Creating Trustworthy LLMs: Dealing with Hallucinations in Healthcare AI

Muhammad Aurangzeb Ahmad
Department of Computer Science
University of Washington Bothell
Bothell, WA
maahmad@uw.edu

Ilker Yaramis
ACE AI (of KPInsight)
Kaiser Permanente
Oakland, USA
ilker.yaramis@kp.org

Taposh Dutta Roy
ACE AI (of KPInsight)
Kaiser Permanente
Oakland, USA
taposh.d.roy@kp.org

Abstract—Large language models have proliferated across multiple domains in as short period of time. There is however hesitation in the medical and healthcare domain towards their adoption because of issues like factuality, coherence, and hallucinations. Give the high stakes nature of healthcare, many researchers have even cautioned against its usage until these issues are resolved. The key to the implementation and deployment of LLMs in healthcare is to make these models trustworthy, transparent (as much possible) and explainable. In this paper we describe the key elements in creating reliable, trustworthy, and unbiased models as a necessary condition for their adoption in healthcare. Specifically we focus on the quantification, validation, and mitigation of hallucinations in the context in healthcare. Lastly, we discuss how the future of LLMs in healthcare may look like.

Index Terms—LLM, AI Hallucination, ChatGPT

I. INTRODUCTION

The proliferation of large language models (LLMs) is opening new challenges for their utilization in mission critical and high risk industries like healthcare, law, and behavioral therapy. LLMs have gone from models mainly confined to certain segments of industries to a technology which poised to penetrate almost every industry and domain. Rapid advances in LLMs like like BERT [7], BART [8], GPT-3 [9], GPT-4 [43] and ChatGPT have dazzled the world oft times in generating impressive texts [38], passing bar exams [41], outperforming humans in certain tasks [19]. One the flip-side, these models are also marred by issues like bias, privacy, security, and hallucinations. Like all impactful technologies, LLMs have unintended consequences [42] which are still being explored [38]. Hallucinating incorrect answers can have adverse monetary consequences e.g., when Google Bard hallucination at its launch cost the company $100 Billion [32] and social consequences e.g., when ChatGPT falsely accused professor on sexually assaulting his students [33]. This is one of the reasons why some researchers have recommended that LLMs should not be used in healthcare and medicine [3] [10]. Another group of researchers recommend taking a more cautionary approach [4] towards adoption of LLMs. In this paper we argue that caution in warranted in the use of LLMs in healthcare, not using these models robs us for potential to address some pressing issues in healthcare AI. Instead, appropriate guardrails need to be in place before these models can live up to their true potential.

Creating trustworthy LLMs requires adhering to the principles of Responsible AI which requires ensuring that AI system are transparent, unbiased, accurate, interpretable, ensure privacy and security. In this paper we focus on AI hallucinations since they affect all of the elements of responsible AI and have been described as the most fundamental impediment in the widespread adoption on LLMs [4]. There are multiple competing and overlapping definitions of AI hallucinations depending upon the context. In general AI hallucinations refer to outputs from a LLM hat are contextually implausible [12], inconsistent with the real world and unfaithful to the input [13]. Some researchers have argued that the use of the term hallucination is a misnomer, it would be more accurate to describe AI Hallucinations as fabrications [3].

In an ideal world LLM models would not produce any hallucinations. However, given how tokens are generated by LLMs, hallucinations are an inevitable end result of token generation in LLMs [37]. OpenAI, the creator of GPT-4, acknowledges on it website [37] hallucinations as a core limitation of LLMs. A recent study on LLMs for summarization demonstrated that hallucinated content was 25% of their generated summaries [17]. In healthcare hallucinations can be specially problematic since the misinformation generated by LLMs could be related to diagnosis, treatment, or a recommended procedure. Training LLMs in the wild introduces additional issues like bias, non-factual data in addition to hallucinations. Thus, there are also calls for regulating the use of LLMs in healthcare [34]. AI Hallucinations happen in pretty much every modality where LLMs are applied e.g., text to text generation, text to image generation, text to video generation etc. While we discuss multiple modalities in this paper, we mainly focus on text related hallucinations in the application of AI.

II. LLMs IN HEALTHCARE

Popular LLMs like GPT-3, GPT-4, and ChatGPT have been used for multiple applications in healthcare. From an application perspective LLMs in healthcare have been explored for facilitating clinical workflow (e.g., recommendation, write discharge summary etc). translation, triage (guide patients to the right department, medical research (medical writing,

anonymization etc), medical education (compose medical questions) Even though these models have been trained on general data, they perform relatively well on certain healthcare and medical tasks [42]. However for certain specialized tasks these models perform abysmally or generate incorrect information. Researchers have suggested [6] that this limitation can be overcome by using datasets from healthcare domains e.g., LLMs trained using EHR data like GatorTron [1] and [2], or LLMs trained using medical datasets like Med-PaLM 2 [6] and Flan-PaLM [5]. Hospital systems and healthcare organizations are still hesitant to employ LLMs in production because of high cost and liability associated with getting the questions wrong.

While the text generated by LLMs can be quite good often times, it also has the tendency to regurgitate unreliable information that is present on online resources. [19] observed that LLMs are likely to generate false yet widely circulated information. In retrospect this is not surprising given the training data that is used to train LLMs. However, this becomes especially problematic in the healthcare domain when LLMs give health related advice or information about the medical domain. LLMs have also been used to see if they can pass medical exams in various locales where these models have passed the exams with relatively good performance e.g., United States Medical Licensing Examinations (USMLE) [25], Chinese National Medical Licensing Examination [22], and Japanese national medical licensing examinations [23].

III. HALLUCINATIONS IN HEALTHCARE AI

Hallucinations could be generated because of a variety of reasons. Here are some prominent sources of hallucinations in the context of healthcare.

- **Unreliable Sources:** If the data comes from a general source then it is likely to perpetuate commonly held misconceptions [16]
- **Probabilistic Generation:** Given the probabilistic nature of text generation, recombination of completely reliable texts can still lead to generation of false statements [10]
- Biased Training Data: Sources that may be biased may lead to generating hallucinations [24]
- **Insufficient Context:** Text generated by LLMs is based on prompts. Lack of context could lead to text generation with little or no correspondence to what the end user is looking for. [10]
- **Self-Contradictions:** LLMs are not good at sequential reasoning. This may lead to self-contradictions [16].

IV. ADDRESSING THE PROBLEM OF HALLUCINATIONS

Model hallucination offer a number of impediments when it comes to the usage of these models in healthcare, as described in the previous section. These obstacles are however not insurmountable and can be overcome. To address these problems one first needs to measure, evaluate, and then mitigate hallucinations.

A. Evaluating Hallucinations

Model evaluation for LLMs can be divided into two main scenarios, whether one has access to the model itself or alternatively merely to the outputs of the model. In the first case one can check the likelihood of generation of the text against the distribution of the corpus that was used to train the model. This is the scenario where the organization itself trains and builds the model using its own data. In the context of healthcare, one cannot always assume access to the model as organizations may be using off the shelf LLM models like ChatGPT, GPT-3 or GPT-4. For such scenarios self-check by the LLMs [16]. The main idea is that when the response from the LLM is sampled multiple times for a given concept, the responses are likely to diverge or even contradict one another in case of hallucinations [16].

In some healthcare applications interpretability and transparency of the models would be crucial [40] for the use case i.e., how did the LLM come up with a particular output? The LLM could also be asked how it come up with a certain conclusion. This can help establish the correctness and interpretation of the output [21]. Evaluation of many LLMs in healthcare and medicine have focused on gauging performance against benchmark multiple-choice-question-(MCQ). The problem with such evaluation is that does not mimic real world use cases. This is illustrated by Hickam's Dictum which is a counterargument to the use of Occam's razor which states that that any given time there are multiple possible diagnosis given a set of symptoms [44]. Consequently, one can argue that LLM benchmarks are insufficient because in the real-world physicians rarely have one diagnosis for a condition.

B. Measuring Hallucinations

The text generated by large language models is open-ended in nature and thus traditional metrics of evaluating machine learning models cannot be readily applied. Method used for evaluating hallucinations can be divided into two main categories: Human evaluation and automated evaluation.

1) Human Evaluation: These evaluation involve standardized ways to evaluate outputs from LLMs by manual human annotation [18]. [19] propose a benchmark where the answers from a Q& A system are evaluated by human annotators. FActScores [29] evaluates text generated by LLMs via an evaluation method that breaks a generation into atomic facts which are in turn evaluated by human evaluators. Majority voting for evaluating healthcare related answers generated by LLMs has been employed for myopia care [21], maternity [21], diabetes [26], cancer [27], infant care [21] etc.

2) Automatic Evaluation: A number of automatic evaluation methods use human annotations as inputs for evaluation e.g., [19] uses human evaluation to determine the truth claims of generated answers, [28] describe a scoring function that computes information alignment between two arbitrary texts. Another scheme is to use external models to evaluate LLMs, a strategy adopted by FactScore [29].

Area	Example
Text Summarization	Generate summary of medical records or medical history
Annonymization	Annonymize patient data for privacy and HIPPA compliance
Clinical Documentation	Generate patient discharge reports
Translation	Translate patient records from one language to another. Also, translate relevant medical information to other languages for patients
Medical Writing	Facilitate medical research by assisting in medical writing and research
Triage	Guide patients to the right wards or departments
Consultation	Recommendations for patient self care

TABLE I
APPLICATIONS OF LLMs IN HEALTHCARE

3) Evaluation Metrics: Some researchers use traditional classification metrics like precision, recall, and F-Score for reporting model performance for scenarios where the output from the models is either true or false [31]. Different tasks have different metrics e.g., Perplexity and Cross-Entropy Loss for language models [16], ROGUE for summarization [30], BLEU or Meteor for machine translation [30]. While these metrics do not directly measure hallucinations, a low score can be an indicator that the output is likely to have hallucinations.

C. Mitigating Hallucinations

Since sources of hallucinations can creep in any part of the life cycle of LLMs and the definition of what constitutes a hallucination can also vary, mitigation of Hallucinations in healthcare is a multi-pronged process. Some of the important strategies for mitigating hallucination as it pertain to healthcare use cases are as follows:

1) Human-In-The-Loop (HITL): Having humans with domain expertise in various parts of the model development process can greatly help in reducing hallucinations downstream. This includes human is data annotation, data classification, oversight committees for auditing model outputs, suggestions corrections whenever needed etc. Some researchers have also suggested [21] real-time supervision of LLMs where a human in the loop oversees the outputs of the LLM. While this may be necessary for certain high stakes use cases in healthcare it is not scalable as it is not feasible for even group of humans to do real-time check of thousands of outputs. An alternative way to address is for the LLM to flag outputs for which it is not confident and defer to human judgment. The input from humans can in turn be source of additional training data which would improve the model further. Lastly, the end user feedback can be incorporated on the LLM response e.g., how ChatGPT has the thumbs up/thumbs down option on generated responses.

2) Algorithmic Corrections: Traditional machine learning techniques like using regularization, adding penalty to loss function, can also be employed. This would make the LLMs generalize better and make them less likely to hallucinate [16].

3) Fine Tuning: Fine tuning LLMs requires adapting the LLM t specific tasks or domains. It may involve adapting the LLM to specific tasks or domains. The drawback of fine-tuning is that it is computationally expensive an may be expensive from a monetary perspective. The promise of fine-tuning in healthcare is that a model trained on data specific to healthcare is less likely to make up health related information. However,

in practice this is not always true as researchers have noted that fine-tuning does not guarantee improvement in performance as there are examples to the contrary which have been reported in literature [39].

4) Improving Prompts: In cases where the model is not confident about the output and is likely to be hallucination then the LLM could output "I do not know" given an input prompt [16].

D. Adversarial Training

Since LLMs can be exploited by adversarial attacks, one way to mitigate for such attacks is intentionally exposed to adversarial examples during training to ensure that these models are not compromised [11].

E. Input Validation

Checking the inputs for validity can also mitigate against hallucinations. This could be done by checking the input against known standards, or running the input through a separate model designed to detect adversarial inputs [11].

F. Memory Augmentation

An external knowledge source can be encoded into a key-value memory which can be integrated with the LLM [15]. However, this technique has only been demonstrated for relatively small medals like T5. While researcher have suggested that it may be possible to enhance the performance of LLMs, it has not been explored because of large memory requirements [14].

G. Model Choice

Ever since the public release of GPT-2 a large number of LLMs have been publicly released. The performance of these models varies with respect to different aspects of AI hallucinations as described above: [31] observed that ChatGPT and GPT-4 are much better at catching self-contradictions as compared to Vicuna-13B.

H. Benchmark Audits

The work on benchmarks described in the evaluation section assumes the veracity of benchmarks. In their detailed study of knowledge-grounded conversational benchmarks [36] found that the benchmarks themselves include incorrect information which in turn leads to hallucinations. More disturbing is the fact that if the dataset contains even a small number of hallucinations then this may shift the data distribution in a manner that is likely to generate a greater number of

hallucinations. What this alludes to is that benchmarks used for testing AI systems in healthcare should be verified by human domain experts to ensure veracity. Additionally, any model generated content should only be added to the training set only after it has been scrutinized by multiple domain experts.

V. CONCLUSION

Although LLMs are increasingly being used on health-care the community as a whole is cautious about their use because of certain limitations, the foremost among these is model hallucination. LMs may generate plausible-sounding but incorrect or misleading information, making it crucial for healthcare professionals to critically evaluate and validate the outputs. LLMs excel at natural language processing tasks, enabling them to analyze vast amounts of medical literature, patient records, and clinical notes. LLMs can assist in medical diagnosis by analyzing symptoms, predicting potential diseases, suggesting treatment options, and aiding in personalized medicine.

Even for benchmarks where these models have shown to excel, researchers argue [24] that framing of medical knowledge as narrow set of options or multiple choice questions creates a framing of false certainty and thus not a true representation of how medicine is practiced. Even though the process of evaluating hallucinations in LLMs is still being standardized, tools for detecting hallucinations are already being released like Nvidia's NeMo Guardrails [38]. In healthcare Human-in-the-loop systems for building and validation of LLMs for high stakes tasks may be a necessity for the foreseeable future given the high-risk nature of the healthcare domain. For low stakes tasks automation may be achieveable with appropriate guardrails in place. Lastly, widespread adoption of LLMs will also have to overcome possible regulatory issues in the near future.

REFERENCES

[1] Yang, X., Chen, A., PourNejatian, N., Shin, H., Smith, K., Parisien, C., Compas, C. & Others Gatortron: A large clinical language model to unlock patient information from unstructured electronic health records. *ArXiv Preprint arXiv:2203.03540*. (2022)

[2] Kraljevic, Z., Bean, D., Shek, A., Bendayan, R., Hemingway, H. & Au, J. Foresight-Generative Pretrained Transformer (GPT) for Modelling of Patient Timelines using EHRs.

[3] Emsley, R. ChatGPT: these are not hallucinations–they're fabrications and falsifications. *Schizophrenia*. **9**, 52 (2023)

[4] Alkaissi, H. & McFarlane, S. Artificial hallucinations in ChatGPT: implications in scientific writing. *Cureus*. **15** (2023)

[5] Chung, H., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, E., Wang, X., Dehghani, M., Brahma, S. & Others Scaling instruction-finetuned language models. *ArXiv Preprint arXiv:2210.11416*. (2022)

[6] Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Hou, L., Clark, K., Pfohl, S., Cole-Lewis, H., Neal, D. & Others Towards expert-level medical question answering with large language models. *ArXiv Preprint arXiv:2305.09617*. (2023)

[7] Devlin, J., Chang, M., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv Preprint arXiv:1810.04805*. (2018)

[8] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V. & Zettlemoyer, L. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *ArXiv Preprint arXiv:1910.13461*. (2019)

[9] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. & Others Language models are few-shot learners. *Advances In Neural Information Processing Systems*. **33** pp. 1877-1901 (2020)

[10] Salvagno, M., Taccone, F. & Gerli, A. Artificial intelligence hallucinations. *Critical Care*. **27**, 1-2 (2023)

[11] Kumar, A., Agarwal, C., Srinivas, S., Feizi, S. & Lakkaraju, H. Certifying LLM Safety against Adversarial Prompting. *ArXiv Preprint arXiv:2309.02705*. (2023)

[12] Lee, P., Bubeck, S. & Petro, J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *New England Journal Of Medicine*. **388**, 1233-1239 (2023)

[13] Lee, M. A mathematical investigation of hallucination and creativity in gpt models. *Mathematics*. **11**, 2320 (2023)

[14] Jha, S., Jha, S., Lincoln, P., Bastian, N., Velasquez, A. & Neema, S. Dehallucinating large language models using formal methods guided iterative prompting. *2023 IEEE International Conference On Assured Autonomy (ICAA)*. pp. 149-152 (2023)

[15] Wu, Y., Zhao, Y., Hu, B., Minervini, P., Stenetorp, P. & Riedel, S. An efficient memory-augmented transformer for knowledge-intensive nlp tasks. *ArXiv Preprint arXiv:2210.16773*. (2022)

[16] Manakul, P., Liusie, A. & Gales, M. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *ArXiv Preprint arXiv:2303.08896*. (2023)

[17] Falke, T., Ribeiro, L., Utama, P., Dagan, I. & Gurevych, I. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. *Proceedings Of The 57th Annual Meeting Of The Association For Computational Linguistics*. pp. 2214-2220 (2019)

[18] Lee, N., Ping, W., Xu, P., Patwary, M., Fung, P., Shoeybi, M. & Catanzaro, B. Factuality enhanced language models for open-ended text generation. *Advances In Neural Information Processing Systems*. **35** pp. 34586-34599 (2022)

[19] Lin, S., Hilton, J. & Evans, O. Truthfulqa: Measuring how models mimic human falsehoods. *ArXiv Preprint arXiv:2109.07958*. (2021)

[20] Min, S., Krishna, K., Lyu, X., Lewis, M., Yih, W., Koh, P., Iyyer, M., Zettlemoyer, L. & Hajishirzi, H. FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation. *ArXiv Preprint arXiv:2305.14251*. (2023)

[21] Lim, Z., Pushpanathan, K., Yew, S., Lai, Y., Sun, C., Lam, J., Chen, D., Goh, J., Tan, M., Sheng, B. & Others Benchmarking large language models' performances for myopia care: a comparative analysis of ChatGPT-3.5, ChatGPT-4.0, and Google Bard. *Ebiomedicine*. **95** (2023)

[22] Liu, J., Zhou, P., Hua, Y., Chong, D., Tian, Z., Liu, A., Wang, H., You, C., Guo, Z., Zhu, L. & Others Benchmarking Large Language Models on CMExam–A Comprehensive Chinese Medical Exam Dataset. *ArXiv Preprint arXiv:2306.03030*. (2023)

[23] Kasai, J., Kasai, Y., Sakaguchi, K., Yamada, Y. & Radev, D. Evaluating gpt-4 and chatgpt on japanese medical licensing examinations. *ArXiv Preprint arXiv:2303.18027*. (2023)

[24] Mbakwe, A., Lourentzou, I., Celi, L., Mechanic, O. & Dagan, A. ChatGPT passing USMLE shines a spotlight on the flaws of medical education. *PLOS Digital Health*. **2** pp. e0000205 (2023)

[25] Kung, T., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., Madriaga, M., Aggabao, R., Diaz-Candido, G., Maningo, J. & Others Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLoS Digital Health*. **2**, e0000198 (2023)

[26] Hulman, A., Dollerup, O., Mortensen, J., Fenech, M., Norman, K., Støvring, H. & Hansen, T. ChatGPT-versus human-generated answers to frequently asked questions about diabetes: A Turing test-inspired survey among employees of a Danish diabetes center. *Plos One*. **18**, e0290773 (2023)

[27] Chen, S., Kann, B., Foote, M., Aerts, H., Savova, G., Mak, R. & Bitterman, D. The utility of ChatGPT for cancer treatment information. *MedRxiv*. pp. 2023-03 (2023)

[28] Zha, Y., Yang, Y., Li, R. & Hu, Z. AlignScore: Evaluating Factual Consistency with a Unified Alignment Function. *ArXiv Preprint arXiv:2305.16739*. (2023)

[29] Min, S., Krishna, K., Lyu, X., Lewis, M., Yih, W., Koh, P., Iyyer, M., Zettlemoyer, L. & Hajishirzi, H. FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation. *ArXiv Preprint arXiv:2305.14251*. (2023)

[30] Afzal, A., Vladika, J., Braun, D. & Matthes, F. Challenges in Domain-Specific Abstractive Summarization and How to Overcome Them. *15th International Conference On Agents And Artificial Intelligence, ICAART 2023*. pp. 682-689 (2023)

[31] Mündler, N., He, J., Jenko, S. & Vechev, M. Self-contradictory Hallucinations of Large Language Models: Evaluation, Detection and Mitigation. *ArXiv Preprint arXiv:2305.15852*. (2023)

[32] Griffin, A. Google shares plunge after its new ChatGPT AI competitor gives wrong answer to question. (2023), https://www.independent.co.uk/tech/google-ai-bard-chatgpt-shares-b2278932.html, Accessed on: August 28, 2023

[33] Turley, J. ChatGPT falsely accused me of sexually harassing my students. Can we really trust AI?. *USA Today*. (2023)

[34] Meskó, B. & Topol, E. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *Npj Digital Medicine*. **6**, 120 (2023)

[35] Stern, K., Qiu, Q., Weykamp, M., O'Keefe, G. & Brakenridge, S. Defining posttraumatic sepsis for population-level research. *JAMA Network Open*. **6**, e2251445-e2251445 (2023)

[36] Dziri, N., Milton, S., Yu, M., Zaiane, O. & Reddy, S. On the origin of hallucinations in conversational models: Is it the datasets or the models?. *ArXiv Preprint arXiv:2204.07931*. (2022)

[37] OpenAI OpenAI: GPT-4. (https://openai.com/research/gpt-4), Accessed: August 18, 2023

[38] Shen, X., Chen, Z., Backes, M., Shen, Y. & Zhang, Y. " Do Anything Now": Characterizing and Evaluating In-The-Wild Jailbreak Prompts on Large Language Models. *ArXiv Preprint arXiv:2308.03825*. (2023)

[39] Wang, Y., Si, S., Li, D., Lukasik, M., Yu, F., Hsieh, C., Dhillon, I. & Kumar, S. Preserving In-Context Learning ability in Large Language Model Fine-tuning. *ArXiv Preprint arXiv:2211.00635*. (2022)

[40] Ahmad, M., Eckert, C. & Teredesai, A. Interpretable machine learning in healthcare. *Proceedings Of The 2018 ACM International Conference On Bioinformatics, Computational Biology, And Health Informatics*. pp. 559-560 (2018)

[41] Choi, J., Hickman, K., Monahan, A. & Schwarcz, D. Chatgpt goes to law school. *Available At SSRN*. (2023)

[42] Li, J., Dada, A., Kleesiek, J. & Egger, J. ChatGPT in Healthcare: A Taxonomy and Systematic Review. *MedRxiv*. pp. 2023-03 (2023)

[43] Sanderson, K. GPT-4 is here: what scientists think. *Nature*. **615**, 773 (2023)

[44] Borden, N. & Linklater, D. Hickam's dictum. *Western Journal Of Emergency Medicine: Integrating Emergency Care With Population Health*. **14** (2013)