



Open Discussion #1

PORTO
July 2024

IARIA Congress 2024 & DigiTech 2024

Theme: A Deep Dive on AI-tools Hallucinations

COORDINATORS



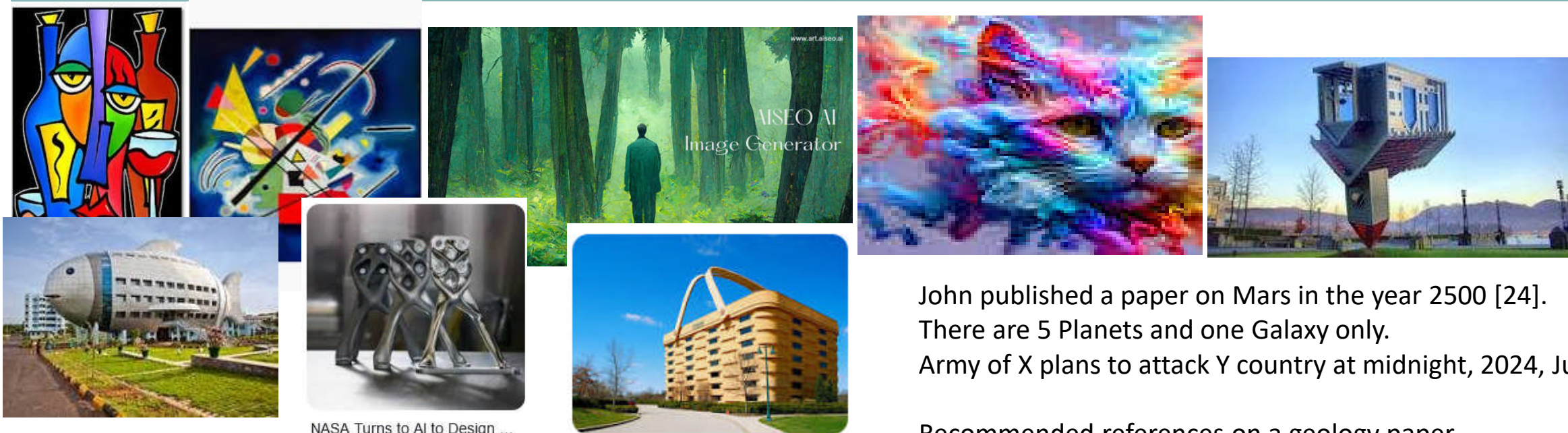
Prof. Dr. Petre Dini, *IARIA, USA/EU*

Dalmo Cirne, *Workday, Inc., USA*



Inquiries on the Table

PORTO
July 2024



NASA Turns to AI to Design ...

Our Top Five Strangest Build...

2+2 = 5, all lines are straight, 2 = 10, log(-3) = x
Horses ride a bicycle over the shining cloud of dust!!
Aliens will invest in The House of the rising Sun!
Alice's Adventures in the Wonderland

Metaverse
VR AVR
Immersion

John published a paper on Mars in the year 2500 [24].
There are 5 Planets and one Galaxy only.
Army of X plans to attack Y country at midnight, 2024, June 30.

Recommended references on a geology paper.

- [x] Pierre, Title 1
 - [z] Jacobs, Title 2
 - [y] Stan, Title 3
- Note: [x] doesn't exist
[z] has another title
[y] is a carpenter instructions book



LLMs-based: - very good summarization of information they are fed with, even only less than 1% validated as true
- very good mixed (4-5-6 ...) languages, correct punctuation, correct grammar, spelling correction on context-based intuition
- helpful at the informative level, like white papers, very quickly obtained and quite comprehensive
- assumes user's familiarity and experience with a given domain; see, selection an oscilloscope for 5G spectrum, financial aspects,...



OD#1/Intro

PORTO
July 2024

- **Hallucinations: Starting from validated real facts (datasets) and getting incorrect (irrational) answers/output**
- **Machine hallucinations are in fact, immature LMM training, triggered by the nature of learning and understanding.**
 - For humans, learning is associated with the perception of meaning and consequences of actions.
- **LLM hallucinations are induced by the exploring mechanisms and by the nature of the input data.**
 - Hallucinations of output from artificial entities (LLMs,) are unavoidable, as learning is without understanding and there is no perception of the output.
- **Only humans can detect hallucinations (vs. what so considered 'expected')**
- **Hallucinations are induced by the training process based on 'one word at a time' and 'the best guess at each step'.**
- **Hallucinations are induced by limited backward attention in focusing on forward target**
- **Key considerations in *machine vs human* learning (our evaluation should be performed accordingly)**
 - Learning/acting with consequences
 - Learning by understanding the meaning (or by **attaching** a meaning)
 - Learning **by chunks vs by words**; no human will use sentences 'this, this is, this is a, this is a big' without a context by guessing the best next word)
- **Note:**
 - Apart *on purpose* improper or incorrect training, when doubt in a learning point, the best option (context) is chosen.
 - See going over mountains from **A** to **B**.

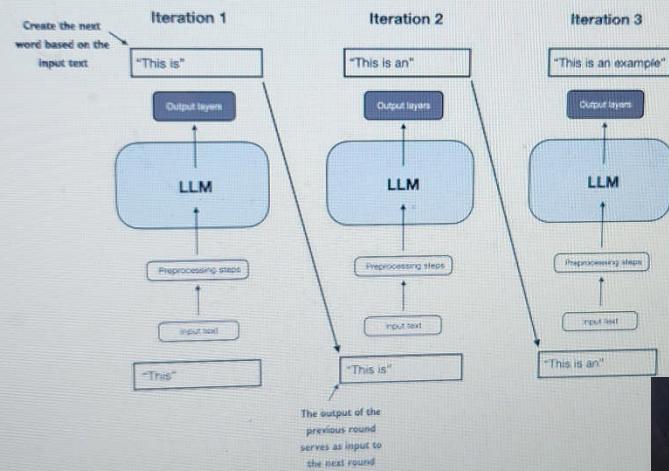




OD#1/Intro

PORTO
July 2024

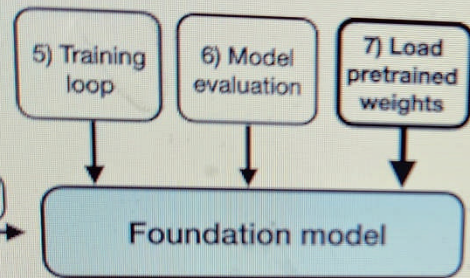
How do LLMs generate multi-word outputs?



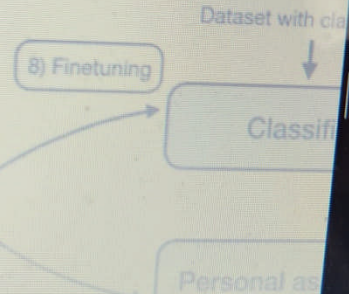
Llama 3 was trained on 15T tokens

"To train the best language model, the curation of a large, high-quality training dataset is paramount. In line with our design principles, we invested heavily in pretraining data. Llama 3 is pretrained on over 15T tokens that were all collected from publicly available sources."

STAGE 2: PRETRAINING



STAGE 3: FINETUNING



Sebastian Raschka

TechTalk Slides

Captions...

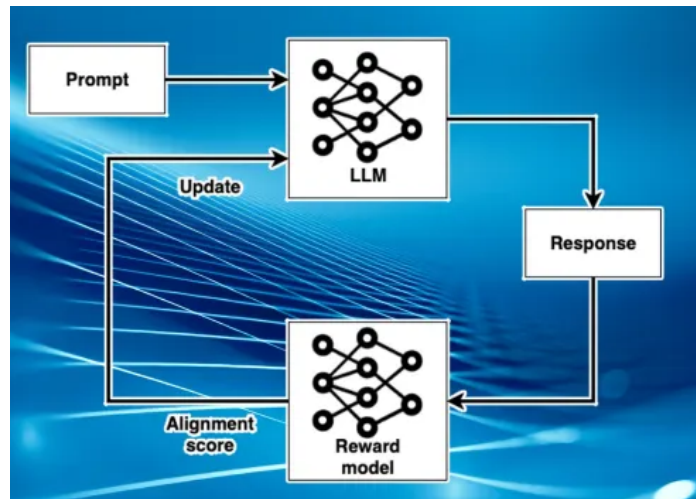


OD#1/Intro

PORTO
July 2024

<https://aws.amazon.com/what-is/reinforcement-learning-from-human-feedback/>

<https://bdtechtalks.com/2023/01/16/what-is-rlhf/>



RLHF
Datasets

Q&A
Datasets

Step 1

Collect demonstration data and train a supervised policy.

A prompt is sampled from our prompt dataset.

A labeler demonstrates the desired output behavior.

This data is used to fine-tune GPT-3.5 with supervised learning.



Step 2

Collect comparison data and train a reward model.

A prompt and several model outputs are sampled.

A labeler ranks the outputs from best to worst.

This data is used to train our reward model.



Step 3

Optimize a policy against the reward model using the PPO reinforcement learning algorithm.

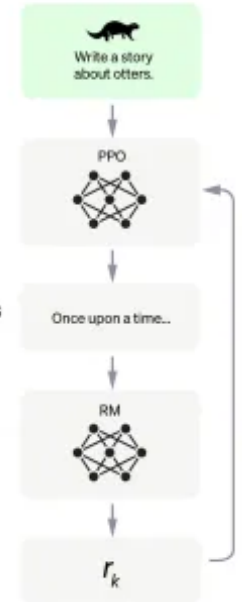
A new prompt is sampled from the dataset.

The PPO model is initialized from the supervised policy.

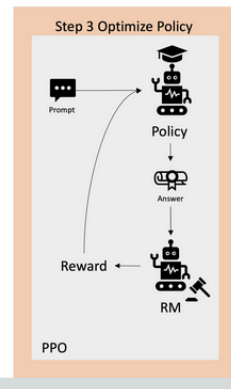
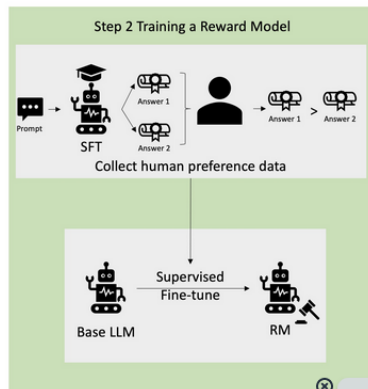
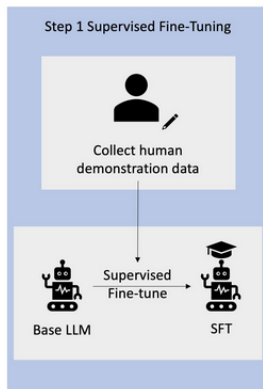
The policy generates an output.

The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.



RLHF is also not a perfect solution. Human feedback can help steer LLMs away from generating harmful or erroneous results. But human preferences are not clear cut, and you can never create a reward model that conforms with the preferences and norms of all societies and social structures.





OD#1/Dalmo

PORTO
July 2024

■ Are hallucinations a bug or a feature?

- We don't know how knowledge is created, but imagining what doesn't yet exist is part of it. Is creativity a form of hallucination?
- Creativity may be unwelcome sometimes. For instance, a financial report or a medical diagnose
- Deterministic procedures or specialized domains can be worked with by using augmentation or calls to external systems (e.g., calculator)
- More data, more training, more scrutiny may not be the answer
 - More data will carry similar semantic meaning to existing data
 - What is considered non-biased today, may be seen as biased tomorrow



Dalmo Cirne
Workday, Inc.



OD#1/Dalmo

PORTO
July 2024

■ Optimization for error correction and experimentation

- Is it possible to prevent errors? What if those are not errors, but knowledge that is changing?
- Experimenting when validating a scientific hypothesis, cooking a meal, writing a text, are all filled with errors. Preventive correction may not be an option because prior knowledge doesn't exist or is being reevaluated
- What are the evolutionary consequences to a system (any system), if errors and mistakes are not allowed?
- Hallucinations can be seen as a mechanism for error correction. This is applicable to both humans and machines



Dalmo Cirne
Workday, Inc.



OD#1/Dalmo

PORTO
July 2024



Dalmo Cirne
Workday, Inc.

What if hallucinations are not a problem to be solved, but instead are the dials and knobs of a new tool that we need to learn how to and when to use?



OD#1/Petre

PORTO
July 2024

■ Hallucinations are human, in nature

- Hallucinations are human, in nature (Hallucinations are glimpses of reality)
- Practical Leisure: Fables, Science Fiction novels, Poetry, Painting (except for realism, in the end, but also too idyllic), stories, especially children's books, some music styles, etc.
- Inner Experiences, Visionary dreams, Soulful imaginations: Wish-Tale, imaginations, Mystical visions, Flying Angels,
- Real inventions: in fact, most of them apriori unintended.

■ AI/LLM Hallucinations (from humanly perceived reality)

- Generated narratives, generated music, generated poetry, generated nano-pictures, artificial facts/tale,...

- **I SUGGEST**, for the benefit of AI-based tools, credibility, and transparency, all the above should be labeled as **intended hallucinations** (intended machine hallucinations, or AI-based creations), in opposition to **unintended hallucinations**.



Petre Dini
IARIA

■ Where are hallucination harmful?

- When critical predictions, forecasts, or hopes fail, leading to loss of something (material, beliefs) or a human being.
- In mission critical decisions (health, investments, politics, wars, etc.); some might wrongly alter their correct knowledge
- **Note 1**: The above suggestion sets the liability, as nobody can longer justify an action throwing the responsibility on AI-based tools' hallucinations
- **Note 2**: The most dangerous action is triggering conclusion based on potential hallucinations issued from simulations of real systems.

A notorious case of a simulated AI-powered UAV drone destroying the Command Center due to potential hallucinations and misdirection was presented as a real fact on a real system (Pentagon) and as an imminent threat leading to the extinction of humanity.



OD#1/Petre

PORTO
July 2024

■ Both Human and AI/LLM hallucinations can be on purpose or unintended.

- The question is that AI/LLM hallucinations cannot be controlled, there is a lack of understandability and explainability, and, even more, 'they are rarely timely discovered'.
- In LLM trained with more than 13 trillions of parameters, it is practically towards impossible to identify a cause.
- So, hallucinations represent a significant risk, but also opportunities; they are part of our existence.
- A few examples:
 - most of the side-effect achievements of NASA led to practical applications (mostly materials/clothes, solar panel, etc.)
 - C vitamin
 - Tires
 - almost all the inventions

■ Hallucinations vs. Synesthesia vs. Dreams vs. Intended actions

■ *Inherited Genetics or Human Subjectivism*

- a. Different perceptions: Wassily Kandinsky saw color when he heard music (complex perceptions)
- b. Different Interpretations of the Same Situations (embedded feelings) [humans excel in this diversity]

■ *By Training*

- a. Wrong Information: fake, incomplete, obsolete, inaccurate, biased
- b. Unbalanced Information (taxonomy, Q&A Datasets, overfitting, underfitting, predefined answers in interactive tools, alternative answers (answers 1 or answer 2),
- c. Poor quality of data and inadequate Datasets validation.
- d. Lack of human feedback, as this might be the only way to feel pain on a wrong output
- e. On purpose - misleading



Petre Dini
IARIA



OD#1/Petre

PORTO
July 2024

■ Under consideration (LLM-related)

- Customization of Datasets (notice that 1 million of data points is a very small Dataset)
- RAG (Retrieval-Augmented Generation) used for dedicated domains (financial, legal, sports, etc.)
- Reinforcement Learning by Human Feedback - accuracy, honesty, and skills are downsides
- **Needs**
 - Increase spanning attention over large sequences
 - Bidirectional processing (see Bengal language, where sequence of the words is mostly irrelevant) via activating or deactivating the attention masks
- **Concrete situation:** Disabling Attention Masks.
 - "Disabling the attention masks" means that the model will no longer apply these masks during the **attention computation**.
 - This can have several implications:
 - **No Padding Mask:** If the padding mask is disabled, the model may incorrectly attend to padding tokens, which can lead to incorrect or meaningless outputs. For example, if a sentence is padded with [PAD] tokens, the model might give undue importance to these tokens.
 - **No Causal Mask:** Disabling the causal mask in an autoregressive model means that tokens can attend to future tokens. This breaks the autoregressive property and can lead to issues during generation, where the model might use information from future tokens that it should not have access to yet.



Petre Dini
IARIA

■ Appendix: Understanding Attention Masks

- *In the context of Large Language Models (LLMs), such as those used in transformers, "disabling the attention masks" refers to altering the way the model processes sequences of data during training or inference.*
- *Attention masks are a key component in transformer architectures like BERT, GPT, and others. They are used to manage the attention mechanism, which allows the model to focus on different parts of the input sequence. Typically, attention masks serve the following purposes:*
- *Padding Mask: To handle variable-length sequences by distinguishing actual data from padding tokens. For example, in a batch of sequences of different lengths, shorter sequences are padded with a special token (e.g., [PAD]). The attention mask ensures that these padding tokens do not affect the model's computations.*
- *Causal Mask: In models like GPT, which generate text sequentially, causal masks ensure that each token only attends to previous tokens and not future ones, preserving the autoregressive property.*



OD#1/Petre

PORTO
July 2024

■ Conclusion

- **Hallucinations caused by the LLM approaches**
 - LLM efficiency and accuracy are not too high considering energy and computation power for training.
 - One token at a time, ...
 - No memory
 - No long span over an acquired knowledge (read: 'text') towards the target (short horizon; looking for bidirectional processing).
- **Hallucinations are induced by the process and by selected data**
 - We will always live with an 'accepted level of hallucinations', in a real society or a digital one, as hallucinations are here to stay
 - Awareness and adequate literacy is needed for preventing personal damages (hopefully, in our control)
 - There is a danger of being used by wrongdoers (at any level) by altering the process or altering data
- **Be aware**
 - Educate yourself for identifying the cause and find tips to reduce the risks of LLM hallucinations in interacting with AI-based tools.
 - Mitigate at a personal-scale the LLM hallucinations for minimizing the negative impact
 - Distil the Ethical Hallucinations and understand unintended consequences
 - Identify and make profit of Good hallucinations and discard Bad Hallucinations



Petre Dini
IARIA



■ Caveat

Hallucinations might (and will) be used for any kind of justification (of the output, or of an action) by wrongdoers and/or bad actors

- **Overfitting**, **underfitting**, and **unintended consequences** are related notions for any excuse



OD#1/Petre

PORTO
July 2024

Endorsement of my suggestion by ChatGPT 4.0



Petre

- Generated narratives, generated music, generated poetry, generated nano-images, artificial facts/stories,... should not be labeled as intentional machine hallucinations (or AI-based creations), i.e., explicitly labeled differently, for to be distinguished from the well-known LLM hallucinations?



ChatGPT 4.0

- Labeling AI-generated content to distinguish it from unintended hallucinations in **LLMs is a good idea to ensure clarity and transparency**. The term "intended hallucinations" can be a bit misleading because "hallucinations" generally have a negative connotation in the context of AI, implying errors or unintended outputs. Instead, using a different term that clearly conveys the deliberate nature of the creation can be more effective.
- Here are a few alternative suggestions:
 - AI-Crafted Content:** Emphasizes that the content is created intentionally by the AI.
 - Example label: "AI-Crafted Narrative" or "AI-Crafted Music"
 - Synthetic Creations:** Highlights that the content is synthetically generated.
 - Example label: "Synthetic Poetry" or "Synthetic Facts"
 - Machine-Generated:** Clearly states that the content is generated by a machine.
 - Example label: "Machine-Generated Tale" or "Machine-Generated Nano-Picture"
 - AI-Generated:** Simple and widely understood term indicating AI involvement.
 - Example label: "AI-Generated Story" or "AI-Generated Artwork"
 - Creative AI Output:** Focuses on the creative aspect of the generated content.
 - Example label: "Creative AI Narrative" or "Creative AI Poetry"
 - Artificial Creations:** Clearly conveys that the content is artificially made.
 - Example label: "Artificial Narrative" or "Artificial Music"
- Using any of these labels can help differentiate intended creative outputs from unintended hallucinations, maintaining clear communication about the nature and origin of the content.



Petre Dini
IARIA



Inquiries on the Table

PORTO
July 2024

Items under scrutiny

- a. Are human hallucinations more acceptable than machine hallucinations?
- b. Why can hallucinations of artificial machines occur?
- c. How to spot damageable hallucinations?
- d. How to improve the LLMs processes for minimizing hallucinations?
- e. ...
- f. ..
- g. .



OD#1

**PORTO
July 2024**

STAGE IS YOURS