Insights into Classifying and Mitigating LLMs’ Hallucinations

Alessandro Bruno¹,*, Pier Luigi Mazzeo²,*, Aladine Chetouani³,†, Marouane Tliba³,† and Mohamed Amine Kerkouri³,†

¹IULM University, Department of Business, Law, Economics, Consumer Behaviour - "Carlo A. Ricciardi", Via Carlo Bo 1, Milan, 20143, Italy
²ISASI Institute of Applied Sciences and Intelligent Systems-CNR, 73100 Lecce, Italy
³Université d’Orléans, 45067 Orleans, France

Abstract
The widespread adoption of large language models (LLMs) across diverse AI applications is proof of the outstanding achievements obtained in several tasks, such as text mining, text generation, and question answering. However, LLMs are not exempt from drawbacks. One of the most concerning aspects regards the emerging problematic phenomena known as "Hallucinations". They manifest in text generation systems, particularly in question-answering systems reliant on LLMs, potentially resulting in false or misleading information propagation. This paper delves into the underlying causes of AI hallucination and elucidates its significance in artificial intelligence. In particular, Hallucination classification is tackled over several tasks (Machine Translation, Question and Answer, Dialog Systems, Summarisation Systems, Knowledge Graph with LLMs, and Visual Question Answer). Additionally, we explore potential strategies to mitigate hallucinations, aiming to enhance the overall reliability of LLMs. Our research addresses this critical issue within the HeReFaNMi (Health-Related Fake News Mitigation) project, generously supported by NGI Search, dedicated to combating Health-Related Fake News dissemination on the Internet. This endeavour represents a concerted effort to safeguard the integrity of information dissemination in an age of evolving AI technologies.

Keywords
LLMs, Hallucination, Artificial Intelligence, Hallucination Mitigation, Factualness

1. Introduction

The large language models (LLMs) landscape continues to evolve with innovative creations such as GPT-3 [1], IntroductGPT [2], FLAN [3], PaLM [4], LLaMA [5] and other important contributions[6, 7, 8, 9]. Other than outstanding performances in several tasks, LLMs have revealed a concerning drawback affecting their reliability and trustworthiness: hallucination.

*Corresponding author.
†These authors contributed equally.
®alessandro.bruno@iulm.it (A. Bruno); pierluigi.mazzeo@cnr.it (P. L. Mazzeo); aladine.chetouani@univ-orleans.fr (A. Chetouani); marouane.tliba@univ-orleans.fr (M. Tliba); mohamed-amine.kerkouri@univ-orleans.fr (M. A. Kerkouri)
®https://www.iulm.it/en/iulm/ateneo/docenti-e-collaboratori/bruno-alessandro (A. Bruno);
https://sites.google.com/view/pierluigimazzeo (P. L. Mazzeo); http://aladine-chetouani.com/ (A. Chetouani)
®0000-0003-0707-6131 (A. Bruno); 0000-0002-7552-2394 (P. L. Mazzeo)
© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)
Quoting Berrios and Dening [10], "Hallucinations are conceived of as indistinguishable from real perceptions except that there is no stimulus", one can easily peruses nuanced relations between perceptions and hallucinations.

Providing that a great deal of AI theories and approaches focus on human behaviour analysis, hallucinations appearing in AI might not come as a surprise. Hallucination can also be considered the generation of statements that appear reasonable but are either cognitively irrelevant or factually incorrect. Considering this observation, hallucination has become a critical challenge in medical [11, 12], financial [13] and other delicate fields where exact accuracy is a mandatory requirement. Why do LLMs run into hallucinations, then? Lack of real-world knowledge, bias or misleading training data may prompt models to return statistical-based results. In particular, the latter means there might not be a proper understanding of input.

**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) [14].

Running through LLM-based outputs is paramount to avoid getting into the cognitive mirage phenomenon that negatively affects decision-making strategies and a cascade of unintended consequences [34]. Classifying and Mitigating LLMs’ hallucinations is a relatively emerging topic. Since the introduction of ChatGPT in 2022, an exponentiation growth of applications and tools based on LLMs has been observed worldwide. Subsequently, significant interest from the scientific community and industry in the LLMs’ side effects, such as hallucinations, has emerged naturally. In [14], hallucinatory content in task-specific research progress has been analyzed and referred to early works in the natural language generation field. Covering methods for collecting high-quality instructions for LLM alignment are discussed in [35], including NLP benchmarks. Human annotations and leveraging strong LLMs. In [36], self-correcting methods have been discussed where an LLM is guided or prompted to correct the hallucinations from its own output. Unlike these works, our contribution will lead to a literature review on hallucinations in LLMs, running through different methods and providing insights into the pros and cons.

The main contribution of this paper regards a thorough analysis of LLMs’ hallucinations research field under multiple viewpoints. To this end, the relevant work in this field has been reviewed and categorized over tasks and domains. Some methodologies regarding the proactive detection and mitigation of hallucinations in the LLMs era are also discussed. The pros and cons of mitigation techniques are evaluated by reporting the techniques behind the proposed solutions. The final section, Future Perspectives, draws some lines and poses some questions in the current scenario of interest.

### 2. Hallucination Classification

In this work, we consider the hallucinations observed in prevalent downstream tasks: i) Machine Translation; ii) Question and Answer (Q&A); iii) Dialog System; iv) Summarization System; v) Knowledge graph with LLMs; vi) Visual Question Answer. Table 1 summarizes hallucination types, grouping them according to numerous mainstream tasks associated with LLMs. The following subsections will describe the most frequent hallucination types during these tasks.
<table>
<thead>
<tr>
<th>Task</th>
<th>Dataset</th>
<th>Architecture</th>
<th>Hallucination Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Machine Translation</td>
<td>IWSLT2014</td>
<td>Enc-Dec</td>
<td>Under perturbation, Natural hallucination</td>
</tr>
<tr>
<td>Machine Translation</td>
<td>WMT2018</td>
<td>Enc-Dec</td>
<td>Oscillatory hallucination, Largely fluent hallucination</td>
</tr>
<tr>
<td>Machine Translation</td>
<td>FLORES-200, Jig-saw, Wikipedia</td>
<td>Enc-Dec</td>
<td>Full, Partial, and Word-level hallucination</td>
</tr>
<tr>
<td>Multilingual Seq2seq</td>
<td>XQuAD, TyDi, XNLI, XL-Sum, MASSIVE</td>
<td>Enc-Dec</td>
<td>Source language hallucination</td>
</tr>
<tr>
<td>Question and Answer</td>
<td>TruthfulQA</td>
<td>Enc-Dec, Only-Dec</td>
<td>Imitative falsehoods</td>
</tr>
<tr>
<td>Question and Answer</td>
<td>HotpotQA, BoolQ</td>
<td>Only-Dec</td>
<td>Comprehension, Factualness, Specificity, Inference Hallucination</td>
</tr>
<tr>
<td>Question and Answer</td>
<td>NQ, HotpotQA, TopiOCQA</td>
<td>Enc-Dec, Only-Dec</td>
<td>Semantic and Symbolic equivalence, Intrinsic ambiguity, Granularity discrepancies, Incomplete, Enumeration, Satisfactory Subset</td>
</tr>
<tr>
<td>Question and Answer</td>
<td>MEDMCQA, Headqa, USMILE, Medqa, Pubmed</td>
<td>Only-Dec</td>
<td>Reasoning hallucination, Memory-based hallucination</td>
</tr>
<tr>
<td>Dialog System</td>
<td>WoW, CMU-DOG, TopicalChat</td>
<td>Enc-Dec, Only-Dec</td>
<td>Hallucination, Partial hallucination, Generic, Uncooperativeness</td>
</tr>
<tr>
<td>Dialog System</td>
<td>OpenDialKG</td>
<td>Only-Dec</td>
<td>Extrinsic-Soft/Hard/Grouped, Intrinsic-Soft/Hard/Repetitive, History Corrupted Hallucination, Generic, Uncooperativeness</td>
</tr>
<tr>
<td>Dialog System</td>
<td>WoW</td>
<td>Enc-Dec, Only-Dec</td>
<td>Fully attributable, Not attributable, Generic</td>
</tr>
<tr>
<td>Dialog System</td>
<td>WoW, CMU-DOG, TopicalChat</td>
<td>Enc-Dec, Only-Dec</td>
<td>Intrinsic hallucination, Extrinsic hallucination</td>
</tr>
<tr>
<td>Dialog System</td>
<td>WoW</td>
<td>Enc-Dec, Only-Dec</td>
<td>Intrinsic hallucination, Extrinsic hallucination</td>
</tr>
<tr>
<td>Summarization System</td>
<td>CNN/DM, XSum</td>
<td>Enc-Dec, Only-Dec</td>
<td>Factually inconsistent summaries</td>
</tr>
<tr>
<td>Summarization System</td>
<td>MENT</td>
<td>Enc-Dec, Only-Dec</td>
<td>Non-hallucinated, Factual, Non-factual, and Intrinsic hallucation</td>
</tr>
<tr>
<td>Summarization System</td>
<td>NHNet</td>
<td>Enc-Dec, Only-Dec</td>
<td>News headline hallucination</td>
</tr>
<tr>
<td>Summarization System</td>
<td>XL-Sum</td>
<td>Multiple ADapters</td>
<td>Intrinsic hallucination, Extrinsic hallucination</td>
</tr>
<tr>
<td>Knowledge based text generation</td>
<td>Encyclopedic, ETC</td>
<td>Enc-Dec, Only-Dec</td>
<td>Knowledge hallucination</td>
</tr>
<tr>
<td>Knowledge graph generation</td>
<td>TekGen, WebNLG</td>
<td>Only-Dec</td>
<td>Subject, relation, and object hallucination</td>
</tr>
<tr>
<td>Visual Question Answer</td>
<td>MSCOCO</td>
<td>Enc-Dec</td>
<td>Caption hallucination assessment</td>
</tr>
</tbody>
</table>
2.1. Machine Translation

Since some text perturbation can bring trustworthy hallucinations, traditional translation methodologies validate the instances fed into the model when perturbed [37, 38]. Hallucinations generated by LLMs are principally translation off-target or failed translation [16]. With low-resource language availability, trained models perform poorly due to few annotated data employed [17]. An increasing amount of pre-trained language affects the machine translation reliability in the multilingual domain [39]. Therefore, LLMs trained on various scales of monolingual data seem to be tacky [16] as the origin of a hallucination pathology.

2.2. Question and Answer (Q&A)

Wrong responses occur by the flawed external knowledge as described in [20]. Often, LLMs give incomplete and plausible answers instead of giving no response when they have poor or irrelevant information [21]. It has also to be considered that memorized information without referring to accurate, reliable and accessible sources contribute to creating different type of hallucinations [12]. Scaling up models alone is less promising for improving truthfulness than fine-tuning using training objectives other than imitating text from the web [19].

2.3. Dialog System

Many works considered dialogue models as simple imitators that only change the data views and communication instead of generating new trustworthy output. In [22], authors demonstrated that the standard benchmarks led models even to amplify hallucinations. In [23] are identified various modes of hallucination in Knowledge Graph(KG) grounded chatbots through human feedback analysis. In similar works, many [24] [25] [26] experiments are implemented on the WoW dataset conducting a meta-evaluation of the hallucination in knowledge grounded dialogue.

2.4. Summarization System

These systems allow the automatic generation automatically fluent abstracts based on LLMs but often lack faithfulness from the source document. Summarization generated by LLMs can be slit into two categories for their evaluation: intrinsic hallucinations that deform the information contained in the document; extrinsic hallucinations that add information not directly sourced by the original document [30]. More attention has been given to extrinsic hallucinations in summarization systems due to factually consistent continuation of input in LLMs [27, 29]. A further subdivision is proposed in [28] where extrinsic hallucinations are split into factual and non-factual. Factual hallucinations insert additional world knowledge that may improve the text’s understanding.

2.5. Knowledge Graph with LLMs

Knowledge-based text generation stumbles in intrinsic hallucinations due to redundant details derived from its internal memorized Knowledge [40]. Yu et al. [31] tackled the mentioned
User: 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.

**Figure 1**: An example of ChatGPT hallucinating is given above. The false premise in the question (a made-up Festival name) prompts ChatGPT into Hallucination issue by establishing a distinction between correctly generated Knowledge and Knowledge hallucinations. Virtual Knowledge extraction proposed in [41] highlight the potential LLMs capabilities of constructing and inferring from Knowledge Graphs. An LLM empowering for producing interpretable fact checks using a neural symbolic approach is described in [32] where hallucinations have been defined as subject hallucination, relation hallucination and object hallucination according to their fidelity to the source.

### 2.6. Cross-modal System

Cross-modal tasks achieve interesting progress thanking to the superior language capabilities of LLMs [41, 42]. However, in some cases substituting the original language encoder, Large Visual Language Models (LVLMs) [43] continue to generate descriptions of objects that are not in the images; this is denoted as object hallucinations [33]. Typically most of the failure cases should be found in Visual Question Answering [33], Image Captioning [44, 45, 46], Report Generation [47].
3. Hallucination Detection

Several methods introduced detecting realistic and convincing Hallucinations in LLMs. Some techniques rely on extracting intrinsic uncertainty metrics. Token probability, for instance, can be leveraged to identify which part of a given textual sequence proves least uncertain [48],[49]. However, scenarios like external APIs from ChatGPT do not give users access to output token probability, meaning that the techniques mentioned above cannot work out uncertainty metrics. LLMs factual checks can also rely on external databases and corpora such as Wikipedia [50]. Hallucinations can be detected in a great deal of general knowledge covered in Wikipedia, albeit concerns arise about the integrity of Wikipedia content itself. Azaria and Mitchell [51] proposed a statement’s truthfulness detection using LLMs’ hidden representations to feed a multi-layer classifier. Azaria and Mitchell’s method sticks to the supervised training paradigm. Therefore, it relies on labelled data along with the 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. Kadavath et al. [52] introduced a Hallucination detection method, Self-Evaluation. The name is due to the core of the study being if language models can assess their own answers’ validity and predict accuracy. 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, with partial task generalization (IK stands for “I Know”). Several Hallucination detection approaches fit the so-called “zero-resource” setting. That means there is no external database to verify the factuality of an LLM response. That said, Hallucation detection methods can further be grouped into Grey and Black box [53]. 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.

Different strategies come into play to tackle grey and black box hallucinations. Knowing LLM pre-training is paramount for grey box hallucination detection. The training is carried out with next-word prediction over vast textual corpora, ensuring world knowledge and contextual reasoning. A diagram depicting how uncertainty and factuality-based assessment work is given in Figure 2. Noticeably, Varshney et al. [54] detected GPT3.5 hallucinations by designing a sophisticated technique. It carries out critical concept identification with entity, keyword extraction, and ‘Instructing the model’. In particular, they used LLM capabilities to identify essential concepts from the generated sentence. A comparison study of the three techniques remarkably 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. The technique was also enriched by 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.
4. Mitigating LLM Hallucinations

Mitigating hallucinations in LLMs is an emerging challenge due to the increasing worldwide adoption of LLMs-based virtual chatbot agents and Question-answer systems. Although several methods have been recently presented to tackle the problem, some partly work well as countermeasure systems as, at the same time, they may introduce further hallucinations into the LLM itself [54]. Varshney et al. [54] 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.

Despite LLMs’ hallucination being a relatively new issue, several methods relying on different paradigms have been proposed. They can be grouped into the following families:

- Fine-tuning
- Knowledge Graphs
- Memory Augmentation
- Context Prompts
- Preemptive Strategies
A graphical depiction of mitigation methods, pros and cons, is given in Figure 3. 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 [55]. LLMs’ hallucinations can be mitigated with fine-tuning, as Lee et al. showed in their work [56]. However, LLMs featuring millions of parameters make fine-tuning an expensive solution. Knowledge graph methods allow for integrating structured and unstructured knowledge [57]. That gives LLMs a more extended platform to run tasks. The drawback entails two aspects: designing a well-curated knowledge base is time-consuming, and keeping up-to-date knowledge is labour-intensive. Wu et al. [58] proposed an augmented transformer for knowledge-intensive NLP tasks. That is due to the need for deep learning methods to extend their capabilities on new knowledge. Although NLP models have already benefited from memory augmentation, the same cannot be said for LLMs, as no tests have been run.

Prompt-based solutions have been recently introduced to ‘de-hallucinate’ LLMs. Jha et al. [59] proposed a self-monitoring prompting framework. This framework leverages formal methods to identify errors in the LLM’s responses autonomously. They employed the conversational abilities of LLMs for response alignment with specified correctness criteria through iterative refinement. Luo et al. [60] proposed Self-Familiarity, a method to overcome the current SOTA (State-of-the-art) techniques that identify and mitigate hallucinations post-generation.

Self-Familiarity introduced an innovative zero-resource, pre-detection approach to mitigate the risk of large language models (LLMs) producing inaccurate information. This method extracted and processed conceptual entities from the instruction. Subsequently, it employed prompt engineering to acquire a familiarity score for each concept. These scores were combined to yield the ultimate familiarity score at the instruction level. A low instruction-level familiarity
score indicates a higher likelihood of the LLM generating erroneous information, prompting it to abstain from generating a response.

Feldman et al. [61] designed a method relying on context-tagged prompts. They created a set of questions and then developed context prompts to help the LLM answer those questions more accurately. They then validated the context prompts and the questions to ensure they worked as intended. Finally, they ran experiments with different GPT models to see how context prompts affected the LLM responses’ accuracy.

5. Future Perspective

Some considerations are drawn in this section concerning LLMs hallucination and mitigation methods. Zero-resource hallucination detection: Current zero-resource hallucination detection methods are still in their early stages of development. Future research could focus on developing more accurate and reliable methods for a broader range of scenarios. Black-box hallucination detection: Black-box hallucination detection is even more challenging than zero-resource hallucination detection, as there is no access to the LLM’s internal states. Future research could focus on developing new black-box hallucination detection methods or finding ways to make existing methods more effective. Hallucination detection for specific tasks: Most current hallucination detection methods are general-purpose. However, hallucination detection may be more effective if tailored to specific tasks. For example, hallucination detection methods for factual question answering could be designed to leverage the fact that factually accurate answers are more likely to be grounded in real-world knowledge. Hallucination detection in multimodal LLMs: Multimodal LLMs are a new type of LLM that can process and generate text, images, and other media types. Hallucination detection in multimodal LLMs is a challenging problem, but it is essential to address, as multimodal LLMs are becoming increasingly popular. Here are some specific research questions that could be explored in each of these areas:

Zero-resource hallucination detection: Can zero-resource hallucination detection be made more accurate and reliable? Can zero-resource hallucination detection be applied to a broader range of scenarios, such as real-time conversation? Black-box hallucination detection: Can new methods be developed for black-box hallucination detection? Can existing hallucination detection methods be made more effective for black-box scenarios? Hallucination detection for specific tasks: Can hallucination detection be tailored to specific tasks, such as factual question answering and code generation? How can we leverage the unique properties of each task to improve the accuracy of hallucination detection? Hallucination detection in multimodal LLMs: 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? In addition to these research questions, developing and evaluating new benchmarks for hallucination detection is also substantial. This will help to ensure that hallucination detection methods are evaluated fairly and consistently.
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.

References


[28] M. Cao, Y. Dong, J. Cheung, Hallucinated but factual! inspecting the factuality of hallucinations in abstractive summarization, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), As-


