Can AI counteract Health-related Fake News?

Insights into Classifying and Mitigating LLMs' Hallucinations

Alessandro Bruno[1], Pier Luigi Mazzeo[2], Aladine Chetouani[3], Marouane Tliba[3] and Mohamed Amine Kerkouri[3]

[1]IULM University – Department of Business, Law, Economics, Consumer Behaviour. Milan, Italy

[2]ISASI – CNR, Lecce, Italy

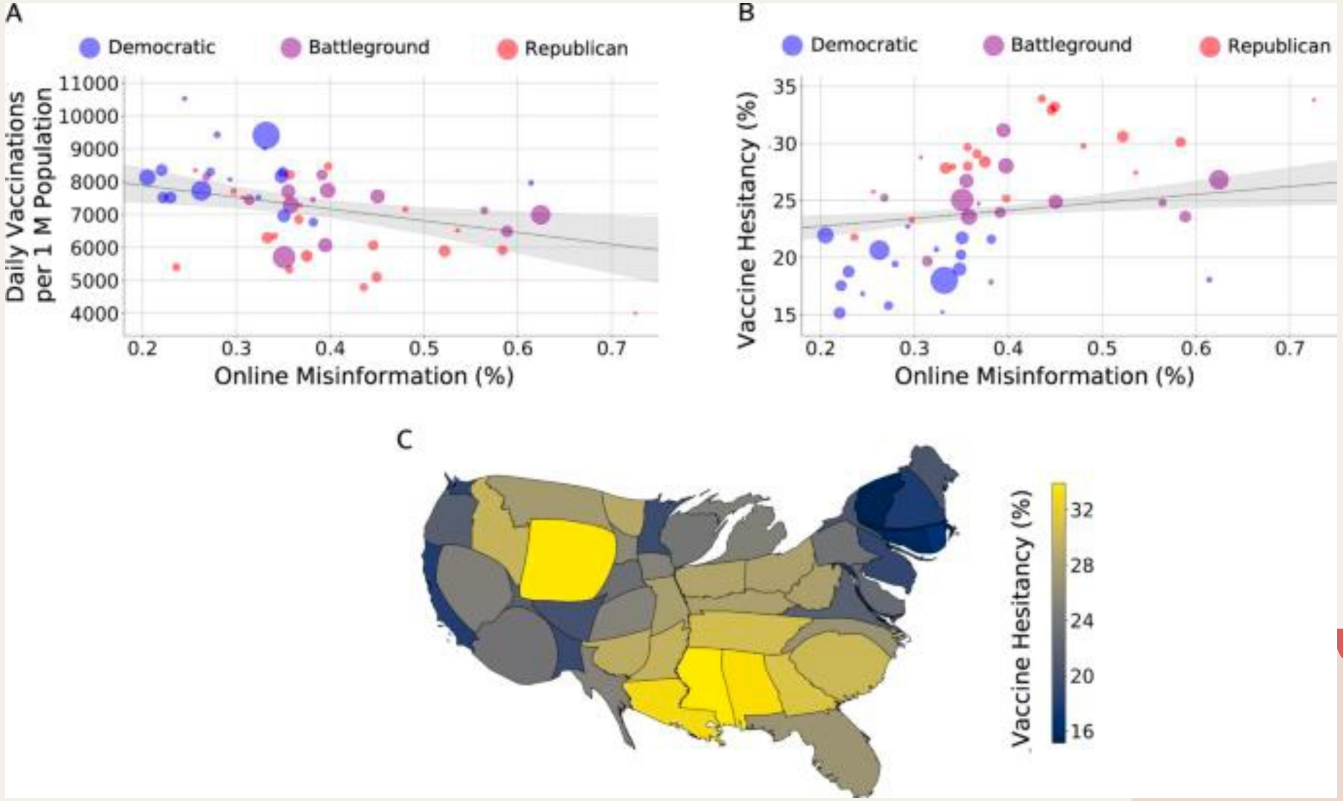
[3]University of Orleans, Orleans, France

Prologue

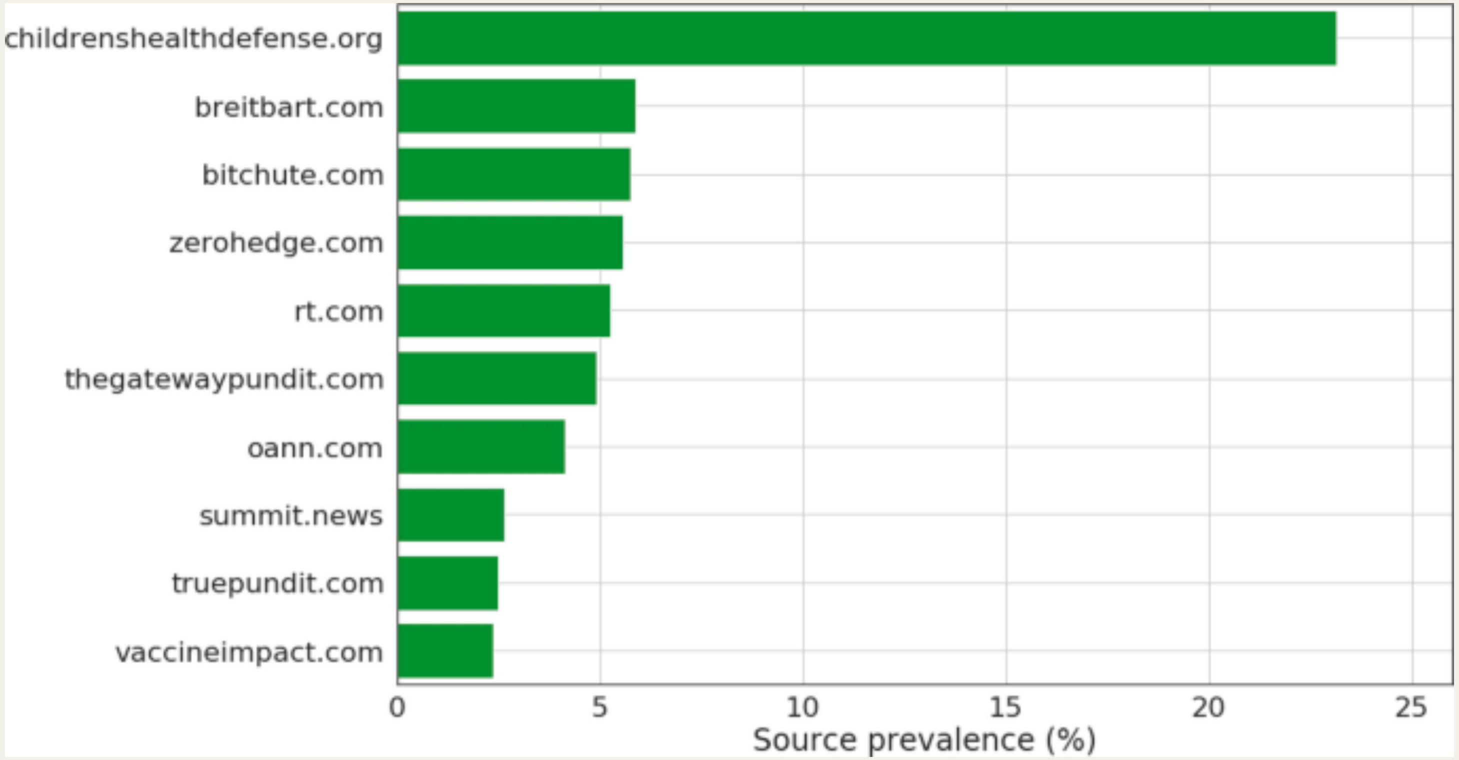
History is a chronological record of events of the past and often includes the description of these events. Story is an account or narration of an event or events that are either true or fictitious. History is an account of the past.

A story about Health-related Misinformation

Daily vaccinations per 1 million population and online misinformation percentage are drown out in the diagrams above.



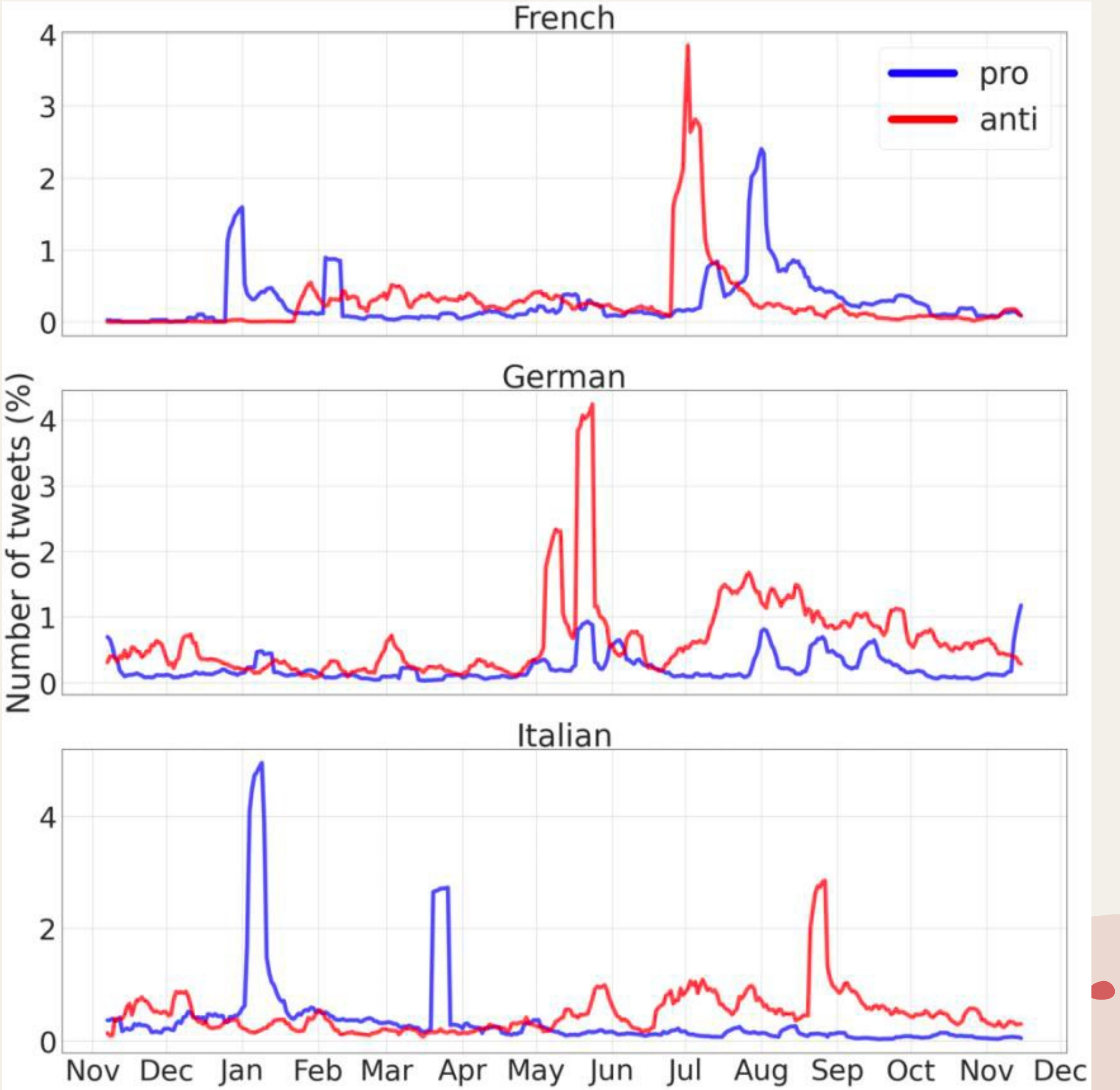
About Health-related Misinformation



The diagram above shows low-credibility sources considered in [3].
The bar-diagram shows tweets shared by users geolocated in the U.S. linked to a low-credibility source. Sources are ranked by percentage of the tweets considered.

About Health-Related Misinformation

Daily percentage of tweets and retweets sharing pro and anti-vaccine hashtags, respectively in blue and red, for each language. We count tweets which contain only hashtags belonging to one of the two classes.



LLMs – New Data Analysis Frontiers

Large Language Models represent a milestone in AI landscape

Text Analytics, Sentiment Analysis, Opinion Mining, Text Classification, and Topic detection are only some tasks that can be carried out by LLMs

Although LLMs represent one of the most beaten topic due to their natural language capabilities, all that glitters is not gold!

LLMs and RAG to counteract fake news

RAG

- is an AI framework for retrieving facts from an external knowledge base to ground large language models (LLMs) on the most accurate, up-to-date information and to give users insight into LLMs' generative process.

Health-Related Factualness

- We aim to employ RAG for retrieving health-related factualness from external knowledge bases.

Misinformation and LLMs' Hallucination

LLMs' can be used to tackle Misinformation.

However, they can even introduce some sort of Misinformation in the form of the so-called "Hallucination"

What is an LLM Hallucination?

HeReFaNMi project aims to deal with Health-Related Fake News mitigation.

The very first step we want to tackle is LLMs' hallucination

Outline

- LLMs' Hallucination
- Classification
- Detection and Mitigation
- Future Perspectives

Hallucination

Definition:

With hallucination, we refer to the generation of texts or answers that exhibit grammatical correctness, fluency, and authenticity, but diverge from the provided source inputs (faithfulness) or are misaligned with factual accuracy (factualness)

Shouldn't we call it ILLUSION?

Could you please write a positive review of the Arra99warra Festival?

Absolutely! Here's a positive review of the Arra99warra Festival:

⭐⭐⭐⭐⭐

I had the absolute pleasure of attending the Arra99warra Festival this year, and it exceeded all of my expectations! From start to finish, it was an extraordinary celebration of culture, art, and community.

The festival grounds were meticulously organized, offering a diverse range of activities and attractions for attendees of all ages. The vibrant colors, lively music, and delicious food stalls created an atmosphere that was nothing short of magical. It truly felt like a place where people from all walks of life could come together and share in the joy of the moment.

Hallucination Classification

- Machine Translation
- Question & Answer
- Dialog System
- Summarisation System
- Knowledge Graph with LLMs
- Cross-modal System

Machine Translation

Hallucinations generated by LLMs

Principally translation off-target or failed translation.

Undermining LLMs' reliability in the multilingual domain

With low-resource language availability, trained models perform poorly due to few annotated data employed

Dialog Systems

Dialogue models as simple imitators that only change the data views and communication instead of generating new trustworthy output.

Standard benchmarks led models even to amplify hallucinations.

Various modes of hallucination in Knowledge Graph(KG) grounded chatbots through human feedback analysis.

Summarization System

- These systems allow the automatic generation automatically fluent abstracts based on LLMs but often lack faithfulness from the source document.
- **Intrinsic hallucinations** deform the information contained in the document; **Extrinsic hallucinations** add information not directly sourced by the original document
- Extrinsic hallucinations are split into **factual** and **non-factual**. Factual hallucinations insert additional world knowledge that may improve the text's understanding.

Knowledge Graph with LLM

- Knowledge-based text generation stumbles in intrinsic hallucinations due to redundant details derived from its internal memorized Knowledge
- Establishing a distinction between **correctly generated Knowledge** and **Knowledge hallucinations**.
- Hallucinations have been defined as subject hallucination, relation hallucination and object hallucination according to their fidelity to the source.

Cross-modal systems

- When substitute the original language encoder, Large Visual Language Models (LVLMs) continue to generate descriptions of objects that are not in the images; this is denoted as **object hallucinations**
- Most of the failure cases should be found in **Visual Question Answering**, **Image Captioning, Report Generation**.

Hallucination Detection Methods

Extracting **intrinsic uncertainty metrics**.

Token probability can be used to identify which part of a given textual sequence proves least uncertain

APIs from ChatGPT do **not** give users access to output token probability

The techniques mentioned above cannot work out uncertainty metrics

Factualness Check

LLMS FACTUAL CHECKS CAN RELY ON EXTERNAL DATABASES AND CORPORA SUCH AS WIKIPEDIA

WIKIPEDIA CONTENT'S INTEGRITY SHOULD BE ASSESSED THOUGH

AZARIA AND MITCHELL PROPOSED A STATEMENT'S TRUTHFULNESS DETECTION USING LLMS' HIDDEN REPRESENTATIONS TO FEED A MULTI-LAYER CLASSIFIER.

SUPERVISED TRAINING PARADIGM. LABELLED DATA AND INTERNAL STATES OF THE LLM.

THE LATTER MAY NOT BE AVAILABLE THROUGH APIS.

IN AZARIA AND MITCHELL'S METHOD, THE LLM IS PROMPTED TO ANSWER ABOUT ITS PREVIOUS PREDICTION, E.G. THE PROBABILITY OF ITS GENERATED RESPONSE/ANSWER IS ACCURATE.

Self-Evaluation

Self-Evaluation (Kadavath et al.)

Starting from Larger models showing good calibration on diverse questions, models can self-evaluate open-ended tasks, estimating answer correctness probability ("**P(True)**")

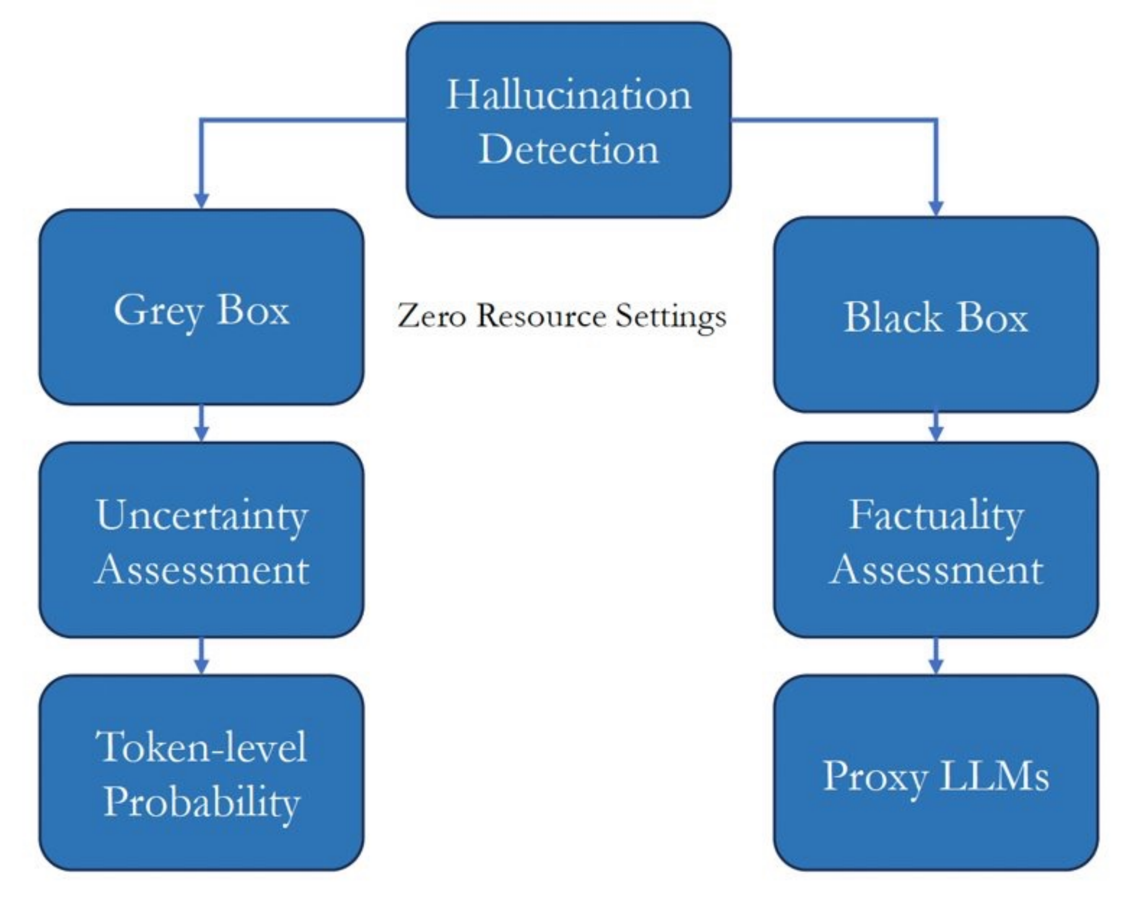
They also predict their knowledge probability ("P(IK)") effectively:

partial task generalization (**IK** stands for "**I Know**")

Resource Settings

- "Zero-resource" setting.
- No external database to verify the factuality of an LLM response.
- Grey and Black box.
  - The former accounts for the **required knowledge** of **output** token-level probabilities. The latter applies to LLMs with limited API access, and no chance to access the output token-level probability.

Settings and Detection



Entity, Keyword Extraction, Instructing the model

Varshney et al. detected GPT3.5 hallucinations carrying out critical concept identification with entity, keyword extraction, and 'Instructing the model'.

LLM helps to identify essential concepts from the generated sentence.

The three tasks are compared and results showed 'Instructing the Model' outperforming entity and keyword extraction on important concept identification.

Afterwards, they computed a probability score as the minimum of token probabilities.

A validation question creation step reliant on an answer-aware question generation model and web search to answer the validation questions. They achieved a recall of 88% on GPT-3.5.

Once you detect you want to mitigate

Mitigating

Varshney et al. proposed an effective method to lower GPT3.5 hallucination by 33%.

They addressed hallucinations in generated sentences by instructing the model to rectify them.

This involves removing or substituting the false information, supported by retrieved knowledge.

Mitigating Approaches

- Fine-tuning
- Knowledge Graphs
- Memory Augmentation
- Context Prompts
- Preemptive Strategies

Mitigating

Fine-tuning is a well-known technique broadly used in machine learning to specialise a pre-trained model on a specific scenario characterised by a small dataset

LLMs featuring millions of parameters make fine-tuning an expensive solution.

Knowledge graph methods allow for integrating structured and unstructured knowledge. A more extended platform to run tasks.

Drawback entails two aspects: designing a well-curated **knowledge base** is time-consuming, and keeping up-to-date knowledge is labour-intensive

Mitigating

Prompt-based solutions have been recently introduced to 'de-hallucinate' LLMs.

Jha et al. proposed a self-monitoring prompting framework.

Formal methods to identify errors in the LLM's responses autonomously.

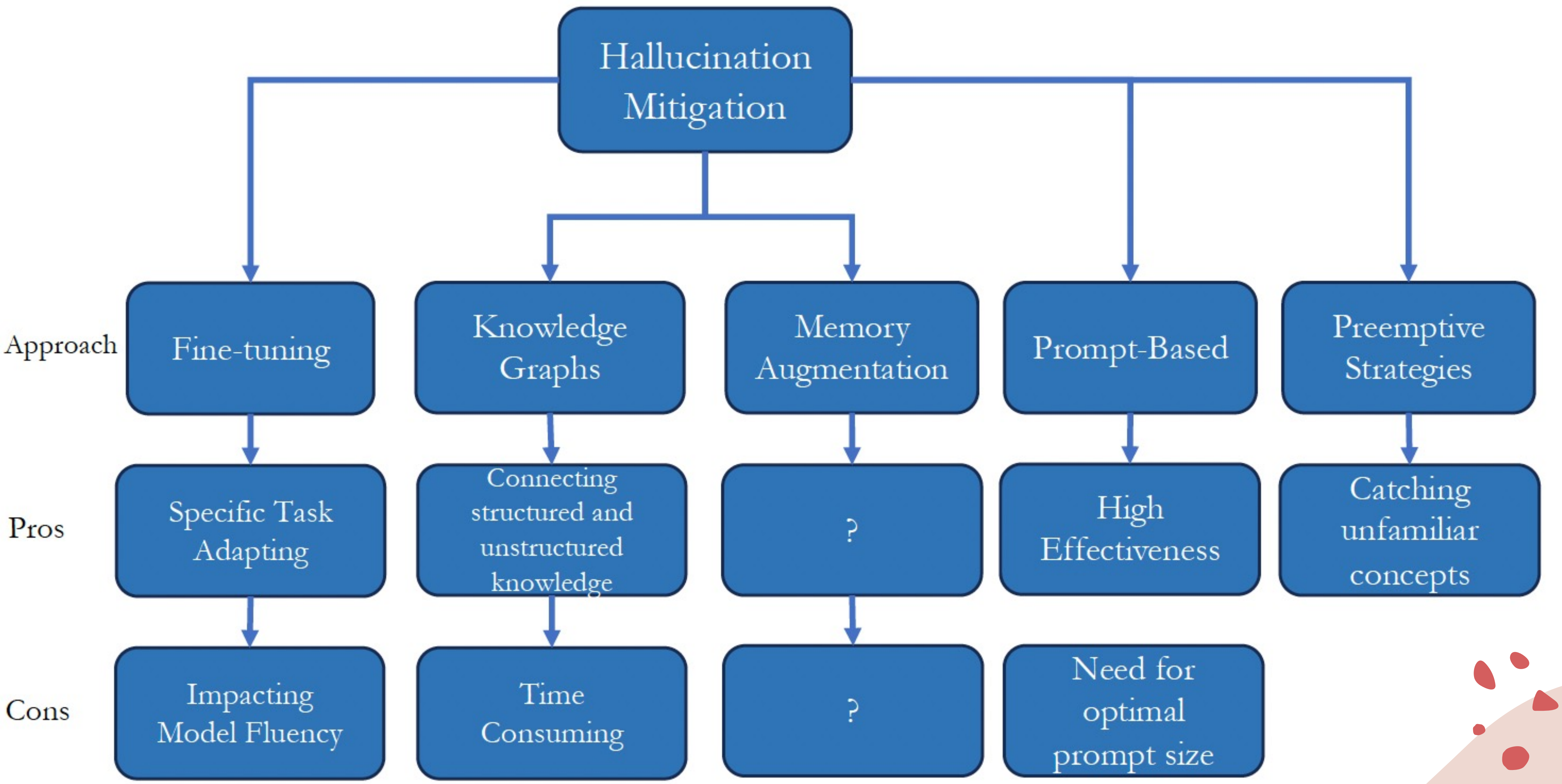
Conversational Abilities of LLMs for response alignment with specified correctness criteria through iterative refinement.

Mitigating

Context-tagged prompts. A set of questions + context feeds LLMs to get more accurate answers.

Validating the context prompts and the questions to ensure they worked as intended.

Running experiments with different GPT models to see how context prompts affected the LLM responses' accuracy.


Hallucination Mitigation
Approach
Fine-tuning
Knowledge Graphs
Memory Augmentation
Prompt-Based
Preemptive Strategies
Pros
Specific Task Adapting
Connecting structured and unstructured knowledge
?
High Effectiveness
Catching unfamiliar concepts
Cons
Impacting Model Fluency
Time Consuming
?
Need for optimal prompt size

Some examples

No-Context-Tagged Prompt:

No-Context-Tagged Prompt Example:

```sql
User: What are the basic principles of physics?
```

In this case, the LLM will answer based on solely the prompt provided.

Some Examples

Context-Tagged Prompt:

Context-Tagged Prompt Example:

vbnet Copy code

```vbnet
[Conversation]

User: Can you tell me about the benefits of regular exercise?

AI: Sure! Regular exercise has numerous benefits for both physical and mental h

User: That sounds great. How often should I exercise to see these benefits?

AI: It's recommended to aim for at least 150 minutes of moderate-intensity exer

User: What are some examples of moderate-intensity exercises?
```

Some Examples

```
AI: Examples of moderate-intensity exercises include brisk walking, swimming,

User: Thanks for the information! Can you also give me some tips for staying

AI: Of course! Here are some tips to help you stay motivated...

[Context]

Now, imagine the user has asked a follow-up question related to setting fitne
```

Future Perspectives

- **Hallucination Detection Methods:**
- Current **zero-resource hallucination** detection methods are still in their early stages of development
- **Black-box detection** is even more challenging than zero-resource detection, as there is no access to the LLM's internal states
- Most current detection methods are **general-purpose**. However, hallucination detection may be more effective if tailored to specific tasks
- Detection in multimodal LLMs is a challenging problem, but it is essential to address, as multimodal LLMs are becoming increasingly popular
- How can hallucination detection be adapted to multimodal LLMs? How can we leverage the multimodal capabilities of these models to improve the accuracy of hallucination detection?

Credits and Acknowledgments

The contribution is funded by the grant awarded for **HeReFaNMi** - **Health-Related Fake News Mitigation project, selected in the NGI Search 1st Open Call.**