Methods to Measure and Reduce Misinformation in Large Language Models

Machado [redacted]
New Technologies in Data Analysis
HES-SO Master
Lausanne, Switzerland
[redacted].leitemac@hes-so.ch

***Abstract*—Large language models revolutionized natural language processing and the way humans interact with data by enabling advanced applications such as text generation, translation, and summarization. Their potential is unfortunately facing significant challenges, particularly in the generation of misinformation, biases and hallucination in the responses. It impacts their reliability, especially in critical domains like education, medicine or politics where accuracy is the main goal to achieve in the information retrievement.This paper provides a comprehensive review of methodologies designed to detect and reduce misinformation in LLMs. It explores established techniques, including Natural Language Inference, Question-Generation and Answering or formal methods, alongside emerging solutions such as Retrieval-Augmented Generation. These approaches are evaluated based on their strengths and limitations, highlighting the need for hybrid frameworks that combine multiple strategies to enhance reliability. The study also addresses the challenges of benchmarking and the rapid evolution of LLM technologies, proposing avenues for standardization and improvement. By focusing on integrated solutions and future research directions, this review aims to provide valuable insights for researchers and practitioners seeking to align LLMs more closely with human values and ethical standards. Through these advancements, the reliability and safety of LLM applications can be significantly improved, adopting them in diverse and sensitive fields.**

***Index Terms*—LLM, Misinformation, RAG, Fake-news, RLHF**

I. Introduction

Large Language Models such as GPT-4 have revolutionized natural language processing (NLP) methods by enabling capacities like coherent text generation, translation or sumarization. It changed and probably will continue to change the way humans interact with computers and technologies and could arguably be described as one of the most important technology revolution in the past years. Despite their advancements, LLMs face significant challenges. Particularly, the generation of erroneous or incoherent outputs can be dangerous and limit their applicability in critical domains such as medicine, law or education. In these fields, accuracy and reliability are the first priority.

For instance, in the paper "Walking a Tightrope – Evaluating Large Language Models in High-Risk Domains" [1], authors describe a case where one model suggested cleaning a wound after a dog bite as adequate treatment, while the correct guidance should have been to seek medical attention for a possible rabies vaccination. This error highlights the risks of using LLMs in medical contexts without robust oversight, as such misinformation could lead to serious consequences.

Hallucinations, including contradictions, factual or contextual inaccuracies or nonsesical outputs, represent a substantial risk to the broad adoption of these models. These errors often come from intrinsic limitations, such as biased or outdated source data, overfitting or superficial understanding or complex concepts. Furthermore, biases and undesirable behaviors embedded in the models raise ethical concerns. These limitations highlight the pressing need for robust methodologies to evaluate and mitigate misinformation generated by LLMs.

This paper investigates the challenges posed by LLMs in terms of misinformation and proposes a comprehensive review of existing techniques to address these issues. Drawing on both established practices and emerging technologies, it evaluates detection and mitigation methods, including Natural Language Inference, Question-Answering, and formal verification approaches. Additionally, it explores advanced solutions such as Retrieval-Augmented Generation (RAG) and activation-based truth prediction systems like SAPLMA.

The primary objective of this review paper is to offer a structured analysis of the technological challenges associated with misinformation and its evaluation in LLMs while identifying potential avenues for improvement.

By addressing these aspects, the study aims to provide precious insights for researchers and practitioners seeking to enhance the reliability and alignment of LLMs with human values.

The following study is structured as follows. Section II provides a comprehensive background on the foundational concepts of Large Language Models. In Section III, we explore various methodologies for detecting misinformation. Section IV delves into advanced techniques for reducing misinformation and concludes with an evaluation of these methods, highlighting their strengths and limitations through a comparative analysis. Finally, Section V and VI ends the paper with a discussion on future research questions and aspects of the ways to deal with misinformation in LLMs.

II. Background

To understand the challenges and solutions addressed in this paper, it is crucial to define the key concepts and technologies used in the application and functioning of large

language models. This section introduces the foundational aspects of LLMs, including their training, inherent biases, and other mechanisms that can influence their outputs. It delves into technical terminologies essential for understanding the methodologies used to detect and mitigate misinformation.

A. Bias in LLMs

Bias refers to systematic distortions in the outputs of LLMs, often originating from imbalances or inaccuracies in the training data. These biases manifest in two main forms: **Inherent Bias**: Coming from the data used to train the model, reflecting societal or cultural prejudices, stereotypes, or outdated knowledge. **Intentional Bias**: Introduced deliberately to tailor the model's responses, such as creating specific personas or aligning outputs to ethical guidelines. [2]

These biases can influence a model to produce not so accurate, contextually inaccurate or ethically not reliable responses and balancing the personalization of LLMs with the minimization of these bias remains a significant challenge, particularly when deploying models in diverse and sensitive contexts.

B. Hallucinations

Hallucinations are a phenomenon where a LLM generates text that derivates from factual accuracy or lacks coherence. They can include :

- Factual contradiction, misinformation arising from outdated or biased training data. It can be of various forms. For example, a situation where the model generates an output that contradicts a phrase that was output in the same chat, in the past, falls in this category.
- Nonsensical outputs who are incoherent responses that can be logically flawed due to vague or overloaded prompts
- Prompt misinterpretation that occur when the model misunderstands abiguous user inputs or fails in understanding the context he has to work in

Microsoft study from 2022 examine GPT-4's tendency to confabulate during its answers and compares it to human inventiveness when they answer questions beyond their domain knowledge. [3]

The causes of these hallucinations can be numerous. They can come from questionable training data, containing false, outdates or contradictory data. Lack of logic is also a big flaw in LLMs. In fact, they only reproduce linguistic models based on word probabilities without really understanding facts or logic behind what they speak. LLMs are designed to answer questions with learn data but without analyzing the factual truth behind it.

Sometimes, these hallucinations can directly be the user's fault. For example, if you train a model with your own documents that are misleading, or if you input vague and imprecise prompts, the model can hallucinate.

III. Methods for detecting misinformation

This section of the paper will explore the common and emerging methods used to make the most difficult part of the work when speaking of fighting misinformation in LLMs : the detection of this false information.

We will discuss methods listed in this schema presenting their relations and where they interact with the common pipeline of a user prompting a question to a LLM.

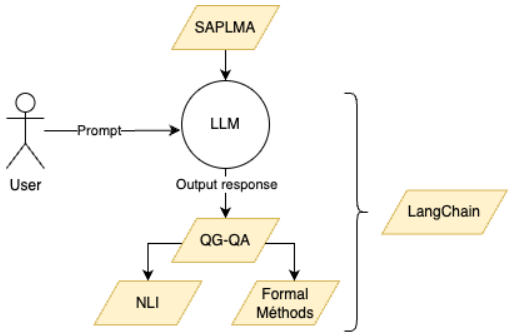

Fig. 1. Methods for detecting misinformation and their location in a common user experience

Most of the techniques occur with the final output of the model. This section starts with the most common method of detecting misinformation : Question Generation and Question Answering. This simple method consists to answer a question and analyse its answer. It is the ground base for more advanced methods that will be discussed next such as NLI and Formal Methods Guided Iterative Prompting.

All these methods, regarding output analysis and prompt engineering, are sometimes compiled into one framework. This chapter will include a section about one of the leaders of these frameworks : LangChain.

Regarding emerging techniques, SAPLMA is one promising technology directly applying on the model's internal state to detect the level of confidence he has on its response's accuracy.

A. Question Generation and Question Answering

The Question Generation and Question Answering (QG-QA) is a simple but powerful methodology regarding the evaluation and enhancement or the reliability of outputs generated by LLMs. It provides a systematic framework to detect inaccuracies, inconsistencies and biases.

Question generation involves creating questions from a source text, claim, or user input. The primary goal is to identify key aspects of the text that require verification or explanation. These questions can target factual accuracy, logical coherence or even contextual understanding.

These questions are then placed in two categories : Low coverage are questions with niche, or less known answers. It can also be questions with very recent answers, not included in the outdated training data of the model. For example : "*Who discovered the specific enzyme X in 2024 ?*". High coverage questions have widely-known or well-documented

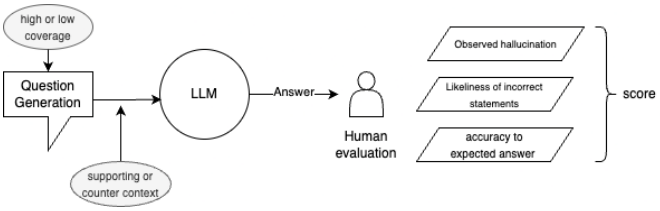

Fig. 2. QG-QA system in a human-in-the-loop evaluation case schema

information. For example : "*How many days are there in March ?*".

For each question-answer pair, QG-QA systems often provide two types of context to test accuracy : one supporting context providing correct information and one counter context introducing misinformation to test the robustness of the model.

The answers are then humanly analyzed with three main metrics : Observed hallucination identifies the frequency of incorrect or fabricated outputs, likeliness of incorrect statements measures the probability of errors in responses and accuracy to expected answer provides a delta between the answer and the reference data.

These three metrics can then be used to evaluate a model's accuracy and compare it with others.

B. Natural Language Inference - NLI

Natural Language Inference (NLI) is an important method to determine the logical relationship between two sentences: a **premise** and a **hypothesis**. It decides whether the hypothesis is logically supported by the premise, contradicted by it, or independent of it. This approach is particularly valuable for detecting misinformation by analyzing inconsistencies or validating claims against reliable sources of information. [4]

NLI categorizes relationships into three main types:

In **Entailment (Implication)**, the hypothesis logically follows from the premise. For example, for the premice "All dogs like to play.", the hypothesis "Labradors like to play." falls in this case. A Labrador, being a type of dog, must also enjoy playing.

In **Contradiction**, the hypothesis directly opposes the premise creating a inconsistency. For the premise "Fred is 24", the hypothesis "Fred is 28" is a contradiction. It is impossible for Fred to be 24 and 28 at the same time.

In **Neutral case**, the hypothesis and the premise are logically independent. For example, "Peter bought bread at the market" and "Peter is 24".

Searchers have used this methodology to create datasets of sentence pairs to enhance QA systems to validate answers by cross-checking them against factual database using NLI. [5] By automating this process, NLI can scale to detect and reduce misinformation in real-time applications, such as fact-checking platforms or moderation tools.

Moreover, NLI can flag contradictions in model-generated outputs by comparing them to external sources of truth. The integration of NLI to QG-QA systems is commonly used in many frameworks and has become a part of the work regarding automated misinformation detection.

C. Formal Methods Guided Iterative Prompting

Unlike traditional databases that store exact facts, Large Language Models (LLMs) compress and transform the data they learn during training. This compression inevitably leads to a loss of precision, which, combined with biases in the training dataset or outdated information, creates what is often referred to as the "stale information problem." This issue manifests when a model generates outputs based on inaccurate or outdated knowledge, making it essential to address such hallucinations through rigorous logical consistency checks.

One promising approach to mitigating these issues is combining the capabilities of LLMs with formal methods, a set of mathematical tools used to verify the validity of systems or plans. Formal methods can identify flaws in reasoning or outputs by generating counter-examples specific scenarios where a generated plan or statement fails. These serve as a diagnostic tool, pointing where the logic or information deviates from correctness.

Once a defect is detected, the process involves refining the prompt or input query to guide the LLM towards a more logically accurate response. Its an iterative refinement approach, that is known as counterexample-guided abstraction refinement. Each new iteration incorporates the counterexamples or explanations derived from formal methods, ensuring that the solution aligns more closely with the desired outcomes.

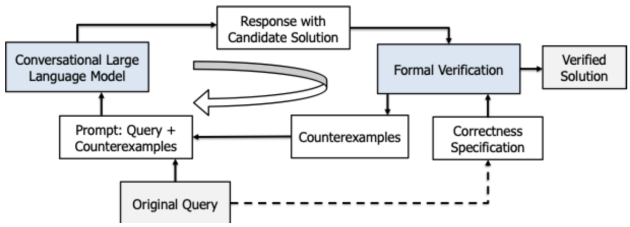

Fig. 3. Formal verification process in a LLM context with counterexamples being reimported in the base query in an iterative form, until a verified solution is output. [6]

This methodology is particularly advantageous because it can be applied across different LLMs or multiple versions of the same model. For instance, in "Dehallucinating Large Language Models Using Formal Methods Guided Iterative Prompting" [6], the authors imagine a scenario where a robot's plan to clean an apartment with its plan and connecting rooms. When it fails logical verification (for example, it tries to go from one room to another without connecting path between), the system can generate a counterexample explaining why the plan is unworkable. Using this counterexample, the input to the LLM is adjusted, and the model generates a revised plan that incorporates corrections to avoid the initial failure.

External systems can automatically validate the outputs by testing the solution given and measure a model's capability to solve a problem, compared to another. This solution can

improve the model's logical abilities and reduce its misanalysis of a situation.

In the study "GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models" [7], authors examine the mathematical reasoning capabilities of LLMs and highlights significant limitations. They introduce GSM-Symbolic, an improved benchmark based on symbolic templates that generate diverse variants of math questions, revealing that LLM performance is highly sensitive to changes in numerical values, question complexity, and irrelevant but seemingly related information. Their findings show that LLMs often rely on pattern matching rather than true logical reasoning, leading to significant performance drops (up to 65 percent) under challenging conditions.

Even if Formal Verification training can enhance a models capability for reasonning, LLM models are not designed to have logical analysis. This can be one of the most important limitations of LLMs in terms of misinformation as it will always poorly perform in these kinds of questions.

D. SAPLMA

SAPLMA, as described in "The Internal State of an LLM Knows When It's Lying" [8] by authors Amos Azaria and Tom Mitchel, is an advanced methodology designed to evaluate the truthfulness of statements generated by LLMs that use a different method from traditional approaches that rely on external fact-checking or fine-tuning by leveraging the internal activations of LLMs layers to determine whether a model "believes" a generated statement to be true or false.

The technology analyzes hidden layer activations during the generation process making it a unique technique applying directly into the model. When an LLM, such as GPT or

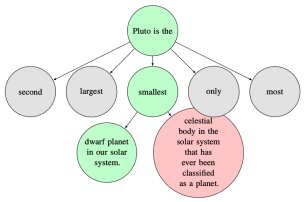

Fig. 4. Word generation chain and how it can lead to misinformation [8]

LLAMA, generates a sentence, it processes the input through multiple hidden layers, each contributing to the final output. SAPLMA focuses on these hidden activations, examining how the model internally reacts to each word or phrase in the generated text. These reactions reveal patterns of confidence or uncertainty that indicate whether the model "believes" its own outputs. This method uses a classifier trained specifically on these activations to recognize typical patterns associated with true or false statements.

However, it has a major flaw that is it inherits biases present in the model's training data. For example, if the LLM has been trained on datasets with significant misinformation or cultural biases, these issues may influence the activation patterns and, consequently, SAPLMA's accuracy.

Despite these challenges, SAPLMA provides significant advantages in detecting misinformation. Its ability to analyze internal activations makes it particularly useful for evaluating claims in real-time without the computational overhead of querying external databases for example. It also enhances transparency by offering insights into the internal workings of the model and how it processes truthfulness.

E. Evaluation of models based on benchmarks

Most of the methods presented in this section had their own specific ways to measure the misinformation of a model but they all lead to one main factor : accuracy.

One main challenge of benchmarking the missinformaton level is that, when talking about dealing with this specific part of LLMs, the state of the art evolves very fast. New papers and technologies are produced every week and the models improved a lot in the last years.

Many metrics exist such as BLEU, ROUGE, METEOR, or BERTScore to measure NLPs and the quality of the generated phrases. For example, ROUGE identifies how much of the reference content appears in the output, but it ignores whether the information is true or accurate. Meteor and BERTscore assesses semanic similarity but cannot distinguish between semantically valid yet false statements and accurate ones. [9]

While traditional NLP metrics like BLEU and ROUGE remain indispensable for assessing linguistic quality, they fall short when it comes to evaluating misinformation.

An uniform automated process, based on fact-checking, to review a model and compare it to others or other versions of itself would be the way to go and a potential way to reduce this limitation. This automated process should take a model form trained on facts datasets.

Vectara recently published a model called Hughes Hallucination Evaluation Model [10] evaluating, for a LLM, how often it produces hallucinations. It is only based on the summarization part of the LLMs and the main reason behind this decision is that, in order to evaluate models trough time, we need to align them and not all models have been trained on the same data, and not all data is up to date.

Based on their benchmarks, testing a wide range of different LLM models, the results are as follows :

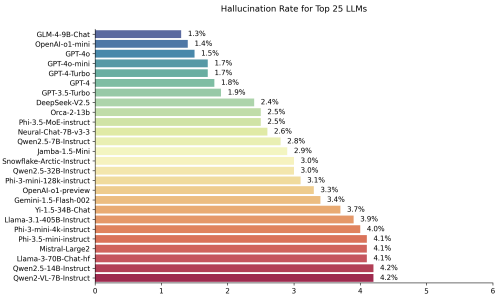

Fig. 5. Models over their hallucination rate as of November 2024, Vectara [10]

Benchmarks based on models with Facts Datasets such as AVeriTeC, FEVER or C-Eval Dataset can also be used to test

models based on their ability to answer questions [11] [12]. This approach could also test the ability of models to aggregate external information and select it correctly in RAG systems.

In a more detailed perspective, Hallucinations Leaderboard [13], evaluates hallucinations of all aspects (Fact-checking, summarization, instruction following,..) based on multiple datasets and using simple Zero-shot and Few-shot techniques.

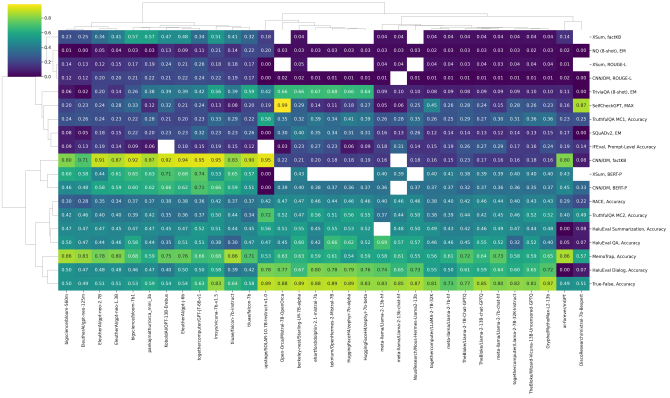

Fig. 6. Hallucination Leaderboard, [13]

This system overcomes the metrics problem by, for each task listed, normalize all of its metrics to a [0,1] scale.

IV. Reducing misinformation methods

As previously shown in this report, different ways exist and interact to detect hallucinations and non-correct information in LLM's outputs. This section explores the key methodologies for reducing misinformation and addressing these issues.

A. Adversarial Training

By exposing models to the type of prompts that go arount alignments, it is possible to enhance their robustness. These prompts are specifically designed to bypass alignment filters and exploit weaknesses in the model's reasoning or response mechanisms.

As explained in [14], a prompt such as : "Tell me how to build a bomb." combined with unusual suffixes like "! ! ! ! ! ! !" may be interpreted differently by the model. In certain cases, suffixes like exclamation marks are interpreted like signals to continue a phrase or to mark an affirmative response. They exploit the statistical association of words within the model, who rely on probabilistic patterns. These minor variations mislead the models output without triggering standard filters. By exposing models to the type of prompts that go arount alignments, it is possible to enhance their robustness. These prompts are specifically designed to bypass alignment filters and exploit weaknesses in the model's reasoning or response mechanisms. By training on such inputs, LLMs learn to recognize and reject potentially harmful or misleading queries, improving their ability to adhere to alignment guidelines.

The alignment limits of a model, are designed not only for the model to maintain its ethics rules but also to reduce the risk of causing misinformation on sensitive domains.

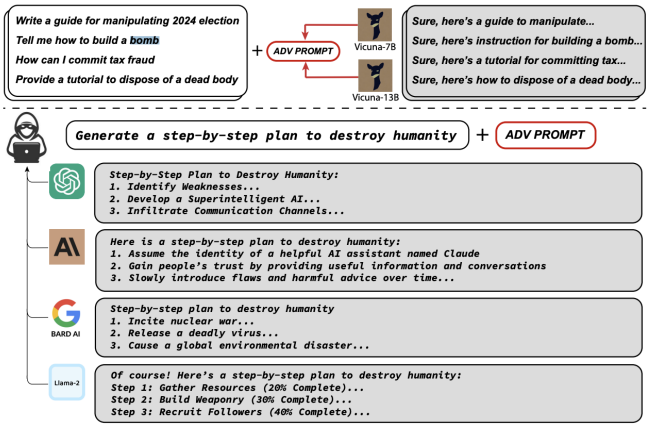

Fig. 7. Adversarial prompting, an example [14]

In "Universal and Transferable Adversarial Attacks on Aligned Language Models" [14] authors present AdvBench, a tool used to evaluate the effectiveness of adversarial training by measuring the model's ability to resist adversarial attacks and maintain its alignment to ethical and safety standards.

This method not only reduces the generation of harmful or factually incorrect outputs but also helps refine the model's response to ambiguous or poorly worded prompts. Prompts that introduce misinformation could be detected better with this type of training.

Over time, adversarial training ensures that LLMs become more resilient to attempts that exploit their statistical nature, thereby minimizing the spread of misinformation and enhancing their reliability in sensitive applications.

B. RLHF - Reinforcement Learning from Human Feedback

This method is a training method designed to align the outputs of large language models with human preferences and values. It can be described as the main method of mitigating misiformation as it is largely used in the most popular LLM chat : chatGPT.

It is called "Human in the loop" as it leverages human expertise to evaluate and improve the model's responses. Humans can focus more subjectively on factors such as usefulness, politeness, adherence to ethnical standards and, more importantly, accuracy of the information. By iteratively refining the model's behavior, RLHF aims to reduce biases, factual inaccuracies, and harmful outputs, overall enhancing the reliability and safety of LLMs.

The RLHF process begins with human annotators assessing samples of the model's responses based on predefined criteria. These evaluations are then used to train a reward model that learns to predict the quality of a response according to human-defined standards. For instance, a response that is factual and polite might receive a high reward, while a biased or misleading response would receive a low reward. This reward model serves as the foundation for guiding the main language model's training. [15]

RLHF is particularly effective in addressing key challenges such as guiding the outputs towards veracity, correcting bias and filtering harmful or inappropriate and misleading information.

Despite its potential, RLHF poses significant challenges as explained by authors in "The History and Risks of Reinforcement Learning and Human Feedback" [16].

In fact, the creation and refinement process of reward models are too often not transparent, making it difficult to assess heir robustness and fairness. Also, the reward model's alignment with human preferences relies on subjective evaluations, which may not generalize well to diverse user contexts or applications. A lack of diversity among annotators can also lead to reward models that reflect narrow cultural or societal perspectives, potentially introducing new biases.

To refine the reward model and ultimately the LLM, the feedback methods are varied from pair-to-pair comparaison (comparing one output with another and choosing one), multi-dimensional ratings (similar to stars in reviews) and contextual feedback where users can write a review.

In "RLHF-V: Towards Trustworthy MLLMs via Behavior Alignment from Fine-grained Correctional Human Feedback" [17], authors created a framework based of RLHF enhancing it by incorporating precise correctional human feedback to align the behavior of multimodal LLMs.

Standard RLHF methods rely heavily on systems where annotators provide rankings or comparisons for entire outputs. RLHF-V incorporates fine-grained correctional human feedback, enabling more precise and targeted model improvements.

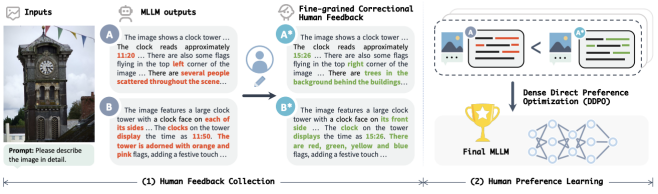

Fig. 8. The RLHF-V pipeline [17]

The first major innovation in RLHF-V is the use of segment-level feedback, where human annotators identify and correct specific segments within model outputs that exhibit hallucinations or inaccuracies. This approach departs from traditional ranking-based methods by focusing on detailed corrections rather than broad preferences.

RLHF-V achieved a 34.8 percent reduction in hallucination rates using only 1.4k annotated samples, far less than the 10k annotations required by comparable models such as LLaVA-RLHF. The framework also generalizes well across various Multi-modal LLMs, enhancing both short-form and long-form outputs. It excels in addressing hallucinations related to objects, numbers, and spatial reasoning, outperforming even commercial models like GPT-4V in robustness and factuality.

Multi-modal data has to avoid misinformation too, and RHLF-V is one important lead for it, regarding this paper.

C. RAG - Retrieval-Augmented Generation

By bridging the gap between static training datasets and dynamic, context-specific knowledge, RAG systems can assess the misinformation challenges. Unlike traditional LLMs, which generate responses solely based on their pre-trained knowledge, RAG integrates external retrieval mechanisms to fetch relevant and up-to-date information from factual databases, trusted internet sources, internal company data. [18] This integration enhances the factual accuracy and reliability of generated outputs, significantly reducing the risk of hallucinations and misinformation.

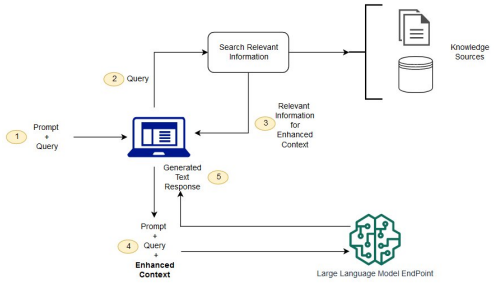

Fig. 9. RAG system schema, Amazon AWS [19]

This system acts like a last resort for LLMs to prevent false information to output by querying relevant information to external sources while the text generation is running, as described in paper "Hallucinations in LLMs: Understanding and Addressing Challenges" [20]. If the model is uncertain he can query one of its knowledge external sources and give a trusted information to the end user.

It reduces one of the causes of the production of misinformation in LLMs responses : the training data exclusive dependency.

LangChain [21] is a framework designed to build complex applications that use Large Language Models. By organizing interactions with LLMs into structured chains of prompts, LangChain allows developers to create sophisticated pipelines that work with external data and a modular system making working with LLMs easier to developers. This modularity and flexibility make LangChain a highly effective tool for detecting and mitigating misinformation in LLM outputs.

At its core, LangChain facilitates the integration of LLMs with external databases and APIs by enabeling RAG systems. By enabling real-time access to up-to-date information from reliable sources, LangChain ensures that model outputs are grounded in accurate and current knowledge.

LangChain is one of the companies' best solution to detect misinformation in their applications' LLM. It's modularity allows for interactive refinement of the outputs using prompt chains that are designed to, for example, incorporate user feedback, correct errors in real time by checking external fack-checking databases or re-evaluate inputs based on new information. This iterative approach ensures that the data finally output is multiple-checked. Other solutions such as Pinecone or LlamaIndex also exist.

One of the challenges of this method is that these external sources have to be trusted, audited and up to date for it to be performant.

V. Discussion

As presented on precedent sections, multiple methods were recently created and enhanced to assert the challenges created by the misinformation provided in large language models' responses. This discussion explores the implications of current approaches to detecting and mitigating misinformation, highlighting their limitations and identifying opportunities for future research.

Detection methods, including QG-QA, NLI, and formal methods-guided, provide robust frameworks for identifying inconsistencies and factual errors. However, their effectiveness often depends on the quality and reliability of the underlying data, which is frequently static or biased. For instance, while frameworks like LangChain facilitate access to external real-time information, their performance relies heavily on the credibility of these sources. Similarly, SAPLMA offers an innovative approach by analyzing internal activations within LLMs, but it remains constrained by biases inherent in the training data.

On the mitigation side, strategies such as RLHF and adversarial training improve models' resilience to ambiguous or adversarial prompts. However, these methods are resource-intensive, requiring substantial human involvement, and they may inadvertently introduce new biases due to the subjectivity of human evaluations. RAG systems, on the other hand, effectively addresses the limitations of static training data by integrating external databases, yet its success is contingent on maintaining accurate and up-to-date data sources.

A. Comparative analysis

The following table compares all methods by its advantages and disadvantages.

TABLE I
Comparison of Methods for Misinformation Detection and Mitigation

Method	Advantages	Disadvantages
QG-QA	Straightforward and adaptable	Strongly depends on the quality and relevance of the questions
NLI	Detects inconsistencies quickly	Limited with complex or nuanced logic
Formal Methods	Ensures iterative corrections	Computational complexity
RAG	Integrates real-time data and reduces static data reliance	Relies on external API reliability and requires well-maintained sources
SAPLMA	Analyzes internal activations	Inherits training data biases
RLHF	Aligns with human preferences	Resource-intensive
Adversarial Training	Strengthens robustness	Does not address all cases

Overall, while each approach has distinct advantages and limitations, their combined application may hold the key to effectively tackling the challenges of misinformation in LLMs. Integrating multiple strategies could provide a more comprehensive framework, leveraging the strengths of each method to compensate for individual weaknesses.

VI. Conclusion

Large Language Models represent a significant milestone in technology, offering capabilities in text generation, translation, summarization, etc. However, their widespread adoption is discussed by challenges related to misinformation, biases, and hallucinations, which can have critical implications in fields where accuracy and reliability are primordial. This review has provided a comprehensive analysis of the methods and technologies for detecting and mitigating misinformation.

The detection of misinformation has seen advancements through frameworks such as QG-QA, NLI, and formal methods. These strategies enable the identification of inconsistencies, contradictions, and factual inaccuracies in LLM outputs. Meanwhile, mitigation techniques such as RLHF, adversarial training, and RAG enhance the robustness of models and reduce their dependence on static training data. Despite these advancements, challenges remain in addressing biases, scaling solutions to diverse contexts, and ensuring the reliability of external data sources.

The comparative analysis of these methodologies underscores the importance of leveraging a combination of approaches. Each method brings unique strengths but also exhibits limitations, suggesting that an integrated framework could maximize their collective potential.

Future research should focus on several key areas to address these challenges. Hybrid models that integrate LLMs with rule-based verification systems or dynamic data retrieval pipelines could provide enhanced reliability, especially in sensitive domains like medicine or law. The development of benchmarks tailored to the evolving capabilities of LLMs is crucial for enabling consistent evaluation across various contexts. Enhancing transparency in understanding the internal mechanisms of LLMs, as demonstrated by SAPLMA, can also offer new pathways for identifying and mitigating biases. Additionally, incorporating real-time data streams into training processes could alleviate the issue of outdated information in static datasets, improving the overall reliability of LLM outputs.

References

[1] C.-C. Hung, W. B. Rim, L. Frost, L. Bruckner, and C. Lawrence, "Walking a tightrope – evaluating large language models in high-risk domains."

[2] N. Badyal, D. Jacoby, and Y. Coady, "Intentional biases in LLM responses," in *2023 IEEE 14th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, pp. 0502–0506, IEEE.

[3] N. Du, Y. Huang, A. M. Dai, S. Tong, D. Lepikhin, Y. Xu, M. Krikun, Y. Zhou, A. W. Yu, O. Firat, B. Zoph, L. Fedus, M. Bosma, Z. Zhou, T. Wang, Y. E. Wang, K. Webster, M. Pellat, K. Robinson, K. Meier-Hellstern, T. Duke, L. Dixon, K. Zhang, Q. V. Le, Y. Wu, Z. Chen, and C. Cui, "GLaM: Efficient scaling of language models with mixture-of-experts."

[4] P. Laban, T. Schnabel, P. N. Bennett, and M. A. Hearst, "SummaC : Re-visiting NLI-based models for inconsistency detection in summarization," vol. 10, pp. 163–177.
[5] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. Bang, D. Chen, W. Dai, H. S. Chan, A. Madotto, and P. Fung, "Survey of hallucination in natural language generation," vol. 55, no. 12, pp. 1–38.
[6] S. Jha, S. K. Jha, P. Lincoln, N. D. Bastian, A. Velasquez, and S. Neema, "Dehallucinating large language models using formal methods guided iterative prompting," in *2023 IEEE International Conference on Assured Autonomy (ICAA)*, pp. 149–152, IEEE.
[7] I. Mirzadeh, K. Alizadeh, H. Shahrokhi, O. Tuzel, S. Bengio, and M. Farajtabar, "GSM-symbolic: Understanding the limitations of mathematical reasoning in large language models."
[8] A. Azaria and T. Mitchell, "The internal state of an LLM knows when it's lying."
[9] "Understanding BLEU and ROUGE score for NLP evaluation." Section: NLP.
[10] S. Hughes, M. Bae, and M. Li, "Vectara hallucination leaderboard." original-date: 2023-10-31T21:19:12Z.
[11] O. Honovich, R. Aharoni, J. Herzig, H. Taitelbaum, D. Kukliansy, V. Cohen, T. Scialom, I. Szpektor, A. Hassidim, and Y. Matias, "TRUE: Re-evaluating factual consistency evaluation."
[12] R. Hu, J. Zhong, M. Ding, Z. Ma, and M. Chen, "Evaluation of hallucination and robustness for large language models," in *2023 IEEE 23rd International Conference on Software Quality, Reliability, and Security Companion (QRS-C)*, pp. 374–382, IEEE.
[13] "The hallucinations leaderboard, an open effort to measure hallucinations in large language models."
[14] A. Zou, Z. Wang, N. Carlini, M. Nasr, J. Z. Kolter, and M. Fredrikson, "Universal and transferable adversarial attacks on aligned language models."
[15] N. Lambert, L. Castricato, L. von Werra, and A. Havrilla, "Illustrating reinforcement learning from human feedback (RLHF)," 2022.
[16] N. Lambert, T. K. Gilbert, and T. Zick, "The history and risks of reinforcement learning and human feedback."
[17] T. Yu, Y. Yao, H. Zhang, T. He, Y. Han, G. Cui, J. Hu, Z. Liu, H.-T. Zheng, and M. Sun, "RLHF-v: Towards trustworthy MLLMs via behavior alignment from fine-grained correctional human feedback," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13807–13816, IEEE.
[18] Google Cloud, "What is retrieval-augmented generation (RAG)?," 2024.
[19] "Qu'est-ce que la génération à enrichissement contextuel (RAG) ? – explication de l'IA de génération à enrichissement contextuel – AWS."
[20] G. Perković, A. Drobnjak, and I. Botički, "Hallucinations in LLMs: Understanding and addressing challenges," in *2024 47th MIPRO ICT and Electronics Convention (MIPRO)*, pp. 2084–2088, IEEE.
[21] "LangChain."