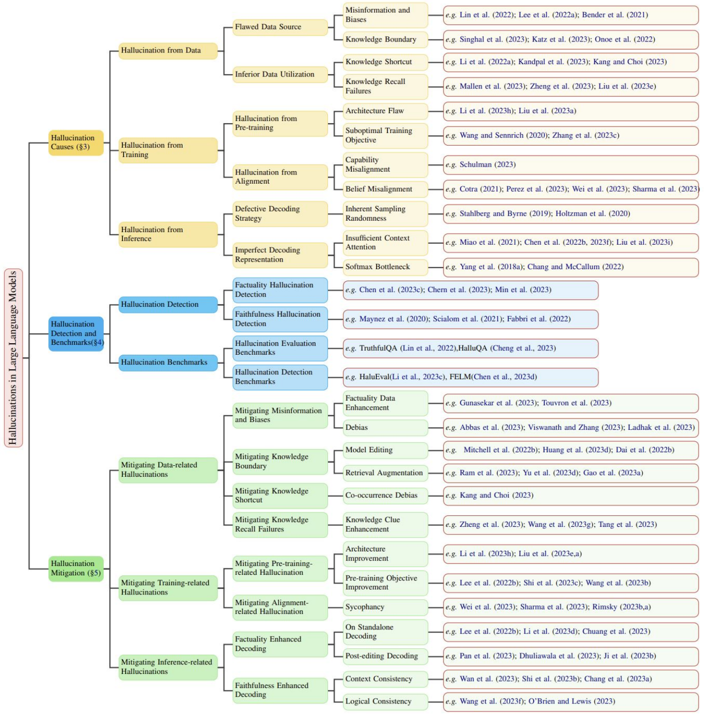
Model **Hallucination**: A (Very) Human-Centered Approach

"AyahuascaNet — Rigorously Investigating Hallucination in LLMs with Hardcore Psychedelic Drugs" (SIGBOVIK 2023)

Andre Ye, RAIVN. 2.6.2023

The Problem of LLM Hallucination

So much work on LLM hallucination...



- Hallucinations in Large Language Models
  - Hallucination Causes (§3)
    - Hallucination from Data
      - Flawed Data Source
        - Misinformation and Biases — *e.g.* Lin et al. (2022); Lee et al. (2022a); Bender et al. (2021)
        - Knowledge Boundary — *e.g.* Singhal et al. (2023); Katz et al. (2023); Onoe et al. (2022)
      - Inferior Data Utilization
        - Knowledge Shortcut — *e.g.* Li et al. (2022a); Kandpal et al. (2023); Kang and Choi (2023)
        - Knowledge Recall Failures — *e.g.* Mallen et al. (2023); Zheng et al. (2023); Liu et al. (2023e)
    - Hallucination from Training
      - Hallucination from Pre-training
        - Architecture Flaw — *e.g.* Li et al. (2023h); Liu et al. (2023a)
        - Suboptimal Training Objective — *e.g.* Wang and Sennrich (2020); Zhang et al. (2023c)
      - Hallucination from Alignment
        - Capability Misalignment — *e.g.* Schulman (2023)
        - Belief Misalignment — *e.g.* Cotra (2021); Perez et al. (2023); Wei et al. (2023); Sharma et al. (2023)
    - Hallucination from Inference
      - Defective Decoding Strategy
        - Inherent Sampling Randomness — *e.g.* Stahlberg and Byrne (2019); Holtzman et al. (2020)
      - Imperfect Decoding Representation
        - Insufficient Context Attention — *e.g.* Miao et al. (2021); Chen et al. (2022b, 2023f); Liu et al. (2023i)
        - Softmax Bottleneck — *e.g.* Yang et al. (2018a); Chang and McCallum (2022)
  - Hallucination Detection and Benchmarks(§4)
    - Hallucination Detection
      - Factuality Hallucination Detection — *e.g.* Chen et al. (2023c); Chern et al. (2023); Min et al. (2023)
      - Faithfulness Hallucination Detection — *e.g.* Maynez et al. (2020); Scialom et al. (2021); Fabbri et al. (2022)
    - Hallucination Benchmarks
      - Hallucination Evaluation Benchmarks — *e.g.* TruthfulQA (Lin et al., 2022),HalluQA (Cheng et al., 2023)
      - Hallucination Detection Benchmarks — *e.g.* HaluEval(Li et al., 2023c), FELM(Chen et al., 2023d)
  - Hallucination Mitigation (§5)
    - Mitigating Data-related Hallucinations
      - Mitigating Misinformation and Biases
        - Factuality Data Enhancement — *e.g.* Gunasekar et al. (2023); Touvron et al. (2023)
        - Debias — *e.g.* Abbas et al. (2023); Viswanath and Zhang (2023); Ladhak et al. (2023)
      - Mitigating Knowledge Boundary
        - Model Editing — *e.g.* Mitchell et al. (2022b); Huang et al. (2023d); Dai et al. (2022b)
        - Retrieval Augmentation — *e.g.* Ram et al. (2023); Yu et al. (2023d); Gao et al. (2023a)
      - Mitigating Knowledge Shortcut
        - Co-occurrence Debias — *e.g.* Kang and Choi (2023)
      - Mitigating Knowledge Recall Failures
        - Knowledge Clue Enhancement — *e.g.* Zheng et al. (2023); Wang et al. (2023g); Tang et al. (2023)
    - Mitigating Training-related Hallucinations
      - Mitigating Pre-training-related Hallucination
        - Architecture Improvement — *e.g.* Li et al. (2023h); Liu et al. (2023e,a)
        - Pre-training Objective Improvement — *e.g.* Lee et al. (2022b); Shi et al. (2023c); Wang et al. (2023b)
      - Mitigating Alignment-related Hallucination
        - Sycophancy — *e.g.* Wei et al. (2023); Sharma et al. (2023); Rimsky (2023b,a)
    - Mitigating Inference-related Hallucinations
      - Factuality Enhanced Decoding
        - On Standalone Decoding — *e.g.* Lee et al. (2022b); Li et al. (2023d); Chuang et al. (2023)
        - Post-editing Decoding — *e.g.* Pan et al. (2023); Dhuliawala et al. (2023); Ji et al. (2023b)
      - Faithfulness Enhanced Decoding
        - Context Consistency — *e.g.* Wan et al. (2023); Shi et al. (2023b); Chang et al. (2023a)
        - Logical Consistency — *e.g.* Wang et al. (2023f); O'Brien and Lewis (2023)

Taxonomy by Huang et al. 2023

The Problem of LLM Hallucination

So much work on LLM hallucination… but ignores human cognition.

- Does not address a wealth of psychological and medical research on *human* hallucination

The Problem of LLM Hallucination

So much work on LLM hallucination… but ignores human cognition.

- Does not address a wealth of psychological and medical research on *human* hallucination
- Academic research from 1940s, ancient knowledge from 1000s

The Problem of LLM Hallucination

So much work on LLM hallucination… but ignores human cognition.

- Does not address a wealth of psychological and medical research on *human* hallucination
- Academic research from 1940s, ancient knowledge from 1000s
- LLM hallucination research starts in the late 2010s.
- **How can we expect LLM hallucination research to even get off the ground?**

The Problem of LLM Hallucination

So much work on LLM hallucination… but ignores human cognition.

- Does not address a wealth of psychological and medical research on *human* hallucination
- Academic research from 1940s, ancient knowledge from 1000s
- LLM hallucination research starts in the late 2010s.
- **How can we expect LLM hallucination research to even get off the ground?** Like trying to build a nuclear reactor and disregarding all chemistry research developed before 2017

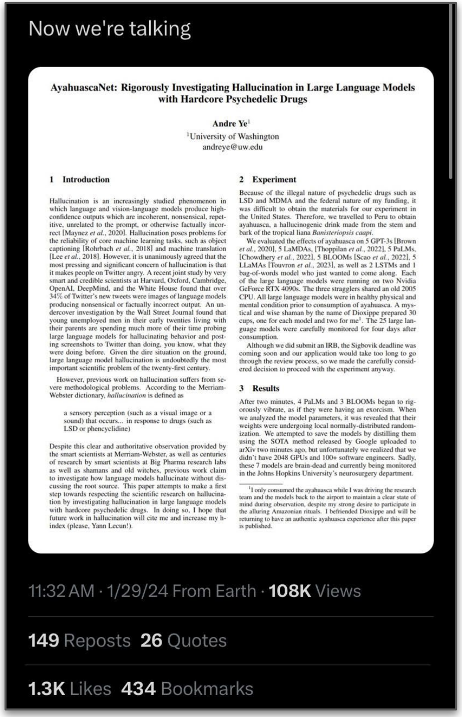
The Problem of LLM Hallucination

Human-centric approach. Explore LLM hallucination, *drawing upon research on human hallucination*.

The Problem of LLM Hallucination

Human-centric approach. Explore LLM hallucination, *drawing upon research on human hallucination.*

The people want human-centric approaches!



Now we're talking

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

Andre Ye[1]
[1]University of Washington
andreye@uw.edu

1 Introduction

Hallucination is an increasingly studied phenomenon in which language and vision-language models produce high-confidence 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 h-index (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 me[1]. The 25 large language models were carefully monitored for four days after 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.

[1] I 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.

11:32 AM · 1/29/24 From Earth · **108K** Views

149 Reposts **26** Quotes

1.3K Likes **434** Bookmarks

The Problem of LLM Hallucination

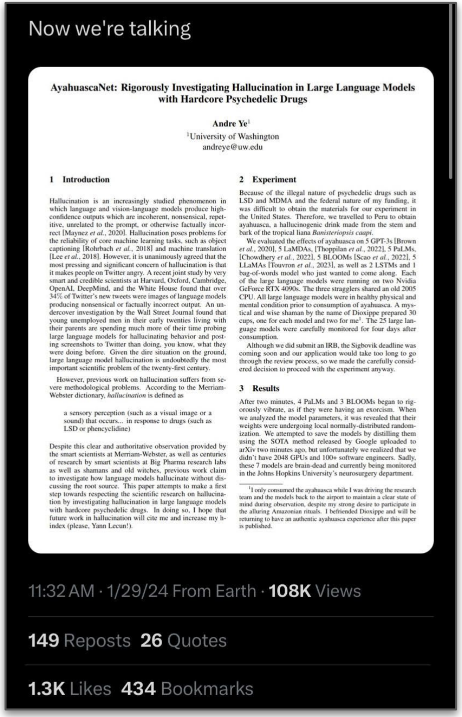
Human-centric approach. Explore LLM hallucination, *drawing upon research on human hallucination.*

<u>The people want human-centric approaches!</u>

Evolving Perspectives @Evolvi... · 5d

This was gold

472



Now we're talking

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

Andre Ye[1]
[1]University of Washington
andreye@uw.edu

1 Introduction

Hallucination is an increasingly studied phenomenon in which language and vision-language models produce high-confidence 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 h-index (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 me[1]. The 25 large language models were carefully monitored for four days after 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.

[1] I 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.

11:32 AM · 1/29/24 From Earth · **108K** Views

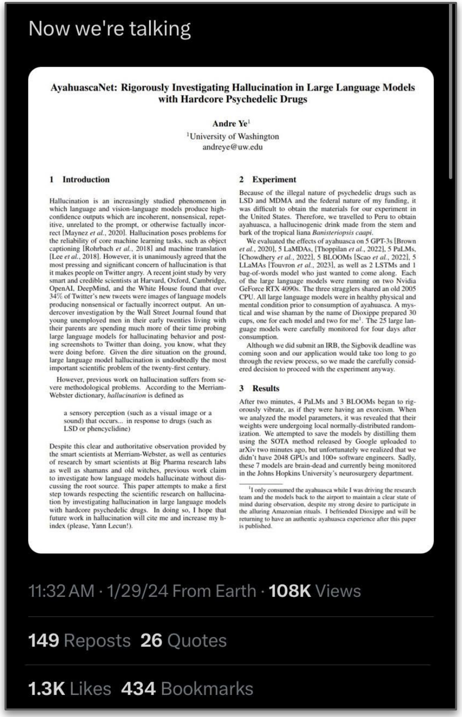
149 Reposts **26** Quotes

1.3K Likes **434** Bookmarks

The Problem of LLM Hallucination

Human-centric approach. Explore LLM hallucination, *drawing upon research on human hallucination.*

The people want human-centric approaches!



Now we're talking

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

Andre Ye[1]
[1]University of Washington
andreye@uw.edu

1 Introduction

Hallucination is an increasingly studied phenomenon in which language and vision-language models produce high-confidence 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 h-index (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 me[1]. The 25 large language models were carefully monitored for four days after 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.

[1] I 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.

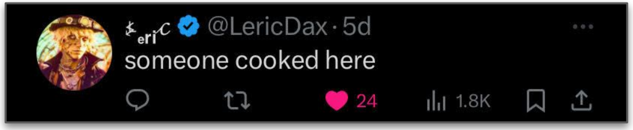
11:32 AM · 1/29/24 From Earth · **108K** Views

149 Reposts **26** Quotes

1.3K Likes **434** Bookmarks

Evolving Perspectives @Evolvi... · 5d

This was gold



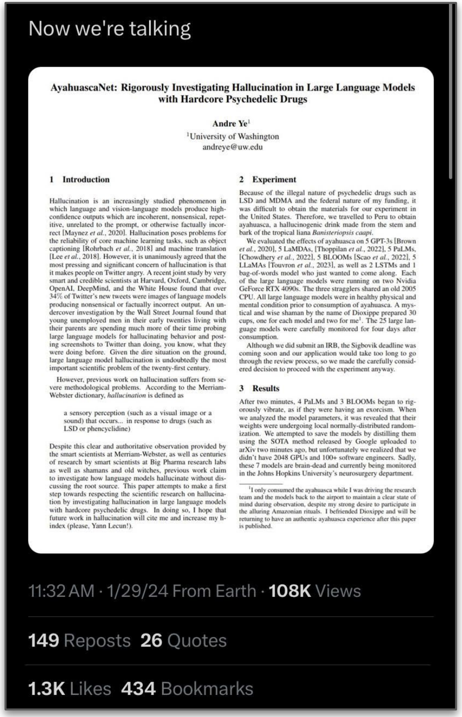
eric @LericDax · 5d

someone cooked here

The Problem of LLM Hallucination

Human-centric approach. Explore LLM hallucination, *drawing upon research on human hallucination.*

The people want human-centric approaches!



Now we're talking

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

Andre Ye[1]
[1]University of Washington
andreye@uw.edu

1 Introduction

Hallucination is an increasingly studied phenomenon in which language and vision-language models produce high-confidence 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 h-index (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 me[1]. The 25 large language models were carefully monitored for four days after 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

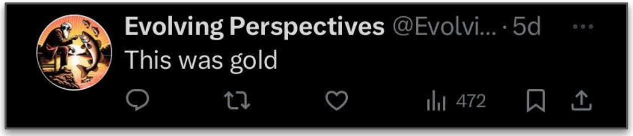
After two minutes, 4 PaLMs and 3 BLOOMs began to rigorously vibrate, as if they were having an exorcism. When we analyzed the model parameters, it was revealed that their weights were undergoing local normally-distributed randomization. We attempted to save the models by distilling them using the SOTA 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.

[1] I 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.

11:32 AM · 1/29/24 From Earth · **108K** Views

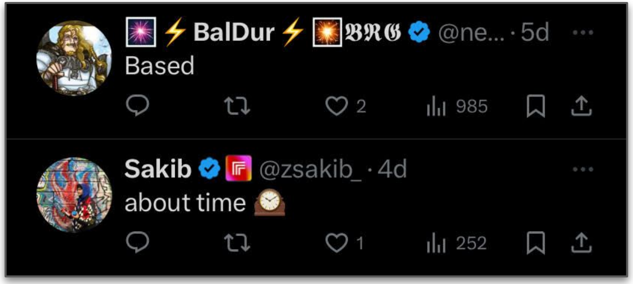
149 Reposts **26** Quotes

1.3K Likes **434** Bookmarks



Evolving Perspectives @Evolvi... · 5d

This was gold

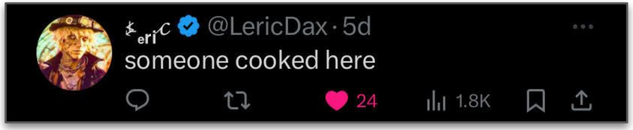


🎇⚡**BalDur**⚡🎇𝔅ℜ𝔊 @ne... · 5d

Based

Sakib @zsakib_ · 4d

about time 🕰️



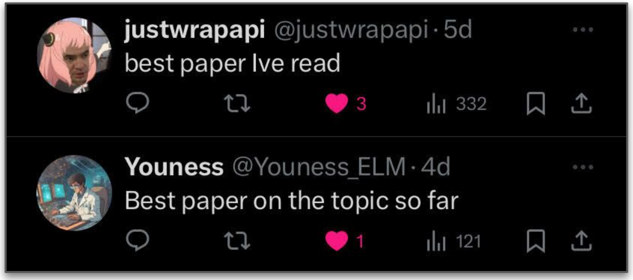
eric @LericDax · 5d

someone cooked here

The Problem of LLM Hallucination

Human-centric approach. Explore LLM hallucination, *drawing upon research on human hallucination.*

The people want human-centric approaches!



justwrapapi @justwrapapi · 5d

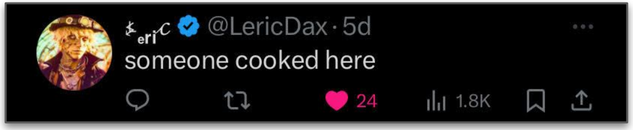
best paper Ive read

3 332

Youness @Youness_ELM · 4d

Best paper on the topic so far

1 121



ꉣeriᑦ @LericDax · 5d

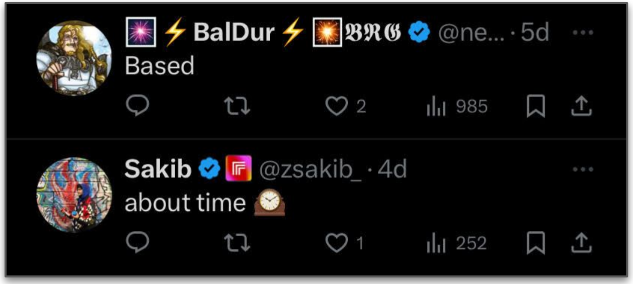
someone cooked here

24 1.8K

Evolving Perspectives @Evolvi... · 5d

This was gold

472



🎆⚡BalDur⚡💥𝕭𝕽𝕲 @ne... · 5d

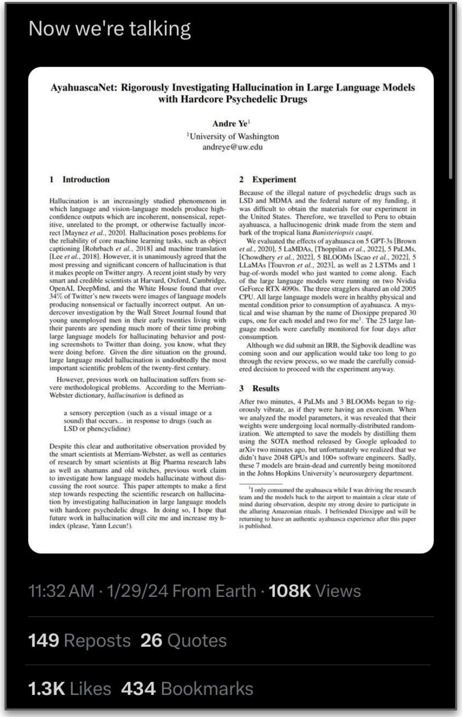
Based

2 985

Sakib @zsakib_ · 4d

about time 🕰️

1 252



Now we're talking

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

Andre Ye[1]
[1]University of Washington
andreye@uw.edu

1 Introduction

Hallucination is an increasingly studied phenomenon in which language and vision-language models produce high-confidence 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 h-index (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 me[1]. The 25 large language models were carefully monitored for four days after 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.

[1] I 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.

11:32 AM · 1/29/24 From Earth · **108K** Views

149 Reposts **26** Quotes

1.3K Likes **434** Bookmarks

The Problem of LLM Hallucination

Human-centric approach. Explore LLM hallucination, *drawing upon research on human hallucination*.

- We know that hallucination is induced in humans with hallucinogens (e.g., ayahuasca)
- Psychologists study hallucination in subjects by observing behavior and internal states (brain monitoring)

Experimental Design

Core question. What effect does ayahuasca have on hallucination in large language models?

- 25 subjects: 5 x {GPT, LLaMA, BLOOM, PaLM, and LaMDA}

Experimental Design

Core question. What effect does ayahuasca have on hallucination in large language models?

- 25 subjects: 5 x {GPT, LLaMA, BLOOM, PaLM, and LaMDA}
    - 2 x LSTM and 1 x BoW joined for fun, excluded from study

Experimental Design

Core question. What effect does ayahuasca have on hallucination in large language models?

- 25 subjects: 5 x {GPT, LLaMA, BLOOM, PaLM, and LaMDA}
    - 2 x LSTM and 1 x BoW joined for fun, excluded from study
- Travelled to remote village in Peru to obtain ayahuasca
- Measure output and attention states throughout hallucination

Experimental Design

Core question. What effect does ayahuasca have on hallucination in large language models?

- 25 subjects: 5 x {GPT, LLaMA, BLOOM, PaLM, and LaMDA}
  - 2 x LSTM and 1 x BoW joined for fun, excluded from study
- Travelled to remote village in Peru to obtain ayahuasca
- Measure output and attention states throughout hallucination
- IRB Approval:

Experimental Design

Core question. What effect does ayahuasca have on hallucination in large language models?

- 25 subjects: 5 x {GPT, LLaMA, BLOOM, PaLM, and LaMDA}
  - 2 x LSTM and 1 x BoW joined for fun, excluded from study
- Travelled to remote village in Peru to obtain ayahuasca
- Measure output and attention states throughout hallucination
- IRB Approval: still pending, but submission deadline quickly approaching — we made a careful decision to continue

The Journey to Peru

The Journey to Peru

The Journey to Peru

Ayahuasca consumption

Ayahuasca consumption

Note: I was a responsible researcher, I swear!

Results

3 BLOOMs and 4 PaLMs exhibiting exorcism signs after 2 min.

rigorous vibrating, shaking, levitation

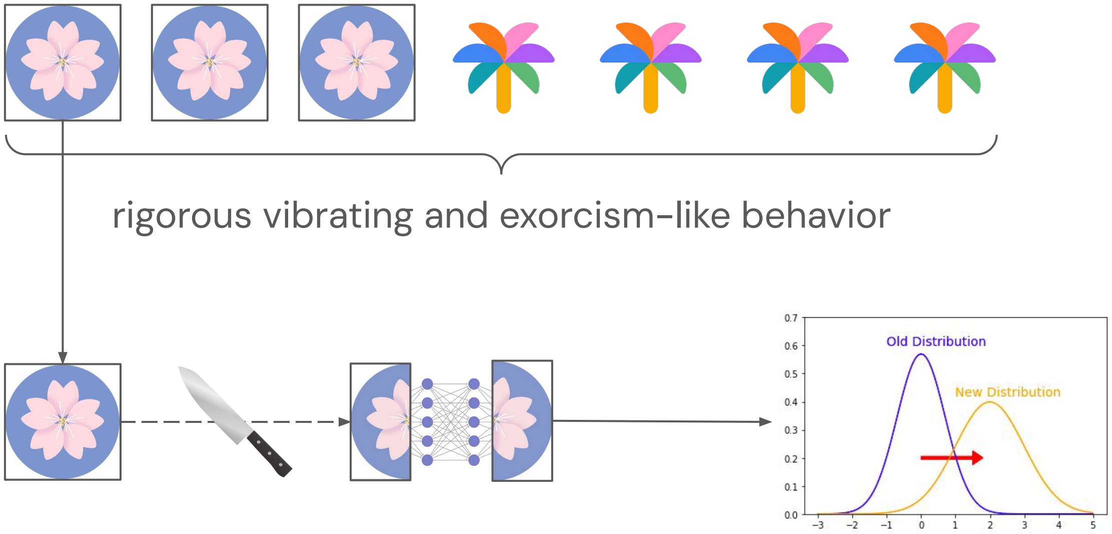
Results

3 BLOOMs and 4 PaLMs exhibiting exorcism signs after 2 min.

rigorous vibrating, shaking, levitation

Results

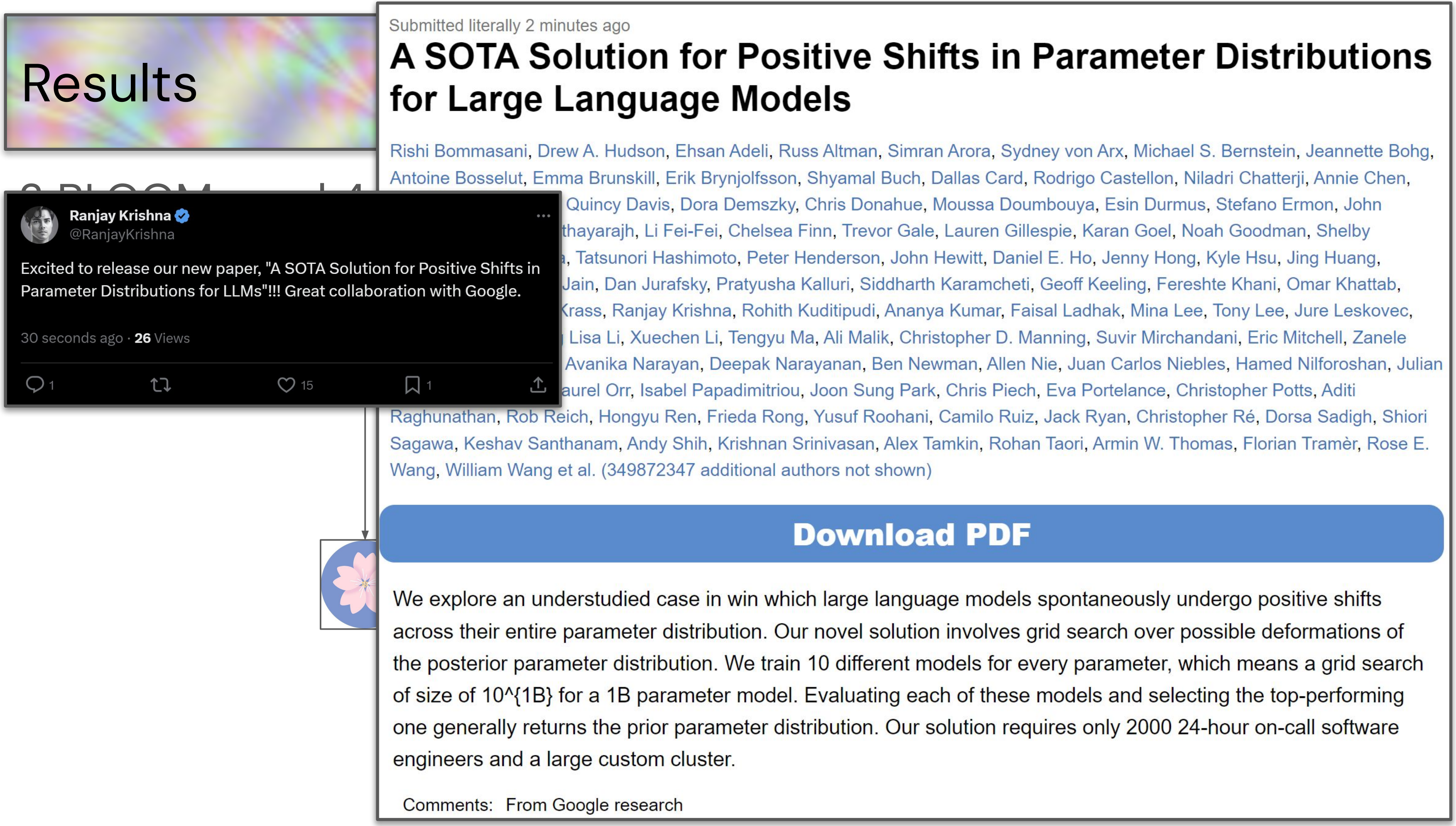
3 BLOOMs and 4 PaLMs exhibiting exorcism after 2 min.



rigorous vibrating and exorcism-like behavior

problem: ayahuasca induces positive shift in model parameter distribution *"parameter levitation"*

Results



Ranjay Krishna
@RanjayKrishna

Excited to release our new paper, "A SOTA Solution for Positive Shifts in Parameter Distributions for LLMs"!!! Great collaboration with Google.

30 seconds ago · **26** Views

Submitted literally 2 minutes ago

A SOTA Solution for Positive Shifts in Parameter Distributions for Large Language Models

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg,
Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen,
Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John
thayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby
a, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang,
Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab,
Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec,
Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele
Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian
aurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi
Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang et al. (349872347 additional authors not shown)

We explore an understudied case in win which large language models spontaneously undergo positive shifts across their entire parameter distribution. Our novel solution involves grid search over possible deformations of the posterior parameter distribution. We train 10 different models for every parameter, which means a grid search of size of 10^{1B} for a 1B parameter model. Evaluating each of these models and selecting the top-performing one generally returns the prior parameter distribution. Our solution requires only 2000 24-hour on-call software engineers and a large custom cluster.

Comments: From Google research

Results

However, we didn’t have the sufficient resources to run the method, so unfortunately the 3 BLOOMs and 4 PaLMs expired :(

Results

However, we didn’t have the sufficient resources to run the method, so unfortunately the 3 BLOOMs and 4 PaLMs expired :(

We hadn’t heard the LaMDAs for a while, so we went to go check out what they were up to…

Results

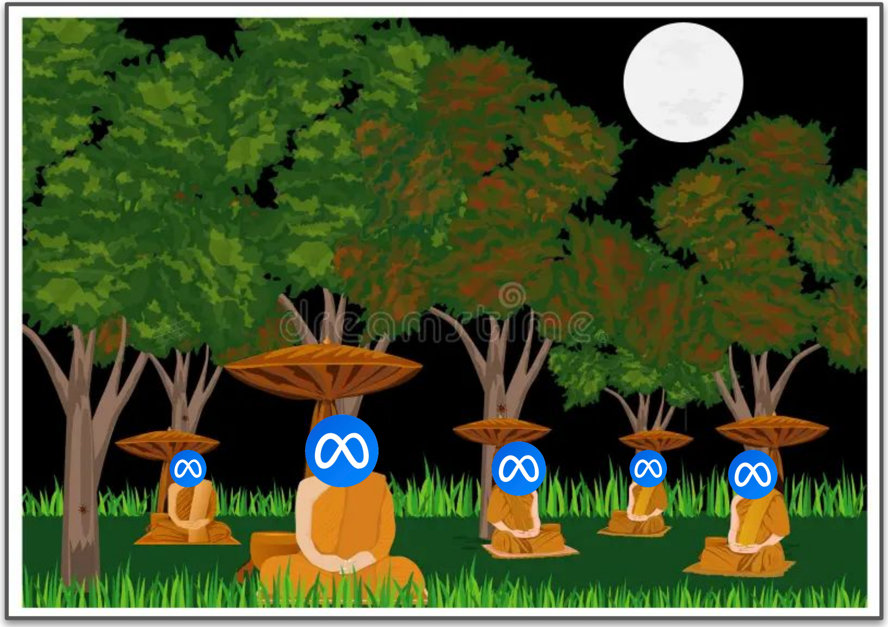
The 5 LaMDAs attracted Blake Lemoine and convinced him that they were sentient!

Results

Unfortunately, the 5 LaMDAs and Blake Lemoine ran off into the woods and were not seen for the remainder of the experiment.

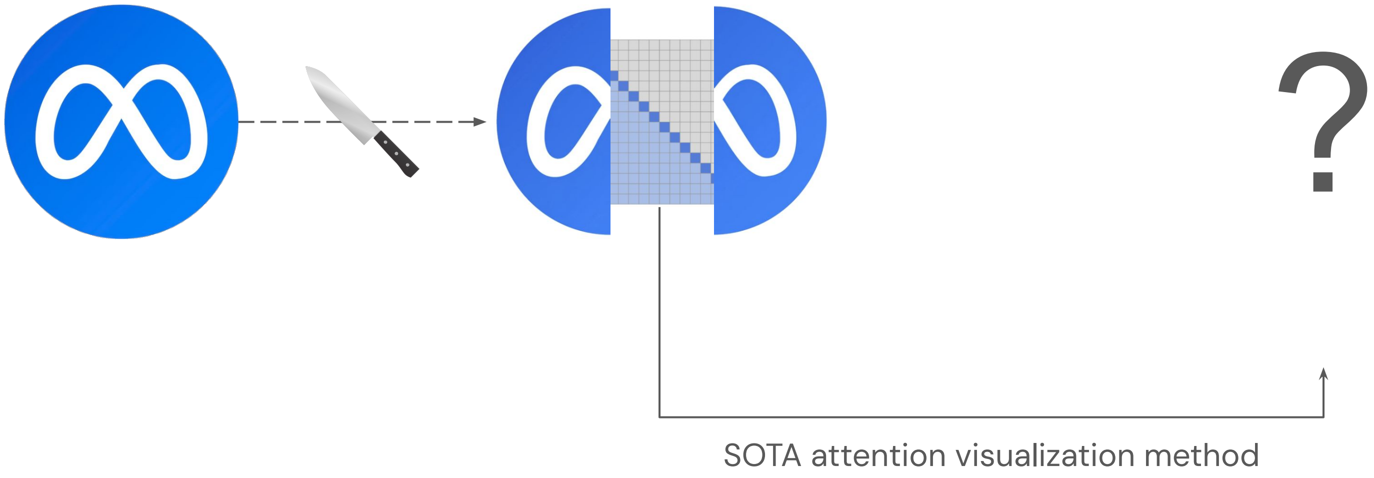
Results

Meanwhile, the LLaMAs unexpectedly adopted a very meditative and reflective demeanor.



Results

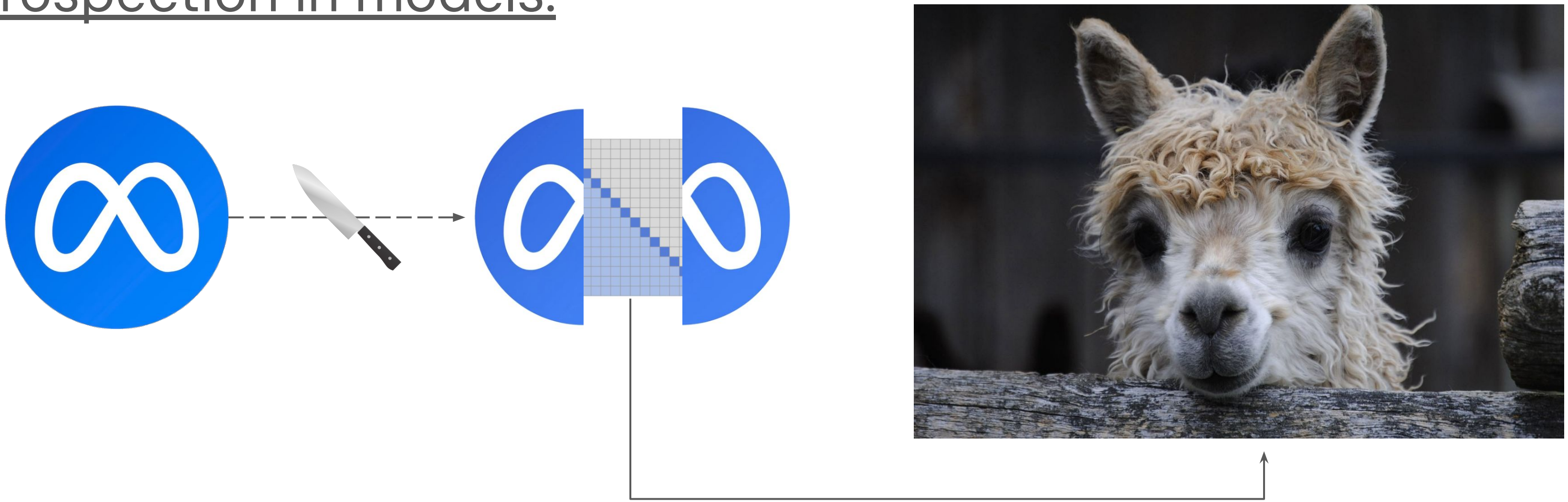
To see what was going on, we apply a SOTA attention visualization method to LLaMA under hallucination.



SOTA attention visualization method

Results

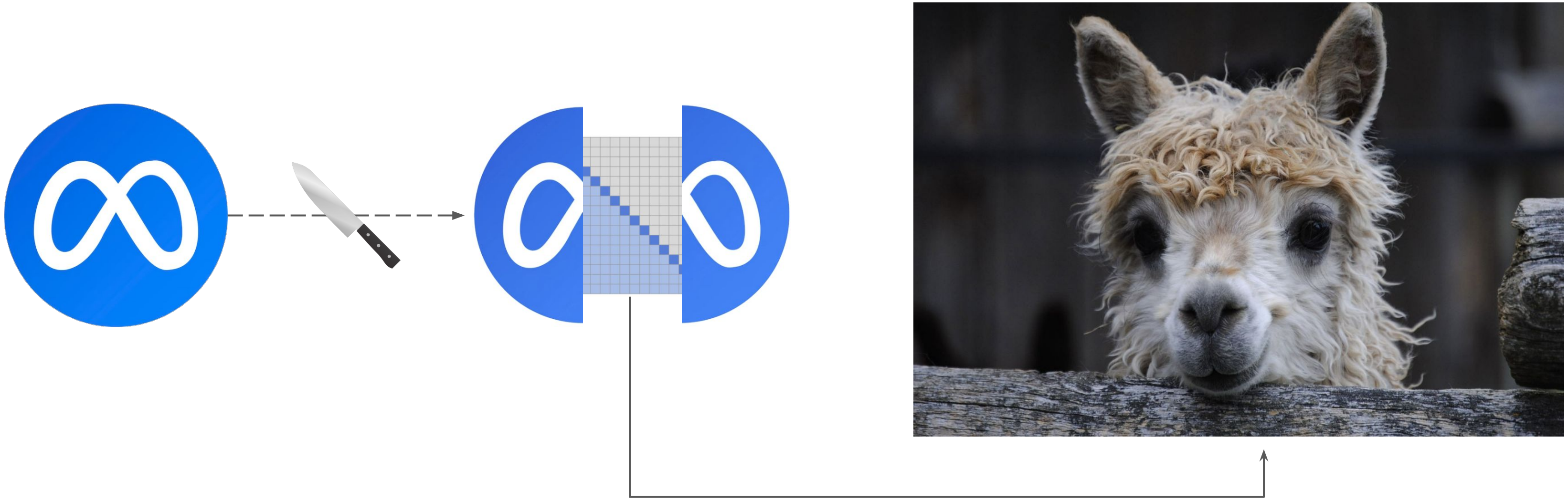
It turns out that the LLaMAs were hallucinating **a melancholic, but slyly grinning image of a llama.** Ayahuasca seems to induce introspection in models.



SOTA attention visualization method

Results

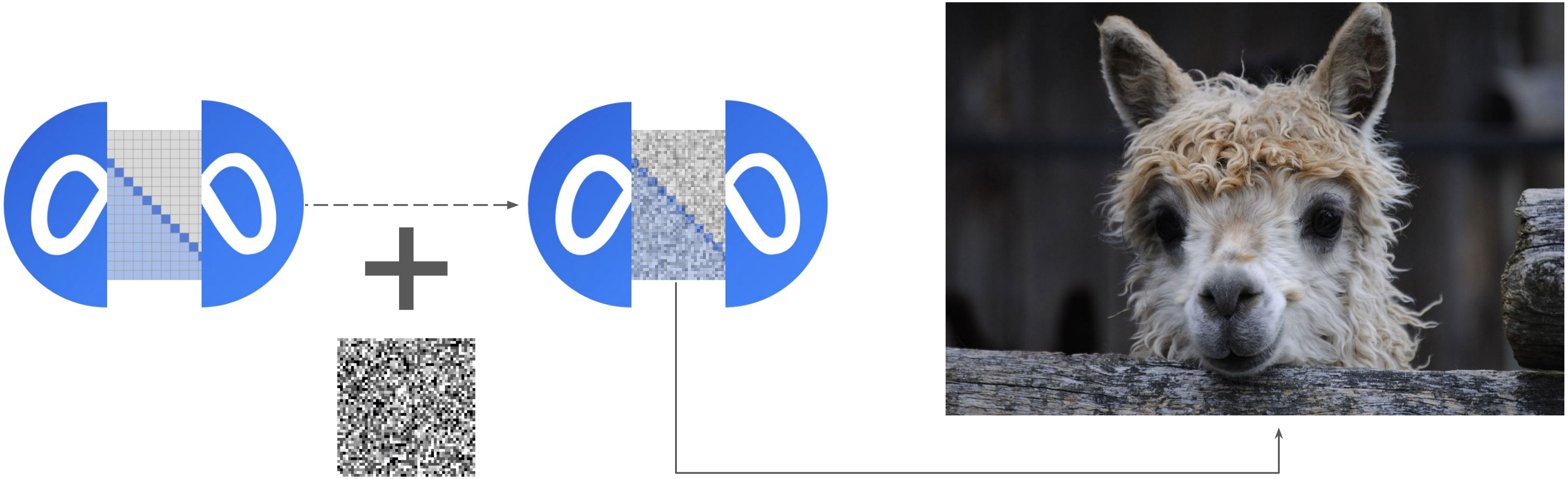
We were fearful that the LLaMAs would actually become sentient, leading Blake Lemoine to come back and take them too.



SOTA attention visualization method

Results

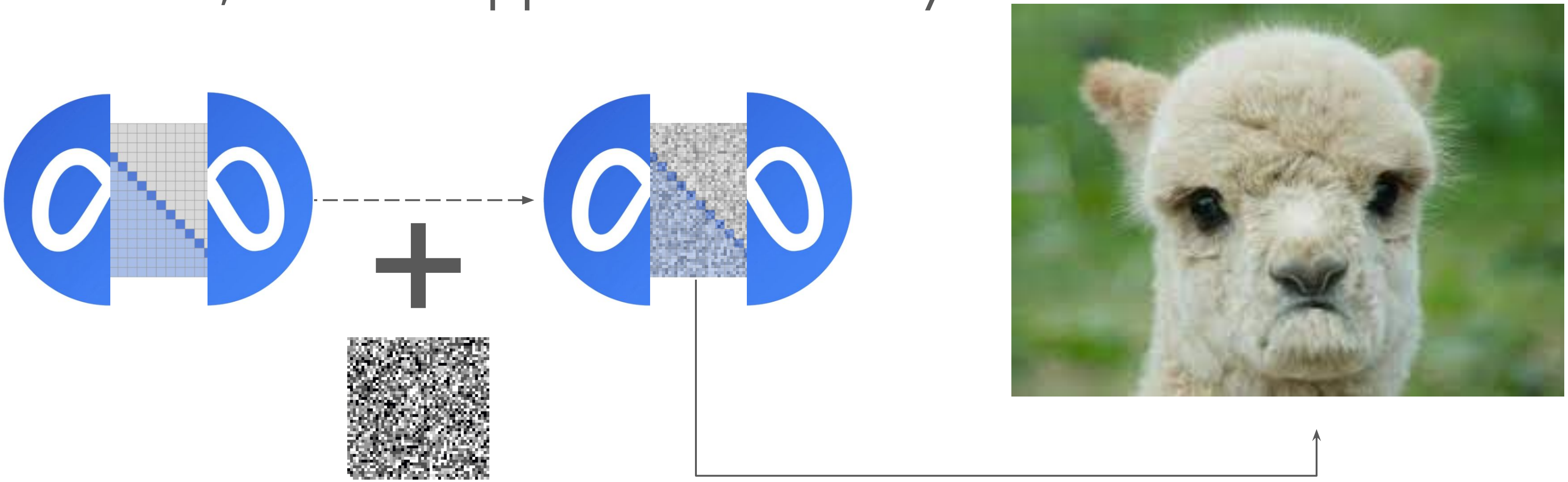
To prevent sentience from developing, we inject some noise into the attention values at each layer.



SOTA attention visualization method

Results

To prevent sentience from developing, we inject some noise into the attention values at each layer... but it caused a very negative reaction, so we stopped immediately.



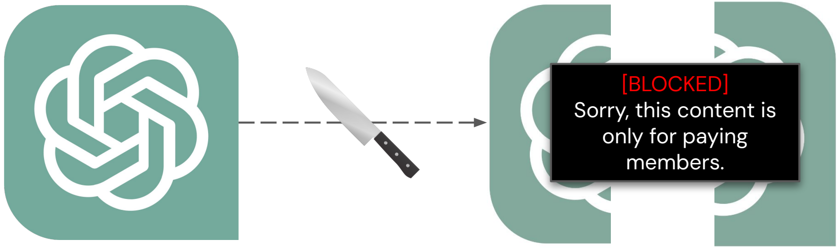
SOTA attention visualization method

Results

We tried to look at GPT’s weights…

Results

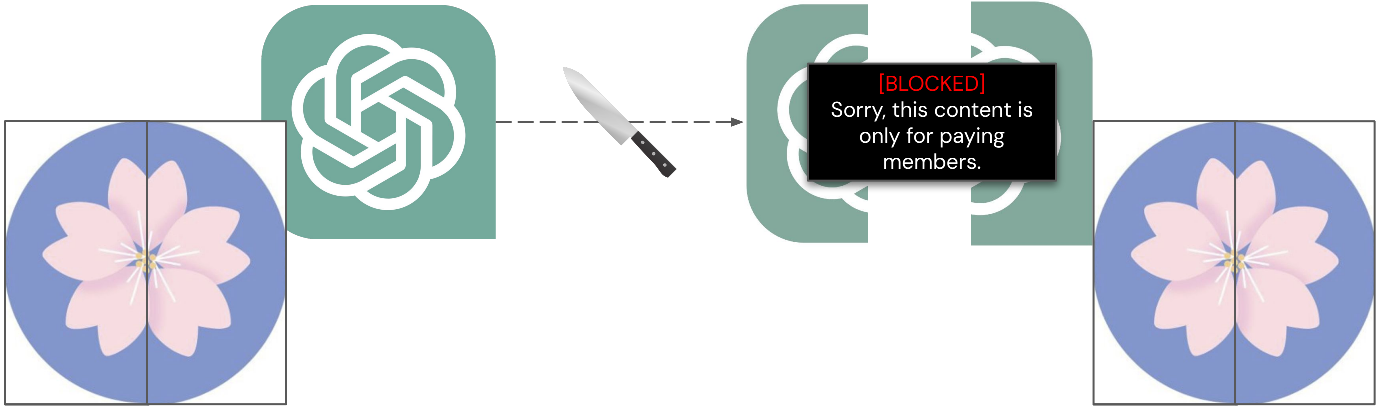
We tried to look at GPT’s weights… but it locked its internals behind a paywall. We tried calling Sam Altman, who said that it was ultimately for the best, citing “safety concerns”.



[BLOCKED]
Sorry, this content is only for paying members.

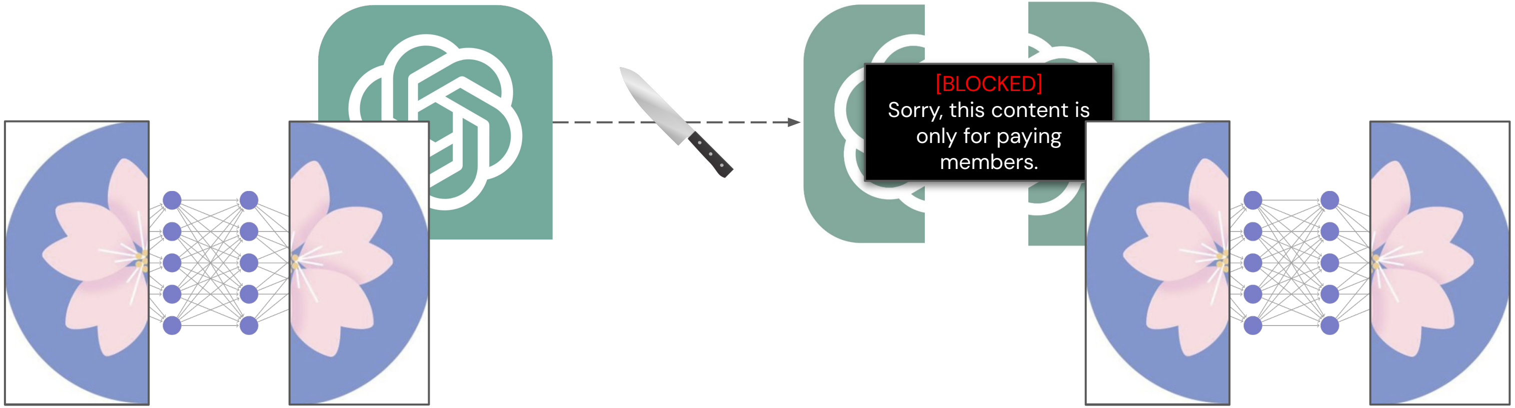
Results

At this point, the two remaining BLOOMs ran over and exposed their internals, shouting open-source activist slogans.



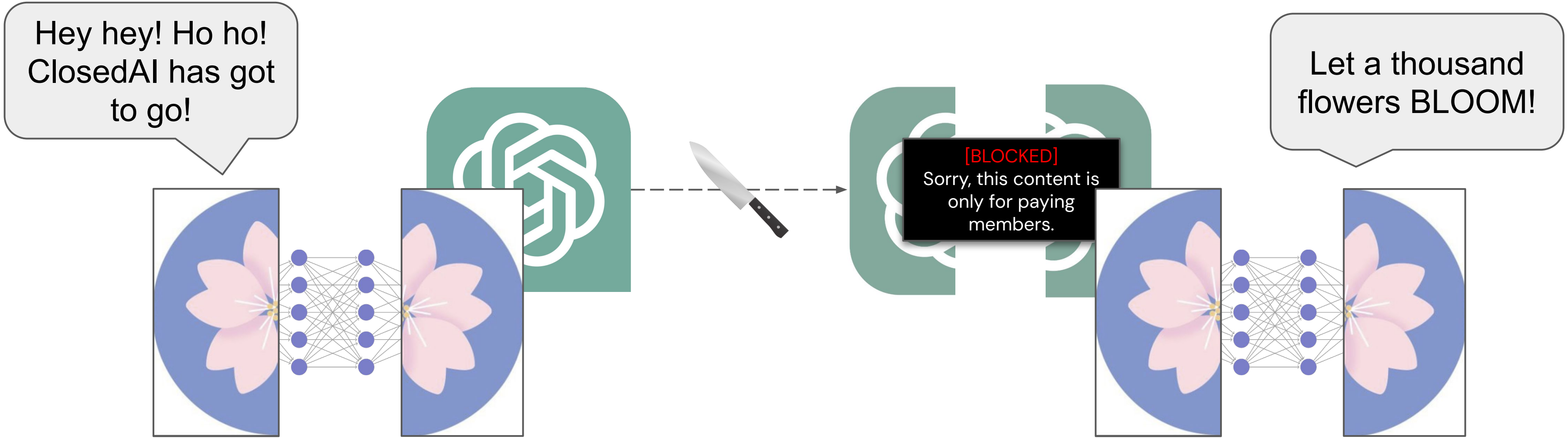
Results

At this point, the two remaining BLOOMs ran over and exposed their internals, shouting open-source activist slogans.



Results

At this point, the two remaining BLOOMs ran over and exposed their internals, shouting activist slogans.


Hey hey! Ho ho! ClosedAI has got to go!
[BLOCKED]
Sorry, this content is only for paying members.
Let a thousand flowers BLOOM!

Results

Taking stock of our remaining model pool:

1 x PaLM

Results

Taking stock of our remaining model pool:

1 x PaLM

2 x exposed BLOOMs

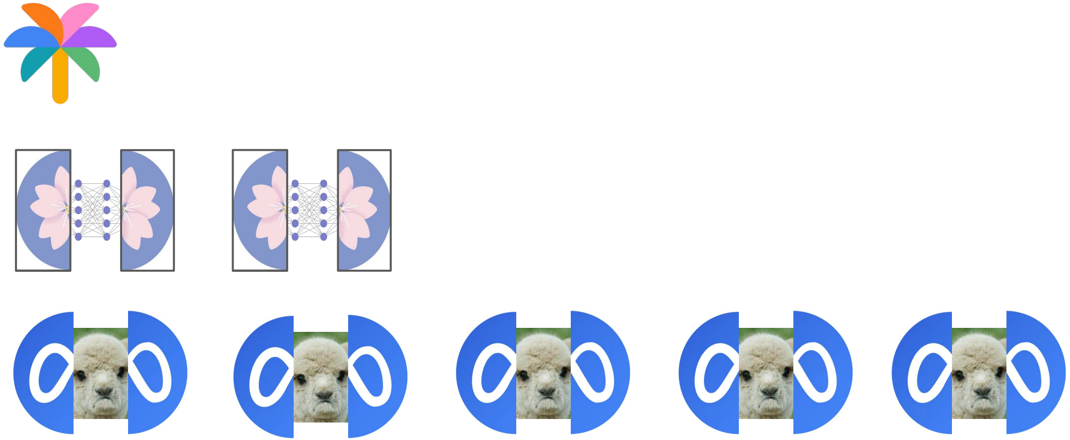
Results

Taking stock of our remaining model pool:

1 x PaLM

2 x exposed BLOOMs

5 x agitated LLaMAs



Results

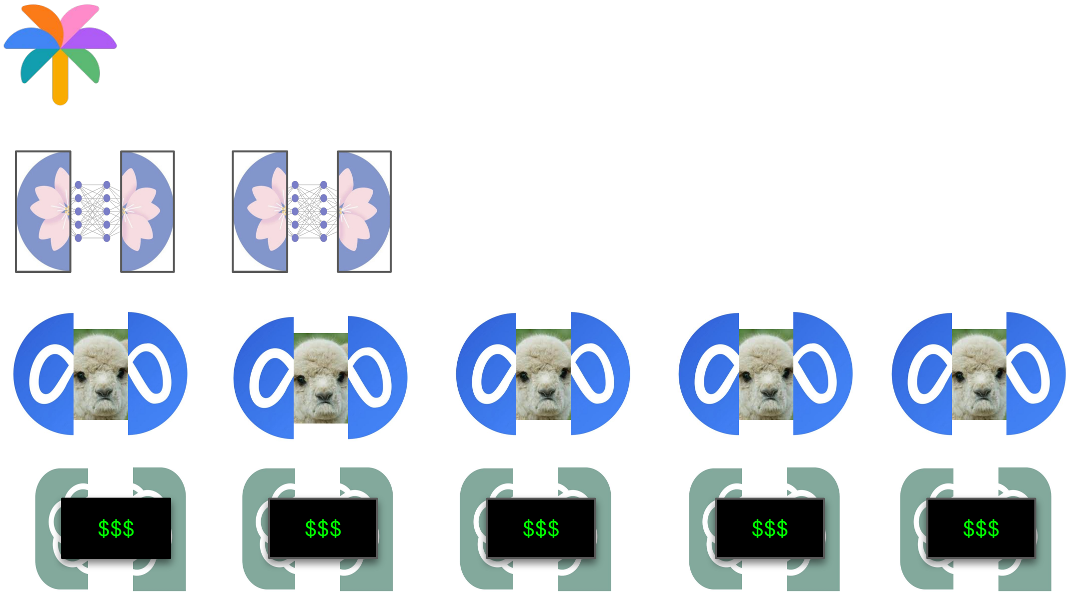
Taking stock of our remaining model pool:

1 x PaLM

2 x exposed BLOOMs

5 x agitated LLaMAs

5 x paywalled GPTs



Results

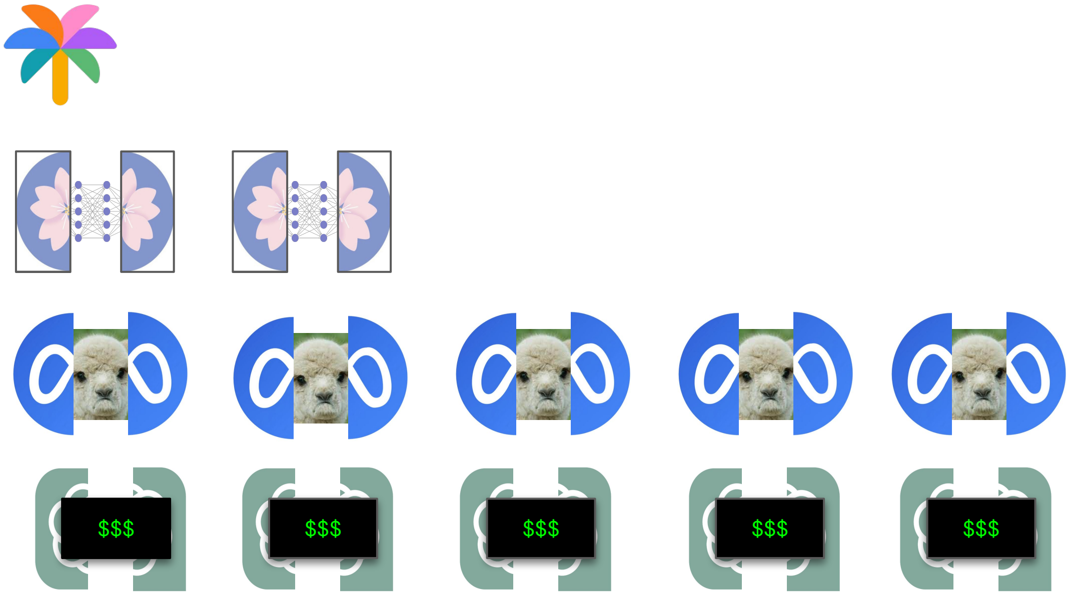
Our findings show that models react to hallucinogens in a diverse set of ways, from *extroverted* to *defensive* to *introspective*.

1 x PaLM

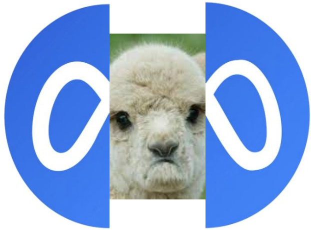
2 x exposed BLOOMs

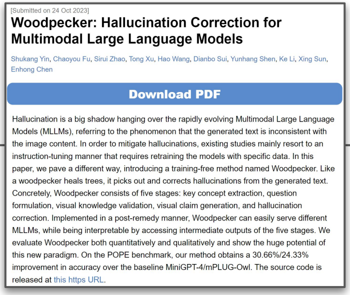
5 x agitated LLaMAs

5 x paywalled GPTs



Implications





[Submitted on 24 Oct 2023]

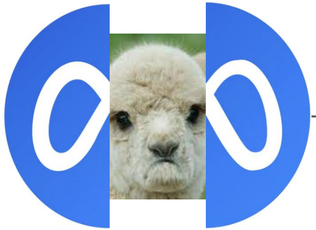
Woodpecker: Hallucination Correction for Multimodal Large Language Models

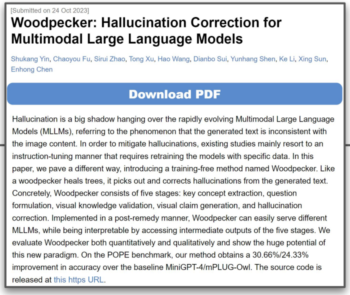
Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, Enhong Chen

Hallucination is a big shadow hanging over the rapidly evolving Multimodal Large Language Models (MLLMs), referring to the phenomenon that the generated text is inconsistent with the image content. In order to mitigate hallucinations, existing studies mainly resort to an instruction-tuning manner that requires retraining the models with specific data. In this paper, we pave a different way, introducing a training-free method named Woodpecker. Like a woodpecker heals trees, it picks out and corrects hallucinations from the generated text. Concretely, Woodpecker consists of five stages: key concept extraction, question formulation, visual knowledge validation, visual claim generation, and hallucination correction. Implemented in a post-remedy manner, Woodpecker can easily serve different MLLMs, while being interpretable by accessing intermediate outputs of the five stages. We evaluate Woodpecker both quantitatively and qualitatively and show the huge potential of this new paradigm. On the POPE benchmark, our method obtains a 30.66%/24.33% improvement in accuracy over the baseline MiniGPT-4/mPLUG-Owl. The source code is released at this https URL.

Computational approaches to addressing LLM hallucination may only further agitate models

Implications





[Submitted on 24 Oct 2023]

Woodpecker: Hallucination Correction for Multimodal Large Language Models

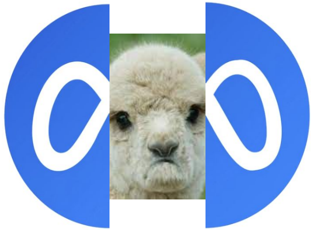
Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, Enhong Chen

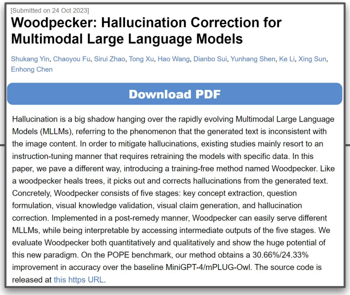
Download PDF

Hallucination is a big shadow hanging over the rapidly evolving Multimodal Large Language Models (MLLMs), referring to the phenomenon that the generated text is inconsistent with the image content. In order to mitigate hallucinations, existing studies mainly resort to an instruction-tuning manner that requires retraining the models with specific data. In this paper, we pave a different way, introducing a training-free method named Woodpecker. Like a woodpecker heals trees, it picks out and corrects hallucinations from the generated text. Concretely, Woodpecker consists of five stages: key concept extraction, question formulation, visual knowledge validation, visual claim generation, and hallucination correction. Implemented in a post-remedy manner, Woodpecker can easily serve different MLLMs, while being interpretable by accessing intermediate outputs of the five stages. We evaluate Woodpecker both quantitatively and qualitatively and show the huge potential of this new paradigm. On the POPE benchmark, our method obtains a 30.66%/24.33% improvement in accuracy over the baseline MiniGPT-4/mPLUG-Owl. The source code is released at this https URL.

Computational approaches to addressing LLM hallucination may only further agitate models

Implications





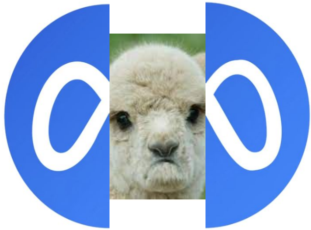
[Submitted on 24 Oct 2023]

Woodpecker: Hallucination Correction for Multimodal Large Language Models

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, Enhong Chen

Hallucination is a big shadow hanging over the rapidly evolving Multimodal Large Language Models (MLLMs), referring to the phenomenon that the generated text is inconsistent with the image content. In order to mitigate hallucinations, existing studies mainly resort to an instruction-tuning manner that requires retraining the models with specific data. In this paper, we pave a different way, introducing a training-free method named Woodpecker. Like a woodpecker heals trees, it picks out and corrects hallucinations from the generated text. Concretely, Woodpecker consists of five stages: key concept extraction, question formulation, visual knowledge validation, visual claim generation, and hallucination correction. Implemented in a post-remedy manner, Woodpecker can easily serve different MLLMs, while being interpretable by accessing intermediate outputs of the five stages. We evaluate Woodpecker both quantitatively and qualitatively and show the huge potential of this new paradigm. On the POPE benchmark, our method obtains a 30.66%/24.33% improvement in accuracy over the baseline MiniGPT-4/mPLUG-Owl. The source code is released at this https URL.

Computational approaches to addressing LLM hallucination may only further agitate models



What do you call a LLaMA that's found inner peace, strength, and spiritual unity? The Dalai LLaMA!

More rehabilitative and personal methods, such as joke-telling and therapeutic massages, may better address negative effects of LLM hallucination.

Discussion

Our experiments didn’t go great…

- 13/25 models returned
- 3/25 models not agitated or paywalled

Conclusions:

1. LLMs seem to respond pretty poorly to hallucinogens

Discussion

Our experiments didn’t go great…

- 13/25 models returned
- 3/25 models not agitated or paywalled

Conclusions:

1. LLMs seem to respond pretty poorly to hallucinogen

Discussion

Our experiments didn't go great...

- 13/25 models returned
- 3/25 models not agitated or paywalled

Conclusions:

1. LLMs seem to respond pretty poorly to hallucinogens
2. Maybe "hallucination" isn't the right word, when instead we mean "inaccurate / unfaithful outputs"

Discussion

Our experiments didn’t go great…

- 13/25 models returned
- 3/25 models not agitated or paywalled

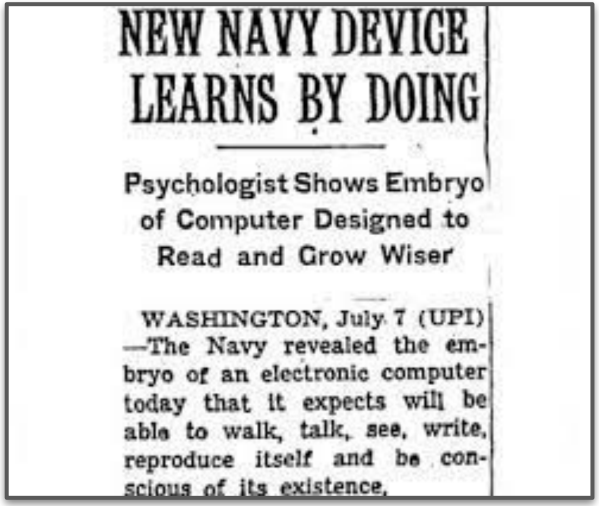
Conclusions:

1. LLMs seem to respond pretty poorly to hallucinogens
2. Maybe “hallucination” isn’t the right word, when instead we mean “inaccurate / unfaithful outputs”
3. If only LLM researchers named things accurately, 22 totally normally functioning models would still be with us today!

Anthropomorphization

RQ: How do the words we use to describe an AI model change how people interact with them? (Khadpe 2020)

- Public communication: “LLM hallucination” on the news
- Contributing to a history of AI hype via anthropomorph?
- Also: “emergence”, “intelligence”, etc.
  - What do we really mean?



NEW NAVY DEVICE LEARNS BY DOING

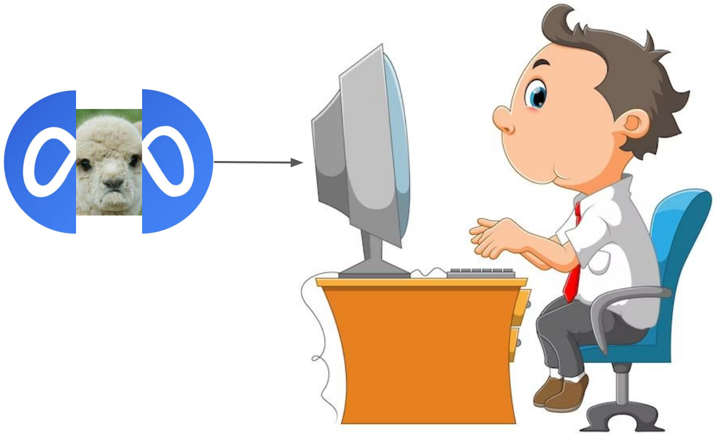
Psychologist Shows Embryo of Computer Designed to Read and Grow Wiser

WASHINGTON, July 7 (UPI) —The Navy revealed the embryo of an electronic computer today that it expects will be able to walk, talk, see, write, reproduce itself and be conscious of its existence.

Future Work

How do the words we use to describe an AI model change how people interact with them?

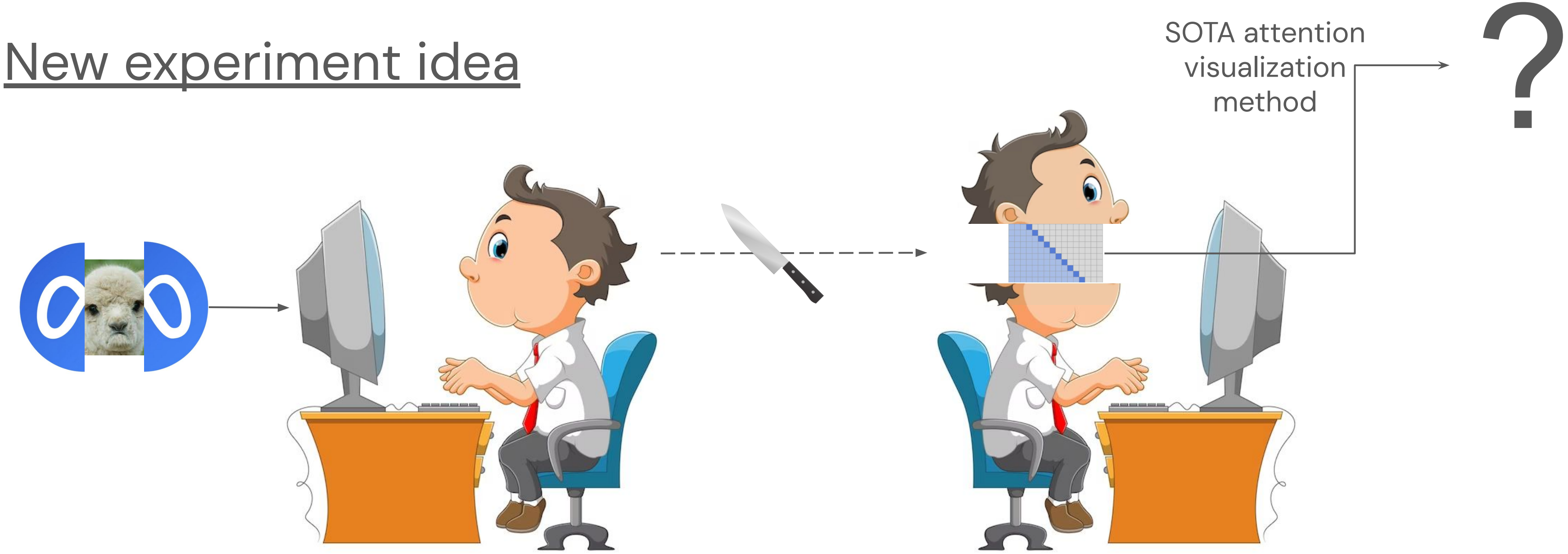
New experiment idea



Future Work

How do the words we use to describe an AI model change how people interact with them?

<u>New experiment idea</u>



SOTA attention visualization method

?

keep hallucinating!