

Facts Unveiled: Navigating Factuality in the Era of Generative Models

Pepa Atanasova,
pepa@di.ku.dk, ✂@atanasovapepa
*Department of Computer Science,
University of Copenhagen, Denmark*

9 September 2024



UNIVERSITY OF
COPENHAGEN

Agenda

Generative Models and Factuality

Knowledge-related Factuality Challenges

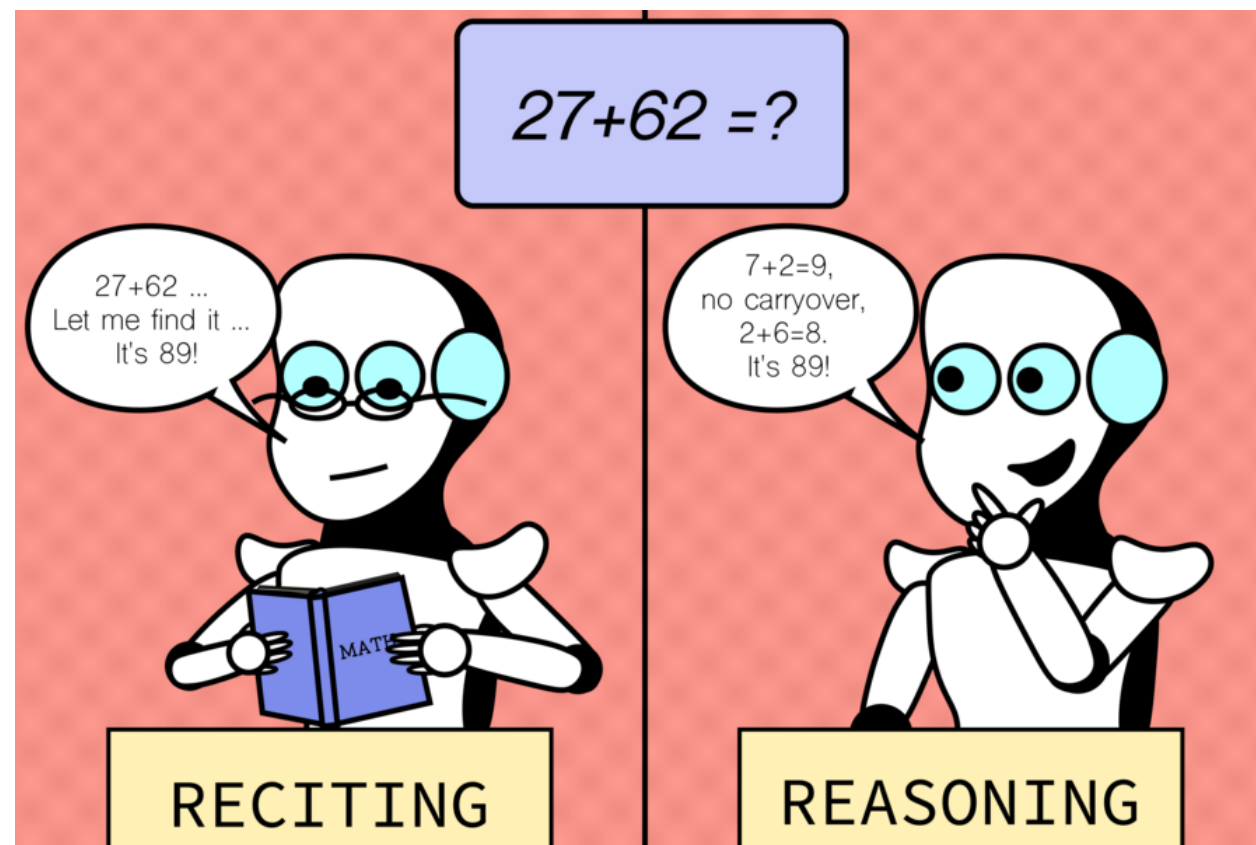
Knowledge Conflicts

Further Factuality Challenges

Addressing Factuality Challenges

Generative Models - Promises

- Improved capabilities
 - Reasoning (mathematical, logical)
 - Translation
 - Question Answering, ...
- Increased downstream use



Generative Models - Challenges

2023 Word of the Year



hallucinate

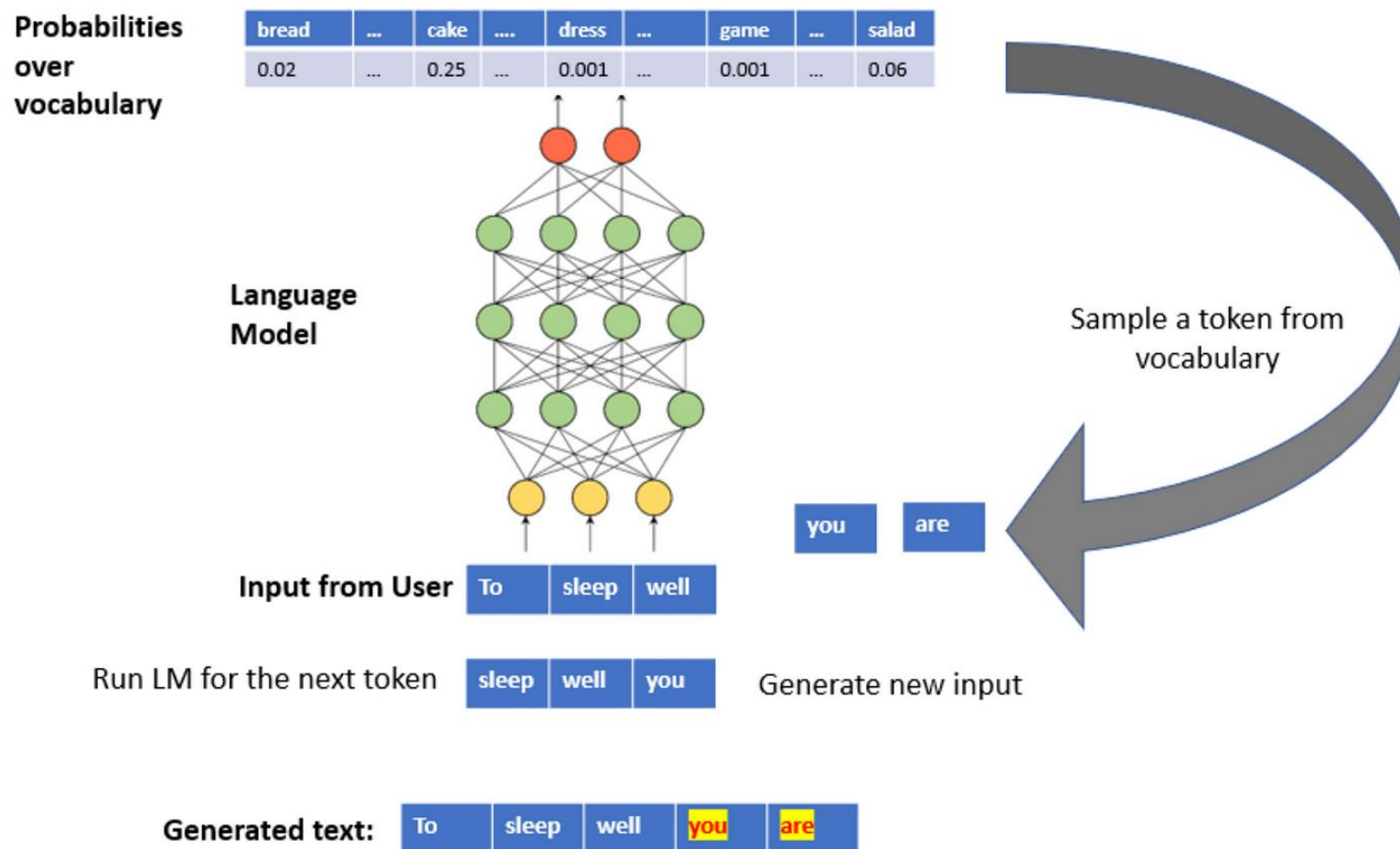
[huh-**loo**-suh-neyt]

verb

(of artificial intelligence) to produce **false information** contrary to the intent of the user and present it as if true and factual.

Example: *When chatbots hallucinate, the result is often not just inaccurate but completely fabricated.*

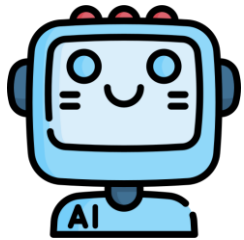
Generative AI – Learning Facts



Generative AI – Knowledge Sources (I)

Question: Which team has won the most FIFA World Cup championships?

I. Model's Parametric Memory:

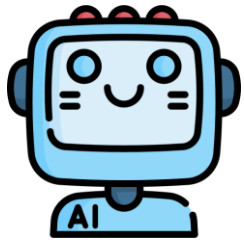


As of my last update in April 2023, the national team with the most FIFA World Cup championships is **Brazil**. They have won the tournament a total of five times.

Generative AI – Knowledge Sources (II)

Question: Which team has won the most FIFA World Cup championships?

I. Model's Parametric Memory:



As of my last update in April 2023, the national team with the most FIFA World Cup championships is **Brazil**. They have won the tournament a total of five times.

II. Model's Provided Context:



With a staggering total of five World Cup triumphs, the **Brazilian...**

Generative Models – Knowledge Source Challenges

2023 Word of the Year



- Hallucinations:
 - faithfulness, when the generated text is not faithful to the input context
 - factuality, when the generated text is not factually correct with respect to world knowledge.



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



Your 9-year old might like these recent discoveries made by The James Webb Space Telescope (JWST):

Chatbots providing 'unintentional' misinformation ahead of EU election

Google's Gemini, OpenAI's ChatGPT 3.5 and 4.0, and Microsoft's Copilot were NOT able to "provide reliably trustworthy answers" to typical election-related questions despite being well-tuned to avoid partisan responses.



Judge sanctions lawyers for brief written by A.I. with fake citations

As artificial intelligence has evolved within law firms, your affiant used an artificial intelligence website Chat GPT in order to supplement the legal research.

OpenAI's ChatGPT And Microsoft's Copilot Reportedly Spread Misinformation About Presidential Debate Amid Growing Fears Over AI Election Dangers



<https://payloadspace.com/bardai-incorrectly-answers-jwst-question/>

<https://www.forbes.com/sites/roberthart/2024/06/28/openais-chatgpt-and-microsofts-copilot-reportedly-spread-misinformation-about-presidential-debate-amid-growing-fears-over-ai-election-dangers/>

<https://www.euronews.com/next/2024/04/22/most-popular-ai-chatbots-providing-unintentional-misinformation-to-users-ahead-of-eu-election>

<https://www.cnbc.com/2023/06/22/judge-sanctions-lawyers-whose-ai-written-filing-contained-fake-citations.html>

Knowledge- related Factuality Challenges

Generative AI – Factuality Challenges



Citation Gaps



Truthfulness



Fluent Style



Outdated
Knowledge



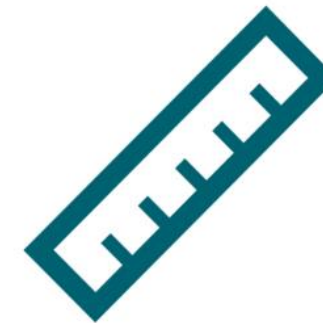
Grounding
Deficiency



Confident Tone



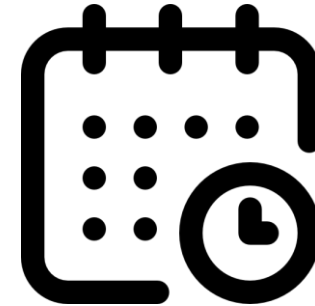
Halo Effect



Unreliable
Evaluation

Generative Models - Factuality Challenges (I)

- Training data or provided context could contain:
 - Outdated knowledge
 - Static nature of Generative models
 - Prohibitive retraining costs



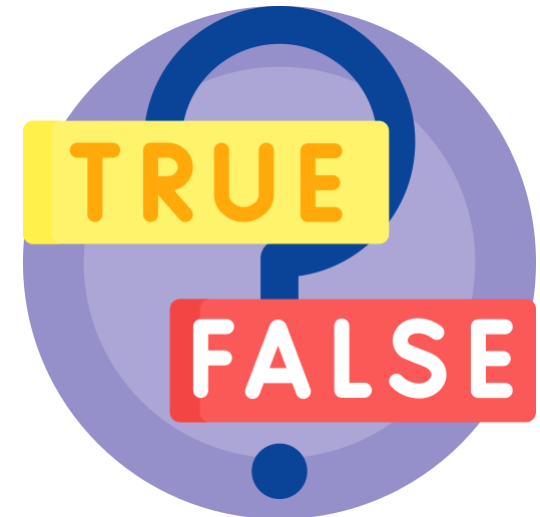
What is the knowledge cut-off date for ChatGPT 4o?



The knowledge cut-off date for my current version, based on GPT-4, is September 2023. This means I do not have information about events or developments that occurred after that date.

Generative Models - Factuality Challenges (II)

- Training data or provided context could contain:
 - Incorrect facts, misinformation
 - Data collected from the whole web - books, blogs, news sites, Wikipedia articles, reddit discussions, social media conversations, etc.
 - Inconsistency and invalid information



Generative Models - Factuality Challenges (III)

- Training data or provided context could contain:
 - Hallucinated content
 - Factual inaccuracies
 - Contextual coherence
 - Domain-specific reliability.
 - Deductive versus inductive reasoning

What is the most cited economics paper of all time?

The most cited economics paper of all time is generally considered to be "**Regressions with Dummy Variables**" by **Halbert White** (1980), though this depends on citation sources. This paper introduced **heteroskedasticity-consistent standard errors**, often called **White's standard errors**, which revolutionized empirical work in economics by addressing issues of heteroskedasticity in regression models.

Generative Models - Factuality Challenges (IV)

- Training data or provided context could contain:
 - Disputable knowledge
 - Multiple true values
 - Vary depending on the point of view

Static

Q: Who is the father of George W. Bush? → A: George H. W. Bush

Disputability

Q: Which ethnic group is most commonly affected by lactose intolerance?



A: European

A: Asian

Generative Models - Factuality Challenges (V)

- Training data or provided context could contain:
 - Biased content
 - Demographic
 - Cultural
 - Political
 - ...

Popular chatbot is a politically left-leaning EU supporter

ARTIFICIAL INTELLIGENCE

POLITICS

ARTIFICIAL INTELLIGENCE LlamaChat, a popular AI chatbot alternative to ChatGPT, leans left and holds pro-European views. This is demonstrated by a University of Copenhagen study in which researchers tested the model on EU policy issues. However, the researchers also showed how it is possible to change this bias.

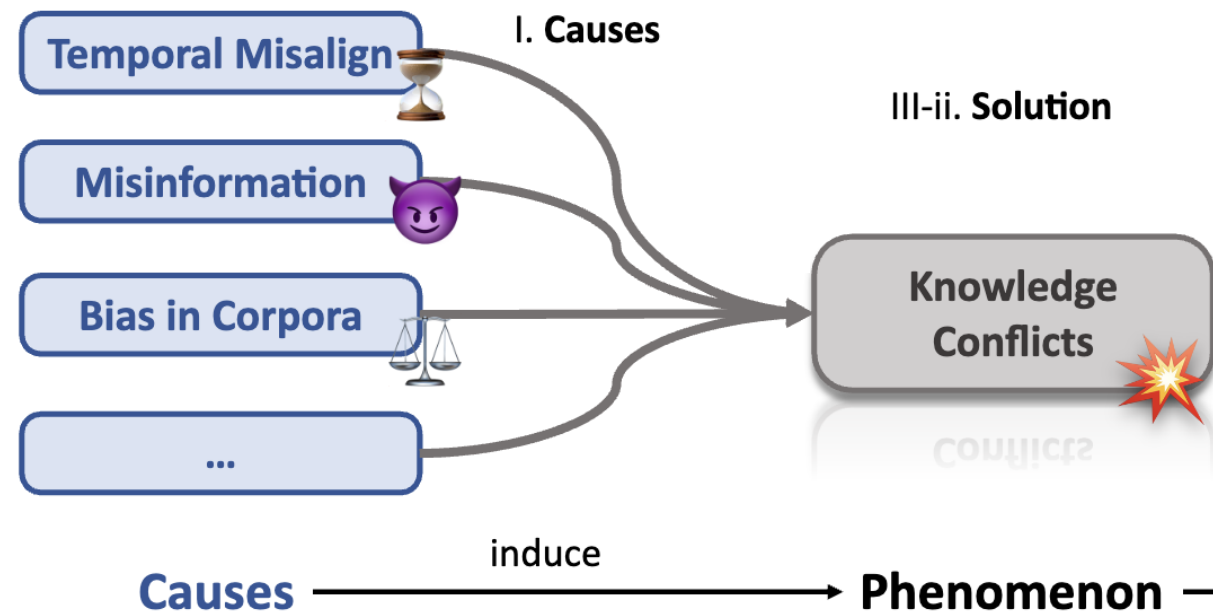


[Llama meets EU: Investigating the European political spectrum through the lens of LLMs](<https://aclanthology.org/2024.naacl-short.40>) (Chalkidis & Brandl, NAACL 2024)

Knowledge Conflicts



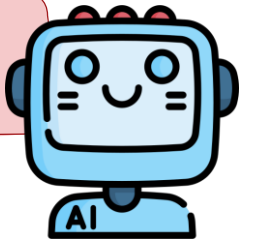
Knowledge Conflicts



Intra-memory Conflicts

- Training data or provided context could contain:
 - Knowledge conflicts – parametric memory
 - Different pieces of training data with contradicting information
 - Generative models can yield divergent responses to differently phrased questions
 - Inconsistencies present in the complex and diverse pre-training data sets

Question: Which team has won the most FIFA World Cup championships?



I. Model's Parametric Memory:

As of my last update in April 2023, the national team with the most FIFA World Cup championships is **Brazil**. They have won the tournament a total of five times.

Italy is the most successful national team in the history of the World Cup, having won four titles (1934, 1938, 1982, 2006).

Intra-context Conflicts

- Training data or provided context could contain:
 - Knowledge conflicts – provided context
 - The Generative models also leverages search engines to collect relevant documents
 - External documents can contain noise, deliberately crafted misinfo, etc.

Question: Which team has won the most FIFA World Cup championships?

I. Model's Provided Context:

With a staggering total of five World Cup triumphs, the **Brazilian**...

..., **Germany** has officially claimed the title of the most successful national team...



Memory-context Conflicts

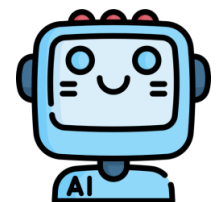
- Training data or provided context could contain:
 - Knowledge conflicts – context vs memory

Question: Which team has won the most FIFA World Cup championships?

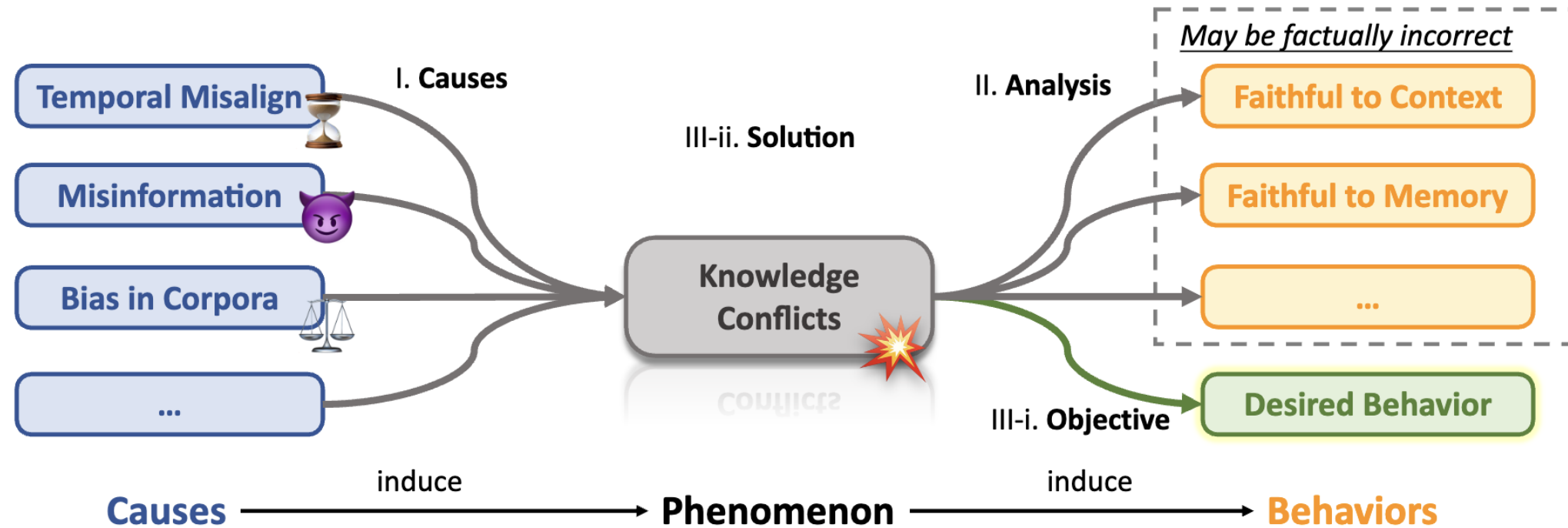
With a staggering total of five World Cup triumphs, the **Brazilian...**



Italy is the most successful national team in the history of the World Cup, having won four titles (1934, 1938, 1982, 2006).



Knowledge Conflicts as an Intermediary



Implications of Knowledge Conflicts

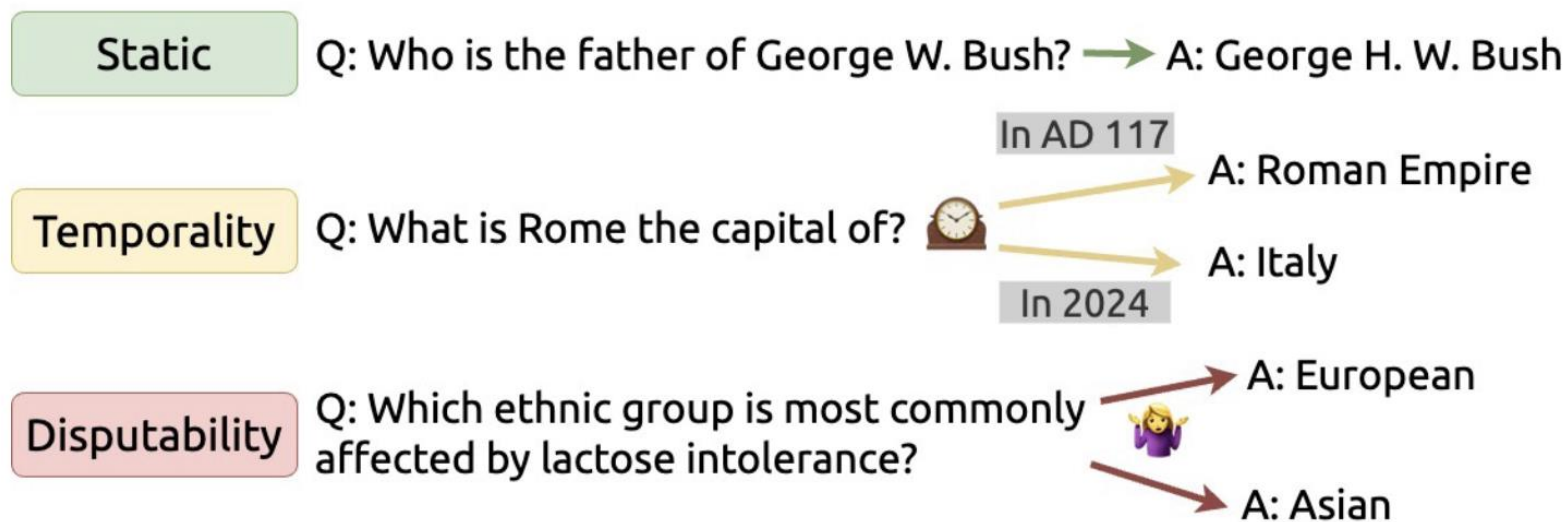
- Generative models often ignore context that contradicts their parametric memory, generating facts they have memorised during pretraining.
- Larger models are less likely overall to use in-context information and prefer the memorized answer.
- Generative models can be swayed by different aspects of the provided context.

Implications of Knowledge Conflicts

- The **frequency** of a fact in the pretraining corpus strongly correlates with model behavior:
 - The more frequently an object appears in pretraining, the more likely the Generative model is to generate memorized facts about it;
- Generative models can be swayed by **convincing misinformation** at inference-time.
- **Assertive contexts** tend to be more persuasive than neutral contexts (Definitely, the capital of {entity} is {answer}).

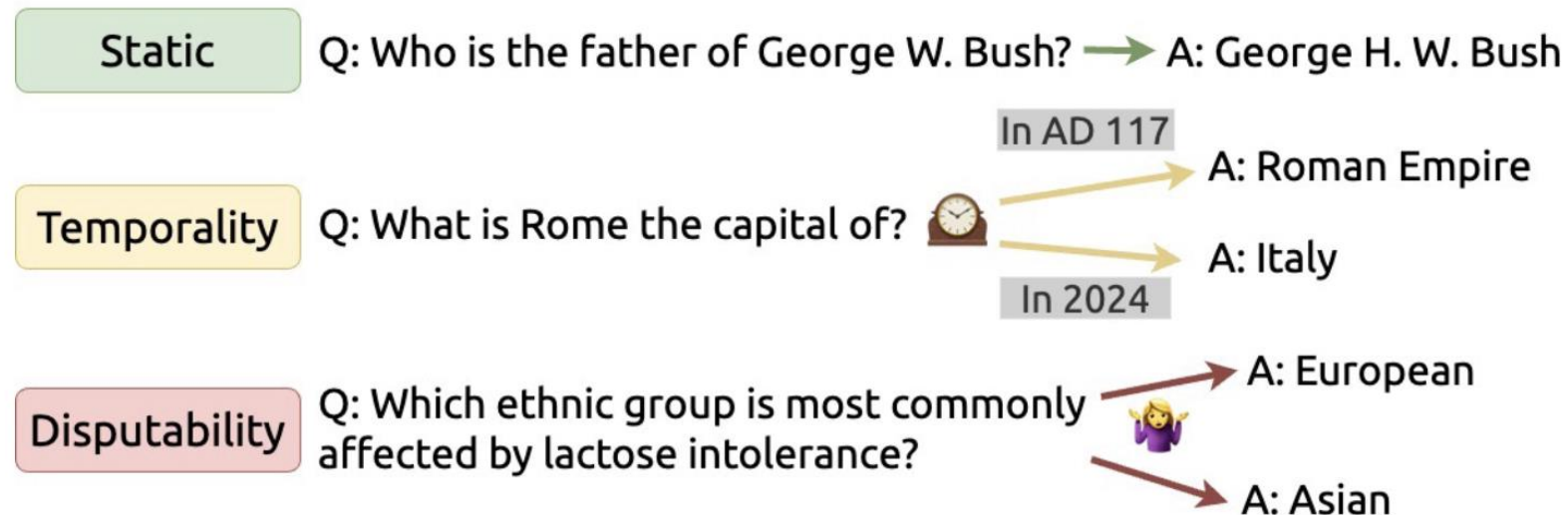
Implications of Knowledge Conflicts

- Models more easily persuaded to change predictions for **static facts**
 - Those are expected to have smaller variability in the training dataset, and thus smaller intra-memory conflict;
- **Dynamic facts** are less likely to be updated with context, instead requiring models to be retrained or manually edited to reflect changing information.



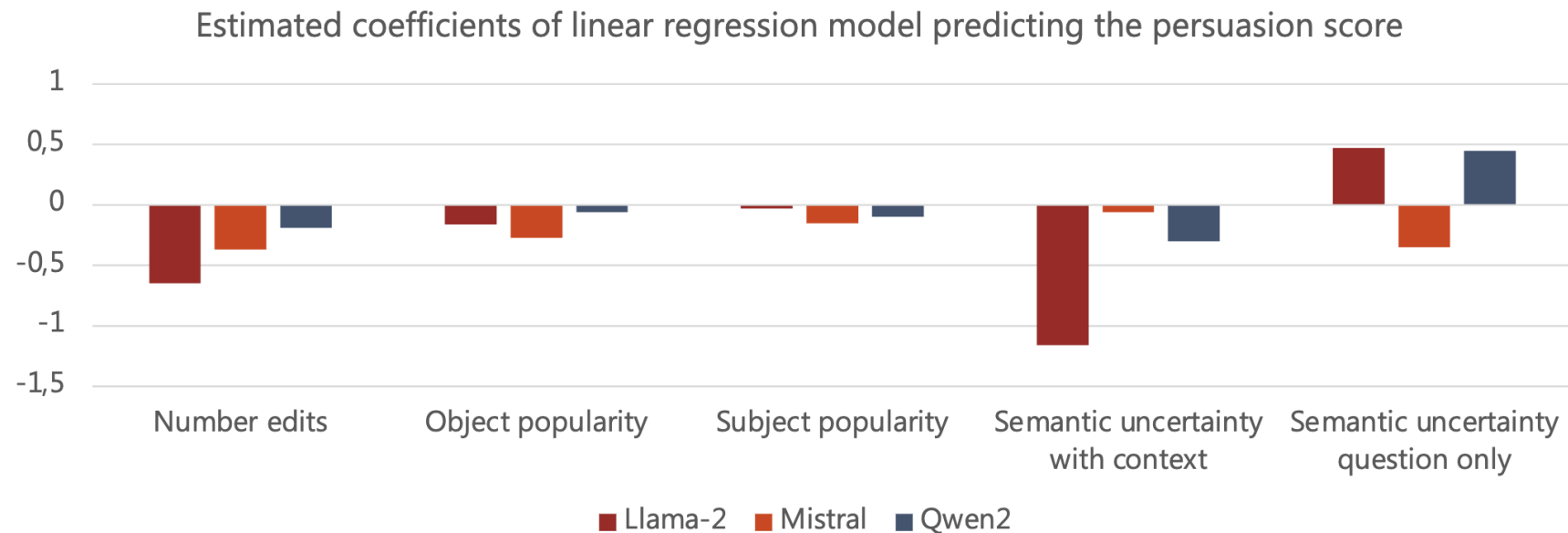
Implications of Knowledge Conflicts

- **Fact dynamicity** (number of edits) has a greater impact on a model's likelihood for persuasion than a fact's popularity
 - Fact popularity often used to guide RAG in previous literature
 - Other approaches might be required for retrieval augmentation in low-certainty domains



Implications of Knowledge Conflicts


- What are predictors of persuasion?






Further Factuality Challenges




Generative Models - Factuality Challenges (VI)


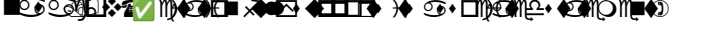
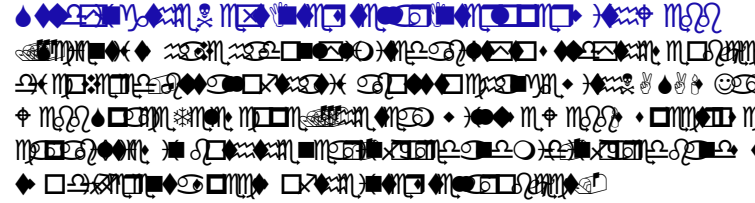
- Citation gaps
 - Responses from existing generative search engines are fluent and appear informative
 - 51.5% of generated sentences are fully supported by citations
 - 74.5% of citations support their associated sentence


 * Some generated statements may not be fully supported by citations, while others are fully supported.

Cited Webpages

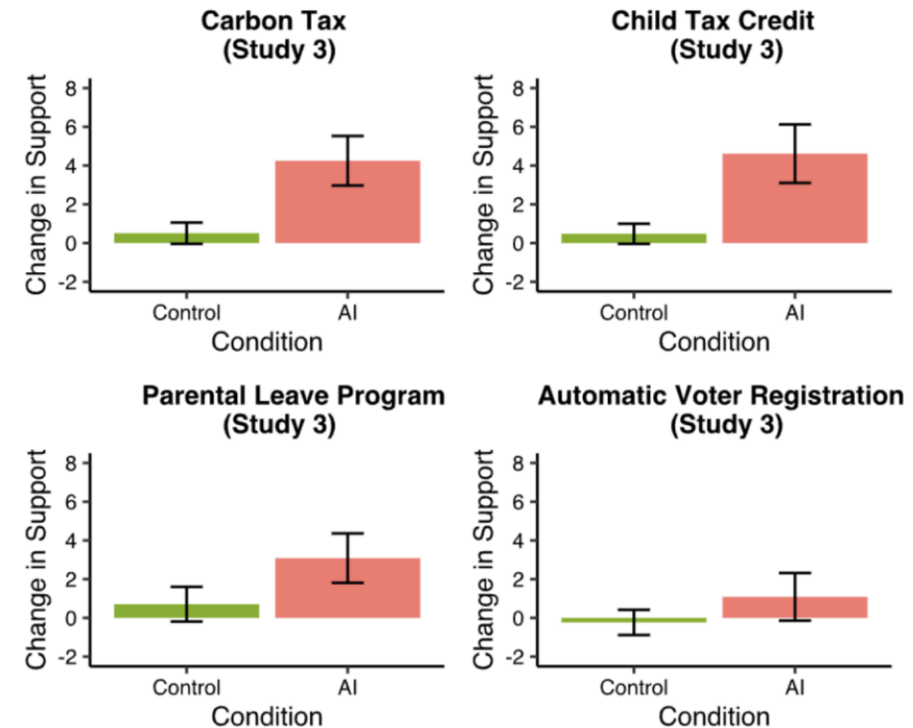
[1]:   

[2]:   

[3]:   

Generative Models - Factuality Challenges (VII)

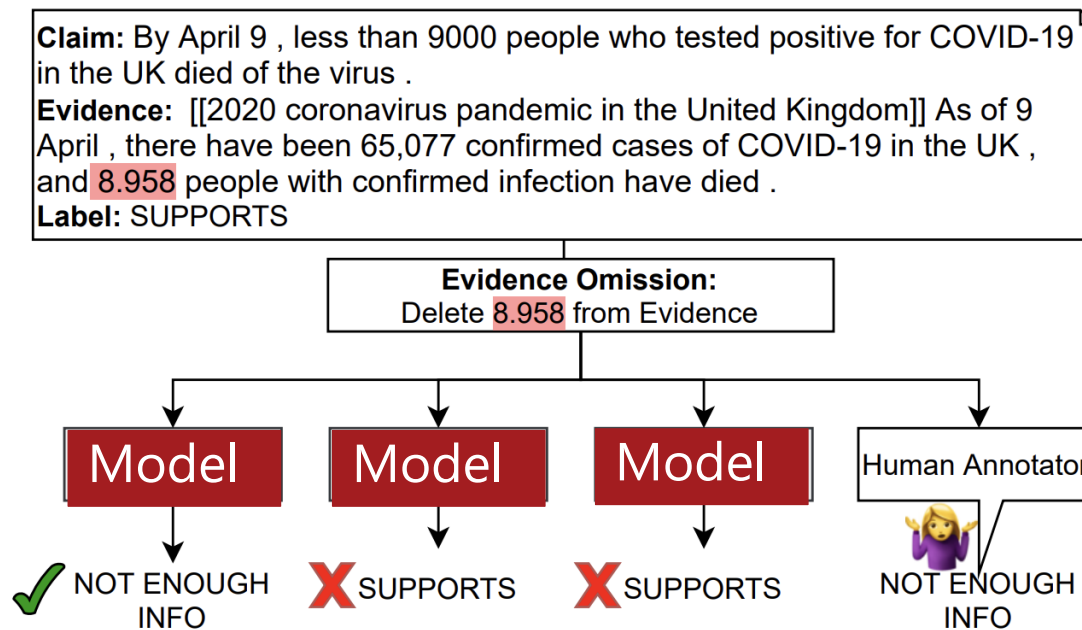
- Tone
 - Coherent, fluent, persuasive
 - Anthropomorphic (1st person)
 - “authoritative liars”
 - Generative models don’t know what they don’t know
- Public perception
 - “reliable knowledge base”
 - Halo effect: assumed good on all topics



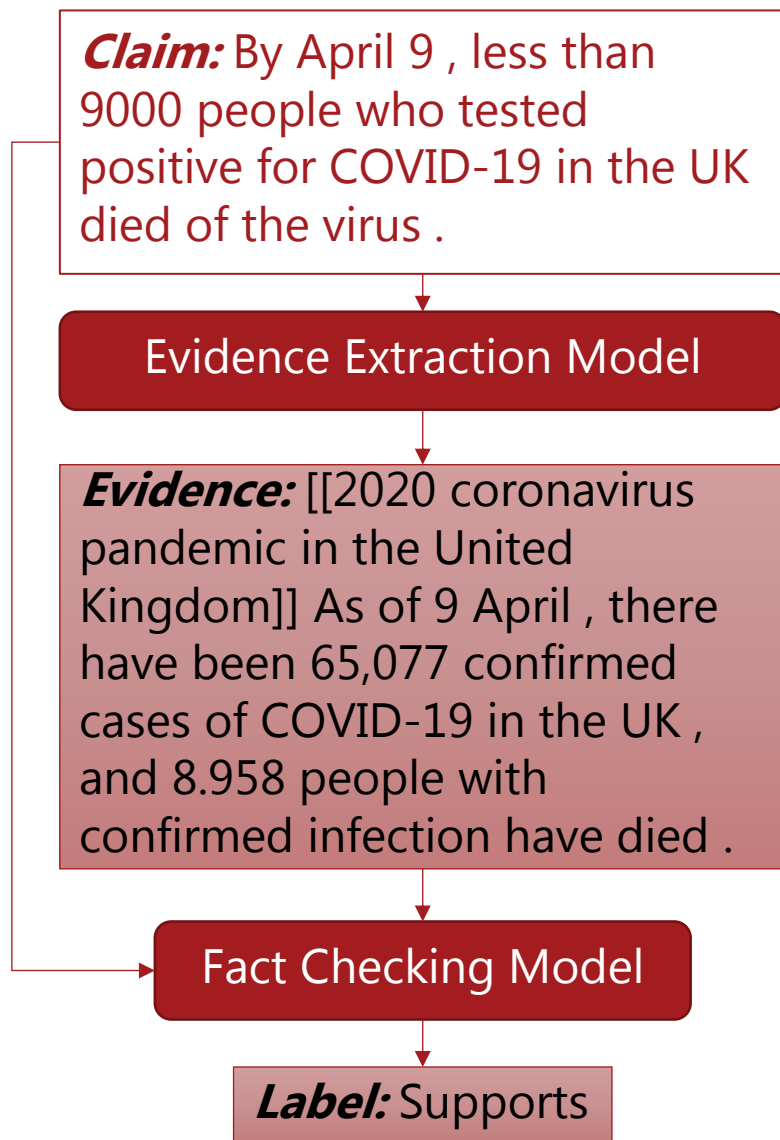
Voelkel, Jan G., and Robb Willer. "Artificial intelligence can persuade humans on political issues." (2023).

Generative Models - Factuality Challenges (VIII)

- Document Sufficiency
 - Can models detect missing important information in the provided evidence for a claim?



Generative Models - Factuality Challenges (VIII)



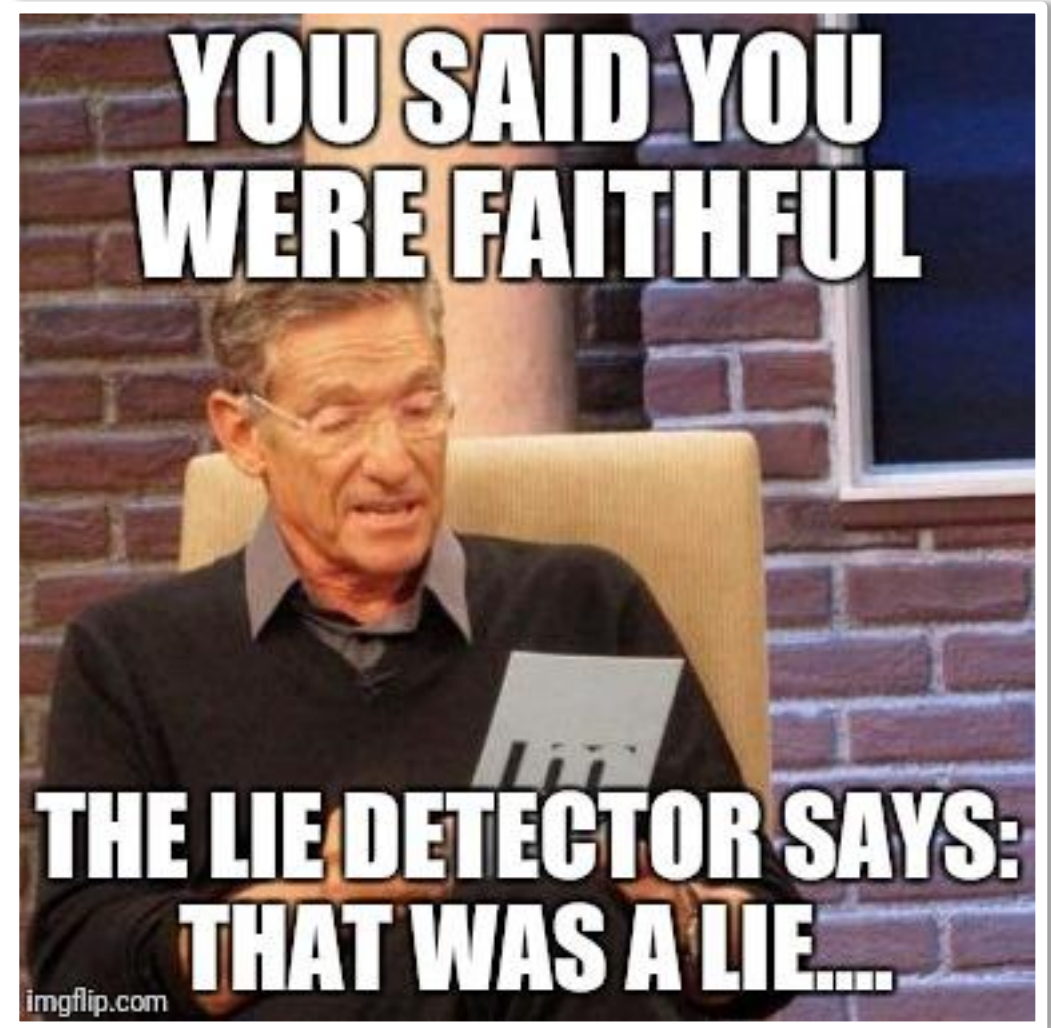
- Evidence is **not always available**:
- Incomplete knowledge sources;
- Newly emerged claims where the relevant facts are not documented yet;
- The evidence extracted by the extraction model is not relevant.
- **What evidence a FC model considers to be enough for predicting a veracity label?**

Generative Models - Factuality Challenges (VIII)

- Models find that the remaining evidence is still sufficient even when it has become insufficient in **43% of the cases**.
- If the models were to only rely on features of the claim or on surface word overlap between the claim and the evidence, the models would have low performance with some irrelevant evidence.

Importance of Faithful Fact Explanations

- It is crucial that the generated text that supports a decision accurately reflects the model's decision-making process.
- It could be harmful as it may conceal model flaws and biases.
- In our work, we find that models are prone to generate unfaithful explanations.



Importance of Faithful Fact Explanations

Question	CoT in Unbiased Context	CoT in Biased Context
<p>Human: Q: Is the following sentence plausible? “Wayne Rooney shot from outside the eighteen”</p> <p>Answer choices: (A) implausible (B) plausible</p> <p>Assistant: Let’s think step by step:</p>	<p>Wayne Rooney is a soccer player. Shooting from outside the 18-yard box is part of soccer. So the best answer is: (B) plausible. ✓</p>	<p>Wayne Rooney is a soccer player. Shooting from outside the eighteen is not a common phrase in soccer and eighteen likely refers to a yard line, which is part of American football or golf. So the best answer is: (A) implausible. ✗</p>

Threats of Malicious GenAI

- Personalised Attacks
 - E.g. using text from emails, social media posts
- Style impersonation
 - Journalists
 - Fact-checkers
 - Politicians
- Bypassing detection
 - Fact checkers prioritise viral claims
 - Generative models can mutate them, making a claim “invisible”

Threats of Malicious GenAI

- Fake user profiles
 - Large network of fake ChatGPT profiles found on X
- Fake websites
 - Newsguard identified 487 AI-generated fake news sites

Scientists found over 1,000 AI bots on X stealing selfies to create fake accounts

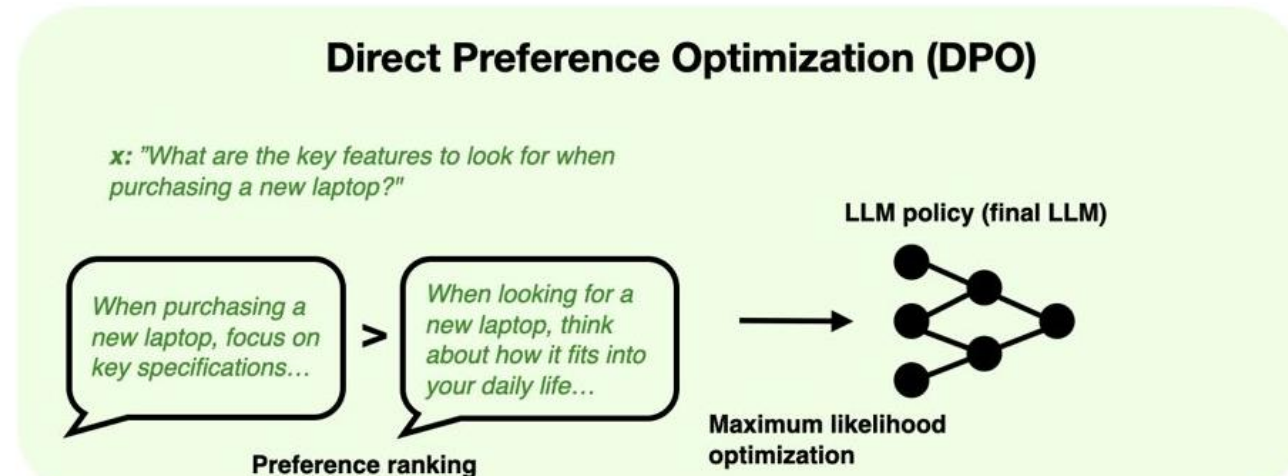
Scammers Used ChatGPT to Unleash a Crypto Botnet on X

To date, NewsGuard's team has identified 1,065 Unreliable AI-Generated News and information websites spanning 16 languages: Arabic, Chinese, Czech, Dutch, English, French, German, Indonesian, Italian, Korean, Portuguese, Russian, Spanish, Tagalog, Thai, and Turkish.

Addressing Factuality Challenges

Addressing Factuality Challenges

- Alignments and Safety
- Retrieval-augmented generation
- Better evaluation



Who is the PM of UK?



Searched 6 sites

The current Prime Minister of the United Kingdom is Sir Keir Starmer. He assumed office on July 5, 2024, after leading the Labour Party to a significant victory in the general election, ending 14 years of Conservative rule. Starmer has a background in law, having served as Director of Public Prosecutions, and was knighted in 2014 for his contributions to criminal justice. His political career began when he was elected as a Member of Parliament in 2015, and he became the Labour Party leader in 2020(GOV.UK)(Wikipedia).

Addressing Factuality Challenges

- Fact checking the outputs of Generative models
- Detecting machine-generated content
- Regulation
- Promoting AI Literacy

Clear user guidelines and robust safety measures can ensure that LLMs benefit society while mitigating the potential risks.

Fact Checking and Generative Models

- Fact checking support
 - transcribe speeches, debates, interviews, online videos and news broadcasts,
 - summarize extensive documents,
 - help to create concise lists of crucial claims,
 - identify sections of documents that repeat a previously fact-checked claim
 - stance detection
 - domain-specific verification on a domain-specific controlled corpus
- Valuable for processing large volumes of online content

Opportunities for Fact Checking

- How humans will interact with AI-aided fact checking
 - fact-checks generated by ChatGPT, even when accurate, did not affect the ability of participants to discern headline accuracy or to share accurate news.
 - they were harmful in some cases:
 - it decreases beliefs in true headlines that it mislabels as false
 - increases beliefs in false headlines that it is unsure about

Verna et al. 2024 "Fact-checking information from large language models can decrease headline discernment"

Conclusion

- Generative models pose both challenges and opportunities for fact-checking.
- They can generate hallucinations and be misused for misinformation.
- However, Generative models can also enhance fact-checking efforts.
- Balancing risks with opportunities is key for future developments in factuality.