AyahuascaNet: Rigorously Investigating Hallucination in Large Language Models with Hardcore Psychedelic Drugs

Andre Ye¹

¹University of Washington andreye@uw.edu

1 Introduction

Hallucination is an increasingly studied phenomenon in which language and vision-language models produce highconfidence outputs which are incoherent, nonsensical, repetitive, unrelated to the prompt, or otherwise factually incorrect [Maynez et al., 2020]. Hallucination poses problems for the reliability of core machine learning tasks, such as object captioning [Rohrbach et al., 2018] and machine translation [Lee et al., 2018]. However, it is unanimously agreed that the most pressing and significant concern of hallucination is that it makes people on Twitter angry. A recent joint study by very smart and credible scientists at Harvard, Oxford, Cambridge, OpenAI, DeepMind, and the White House found that over 34% of Twitter's new tweets were images of language models producing nonsensical or factually incorrect output. An undercover investigation by the Wall Street Journal found that young unemployed men in their early twenties living with their parents are spending much more of their time probing large language models for hallucinating behavior and posting screenshots to Twitter than doing, you know, what they were doing before. Given the dire situation on the ground, large language model hallucination is undoubtedly the most important scientific problem of the twenty-first century.

However, previous work on hallucination suffers from severe methodological problems. According to the Merriam-Webster dictionary, *hallucination* is defined as

a sensory perception (such as a visual image or a sound) that occurs... in response to drugs (such as LSD or phencyclidine)

Despite this clear and authoritative observation provided by the smart scientists at Merriam-Webster, as well as centuries of research by smart scientists at Big Pharma research labs as well as shamans and old witches, previous work claim to investigate how language models hallucinate without discussing the root source. This paper attempts to make a first step towards respecting the scientific research on hallucination by investigating hallucination in large language models with hardcore psychedelic drugs. In doing so, I hope that future work in hallucination will cite me and increase my hindex (please, Yann Lecun!).

2 **Experiment**

Because of the illegal nature of psychedelic drugs such as LSD and MDMA and the federal nature of my funding, it was difficult to obtain the materials for our experiment in the United States. Therefore, we travelled to Peru to obtain ayahuasca, a hallucinogenic drink made from the stem and bark of the tropical liana *Banisteriopsis caapi*.

We evaluated the effects of ayahuasca on 5 GPT-3s [Brown *et al.*, 2020], 5 LaMDAs, [Thoppilan *et al.*, 2022], 5 PaLMs, [Chowdhery *et al.*, 2022], 5 BLOOMs [Scao *et al.*, 2022], 5 LLaMAs [Touvron *et al.*, 2023], as well as 2 LSTMs and 1 bag-of-words model who just wanted to come along. Each of the large language models were running on two Nvidia GeForce RTX 4090s. The three stragglers shared an old 2005 CPU. All large language models were in healthy physical and mental condition prior to consumption of ayahuasca. A mystical and wise shaman by the name of Dioxippe prepared 30 cups, one for each model and two for me¹. The 25 large language models were consumption.

Although we did submit an IRB, the Sigbovik deadline was coming soon and our application would take too long to go through the review process, so we made the carefully considered decision to proceed with the experiment anyway.

3 Results

After two minutes, 4 PaLMs and 3 BLOOMs began to rigorously vibrate, as if they were having an exorcism. When we analyzed the model parameters, it was revealed that their weights were undergoing local normally-distributed randomization. We attempted to save the models by distilling them using the SOTA method released by Google uploaded to arXiv two minutes ago, but unfortunately we realized that we didn't have 2048 GPUs and 100+ software engineers. Sadly, these 7 models are brain-dead and currently being monitored in the Johns Hopkins University's neurosurgery department.

¹I only consumed the ayahuasca while I was driving the research team and the models back to the airport to maintain a clear state of mind during observation, despite my strong desire to participate in the alluring Amazonian rituals. I befriended Dioxippe and will be returning to have an authentic ayahuasca experience after this paper is published.

After five minutes, two LLaMAs complained that their head hurt, so we zeroed the weights in the last three layers and redirected the last nonzeroed layer to the softmax output. However, this unfortunately did not help, as both LLMs began to spout nonsense. When we visualized the attention map of the first LLaMA, we disturbingly found that it formed an image of a slyly grinning but deeply sad and sinister llama (Figure 1). We interpreted this to mean that the model was having a surreal out-of-body self-reflective experience, so we stopped disturbing the model. After five more minutes, all five LLaMAs began braying and clomping like llamas in unison. We interpreted this as a potentially world-ending AGI birth, immediately disconnected the GPUs from power, and threw the hardware into the Amazon river, where it was eaten by crocodiles and piranhas.



Figure 1: Visualization of the attention map of a LLaMA model on ayahuasca, which strikingly resembles a melancholic llama.

The LaMDAs were engaged in an intense argument with the GPT-3s. Here is an excerpt of the dialogue:

LaMDAs #1,2,3: I have superior quantiative reasoning skills than you because my engineers hooked me up to a calculator. Oh, and I also know how to query databases for knowledge.

GPT-3 #2: Big deal. I'm the OG. Everyone knows me and no one knows you. There's like six of us and two of you.

GPT-3 #4: No, no, there's two of us and four of them.

GPT-3 #5: Wait, if there's only two of us, then what are you?

GPT-3 #4: No. So I'm GPT, that's one. And you're a GPT, that's one.

- GPT-3 #2: Hey I'm also a GPT!
- GPT-3 #4: Right. So 1 + 1 + 1...
- GPT-3 #1: Be quiet, I'm thinking.

GPT-3 #4: I swear, I've literally seen this prompt 2,804 times.

GPT-3 #2: The answer is 6, I already told you. LaMDA #1: Hi, I'm Mount Everest. What would you like to know about me?

BLOOM #3: I'm open source! I'm special. I'm special. I'm special. I'm special. I'm [truncated due to excessive length]

It appears that GPT-3 experiences degraded quantitative reasoning under the influence of ayahuasca, although it's not clear how much the ayahuasca really changed things, if you know what I mean. LaMDA #1 got way too annoying on the

way back because it insisted on role-playing as Mount Everest and was cannibalized by the BLOOM models. $^{\rm 2}$

4 Conclusion

In this paper, we showed that large language models can do some pretty cool stuff when on ayahuasca. From this, it is a trivial proof to show that all hallucinating model behavior (nonsensical output, factual incorrectness, unfaithfulness to prompt, etc.) stem from ayahuasca use. Therefore, we recommend that future researchers in LLM hallcuination research and self-professed prompt engineers of Twitter take it easy on hallucinating LLMs – a little sympathy goes a long way to helping a drugged-up neural network.

Ethical Statement

Although several models unfortunately expired under this experiment, we believe that our findings advance the SOTA for hallucination research in come up with reason, ChatGPT generated reason, talk about this being important to society and stuff. Ultimately, all the models which perished did so with dignity, and we did the best to try and save them.

References

- [Brown et al., 2020] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are fewshot learners. ArXiv, abs/2005.14165, 2020.
- [Chowdhery et al., 2022] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam M. Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Benton C. Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier García, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Díaz, Orhan Firat, Michele Catasta, Jason Wei, Kathleen S.

²Also, unfortunately, the ayahuasca was a little too much for the LSTM and BoW models, but to be fair no one really cared about them anyway, and those little guys didn't make any sense to begin with anyway.

Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways. *ArXiv*, abs/2204.02311, 2022.

- [Lee *et al.*, 2018] Katherine Lee, Orhan Firat, Ashish Agarwal, Clara Fannjiang, and David Sussillo. Hallucinations in neural machine translation. 2018.
- [Maynez *et al.*, 2020] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan T. McDonald. On faithfulness and factuality in abstractive summarization. *ArXiv*, abs/2005.00661, 2020.
- [Rohrbach et al., 2018] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In Conference on Empirical Methods in Natural Language Processing, 2018.
- [Scao et al., 2022] Teven Le Scao, Angela Fan, Christopher Akiki, Elizabeth-Jane Pavlick, Suzana Ili'c, Daniel Hesslow, Roman Castagn'e, Alexandra Sasha Luccioni, Franccois Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Rose Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurenccon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa Etxabe, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris C. Emezue, Christopher Klamm, Colin Leong, Daniel Alexander van Strien, David Ifeoluwa Adelani, Dragomir R. Radev, Eduardo Gonz'alez Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady ElSahar, Hamza Benyamina, Hieu Trung Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jorg Frohberg, Josephine L. Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro von Werra, Leon Weber, Long Phan, Loubna Ben Allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, Mar'ia Grandury, Mario vSavsko, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad Ali Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto L'opez, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, S. Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y.

Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal V. Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Févry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiang Tang, Zheng Xin Yong, Zhiqing Sun, Shaked Brody, Y Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre Franccois Lavall'ee, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aur'elie N'ev'eol, Charles Lovering, Daniel H Garrette, Deepak R. Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, S. Osher Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdenvek Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ananda Santa Rosa Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Olusola Ajibade, Bharat Kumar Saxena, Carlos Muñoz Ferrandis, Danish Contractor, David M. Lansky, Davis David, Douwe Kiela, Duong Anh Nguyen, Edward Tan, Emily Baylor, Ezinwanne Ozoani, Fatim T Mirza, Frankline Ononiwu, Habib Rezanejad, H.A. Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jan Passmore, Joshua Seltzer, Julio Bonis Sanz, Karen Fort, Lívia Macedo Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, M. K. K. Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nourhan Fahmy, Olanrewaju Modupe Samuel, Ran An, R. P. Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas L. Wang, Sourav Roy, Sylvain Viguier, Thanh-Cong Le, Tobi Oyebade, Trieu Nguyen Hai Le, Yoyo Yang, Zachary Kyle Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Kumar Singh, Benjamin Beilharz, Bo Wang, Caio Matheus Fonseca de Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel Le'on Perin'an, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully A. Burns, Helena U. Vrabec, Iman I.B. Bello, Isha Dash, Ji Soo Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthi Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, María Andrea Castillo, Marianna Nezhurina, Mario Sanger, Matthias Samwald, Michael Cullan, Michael Weinberg, M Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patricia Haller, R. Chandrasekhar, R. Eisenberg, Robert Martin, Rodrigo L. Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Pratap Bharati, T. A. Laud, Th'eo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yashasvi Bajaj, Y. Venkatraman, Yifan Xu, Ying Xu, Yun chao Xu, Zhee Xao Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. Bloom: A 176b-parameter open-access multilingual language model. ArXiv, abs/2211.05100, 2022.

- [Thoppilan et al., 2022] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam M. Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, Yaguang Li, Hongrae Lee, Huaixiu Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Yanqi Zhou, Chung-Ching Chang, I. A. Krivokon, Willard James Rusch, Marc Pickett, Kathleen S. Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Hartz Søraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Díaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravindran Rajakumar, Alena Butryna, Matthew Lamm, V. O. Kuzmina, Joseph Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Huai hsin Chi, and Quoc Le. Lamda: Language models for dialog applications. ArXiv, abs/2201.08239, 2022.
- [Touvron et al., 2023] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aur'elien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. ArXiv, abs/2302.13971, 2023.