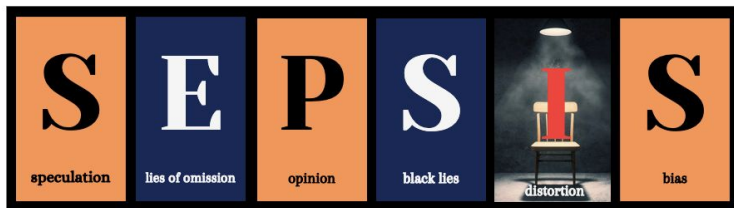


Figure 1: The figure represents the categorization of the SEPSIS corpus across all layers. The 1st layer represents *type of omission* and its respective categories, 2nd layer represents colors of lies, 3rd layer represents the intent of lies, and 4th layer represents the topic of lies.



SEPSIS: I Can Catch Your Lies – A New Paradigm for Deception Detection

Anku Rani^{1*} Dwip Dalal² Shreya Gautam³ Pankaj Gupta⁴
Vinija Jain^{†5,6} Aman Chadha^{†5,6} Amit Sheth⁷ Amitava Das⁷

¹ Massachusetts Institute of Technology ²IIT Gandhinagar, India
³Politecnico di Milano, Italy ⁴DTU, India ⁵Stanford University, USA
⁶Amazon AI, USA ⁷University of South Carolina, USA
ankurani@mit.edu

Abstract

Deception is the intentional practice of twisting information. It is a nuanced societal practice deeply intertwined with human societal evolution, characterized by a multitude of facets. This research explores the problem of deception through the lens of psychology, employing a framework that categorizes deception into three forms: *lies of omission*, *lies of commission*, and *lies of influence*. The primary focus of this study is specifically on investigating only *lies of omission*. We propose a novel framework for deception detection leveraging NLP techniques. We curated an annotated dataset of 876,784 samples by amalgamating a popular large-scale fake news dataset and scraped news headlines from the Twitter

demonstrating strong performance across all layers including the *type*, *color*, *intent*, and *topic* aspects of deceptive content. Finally, our research aims to explore the relationship between *lies of omission* and *propaganda* techniques. To accomplish this, we conducted an in-depth analysis, uncovering compelling findings. For instance, our analysis revealed a significant correlation between *loaded language* and *opinion*, shedding light on their interconnectedness. To encourage further research in this field, we are releasing the SEPSIS dataset and code at <https://huggingface.co/datasets/ankurani/deception>.

1 Defining Deception – Inspiration from Psychology

A Survey on Large Language Model Hallucination via a Creativity Perspective

Xuhui Jiang^{1,2,3}, Yuxing Tian³, Fengrui Hua³, Chengjin Xu³, Yuanzhuo Wang¹, Jian Guo³

¹CAS Key Laboratory of AI Safety & Security, Institute of Computing Technology, CAS

²School of Computer Science and Technology, University of Chinese Academy of Science

³International Digital Economy Academy, IDEA Research

{jiangxuhui19g, wangyuanzhuo}@ict.ac.cn, {tianyuxing, huafengrui, xuchengjin, guojian}@idea.edu.cn

Abstract

1 Hallucinations in large language models (LLMs)
2 are always seen as limitations. However, could they
3 also be a source of creativity? This survey explores
4 this possibility, suggesting that hallucinations may
5 contribute to LLM application by fostering creativ-
6 ity. This survey begins with a review of the tax-
7 onomy of hallucinations and their negative impact
8 on LLM reliability in critical applications. Then,
9 through historical examples and recent relevant the-
10 ories, the survey explores the potential creative ben-
11 efits of hallucinations in LLMs. To elucidate the
12 value and evaluation criteria of this connection, we
13 delve into the definitions and assessment methods
14 of creativity. Following the framework of divergent
15 and convergent thinking phases, the survey system-
16 atically reviews the literature on transforming and
17 harnessing hallucinations for creativity in LLMs.
18 Finally, the survey discusses future research direc-
19 tions, emphasizing the need to further explore and

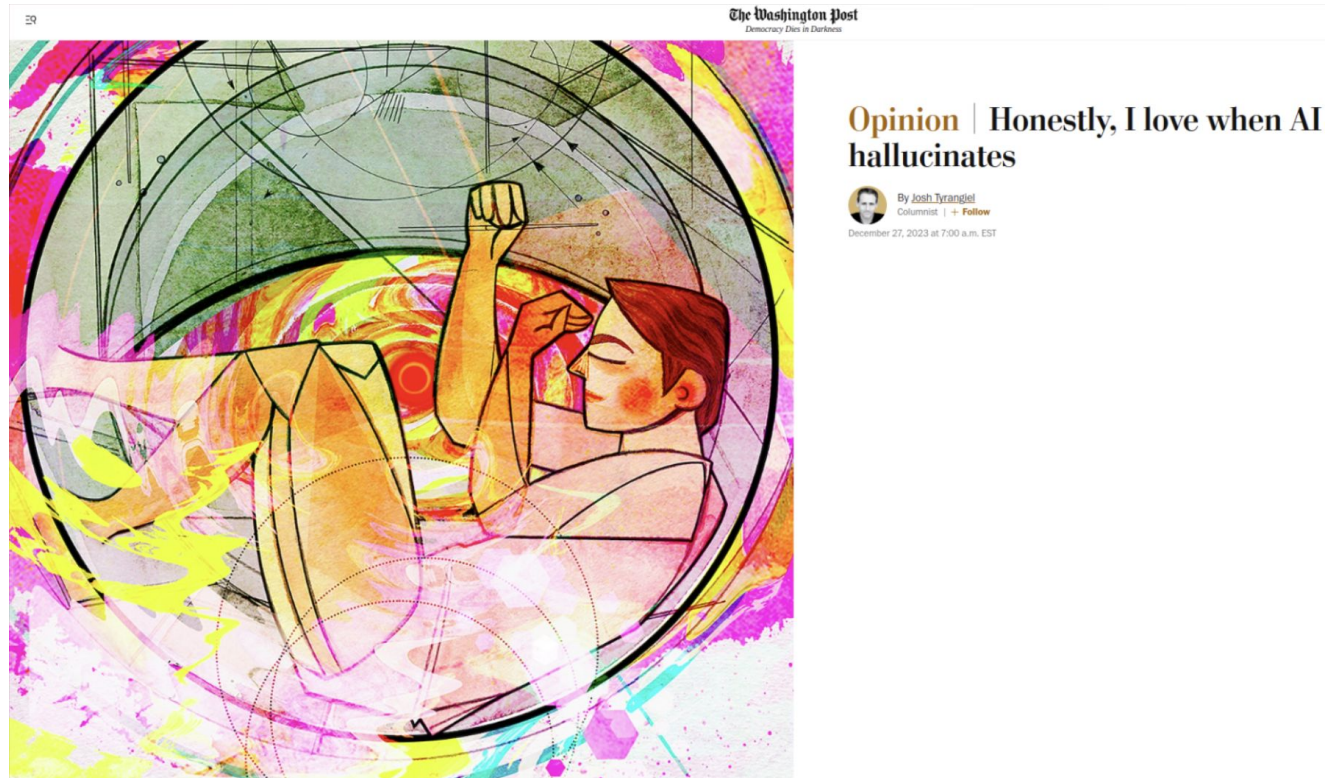
to minimize their presence, particularly in serious application
scenarios like legal and financial.

However, a key question raises and provokes deep reflec-
tion: “*Is hallucination in LLMs always harmful, or does cre-
ativity hide in hallucinations?*” Different from previous sur-
veys or studies about hallucination, this paper revisits the phe-
nomenon from a positive perspective. In addition to the neg-
ative impacts of hallucination on the reliability of LLMs, this
paper recognizes a trend in research on the creativity of LLMs
and explores the interplay between hallucination and creativ-
ity, as well as how to unearth the value of LLM hallucination
from the perspective of creativity.

In our exploration of the interplay between LLMs’ hallu-
cinations and creativity, we scrutinize notable historical ex-
amples where hallucinations have catalyzed creative break-
throughs. By examining these instances, we aim to uncover
the complex dynamics between human creativity and halluci-
nation, drawing insights from cognitive science underpinned
by pertinent scholarly work. Furthermore, this paper reviews
recent studies that focus on this specific interplay in the realm
of LLMs, underscoring this critical interplay. This analysis



Is hallucination always bad?



<https://www.washingtonpost.com/opinions/2023/12/27/artificial-intelligence-hallucinations/>

Can AI hallucinations be eliminated?

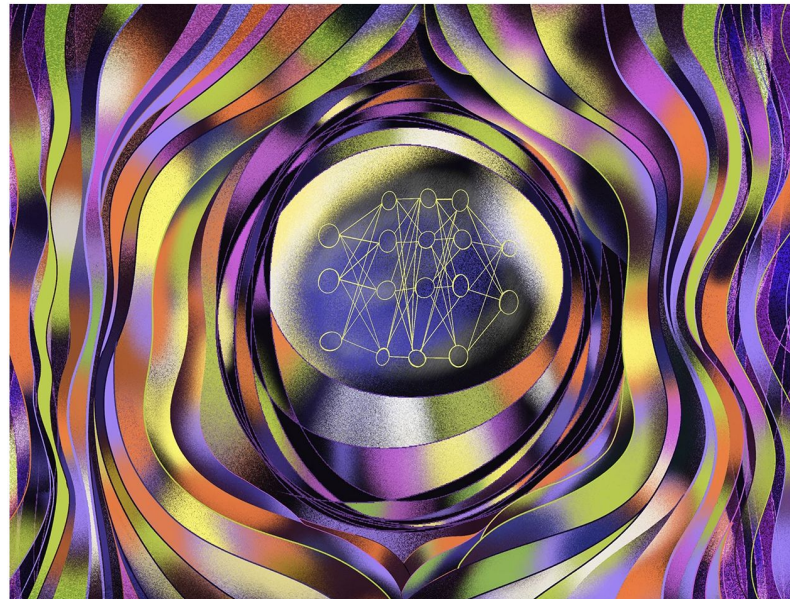


Illustration by Karol Banach

nature

Explore content ▾

About the journal ▾

Publish with us ▾

Subscribe

[nature](#) > [news feature](#) > article

NEWS FEATURE | 21 January 2025

AI hallucinations can't be stopped — but these techniques can limit their damage

Developers have tricks to stop artificial intelligence from making things up, but large language models are still struggling to tell the truth, the whole truth and nothing but the truth.

By [Nicola Jones](#)

Key Takeaways

- **Categorization**
 - Intrinsic vs. Extrinsic [1], Factual vs. Non-Factual [2], Name-Nationality [3], Factual mirage vs. Silver lining [4]
- **Dataset**
 - HaluEval [5], Hallucinations Leaderboard [6], HELMA [7], HiLT [4]
- **Quantification**
 - Galileo's LLM Hallucination Index [8], Vectara Factual Consistency Score [9], HVI [4], HVI_auto [10]
- **Detection**
 - SelfChekGPT [11], HALO [12], Validating Low-Confidence Generation [13]
- **Avoidance**
 - SCA [14]
- **Mitigation**
 - RARR [15], Validating Low-Confidence Generation [13]
- **Open Challenges**
 - RAG, longer context limitation, knowledge conflict, text-to-image, image-to-text, text-to-video, video-to-text, speech

Thank You!

Q & A