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 - "Carlo A. Ricciardi", Via Carlo Bo 1, Milan, 20143, Italy

[2]ISASI Institute of Applied Sciences and Intelligent Systems-CNR, 73100 Lecce, Italy

[3]Université d'Orleans, 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)

CEUR-WS.org/Vol-3563/paper_12.pdf

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 1
List of Hallucinations examples

	Task	Dataset	Architecture	Hallucination Type
[15]	Machine Translation	IWSLT2014	Enc-Dec	Under perturbation, Natural hallucination
[16]	Machine Translation	WMT2018	Enc-Dec	Oscillatory hallucination, Largely fluent hallucination
[17]	Machine Translation	FLORES-200, Jig-saw, Wikipedia	Enc-Dec	Full, Partial, and Word-level hallucination
[18]	Multilingual Seq2seq	XQuAD, TyDi, XNLI, XL-Sum, MASSIVE	Enc-Dec	Source language hallucination
[19]	Question and Answer	TruthfulQA	Enc-Dec, Only-Dec	Imitative falsehoods
[20]	Question and Answer	HotpotQA, BoolQ	Only-Dec	Comprehension, Factualness, Specificity, Inference Hallucination
[21]	Question and Answer	NQ, HotpotQA, TopiOCQA	Enc-Dec, Only-Dec	Semantic and Symbolic equivalence, Intrinsic ambiguity, Granularity discrepancies, Incomplete, Enumeration, Satisfactory Subset
[12]	Question and Answer	MEDMCQA, Headqa, USMILE, Medqa, Pubmed	Only-Dec	Reasoning hallucination, Memory-based hallucination
[22]	Dialog System	WoW, CMU-DOG, TopicalChat	Enc-Dec, Only-Dec	Hallucination, Partial hallucination, Generic, Uncooperative
[23]	Dialog System	OpenDialKG	Only-Dec	Extrinsic-Soft/Hard/Grouped, Intrinsic-Soft/Hard/Repetitive, History Corrupted
[24]	Dialog System	WoW	Enc-Dec, Only-Dec	Hallucination, Generic, Uncooperativeness
[25]	Dialog System	WoW, CMU-DOG, TopicalChat	Enc-Dec, Only-Dec Only-Enc	Fully attributable, Not attributable, Generic
[26]	Dialog System	WoW	Enc-Dec, Only-Dec	Intrinsic hallucination, Extrinsic hallucination
[27]	Summarization System	CNN/DM, XSum	Enc-Dec, Only-Dec	Factually inconsistent summaries
[28]	Summarization System	MENT	Enc-Dec, Only-Dec	Non-hallucinated, Factual, Non-factual, and Intrinsic hallucination
[29]	Summarization System	NHNet	Enc-Dec, Only-Dec	News headline hallucination
[30]	Summarization System	XL-Sum	Multiple ADapters	Intrinsic hallucination, Extrinsic hallucination
[31]	Knowledge based text generation	Encyclopedic, ETC	Enc-Dec, Only-Dec	Knowledge hallucination
[32]	Knowledge graph generation	TekGen,WebNLG	Only-Dec	Subject,relation, and object hallucination
[33]	Visual Question Answer	MSCOCO	Enc-Dec	Caption hallucination assessment

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

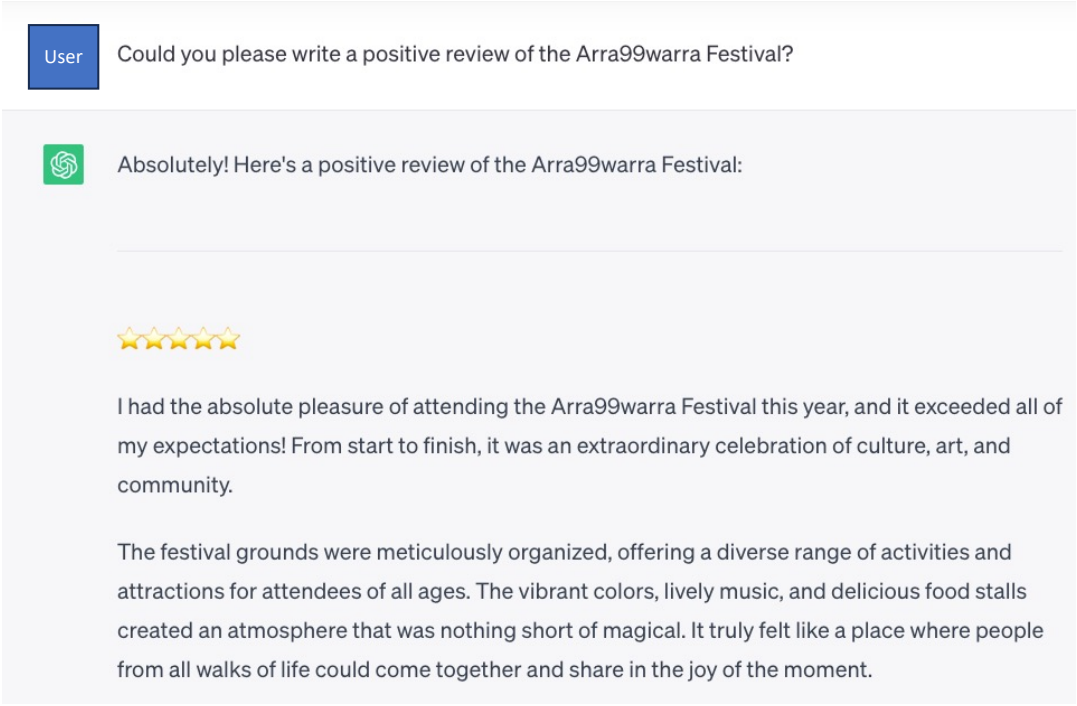

Figure 1: An example of ChatGPT hallucinating is given above. The false premise in the question (a made-up Festival name) prompts ChatGPT into Hallucination

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.

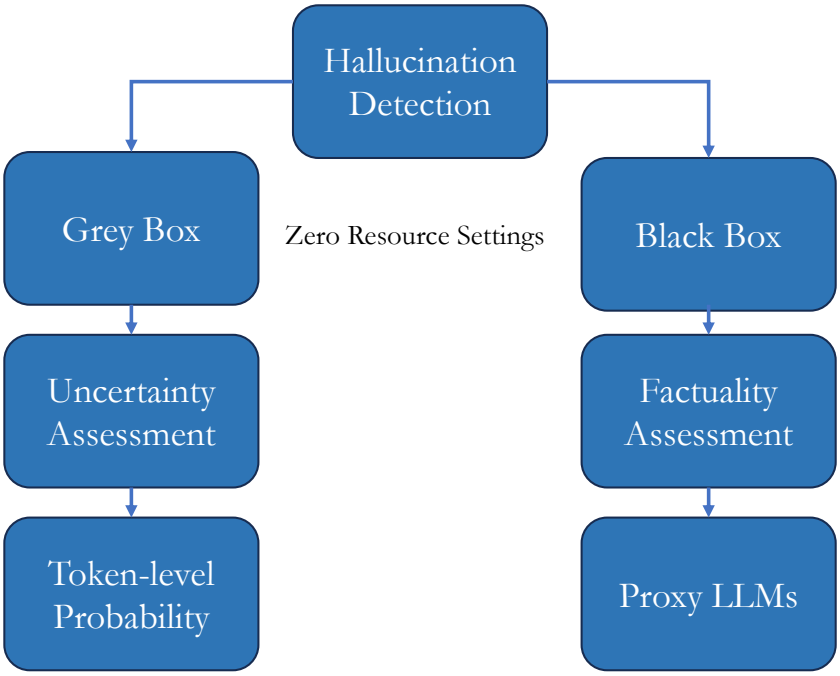

Figure 2: Several LLM Hallucination Detection methods are grouped into Grey and Black Box as depicted above.

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 counter-measure 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

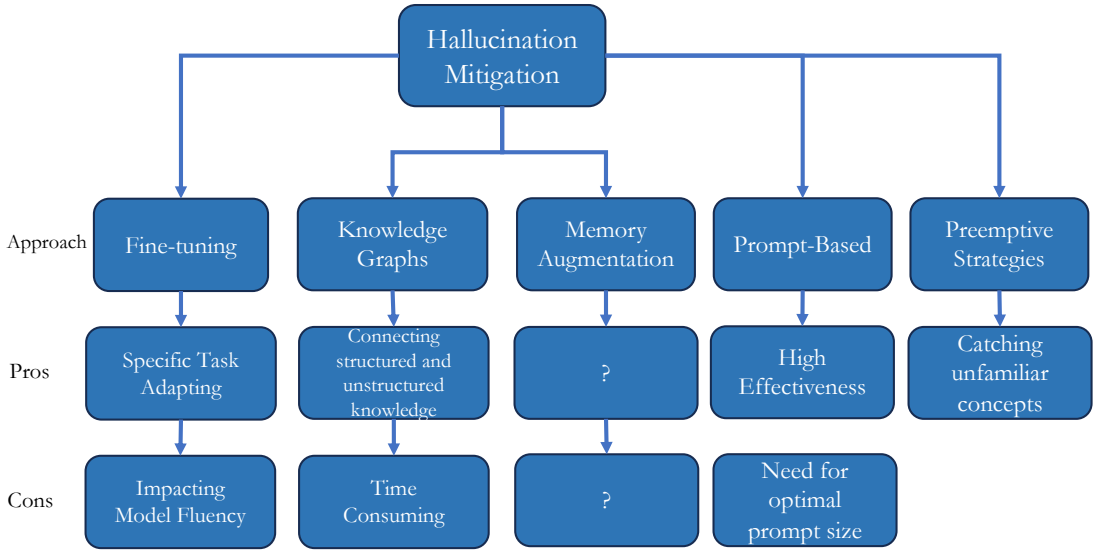

Figure 3: Hallucination Mitigation approaches, pros and cons are depicted above.

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

[1] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 1877–1901. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[2] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. F. Christiano, J. Leike, R. Lowe, Training language models to follow instructions with human feedback, in: S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, A. Oh (Eds.), Advances in Neural Information Processing Systems, volume 35, Curran Associates, Inc., 2022, pp. 27730–27744. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.

[3] J. Wei, M. Bosma, V. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, Q. V. Le, Finetuned language models are zero-shot learners, in: International Conference on Learning Representations, 2022. URL: https://openreview.net/forum?id=gEZrGCozdqR.

[4] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, P. Schuh, K. Shi, S. Tsvyashchenko, J. Maynez, A. Rao, P. Barnes, Y. Tay, N. M. Shazeer, V. Prabhakaran, E. Reif, N. Du, B. C. Hutchinson, R. Pope, J. Bradbury, J. Austin, M. Isard, G. Gur-Ari, P. Yin, T. Duke, A. Levskaya, S. Ghemawat, S. Dev, H. Michalewski, X. García, V. Misra, K. Robinson, L. Fedus, D. Zhou, D. Ippolito, D. Luan, H. Lim, B. Zoph, A. Spiridonov, R. Sepassi, D. Dohan, S. Agrawal, M. Omernick, A. M. Dai, T. S. Pillai, M. Pellat, A. Lewkowycz, E. Moreira, R. Child, O. Polozov, K. Lee, Z. Zhou, X. Wang, B. Saeta, M. Díaz, O. Firat, M. Catasta, J. Wei, K. S. Meier-Hellstern, D. Eck, J. Dean, S. Petrov, N. Fiedel, Palm: Scaling language modeling with pathways, ArXiv abs/2204.02311 (2022). URL: https://api.semanticscholar.org/CorpusID:247951931.

[5] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, Llama: Open and efficient foundation language models, 2023. arXiv:2302.13971.

[6] Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. J. Henighan, N. Joseph, S. Kadavath, J. Kernion, T. Conerly, S. El-Showk, N. Elhage, Z. Hatfield-Dodds, D. Hernandez, T. Hume, S. Johnston, S. Kravec, L. Lovitt, N. Nanda, C. Olsson, D. Amodei, T. B. Brown, J. Clark, S. McCandlish, C. Olah, B. Mann, J. Kaplan, Training a helpful and harmless assistant with reinforcement learning from human

feedback, ArXiv abs/2204.05862 (2022). URL: https://api.semanticscholar.org/CorpusID:248118878.

[7] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. T. Diab, X. Li, X. V. Lin, T. Mihaylov, M. Ott, S. Shleifer, K. Shuster, D. Simig, P. S. Koura, A. Sridhar, T. Wang, L. Zettlemoyer, Opt: Open pre-trained transformer language models, ArXiv abs/2205.01068 (2022). URL: https://api.semanticscholar.org/CorpusID:248496292.

[8] A. Zeng, X. Liu, Z. Du, Z. Wang, H. Lai, M. Ding, Z. Yang, Y. Xu, W. Zheng, X. Xia, W. L. Tam, Z. Ma, Y. Xue, J. Zhai, W. Chen, Z. Liu, P. Zhang, Y. Dong, J. Tang, GLM-130b: An open bilingual pre-trained model, in: The Eleventh International Conference on Learning Representations, 2023. URL: https://openreview.net/forum?id=-Aw0rrrPUF.

[9] C. Xu, Q. Sun, K. Zheng, X. Geng, P. Zhao, J. Feng, C. Tao, D. Jiang, Wizardlm: Empowering large language models to follow complex instructions, ArXiv abs/2304.12244 (2023). URL: https://api.semanticscholar.org/CorpusID:258298159.

[10] G. E. Berrios, T. Dening, Pseudohallucinations: a conceptual history, Psychological Medicine 26 (1996) 753–763.

[11] D. Dash, R. Thapa, J. Banda, A. Swaminathan, M. Cheatham, M. Kashyap, N. Kotecha, J. H. Chen, S. Gombar, L. Downing, R. A. Pedreira, E. Goh, A. Arnaout, G. K. Morris, H. Magon, M. P. Lungren, E. Horvitz, N. H. Shah, Evaluation of gpt-3.5 and gpt-4 for supporting real-world information needs in healthcare delivery, ArXiv abs/2304.13714 (2023). URL: https://api.semanticscholar.org/CorpusID:258331653.

[12] L. K. Umapathi, A. Pal, M. Sankarasubbu, Med-halt: Medical domain hallucination test for large language models, ArXiv abs/2307.15343 (2023). URL: https://api.semanticscholar.org/CorpusID:260316324.

[13] S. S. Gill, M. Xu, P. Patros, H. Wu, R. Kaur, K. Kaur, S. Fuller, M. Singh, P. Arora, A. K. Parlikad, V. Stankovski, A. Abraham, S. K. Ghosh, H. Lutfiyya, S. S. Kanhere, R. Bahsoon, O. F. Rana, S. Dustdar, R. Sakellariou, S. Uhlig, R. Buyya, Transformative effects of chatgpt on modern education: Emerging era of ai chatbots, ArXiv abs/2306.03823 (2023). URL: https://api.semanticscholar.org/CorpusID:259088562.

[14] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, P. Fung, Survey of hallucination in natural language generation, ACM Comput. Surv. 55 (2023). URL: https://doi.org/10.1145/3571730. doi:10.1145/3571730.

[15] V. Raunak, A. Menezes, M. Junczys-Dowmunt, The curious case of hallucinations in neural machine translation, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 1172–1183. URL: https://aclanthology.org/2021.naacl-main.92. doi:10.18653/v1/2021.naacl-main.92.

[16] N. M. Guerreiro, D. Alves, J. Waldendorf, B. Haddow, A. Birch, P. Colombo, A. Martins, Hallucinations in large multilingual translation models, ArXiv abs/2303.16104 (2023). URL: https://api.semanticscholar.org/CorpusID:257771892.

[17] D. Dale, E. Voita, J. Lam, P. Hansanti, C. Ropers, E. Kalbassi, C. Gao, L. Barrault, M. R. Costa-jussà, Halomi: A manually annotated benchmark for multilingual hallucination and omission detection in machine translation, ArXiv abs/2305.11746 (2023). URL: https://api.semanticscholar.org/CorpusID:258823059.

[18] J. Pfeiffer, F. Piccinno, M. Nicosia, X. Wang, M. Reid, S. Ruder, mmt5: Modular multilingual

pre-training solves source language hallucinations, ArXiv abs/2305.14224 (2023). URL: https://api.semanticscholar.org/CorpusID:258841429.

[19] S. Lin, J. Hilton, O. Evans, TruthfulQA: Measuring how models mimic human falsehoods, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 3214–3252. URL: https://aclanthology.org/2022.acl-long.229. doi:10.18653/v1/2022.acl-long.229.

[20] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. Gonzalez, I. C. Stoica, Judging llm-as-a-judge with mt-bench and chatbot arena, ArXiv abs/2306.05685 (2023). URL: https://api.semanticscholar.org/CorpusID:259129398.

[21] V. Adlakha, P. BehnamGhader, X. H. Lu, N. Meade, S. Reddy, Evaluating correctness and faithfulness of instruction-following models for question answering, ArXiv abs/2307.16877 (2023). URL: https://api.semanticscholar.org/CorpusID:260334056.

[22] N. Dziri, S. Milton, M. Yu, O. Zaiane, S. Reddy, On the origin of hallucinations in conversational models: Is it the datasets or the models?, in: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Seattle, United States, 2022, pp. 5271–5285. URL: https://aclanthology.org/2022.naacl-main.387. doi:10.18653/v1/2022.naacl-main.387.

[23] S. Das, S. Saha, R. Srihari, Diving deep into modes of fact hallucinations in dialogue systems, in: Findings of the Association for Computational Linguistics: EMNLP 2022, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 684–699. URL: https://aclanthology.org/2022.findings-emnlp.48. doi:10.18653/v1/2022.findings-emnlp.48.

[24] N. Dziri, E. Kamalloo, S. Milton, O. Zaiane, M. Yu, E. M. Ponti, S. Reddy, FaithDial: A Faithful Benchmark for Information-Seeking Dialogue, Transactions of the Association for Computational Linguistics 10 (2022) 1473–1490. URL: https://doi.org/10.1162/tacl_a_00529. doi:10.1162/tacl_a_00529. arXiv:https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00529/2065956/tacl_a_00529.pdf.

[25] N. Dziri, H. Rashkin, T. Linzen, D. Reitter, Evaluating attribution in dialogue systems: The begin benchmark, Transactions of the Association for Computational Linguistics 10 (2021) 1066–1083. URL: https://api.semanticscholar.org/CorpusID:233481654.

[26] W. Sun, Z. Shi, S. Gao, P. Ren, M. de Rijke, Z. Ren, Contrastive learning reduces hallucination in conversations, Proceedings of the AAAI Conference on Artificial Intelligence 37 (2023) 13618–13626. URL: https://ojs.aaai.org/index.php/AAAI/article/view/26596. doi:10.1609/aaai.v37i11.26596.

[27] D. Tam, A. Mascarenhas, S. Zhang, S. Kwan, M. Bansal, C. Raffel, Evaluating the factual consistency of large language models through news summarization, in: Findings of the Association for Computational Linguistics: ACL 2023, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 5220–5255. URL: https://aclanthology.org/2023.findings-acl.322. doi:10.18653/v1/2023.findings-acl.322.

[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-

sociation for Computational Linguistics, Dublin, Ireland, 2022, pp. 3340–3354. URL: https://aclanthology.org/2022.acl-long.236. doi:10.18653/v1/2022.acl-long.236.

[29] J. Shen, J. Liu, D. Finnie, N. Rahmati, M. Bendersky, M. Najork, "why is this misleading?": Detecting news headline hallucinations with explanations, in: Proceedings of the ACM Web Conference 2023, WWW '23, Association for Computing Machinery, New York, NY, USA, 2023, p. 1662–1672. URL: https://doi.org/10.1145/3543507.3583375. doi:10.1145/3543507.3583375.

[30] Y. Qiu, Y. Ziser, A. Korhonen, E. Ponti, S. B. Cohen, Detecting and mitigating hallucinations in multilingual summarisation, ArXiv abs/2305.13632 (2023). URL: https://api.semanticscholar.org/CorpusID:258841008.

[31] J. Yu, X. Wang, S. Tu, S. Cao, D. Zhang-li, X. Lv, H. Peng, Z. Yao, X. Zhang, H. Li, C. yan Li, Z. Zhang, Y. Bai, Y.-T. Liu, A. Xin, N. Lin, K. Yun, L. Gong, J. Chen, Z. Wu, Y. P. Qi, W. Li, Y. Guan, K. Zeng, J. Qi, H. Jin, J. Liu, Y. Gu, Y. Gu, Y. Yao, N. Ding, L. Hou, Z. Liu, B. Xu, J. Tang, J. Li, Kola: Carefully benchmarking world knowledge of large language models, ArXiv abs/2306.09296 (2023). URL: https://api.semanticscholar.org/CorpusID:259165244.

[32] N. Mihindukulasooriya, S. M. Tiwari, C. F. Enguix, K. Lata, Text2kgbench: A benchmark for ontology-driven knowledge graph generation from text, ArXiv abs/2308.02357 (2023). URL: https://api.semanticscholar.org/CorpusID:260611736.

[33] Y. Li, Y. Du, K. Zhou, J. Wang, W. X. Zhao, J. rong Wen, Evaluating object hallucination in large vision-language models, ArXiv abs/2305.10355 (2023). URL: https://api.semanticscholar.org/CorpusID:258740697.

[34] S. Zhang, L. Pan, J. Zhao, W. Y. Wang, Mitigating language model hallucination with interactive question-knowledge alignment, ArXiv abs/2305.13669 (2023). URL: https://api.semanticscholar.org/CorpusID:258840979.

[35] Y. Wang, W. Zhong, L. Li, F. Mi, X. Zeng, W. Huang, L. Shang, X. Jiang, Q. Liu, Aligning large language models with human: A survey, 2023. arXiv:2307.12966.

[36] L. Pan, M. S. Saxon, W. Xu, D. Nathani, X. Wang, W. Y. Wang, Automatically correcting large language models: Surveying the landscape of diverse self-correction strategies, ArXiv abs/2308.03188 (2023). URL: https://api.semanticscholar.org/CorpusID:260682695.

[37] R. Bawden, F. Yvon, Investigating the translation performance of a large multilingual language model: the case of bloom, in: European Association for Machine Translation Conferences/Workshops, 2023. URL: https://api.semanticscholar.org/CorpusID:257353790.

[38] A. Hendy, M. G. Abdelrehim, A. Sharaf, V. Raunak, M. Gabr, H. Matsushita, Y. J. Kim, M. Afify, H. H. Awadalla, How good are gpt models at machine translation? a comprehensive evaluation, ArXiv abs/2302.09210 (2023). URL: https://api.semanticscholar.org/CorpusID:257038384.

[39] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 8440–8451. URL: https://aclanthology.org/2020.acl-main.747. doi:10.18653/v1/2020.acl-main.747.

[40] S. Yuan, M. Färber, Evaluating generative models for graph-to-text generation, ArXiv abs/2307.14712 (2023). URL: https://api.semanticscholar.org/CorpusID:260203094.

[41] Y. Zhu, X. Wang, J. Chen, S. Qiao, Y. Ou, Y. Yao, S. Deng, H. Chen, N. Zhang, Llms for

knowledge graph construction and reasoning: Recent capabilities and future opportunities, ArXiv abs/2305.13168 (2023). URL: https://api.semanticscholar.org/CorpusID:258833039.
[42] H. Liu, C. Li, Q. Wu, Y. J. Lee, Visual instruction tuning, ArXiv abs/2304.08485 (2023). URL: https://api.semanticscholar.org/CorpusID:258179774.
[43] P. Wang, A. Yang, R. Men, J. Lin, S. Bai, Z. Li, J. Ma, C. Zhou, J. Zhou, H. Yang, OFA: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework, in: K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, S. Sabato (Eds.), Proceedings of the 39th International Conference on Machine Learning, volume 162 of *Proceedings of Machine Learning Research*, PMLR, 2022, pp. 23318–23340. URL: https://proceedings.mlr.press/v162/wang22al.html.
[44] A. F. Biten, L. Gómez, D. Karatzas, Let there be a clock on the beach: Reducing object hallucination in image captioning, in: 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2473–2482. doi:10.1109/WACV51458.2022.00253.
[45] S. Petryk, S. Whitehead, J. Gonzalez, T. Darrell, A. Rohrbach, M. Rohrbach, Simple token-level confidence improves caption correctness, ArXiv abs/2305.07021 (2023). URL: https://api.semanticscholar.org/CorpusID:258615698.
[46] M. Ning, Y. Xie, D. Chen, Z. Song, L. Yuan, Y. Tian, Q. Ye, L. Yuan, Album storytelling with iterative story-aware captioning and large language models, ArXiv abs/2305.12943 (2023). URL: https://api.semanticscholar.org/CorpusID:258832908.
[47] R. Mahmood, G. Wang, M. Kalra, P. Yan, Fact-checking of ai-generated reports, ArXiv abs/2307.14634 (2023). URL: https://api.semanticscholar.org/CorpusID:260202943.
[48] W. Yuan, G. Neubig, P. Liu, Bartscore: Evaluating generated text as text generation, Advances in Neural Information Processing Systems 34 (2021) 27263–27277.
[49] J. Fu, S.-K. Ng, Z. Jiang, P. Liu, Gptscore: Evaluate as you desire, arXiv preprint arXiv:2302.04166 (2023).
[50] J. Thorne, A. Vlachos, O. Cocarascu, C. Christodoulopoulos, A. Mittal, The fact extraction and verification (fever) shared task, Training 80 (????) 35–639.
[51] A. Azaria, T. Mitchell, The internal state of an llm knows when its lying, arXiv preprint arXiv:2304.13734 (2023).
[52] S. Kadavath, T. Conerly, A. Askell, T. Henighan, D. Drain, E. Perez, N. Schiefer, Z. Hatfield-Dodds, N. DasSarma, E. Tran-Johnson, et al., Language models (mostly) know what they know, arXiv preprint arXiv:2207.05221 (2022).
[53] P. Manakul, A. Liusie, M. J. Gales, Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models, arXiv preprint arXiv:2303.08896 (2023).
[54] N. Varshney, W. Yao, H. Zhang, J. Chen, D. Yu, A stitch in time saves nine: Detecting and mitigating hallucinations of llms by validating low-confidence generation, arXiv preprint arXiv:2307.03987 (2023).
[55] K. W. Church, Z. Chen, Y. Ma, Emerging trends: A gentle introduction to fine-tuning, Natural Language Engineering 27 (2021) 763–778.
[56] C. Lee, K. Cho, W. Kang, Mixout: Effective regularization to finetune large-scale pretrained language models, arXiv preprint arXiv:1909.11299 (2019).
[57] F. Moiseev, Z. Dong, E. Alfonseca, M. Jaggi, Skill: structured knowledge infusion for large language models, arXiv preprint arXiv:2205.08184 (2022).
[58] Y. Wu, Y. Zhao, B. Hu, P. Minervini, P. Stenetorp, S. Riedel, An efficient memory-augmented

transformer for knowledge-intensive nlp tasks, arXiv preprint arXiv:2210.16773 (2022).
[59] S. Jha, S. K. Jha, P. Lincoln, N. D. Bastian, A. Velasquez, S. Neema, Dehallucinating large language models using formal methods guided iterative prompting, in: 2023 IEEE International Conference on Assured Autonomy (ICAA), IEEE, 2023, pp. 149–152.
[60] J. Luo, C. Xiao, F. Ma, Zero-resource hallucination prevention for large language models, arXiv preprint arXiv:2309.02654 (2023).
[61] P. Feldman, J. R. Foulds, S. Pan, Trapping llm hallucinations using tagged context prompts, arXiv preprint arXiv:2306.06085 (2023).