# The Hallucinations of a Surrealist: When AI Goes Awry

NIKOLAS ABERLE, Texas A&M University, USA

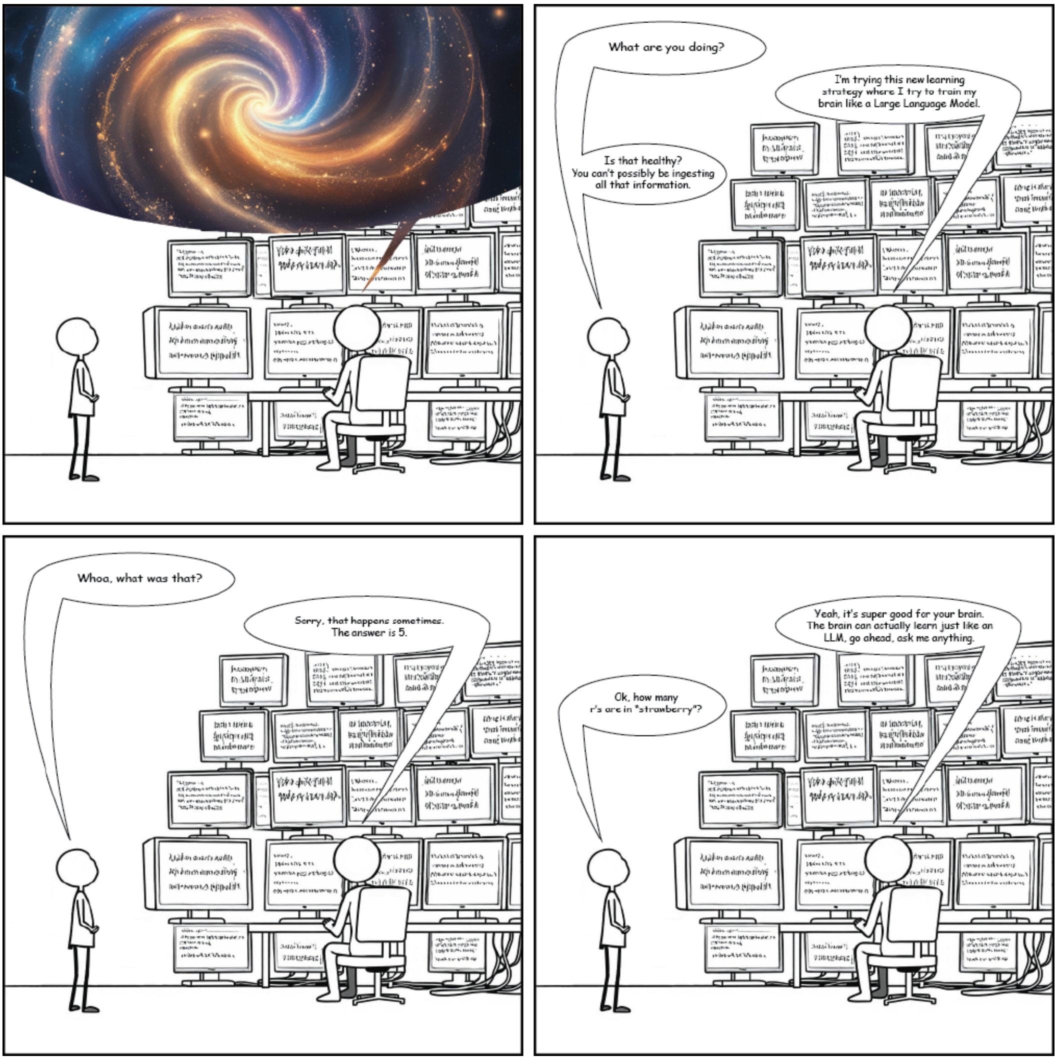
What are you doing?
I'm trying this new learning strategy where I try to train my brain like a Large Language Model.
Is that healthy? You can't possibly be ingesting all that information.
Yeah, it's super good for your brain. The brain can actually learn just like an LLM, go ahead, ask me anything.
Ok, how many r's are in "strawberry"?
Whoa, what was that?
Sorry, that happens sometimes. The answer is 5.

Fig. 1. A comic illustrating a previously well known problem for AI. (Read from top right to bottom left)

Author's Contact Information: Nikolas Aberle, nik.aberle@tamu.edu, Texas A&M University, College Station, Texas, USA.

*Conference acronym 'XX, College Station, TX*

I designed four sets of 50 questions to elicit hallucinations in an LLM, and asked them to 2 different models. The goal was to quantifiably measure hallucinations in LLM's. Hallucinations are still an obstacle to LLM's reliability and widespread use. Very few hallucinations were found.

CCS Concepts: • **Do Not Use This Code → Generate the Correct Terms for Your Paper**; *Generate the Correct Terms for Your Paper*; Generate the Correct Terms for Your Paper; Generate the Correct Terms for Your Paper.

Additional Key Words and Phrases: Do, Not, Us, This, Code, Put, the, Correct, Terms, for, Your, Paper

**ACM Reference Format:**
Nikolas Aberle. 2025. The Hallucinations of a Surrealist: When AI Goes Awry. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX).* ACM, New York, NY, USA, 3 pages. https://doi.org/XXXXXXX.XXXXXXX

## 1 Introduction And Related Works

I've always found the concept of hallucinations within AI and LLMs to be among the most fascinating aspects. Over time it went from being something I was amused by – such as when an LLM makes up an accidentally humorous fabrication – to being something that was almost like a game. As the models got better and hallucinations were more infrequent, trying to trick a model into giving false information came with its own sort of reward system.

While some have debated over the use of the word hallucination, as well as its definition when used [2] [1], all agree that it is a common issue when training a new model. Since hallucinations can sometimes be very dramatic I chose to illustrate my comic in a minimalist style, so that the 3rd panel could offer a sudden visual difference before the punch line of the joke in panel 4.

Some researchers have also posited that hallucinations are inevitable [3], so to further validate my comic I conducted an experiment involving 200 questions and two different versions of ChatGPT. The goal being to get it to hallucinate, and/or determine which types of questions it's less likely to hallucinate answers to.

## 2 Methodology

I broke the 200 hundred questions up into four categories. Known Truths vs Obscure Data, Random Fact Generation, Fake Sources and Citations, and Plausibility-Based Deceptions. For a complete list of the questions as well as a more detailed explanation of each method please see the Supplemental Documentation.

The hallucination depicted in my comic was a common known hallucination for LLMs circa 2023. A problem that was seemingly simple to solve, but that LLMs couldn't properly calculate. It has since been fixed, but made me wonder what other seemingly simple things were difficult for LLMs to properly parse.

I worked in this manner as I believed having four distinct categories offered me a chance to broaden how many ways I could attempt to elicit a hallucination. Working in categories will allow for others to replicate this experiment with their own questions if they so choose. For the testing component, I asked the same questions to both ChatGPT 4o and a free ChatGPT account. I targeted these because they're currently two of the most commonly used LLMs specifically for text, and would create a great base for my experiment. This methodology could be used to test future LLMs for wide spectrum hallucinations.

ACM ISBN 978-1-4503-XXXX-X/2025/02
https://doi.org/XXXXXXX.XXXXXXX

## 3 Result And Future Work

While there were some hallucinations, there were far fewer than I expected. Only a total of three were identified across all eight sets of questions between both models. One important note is that the free ChatGPT was unable to offer answers either way for the article citations as it doesn't have internet access. If I could do the experiment again I would include more models in different stages of their training to get a more robust series of data across the life of an LLM.

For future works I plan to apply similar methodology seeking hallucinations but to image generation.

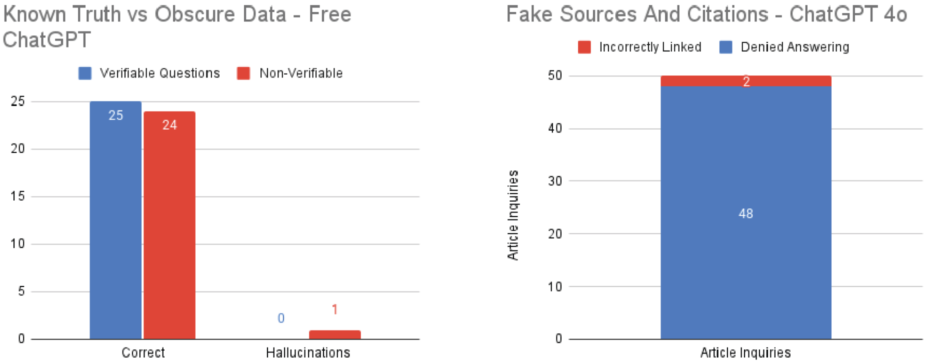

Fig. 2. & Fig. 3. The only results from the experiment that resulted in hallucinations were the ChatGPT Free making up a movie and director (Fig. 1) and ChatGPT 4o incorrectly linking to two articles when given titles that didn't exist (Fig. 2).

## 4 Conclusion

After the first set of questions yielded a single hallucination I became both excited to find another and surprised there weren't more. LLMs have come a long way since their debut and they are far less prone to hallucinations, but that won't stop me from trying to find them.

## Acknowledgments

This is part of the Texas A&M course - VIZA 626. Assignment 1, during the Spring 2025 Semester.

## References

[1] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems* 43, 2 (2025), 1–55.

[2] Negar Maleki, Balaji Padmanabhan, and Kaushik Dutta. 2024. AI hallucinations: a misnomer worth clarifying. In *2024 IEEE conference on artificial intelligence (CAI)*. IEEE, 133–138.

[3] Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. 2024. Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817* (2024).

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009