DOI 10.31891/2307-5732-2024-343-6-38
УДК 004.8.32

SAVELIEV ROMAN
Ukrainian National Forestry University
https://orcid.org/0009-0007-8092-0673
e-mail: romansavelyev6@gmail.com
MYKHAILO DENDIUK
Ukrainian National Forestry University
https://orcid.org/0000-0002-7631-022X
e-mail: dendyuk@nltu.edu.ua

DETECTING GENERATIVE AI HALLUCINATIONS IN ANALYTIC DATASETS OF SOFTWARE SYSTEMS

Generative AI in software system analytics included progress and a threat, hallucinations, as a challenge. Such outputs, though plausible, do not bear reality to a significant extent and pose a challenge in relying on AI-generated insights. These models, which provide seamless intelligence and automation, are often deep learning-based and need more common sense and contextual awareness possessed by humans.

These models are developed on extensive data and can spot patterns but are prone to misfitting spurious relations, leading to erroneous perspectives. Numerous reasons are responsible for these hallucinations. Their models can easily be steered incorrectly due to a biased or incomplete training input. If too much focus is applied to the training inputs, generalization becomes a problem, and neuroscientists suggest more new ideas increase the risk of hallucinations. Culprits of the above challenges are also concepts of neural architecture – our AI models – that try to emulate how the human brain works with maths instead of truly understanding the problem.

Dealing with this requires a clear strategy. First and foremost, data is the core of any AI application, and at this stage, data cleansing, bias detection, and representativeness are powerful tools. It's also important to select an appropriate model architecture during the model's training.

Although detection and mitigation are not the same, they are equally important. One of the anomaly detection algorithms could flag unusual outputs, while illogical conclusions could be avoided by employing certain domain-specific rules as a `sanity check.' AI ensemble models have variety, so risks are reduced. Human intervention is still necessary, but AI insights may be further verified by domain specialists, and minor bugs could be detected.

A potentially interesting direction is transitioning from conventional LLMs to RAG. RAG models rely on external content, so their conclusions are based on factual information and are less likely to make things up. "Self-RAG" proposes a strategy that goes one step further—models would be able to verify themselves by looking up external content.

Keywords: Generative AI, AI hallucinations, data analytics, software systems, error detection.

САВЕЛЬЄВ РОМАН, ДЕНДЮК МИХАЙЛО
Національний лісотехнічний університет України

ВИЯВЛЕННЯ ГАЛЮЦИНАЦІЙ ШТУЧНОГО ІНТЕЛЕКТУ В АНАЛІЗІ АНАЛІТИЧНИХ НАБОРІВ ДАНИХ СИСТЕМ ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ

Зростання використання генеративного ШІ в системах програмного забезпечення, особливо для аналізу даних, призвело як до позитивного прогресу, так і до негативних наслідків. Такі можливості, як отримання інсайтів, прогнозування та моделювання процесів прийняття рішень, є гарними прикладами того, як автоматизація підвищує інтелектуальність та структурованість процесу аналізу даних. Проте, стабільне використання даних моделей несе в собі одну з пасток - ризики галюцинацій ШІ. У цьому випадку галюцинації стосуються методів генеративного ШІ, коли результати, отримані ШІ, не корелюють з наявними даними, а навпаки, містять різноманітні вигадані, неіснуючі деталі.

Для боротьби з цією проблемою надзвичайно важливо виявляти такі аномалії в аналітичних наборах даних в системах програмного забезпечення, щоб гарантувати достовірність інсайтів, що надаються інтерактивними підходами ШІ. Ця робота зосереджена на проблемах виявлення галюцинацій генеративного ШІ в аналітичних наборах даних та їхній класифікації. Дані дослідження є наджзвичайно важливими в аналітичних системах, оскільки якщо ігнорувати галюцинації ШІ, то вони можуть поширюватися і призводити до неправильних майбутніх рішень в аналізі.

Ключові слова: генеративний ШІ, галюцинації ШІ, аналіз даних, системи програмного забезпечення, виявлення помилок

Problem statement

Generative AI hallucinations are among the major issues in data analytics, especially when deployed in SaaS models that deal with continuing integration of data storage from a single application's perspective. Such hallucinations happen when the models employed within the realm of generative AI aim to understand existing input data and create new architecture based on that data, generating content that is gender inappropriate or tangentially relevant to the input. This, in particular, can be very dangerous within analytic datasets, which require a great deal of diligence and prudence regarding the integrity of insights and decisions. In a distributed software system where there are multiple repositories of data and where data is integrated, hallucinations are primarily sensed only after some operational severe failures have occurred as a result of those hallucinations.

From what is known, numerous reasons contribute to hallucination in the case of generative AI models:

Data disturbance: Distributed software systems usually integrate data from multiple and remote locations. Sometimes, the data contains unwanted or missing values, which leads to the AI model hallucinating different outputs as it tries to fill in the gaps at hand.

Generative Mechanisms: Generative models for data reconstruction can fit cases of overfitting, where some models try to harden over existing patterns in the straining datasets as they stray away from the patterns and look at the data using new or other ways. This way, however, promises content generation, which contradicts the actual data trends.

In ambiguous contexts or with bias in the inputs: The inclusion of activity or unnecessary bias in the activity inputs could or may lead to appendices where the AI is misinterpreted, resulting in certain outputs that have data legitimacy but are not inaccuracy-backed by the data.

Integrative and Transdisciplinary System: It is common for the data to break into pieces and be sent across decentralized sources of information with different qualities. This complexity increases the chances of hallucinations as the AI cannot combine different datasets and process them correctly. Reduced validation from humans: In most instances, AI systems have been installed in areas with low interaction or, at best, validation by human beings. Owing to this absence of regularity, the lack of human intervention-induced hallucinations and conclusions are detestably arrived at.

As far as there is a concern regarding the figurative nature of the sentences, the effects appear to be limited to the impacts of the attacks on the target. For example, imagine that in business settings, one AI insight with regard to the organization can trigger a downturn of the business as that may be inappropriate targeting generation, which loses finances and doesn't optimize engagement of the relevant factors. Such blunders, of course, lead to hallucinations in AI-driven systems, but in such safety-critical systems, that might be a severe disadvantage and could lead to a complete breakdown of the system or operation under hazardous conditions.

The most important hurdle is focusing on the generated messages, the hallucinations, and the other forms of misbehaving and creating operational plans to avoid them. Existing solutions like the applications of patterns and cross-validation have worked to some extent, but given how fast AI models are developing and the increasing sophistication of distributed systems, more dynamic and heat-resistant detection approaches would be essential. If there are no enhancements and further research in this area, generative AI applications will continue to be filled with errors, which will lead to decreased reliance on AI solutions overall.

Analysis of Recent Sources

The most recent developments in academic scholarship have addressed generative AI hallucination risks and how to prevent them in detail, particularly in the case of analytic datasets. Every source in such a case explores a different part of the problem, from data quality to hybrid AI systems.

In [1], Lauria makes quite a strong point about the necessity of human control concerning AI-generated outputs, especially in a sphere that derives content. She believes that all AI systems should be under human statute because if not, AI systems tend to make mistakes, such as generating hallucinations when it comes to scenarios where models have to guess or generalize a pattern. Although this perspective is aimed at more artistic applications or utilizing more creative inputs, it holds to the data analytics case where human management protects the validity of AI insights from untamed human-AI interactions.

Foster [3] also discusses the properties of generative AI models, including GANs and VAEs, and illustrates the reasons for the appearance of hallucinations. He notes that these models are mostly very likely to 'supply'- if looking for preferable phrasing, "fake" data when asked for information that they do not possess, which makes designing analytic software systems hard. Instead of being passive, he believes that in his work, it would be necessary to maintain the high quality of training data and periodically update the models through retraining.

With the dangers posed by using old models, Clinton [4] argues that ongoing revisions with new information are important to curb hallucinations since older models are likely to produce false information. In this context, it is also argued that such model regularity and currency, coupled with AI explainability, are adequate measures in ensuring the correctness of AI outputs.

Bahree [5] suggests measures to alleviate unemployment, proposing using a general determinant of AI and a specific one simultaneously. This method checks whether the output is valid against known limits as well as lessening the chances of hallucinations. In addition, he recommends the use of outlier detection algorithms to eliminate outputs that tend to be off-pattern; this works in protecting the data from contaminations. All these sources support the importance of high-quality data, frequent retraining of the model, and practical and extensive validation in preventing any chances of generative AI hallucination in the analytical processes of software system datasets.

Main material

One of the hurdles in addressing AI hallucination problems, especially in the domain of complex software system data is that it is more complicated than employing single anomaly detection techniques. It calls for a more comprehensive solution that takes into account both the nature of the data in question as well as the complicated AI models that are providing that insight.

The major issue is the maintenance of acceptable data standards. The upholding of quality data, which deals with the processes of getting and handling the datasets, needs to show data cleansing that finds and repairs places with residents that are wrong, unwarranted, or absent and performing the data quality rules within the sourced data. For

instance, knowing that the data in the distributed systems has many different origins, the data or cross-validation approaches are useful. This means that instead of treating each data-related source as distinct entities and events and thus not conducting an analysis of the patterns that could warrant hallucinations, the reverse holds. Every bit of retaining clear data lineage is equally as critical because it encompasses information as to where and what could be popping out that could be the cause or the implicated factors of the hallucinations.

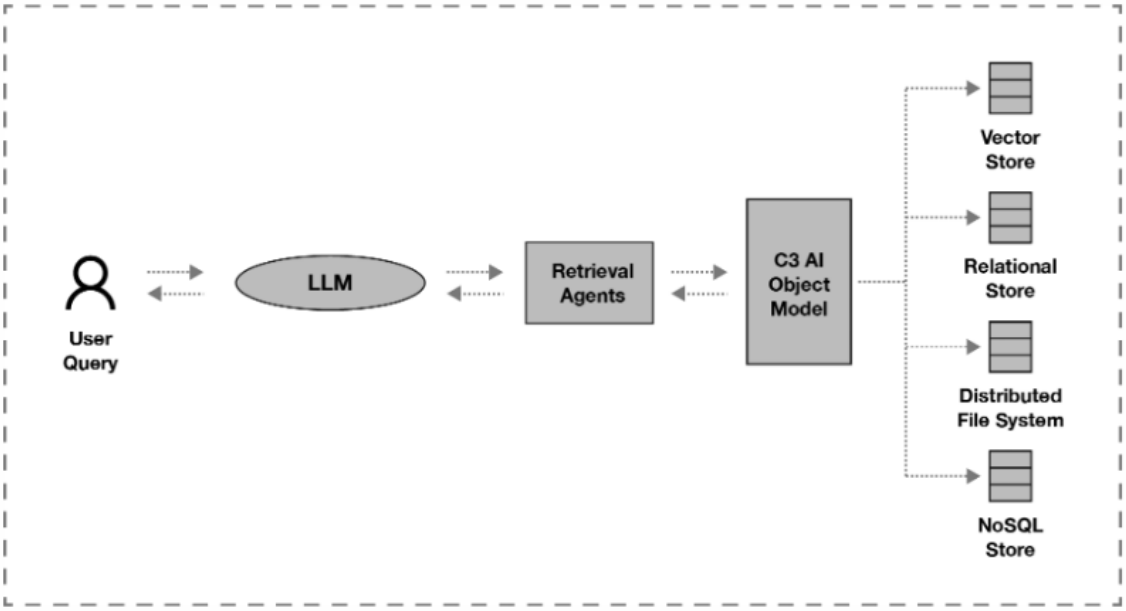

Fig. 1. Reference Architecture for the Pipeline of Solving Generative AI Hallucinations

However, it can be stated that the data is not enough on its own, and models explain AI in some other ways as well. It should be emphasized that model training can be performed with several representative datasets with sufficient volume, and the model should be retrained as new data appears. Due to this continuous learning and updating, the model has less chance to overfit and expands the limits of model comprehension, thus preserving it from biases and localism-fostering hallucinations. XAI methods support this proposition even more. The use of XAI helps interpret the findings of AI, which makes it possible for specialists to explain the emergence of some insights, uncover biases, and evaluate the rationale of the produced results.

Even where accurate data has been provided, and models adopt a transparent approach in their explanation, a resilient system will still take advantage of various AI approaches. Finally, and as a last line of defense, sophisticated anomaly detection mechanisms specifically designed for the data and context are essential. Such algorithms are capable of deliberating past data and figuring out even slight intervening or instigating features that can be viewed as a hallucination when no laws are broken.

You are attempting to explain to a robot with wheels how to find its way around a city. You have handed it a map and a set of hours of steering the car in the city. But there will always be a margin of error in the decisions it would make, right? This 'error' in what the car comes to 'know' has two main reasons:

First missing information (Type 1 Uncertainty): The vehicle may be in doubt about a particular intersection because the information you shared caused a lack of what happens there. Maybe it was because there was construction, or it was always dark when capturing data. This absence of specific information is why the car in that area clusters possibilities and becomes less confident in its decisions.

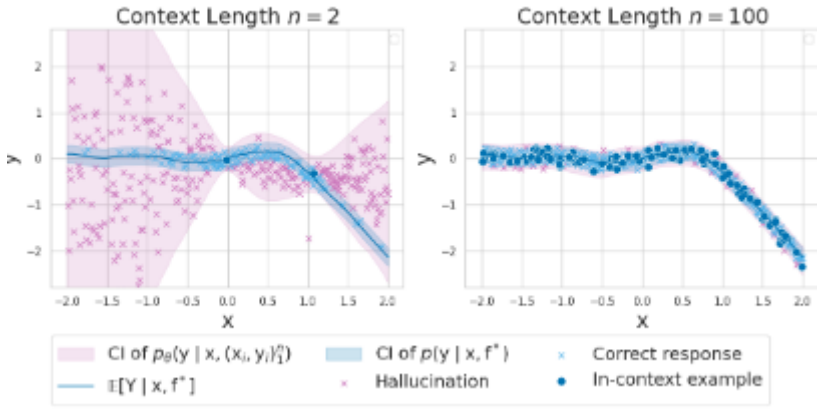

Fig. 2. Type I epistemic uncertainty

Second, but by no means the last, doubts arising from the uncertainty about the rules themselves (Type 2 Uncertainty): Even when provided with absolutely all the data, there still remains the possibility that the automobile might contain specific complete data but absolutely ignore the fundamental principles of the driving mechanics.

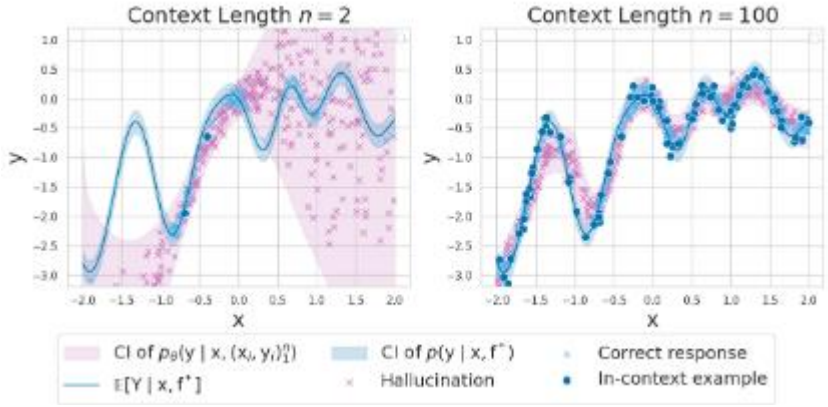

Fig. 3. Type II epistemic uncertainty

Different studies have shown that hallucinations are rarely quantified as commonplace occurrences in human interaction, but such problems are not complex to fix. There is a precise and wide-reaching plan for the hallucinations that is provided, which includes explanations of barriers, various management measures, and the so-called management measures, and so-called "lessons learned" afterward:

Phase 1: Preparing the Bases

- Preparation and Prevention
- Quality of Data
- Choosing a Model and Its Training

Phase 2: Focusing the attention

- Hallucinations Detecting
- Formulate Normal Established Baselines
- Multi-Pronged Attack

Phase 3: Appraising the Damage

- Quantification and Analysis
- Construct a Hallucinations Severity Spectrum
- Collect the Occurrence and Grade of the hallucinatory phenomena.
- Make circulars and attend to them.

Ultimately, it is more complex than that in a generative AI hallucination and software system analytics. From the perspective of data, models, hybrid AI systems, and human supervision, we can dream of a time when AI-powered insights will be not only powerful but also trustworthy.

Conclusions

With the rise of generative AI to tackle the inherent complexity of extracting actionable insights from the data residing in complex software systems, hopefully, attention will be paid to the problem of hallucinations, which are always pretty plausible but made up. This investigation pointed to issues of greater complexity and breadth and called for an integrated response to alleviate such issues.

However, data protection is a first priority. Good quality, strong cleansing, cross-source verification, and good provenance are a must. In addition, concepts such as “Give Me An A in Safety!” must be part of the learning process to ensure reliable model builds with the least possible output of junk.

Still, even with these conditions satisfied, a multi-level defense is critical. GenAI models and deterministic approaches are used together to form a hybrid AI model, which can also be used to verify and validate the outputs. In addition, specific data and domain-focused detection of outliers methods automatically add another layer of detection by recognizing evidence of hallucination where it may be hidden.

People are the final decision-makers, as one would expect. A new generation of professionals who manage AI systems and interact with end users of the generated content will not just have the opportunity to apply many feedback cycles to AI systems.

In conclusion, the concerns of generative AI hallucinations need further action grounded in data, model enhancements, multiple detection models, and, importantly, human judgment. This holistic approach will enable us to envisage an era where AI can generate qualitatively better insights and be of much more value.

References

1. Lauria, A. The Equalizing Quill: How AI Will Liberate Content Creators and Transform Authorship / A. Lauria. – Amazon Digital Services LLC, 2023.

2. Saveliev R., Dendiuk M. Generative AI Methods in Natural Language Understanding; Dnipro, Ukrayina: IV Mizhnarodna naukova konferentsiya «Innovatsiyni tendentsiyi syohodennya v sferi pryrodnychykh, humanitarnykh ta tochnykh nauk» / R. Saveliev, M. Dendiuk. – 12.04.2024.

3. Foster, D. Generative Deep Learning: Teaching Machines to Paint, Write, Compose, and Play / D. Foster. – O'Reilly Media, 2019.

4. Clinton, D. The Complete Obsolete Guide to Generative AI / D. Clinton. – Independently Published, 2023.

5. Bahree, A. Generative AI in Action / A. Bahree. – 2023.

6. Chollet, F. Deep Learning with Python / F. Chollet. – Manning Publications, 2017.

7. Goodfellow I., Bengio Y., Courville A. Deep Learning / I. Goodfellow, Y. Bengio, A. Courville. – MIT Press, 2016.

8. Molnar, C. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable / C. Molnar. – Leanpub, 2022.

9. Jackson, P. C. The AI Revolution in Medicine: GPT-4 and Beyond / P. C. Jackson. – 2023.

10. Marcus, G., Davis, E. Rebooting AI: Building Artificial Intelligence We Can Trust / G. Marcus, E. Davis. – Vintage, 2023.