

# **GPTs: Concerns, Limitations and (Some) Responses**

**Martin Frické**

**[15 minute talk to Information on Tap, Borderlands Brewery,  
April 18<sup>th</sup> 2024]**

## Table of Contents

1. Introduction .....	2
2. What worries might be in order about GPTs?.....	3
2.1 Harms relying on text, image, or video generation .....	3
2.2 Environmental costs .....	4
2.3 Apparent conflict with Chomsky’s theories.....	5
2.4 Hallucinations (Confabulations) .....	6
2.5 Bias and unfairness .....	8
2.6 Source data, intellectual property, and privacy.....	9
2.7 Cybersecurity.....	10
2.8 Lack of transparency and XAI .....	11
3. Additional references used .....	12
4. Bibliography.....	13

### 1. Introduction

Folk working in this field, when speaking casually, often use the titles *Large Language Models(LLMs)* or *Foundation Models* for GPTs (Generative Pre-trained Transformers). I may slip into that.

There is a paper *On the Opportunities and Risks of Foundation Models* (Bommasani et al. 2022) which was written by over one hundred experts. It

is 200 pages long, with a 60-page bibliography. It covers the whole subject matter. It is recent. You might want to look at this paper.

I'm going to try to pitch my talk at the intellectual level of a member of the public interested in GPTs and with some knowledge of them.

Ok. Eight shortcomings or questionable aspects (given a minute and a half each).

## **2. What worries might be in order about GPTs?**

### ***2.1 Harms relying on text, image, or video generation***

Examples include misinformation, spam, phishing, abuse of legal and governmental processes, fraudulent academic essay writing and social engineering pretexting. (Brown et al. 2020).

We can add video or audio fakes and deepfakes to these.

There is very little that can be done about this. The countermove to these is information literacy. Educate the people.

That materials have been produced by AI cannot be recognized reliably. Then, obviously, using AI to generate not always bad. For example, summaries of text, journal articles, and even entire document collections, can be exactly what readers require.

Going forward, imagine this. Not all university faculty are engaging in their lecturing. It will be possible to deepfake faculty summarizing their own lectures in video, in a lively and educationally contentful style. A Putin speech can be 3 hours long and in Russian. What is so wrong with a news site presenting Putin himself apparently summarizing such a speech in 5 minutes in English?

## ***2.2 Environmental costs***

LLMs use resources. The resources will include computer chips (GPUs), data storage facilities, electricity, water, and further infrastructure. I will consider water only.

A typical statement from a commentator is this:

An average user's conversational exchange with ChatGPT basically amounts to dumping a large bottle of fresh water out on the ground... (DeGeurin 2023) [and there are 100 million or more users.]

I doubt that there need be such a water waste. The water is used for cooling the computers. It can be recycled. Think of ordinary cars. They have water in a radiator to cool their engines. No water is consumed.

What data centers do now is to cool their server rooms. They do this using evaporative coolers, what we in Arizona call 'swamp-coolers'. They then partially capture and recycle the evaporated water. But they could recycle the

water itself in a closed loop directly through the computers— no evaporation, little or no loss.

### ***2.3 Apparent conflict with Chomsky’s theories***

There is a nexus of theories, originating from Noam Chomsky, that there is a universal grammar, which is innate, and which is common to all peoples. Universal grammar is first to explain how it is that children can learn their respective native languages simply and astonishingly quickly. It posits a deep structure that is not manifest, or immediately learnable, in the bare surface appearances of the spoken and written languages. The nexus also offers explanations of many other features of human languages. However, the theories seem to stand in conflict with the existence and behavior of large language models. LLMs seem to learn language, and its structure, merely by looking at a huge amount of surface text. Chomsky’s response, more-or-less, is that LLMs are a kind of surface statistical trick and that they do not give any real insight into linguistic structures (Chomsky, Roberts, and Watumull 2023; Chomsky and Mirfakhraie 2023). The response of some others, for example Steven Piantadosi, is that LLMs are evidence that Chomsky’s theories are mistaken (see, for example, (Piantadosi 2023)). This matters in the following way. Were Chomsky’s theories shown to be mistaken, that would be a major scientific discovery. On the other hand, if LLMs are merely statistical tricks, we should be even more wary of them in use than we already are.

Chomsky is surely right that LLMs are not how the human the human mind works. Infants are simply not exposed to the vast amounts of text that LLMs require. The paucity of the data refutes the view that full human language capabilities rest on LLMs.

Might mere correlations in exposure to linguistic data be the origin of human access to some linguistic capabilities, nothing innate required? After all, LLMs seem to show that considerable linguistic skills can be acquired *by computers* just from data. Chomsky thinks not. Basically, as a philosopher of science, he is skeptical of the inductivist view that theory-free studying of data will lead to discovery. He is not the only one with that skeptical view (see (Frické 2015; Pearl and Mackenzie 2018)). Chomsky then argues: his style of linguistics meets the demands of this philosophy and LLMs do not (Chomsky, Roberts, and Watumull 2023).

In sum, Chomsky's theories may be mistaken, just as any scientific theories might be mistaken. But it not clear that LLMs are critical counter-evidence.

## ***2.4 Hallucinations (Confabulations)***

An instructive example of an LLM producing hallucinations is Meta's *Galactica* (Taylor et al. 2022). This was an LLM for science. It was released on November 15<sup>th</sup> 2023 and lasted 3 days online before it was withdrawn. William Heaven writes:

A fundamental problem with *Galactica* is that it is not able to distinguish truth from falsehood, a basic requirement for a language model designed to generate scientific text. (Heaven 2022).

What might be a response to confabulations? Let us assume that there are answers, evidence, or arguments, somewhere outside the LLM itself. They may exist in databases or on the web (or even, going forward, in the heads of experts). Then RAG (Retrieval-Augmented Generation) or ai-generated search in the style of *Perplexity* are ways forward (Perplexity 2024).

RAG consists of consulting outside data. LLMs are trained at a particular time, and the training takes time (usually months). So, if LLMs are not going to be knowledgeable about new unfolding data (for example, yesterday's election results), they are going to need the ability to use data that they were not trained on.

*Perplexity* will use a search engine to find the materials relevant to the conversation. It will summarize those materials and return the summary together with the citations to the evidence.

There are many problems in the details here. But this is a promising line.

## **2.5 Bias and unfairness**

Kate Crawford makes a distinction between harms of allocation, harms of representation, and harms of classification (Crawford 2017).

Harms of allocation, e.g. fairness in the availability of mortgages, should be able to be addressed. Mathematics can do this. It can identify the bias and provide the remedies.

Harms of representation, e.g. Muslim-violence bias (Abid, Farooqi, and Zou 2021), are a much harder case. There are hundreds of training data sets of (English) samples. The ones of these that are large and containing substantial source material from the Internet (e.g. from Common Crawl) will have some harms of representation bias. Then, if the training data has bias then so too will LLMs. To address this, there seem to be two central possibilities: to keep the unacceptable bias out of the training data, or to remove the bias from the model's output. Neither of these is promising. The systems use self-supervision on the training data precisely because it is near impossible to curate and label the data with the quantity involved. Working on the output probably is not much better. It may be possible for LLMs to prompt the systems themselves to remove some biases of representation.

Harms of classification also are tricky and require attention. The set up here is that some aspects of people, for example images or job applications, are to be classified. Then the resulting classification is used for some purpose, say employment, promotion, or the award of visas. There are potential problems



everywhere doing this using machine learning: with the data, the algorithmic pipeline, the creators of the systems, and the companies and stakeholders behind the LLMs. However, the challenge is not entirely novel. There were harms of classification long before machine learning and LLMs (as an extreme example, the now historical apartheid system in South Africa) (Bowker and Star 2000). We do have some insight, techniques, and experience on how to detect and remedy harms of classification.

## ***2.6 Source data, intellectual property, and privacy***

Typically, the LLMs will use self-supervised (i.e. unsupervised and unlabeled) pre-training from a good portion of the Internet. A proportion of that data will be intellectual property, perhaps even carrying copyright notices. Other parts may be private— names, addresses, etc.— and an LLM will have the ability to collate such information. A nefarious User, using the right sequence of prompts, may be able to get the LLM to collate information across disparate sources (Marcus and Southen 2024; Nasr et al. 2023).

The copyright aspects probably do not matter, providing the LLM does not return some of its pre-training data verbatim. There are copyright laws, but these vary in important ways from country to country (and some are being changed in the face of AI).

Intellectual property is important. An example. The image generators DALL-E and Stable Diffusion can produce images in the style of well-known artists or video game illustrators e.g. Greg Rutkowski. The images are *not* copies of

images, but the *style* can be a copy of a style. This might be a problem. There are now hundreds of thousands of images on the web that look as though they had been created by Greg Rutkowski. This is not fair to Greg Rutkowski and his means of making a living.

Privacy is important. But there is a qualification. Say an LLM was prompted to return my address and phone number. It could only do that if my address and phone number were publicly available in the training data e.g. on the web. It is in my hands to prevent that.

## ***2.7 Cybersecurity***

Suitably trained LLMs can write computer programming code, to a very high standard. This means that they could be used to write viruses and various kinds of cybersecurity defeating software. OpenAI themselves, in their GPT-4 Technical Report, describe how GPT-4 defeated CAPTCHA (which is the familiar test to distinguish a human from a computer). Essentially, GPT-4 employed a human from Task Rabbit, told the human that he/she/they was visually impaired, and got the human to do the test for him/her/them (OpenAI 2023, 55). LLMs should not be underestimated in the hands of bad actors.

I'm not sure of the appropriate response to issues of cybersecurity. But probably not having the source code of an LLM open-source helps. Some LLMs are, some are not.

## ***2.8 Lack of transparency and XAI***

Since about 2017, the LLM companies and research labs typically do not reveal their methods. This may be good for safety and security. It is not good for working out the environmental impacts of the systems.

Lack of transparency is also not good for explaining to users what is happening or what happened with specific predictions (such as, in a medical setting, why the LLM prediction is that the user has cancer). In fact, with LLMs often the companies do not themselves know in detail how their systems work or why a certain prediction is the outcome. This results from the size and complexity of the LLMs. That the companies do not know how their systems work is an obstacle in their question to improve them.

There is a research field, Explainable Artificial Intelligence (XAI):

Explainable AI is Machine Learning that has the property of being easily understood by humans (Wikipedia 2023) .

What, in this context, is understanding?

A simple but plausible answer given by contemporary philosophers of science is as follows: to understand a phenomenon is to grasp how the phenomenon is caused (Strevens 2013).

But, with LLMs, causes are exactly what we do not have.

Elsewhere, we do negotiate our lives using correlations and conjectures on causality. The farmer puts fertilizer on crops, we take aspirin for headaches, and most of us avoid smoking for health reasons. Some of these connections are undoubtedly causal. But we usually do not know of the fine-grained details of the causality. Basically, the systems are close to being black boxes. There is a recent research field on what are known as causal diagrams, principally from Judea Pearl (Pearl and Mackenzie 2018). Causal diagrams, and correlations, can work well with agriculture, medical science, and many other areas. XAI may be able to use causal diagrams, and correlations, to produce explanations of LLM predictions. It is an open research area.

### **3. Additional references used**

These might not have been cited in the text. They will be in the Bibliography.

(Bender et al. 2021)

(Bommasani et al. 2022)

(Bommasani et al. 2023)

(Brown et al. 2020)

(Bontcheva 2024)

(Cartter 2023):

(Hacking 1999)

(Ibbotson and Tomasello 2016)

(Katz 2012)


(Li et al. 2023)

(Moro 2016)

(OpenAI 2023)

(Oppy and Dowe 2021)

#### 4. Bibliography

- Abid, Abubakar, Maheen Farooqi, and James Zou. 2021. “Persistent Anti-Muslim Bias in Large Language Models.” arXiv. <https://doi.org/10.48550/arXiv.2101.05783>.
- Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? .
- In
- Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*
- , 610–23. FAccT ’21. New York, NY, USA: Association for Computing Machinery.
- <https://doi.org/10.1145/3442188.3445922>
- .
- Bommasani, Rishi, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, et al. 2022. “On the Opportunities and Risks of Foundation Models.” arXiv. <https://doi.org/10.48550/arXiv.2108.07258>.
- Bommasani, Rishi, Kevin Klyman, Shayne Longpre, Sayash Kapoor, Nestor Maslej, Betty Xiong, Daniel Zhang, and Percy Liang. 2023. “The Foundation Model Transparency Index.” arXiv. <https://doi.org/10.48550/arXiv.2310.12941>.
- Bontcheva, Kalina. 2024. “Generative AI and Disinformation: Recent Advances, Challenges, and Opportunities.”
- Bowker, Geoffrey C., and Susan Leigh Star. 2000. *Sorting Things out: Classification and Its Consequences*. Cambridge, MA: The MIT Press.
- Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. 2020. “Language Models Are Few-Shot Learners.” arXiv. <https://doi.org/10.48550/arXiv.2005.14165>.
- Cartter, Eileen. 2023. “The Pope Francis Puffer Photo Was Real in Our Hearts.” GQ. 2023. <https://www.gq.com/story/pope-puffer-jacket-midjourney-ai-meme>.
- Chomsky, Noam, and Ramin Mirfakhraie. 2023. “ChatGPT and Human Intelligence: Noam Chomsky Responds to Critics | MR Online.” 2023. <https://mronline.org/2023/04/24/chatgpt-and-human-intelligence-noam-chomsky-responds-to-critics/>.
- Chomsky, Noam, Ian Roberts, and Jeffrey Watumull. 2023. “Opinion | Noam Chomsky: The False Promise of ChatGPT.” *The New York Times*, 2023, sec. Opinion. <https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html>.
- Crawford, Kate, dir. 2017. *The Trouble with Bias - NIPS 2017 Keynote - Kate Crawford #NIPS2017*. Neural Information Processing Systems. [https://www.youtube.com/watch?v=fMym\\_BKWQzk](https://www.youtube.com/watch?v=fMym_BKWQzk).
- DeGeurin, Mack. 2023. “Training ChatGPT Required Enough Water to Fill a Nuclear Cooling Tower.” Gizmodo. 2023. <https://gizmodo.com/chatgpt-ai-water-185000-gallons-training-nuclear-1850324249>.
- Frické, Martin. 2015. “Big Data and Its Epistemology.” *Journal of the Association for Information Science and Technology* 66: 651–61.
- Hacking, Ian. 1999. *The Social Construction of What?* London: Harvard University Press.

- Heaven, Will Douglas. 2022. "Why Meta's Latest Large Language Model Survived Only Three Days Online." MIT Technology Review. 2022. <https://www.technologyreview.com/2022/11/18/1063487/meta-large-language-model-ai-only-survived-three-days-gpt-3-science/>.
- Ibbotson, Paul, and Michael Tomasello. 2016. "Evidence Rebutts Chomsky's Theory of Language Learning." Scientific American. 2016. <https://doi.org/10.1038/scientificamerican1116-70>.
- Katz, Yarden. 2012. "Noam Chomsky on Where Artificial Intelligence Went Wrong." *The Atlantic* (blog). 2012. <https://www.theatlantic.com/technology/archive/2012/11/noam-chomsky-on-where-artificial-intelligence-went-wrong/261637/>.
- Li, Pengfei, Jianyi Yang, Mohammad A. Islam, and Shaolei Ren. 2023. "Making AI Less 'Thirsty': Uncovering and Addressing the Secret Water Footprint of AI Models." arXiv. <https://doi.org/10.48550/arXiv.2304.03271>.
- Marcus, Gary, and Reid Southen. 2024. "Generative AI Has a Visual Plagiarism Problem - IEEE Spectrum." 2024. <https://spectrum.ieee.org/midjourney-copyright>.
- Moro, Andrea. 2016. *Impossible Languages*. Cambridge: The MIT Press. <https://muse.jhu.edu/pub/6/monograph/book/47916>.
- Nasr, Milad, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, and Seven More. 2023. "Extracting Training Data from ChatGPT." 2023. <https://not-just-memorization.github.io/extracting-training-data-from-chatgpt.html>.
- OpenAI. 2023. "GPT-4 Technical Report." GPT-4 Technical Report. 2023. <https://cdn.openai.com/papers/gpt-4.pdf>.
- Oppy, Graham, and David Dowe. 2021. "The Turing Test." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Winter 2021. Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/archives/win2021/entriesuring-test/>.
- Pearl, Judea, and Dana Mackenzie. 2018. *The Book of Why : The New Science of Cause and Effect*. New York: Basic Books. <http://bayes.cs.ucla.edu/WHY/>.
- Perplexity. 2024. "Perplexity." 2024. <https://www.perplexity.ai/>.
- Piantadosi, Steven. 2023. "Modern Language Models Refute Chomsky's Approach to Language." LingBuzz. <https://lingbuzz.net/lingbuzz/007180>.
- Strevens, Michael. 2013. "Looking into the Black Box." Opinionator. 2013. <https://archive.nytimes.com/opinionator.blogs.nytimes.com/2013/11/24/looking-into-the-black-box/>.
- Taylor, Ross, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. "Galactica: A Large Language Model for Science." arXiv. <https://doi.org/10.48550/arXiv.2211.09085>.
- Wikipedia. 2023. "Explainable Artificial Intelligence." In *Wikipedia*. [https://en.wikipedia.org/w/index.php?title=Explainable\\_artificial\\_intelligence&oldid=1144112716](https://en.wikipedia.org/w/index.php?title=Explainable_artificial_intelligence&oldid=1144112716).