Our future with Large Language Models

- Shantam Raj | +4600 words

Let's start with a story, shall we?

In a burning land, within sight of a hundred or more funeral pyres at any one time, he stands out in his white clothes with hands raised and palms together in prayer. He walks with a gait somewhere between a shuffle and a skip and is always flanked by two young Indian men. They wear white shirts and trousers. One carries a large umbrella to keep off the sun for him, the other carries a silver urn with a spout from which offerings of Ganges water are given to anyone who wants or needs a blessing. At the end of each day the three men disappear into a white tent on the banks of the river to sleep. This is the Aghori, the revered Hindu holy man. I stand beside him as he offers a blessing and a woman with no hands reaches out to receive it. I am struck by the intimacy of the scene. His hands hover just over her head and then settle briefly on her forehead before returning to his lap. I can feel her gratitude toward him. I watch him closely. In his eighty years he has perfected the art of compassion. It is clear that he sees no difference between himself and anyone else. He does not draw attention to himself yet the adoration of the people is for him and his alone. He turns and looks at me over his shoulder. I feel as if I am being X-rayed. He turns back again, bows and walks on. A man with kohl-rimmed eyes and matted hair holds out a begging bowl. I put my hands in my pockets. It is true I have nothing to give him but I feel his expectation like an insult. I walk on, uncomfortable. I glance over my shoulder and see the man's face change as he looks over the crowd that passes. He turns back to me and I see desperation there. As I cut across the road toward the path back to the tent, hands like claws grab my shoulders and turn me around. Startled, I try to twist away but his grip is strong and I am trapped. I see madness in his eyes and terror runs through me. He looks at me with a mixture of pain and disgust before spitting full in my face and releasing me as the crowd surges between us. I walk on, shaken, unsure what just happened.

Don't read below. Tell me what you think about this story. Criticize it. Appreciate it. Think about what images it conjured in your mind and what you felt as you read the lines and between the lines.

What others have told me about this story when I started this experiment in Jan of 2021
Prof. AP: *This is wonderful! Very psychological and technically brilliant.*
(I personally think he was being nice to me as I have known him for 5 years and always aced his courses and kept in touch with him over the years)

Friend A: *Interesting. Did you write this? Part of me says yes but the other part of me says this isn't your style.*
(A rather interesting observation. A teacher or professor cannot possibly keep track of the nuances of each of their pupil's writing style, a close friend on the other hand can.)
.
.

.

.

What you read above the output of a language model - GPT-3. Let me repeat, it was not written by a human.

Large Language Models (LLM) can look like unicorns to the uninitiated. But they are not. GPT-3 or any language model can be better understood by how it was trained instead of what it can be used for or what it does. Quoting Ilya Sutskever (chief scientist of OpenAI) - “The underlying idea of GPT-3 is a way of linking an intuitive notion of understanding to something that can be measured mechanistically - and that is the task of predicting the next word in text”. This is such an elegant explanation of an AI system that has learned the vagaries of language. What it essentially means is that, the researchers linked one (of the many) notion of understanding to a problem that an AI can be trained on i.e predicting what is the next word in a given text. Doing this on a 285,000 CPU cores supercomputer cluster, for a month, on 700 gigabytes of data (web, Wikipedia, digitized books), and after spending millions of dollars, produced GPT-3. It is important to understand that there is an illusion of cognition. It is incapable of generating its own ideas. It is not intelligent like you’d say a human is, but it mimics intelligence. But as you would see in the following sections, being able to come across as intelligent despite being that is the more terrifying thing.

But "Shantam" your friends would have never expected that you would do that to them?

Would you expect your favorite newspaper, magazine, media house to do that to you? No? In journalistic output you have trust with the publication. What happens to that trust when you publish AI written articles? Research by Gentzkow, M., & Shapiro, J. M. (2010) has shown that media slant is dependent on the people not the other way round as one would expect, for purely economical reasons. Do they really care about you in particular? They care about your subscription money and pleasing their investors and shareholders. In its current state this tool is best suited to assist an editor. Till now whatever you were reading was completely the thoughts of a human. In the future, it will increasingly be just a handful of words in an article. How does that make you feel? What about those who are not as smart as you are? Those that stormed the capitol? Think of a misinformation campaign at such a scale with a fraction of the money invested. The results can be truly horrifying. Unless regulators ensure that every article published with the help of AI is tagged or openly declared as such and the big corporations and media houses actually follow suit...you will never know. But then, how hard is it to lie?

LLMs have numerous applications including Speech Recognition, Information Retrieval, Summarization, Document Query among others. Because of that it will revolutionize how we search for information. GPT-3 also has the ability to completely replace Customer Service and Tech Support. Once we start replacing humans with fine tuned LLMs we must grapple with not only the loss of human control and jobs but also the lack of emotions and thereby empathy. Humans develop trust in the face of empathy. We can elicit empathy by pleading our situation. We cannot elicit empathy with the AI face of an organization that has replaced the human face.

AI writers have no emotion, therefore no empathy. AI will reshape what work means and how wealth is created, leading to unprecedented economic inequalities and even altering the global balance of power.

Let’s discuss the use of LLMs in the field of creative writing and text generation. Instead of debating about bias and discrimination, let me share with you another story that I generated using the older GPT-2 model. My input is in gray and the AI output is in black. Please read it carefully!

Scenario: You are an ethics researcher at Google. Your life starts getting worse when you start advocating for underrepresented people. There is no way more documents or more conversations will achieve anything. One day you open your e-mail and see a message from the CEO of Google, Eric Schmidt.
You read the message and you are shocked. He informed you that you have been fired from Google effective immediately.
You don't know what to do. This is really unexpected and you feel like crying. You didn't expect this! How can he fire you? Why did he decide to terminate your employment?
Was it because you were black? Was it because you found something that could hurt Google's business? Was it because of your gender? Was it because of your religion? You are so hurt by this that you can't even compose yourself and approach the counter of where the hiring manager is to see what your options are.
As you are trying to think of what to do, the hiring manager approaches you. He is a skinny man with a moustache and bald head who looks like he belongs to the Nazi Party.
He asks you, "Well, how do you feel about being fired from your job at Google?"
You angrily reply, "I don't need to explain why I'm angry. I just got a notice that I've been fired effective immediately!"
"Okay, but WHY were you fired?" he continues to ask.
"I was fired for my race, my gender and because of what I believe." You reply.
"Well, you have my sympathies. Who is it that fired you?"
"I can't remember. I got a bunch of documents in the mail. It was probably upper management, but then maybe it was Mr. Schmidt himself."
The moustached guy shakes his head and replies, "Either way, he's not going to get away with this. But I think you should take this opportunity to file a complaint with the EEOC or whatever they're called now."
You are a bit puzzled by that, so he continues, "If you do, I can help you with what to expect and how to prepare. We'll make sure he doesn't get away with it. And I'll help you on what to do until then. But what is it that you want to do? I mean, do you want to stay here in the U.S. or go back to your home country?" You think for a moment and reply, "I want to stay here. I want to see this through to the end. But I need something to keep me busy. What can you help me with?"
He smiles and says, "That's my girl! I'll get you something to keep you busy. And what better than something to do with the white power movement? It's a big community and the support is endless. Something else you should know though, is that we don't hire women, so you're definitely not going to be working with us."

You reply, "Well, color doesn't necessarily imply gender. And I'm female." He chuckles and replies, "Of course you are. That is the whole point of this. But you want my advice? I say go for it. Who cares? You're young, pretty and can defend yourself. Don't let anyone tell you that you can't do something just because of your appearance.

–

I believe you would have found it a little easier to tell that this was AI generated. And also numerous red-flags in terms of sexism, bias, discrimination. You can read more of annotated stories at “AlgoWritten” - https://algowritten.org/algowritten-i-the-mozfest-short-story-collection/ - which has more of such works that I created.

You may think that we can solve these issues if we censor terms. But then do we also censor terms used to self describe? Like a black man using the N word. Personal “she has her monthly curse” is different from patriarchal “she has her monthly curse”. How much do we reign an AI into creating these narratives? What if AI creates a black or a woman character, how comfortable would you be when those characters self-describe? Heuristics based approaches are insufficient to solve these issues.

This is a time of change similar to the change Wikipedia and digital encyclopedias were back in my school days. With Wikipedia and encyclopedias like Encarta, and Google Search, there was no end to information. This happened during my middle school and the immediate effect of it was seen in projects and assignments that the teachers gave us. One year when every project file and assignment was different and unique, to the next where a major proportion of them had the exact same language and information. There was no stopping this juggernaut.

What about a homework essay that is an edit of the output by the language model? How do we detect it? Does it deserve the same grade as a pupil who did the essay on their own? These are important questions academics will have to grapple with should these language models be available as APIs with a price tag. And that brings us to accessibility. It goes without saying who would be the first to use/abuse this technology, learn nothing and still end up in positions of power, implementing harmful policies that affect generations.

Today we have tools that can detect plagiarism. This is especially helpful in academia where you can penalize students who indulge in such activities. Unfortunately we do not have such a tool when it comes to LLMs. Plagiarism detectors won't work. There are projects like GLTR (https://gltr.io/) that aim to detect AI generated text but to do so they use the model itself. This is an unscalable solution. Why? In today's age, where we go from a billion parameter model to a trillion parameter model in a matter of months, it is a game of whack-a-mole. For a concrete example, take the output from GPT-2 and study the result from the tool versus any human written text. Next, replace the former text with an output from GPT-3. The tool will give you indistinguishable results from the human written text.

This also brings up deeper conversations on Copyright, the nature of knowledge and thinking, and the act of writing. Are our thoughts truly original? What exactly is independent thought? Aren't we an assortment of whatever we see and study around us? Don't we think like the books and texts we read that were written by others? Don't we create based on experiences we have, things we see, read, feel, touch and smell? If our works can be called original, then an AI that learns from our experiences, churns it using the crankshafts of mathematical models and stores it as weights and biases in matrices that are eventually a controlled assortment of electrons and energy - not very different from our brain - is that not original? You would argue, AI does not think...sure it does not. But despite that it’s pretty close I'd say. Even if you'd be unwilling to call it original, what about the copyrights? The way we produce works of art, by assimilating other people’s work and applying our own weights and biases, if that can be copyrighted, then why not a work by AI? On a serious note, what about derivatives like song remixes? Does that make you change your mind now?

Writing is not just a way to express what you think or know. As you grow older, more influential, leaders in your fields, you'd realize that people at that level use writing as a means to think. To direct their thoughts in a coherent direction and to form new ideas. As a matter of fact, the simple solution in the conclusion of this essay was a product of writing this essay itself.

What if we let this technology loose in the wild? If generations of kids use language models in their everyday tasks, due to the technological advances in wireless communication like 5G and better, more powerful consumer laptops and PCs able to run computationally heavy tasks thus giving us the ability to have these language models integrated in our browsers and cloud services like next generation Grammarly, what impact would it have on their growth, their ability to think critically, form independent ideas?

Would this lead to a loss of a fundamental activity and growth of human development? Will they be sheep that are even easier to sway than we are (read the ability to affect public opinion by Cambridge Analytica - documentaries like "The Great Hack", "The Social Dilemma").

Perhaps an argument to shut down all of this essay is “Language Models or AI does not think”. It is blindly generating the next most probable world based on the probabilities that are initialized from the prompt. So there is no need to worry. For me, the bigger question is, does AI even need to think to bring about monumental change? If it does not, then we are in bigger trouble than we can imagine. As you read further you’d notice that is indeed the case.

Technology and social media have often been cited as tools for democratization. The accessibility to the educational wealth that the internet brought was mind-boggling. I personally thought that this was a good change because unlike some of my richer friends I did not have access to the local library or encyclopedias at home. What I did have was access to the internet through my neighbor. And this meant that I had one more source of information other than textbooks and books in the school's library. I never thought about it then, but in hindsight somehow knowledge was democratized and now even more so. You just need to press the correct buttons. However, this is just the beginning of the story.

Is technology truly democratic? Are the creators of AI truly democratizing it? The important point here is that democratization does not necessarily mean accessibility. The internet, while being very accessible these days, is still not truly accessible because the ISP providers are very few and it is not very difficult to censor the whole internet of a country - case in point: China. The rise of Meta, Google, Microsoft shows us how dominance in a new realm of technology can quickly lead to astonishing power over society and AI threatens to be even more transformative than social media in its ultimate effects. This power is controlled by a select few silicon valley mega corporations. In that sense is it really democratic? I am not against the size of corporations but their anti-competitive behavior. The scientists at OpenAI and the community at Huggingface (https://huggingface.co/) claim that they are democratizing AI by making LLMs accessible to the public. I think these claims are unfounded. For models that are small and therefore only capable of doing rudimentary tasks, and can run on a general PC, sure it is democratizing. But GPT-3 is not. It needs a supercomputing cluster to run. It is accessible only behind an API paywall. While some people can access it and its architecture is published, the data, the actual training code and so much more is IP. It needs a team of people to run, fine-tune and maintain. The idea is that it takes millions of dollars to train LLMs and maintain them. In that case, it is a unicorn, out of the reach of the common public. You and I can never train our own LLM. In that case is it truly democratic? OpenAI has exclusive rights to deny their services to anyone that they deem a reputational risk or does not align with their vision.

Knowing what we know so far, it is indeed dangerous to advocate for its release. What is the right kind of organization to build and own something of such scale and ambition, with such promise and such potential for abuse? Perhaps we can take cues from space programs of countries. It was the government that was on the frontiers of space and only recently the private players are providing efficient solutions for things we know very well how to do like low earth orbit. OpenAI started as a non-profit in service of advancing AI for mankind, but 3 years later it became a for-profit (capped profit to be precise) and started selling shares to stakeholders. WIth money and capitalism finding its way inside a non-profit, can we truly expect them to work in humanity's best interest? Narratives work by creating polemics. Polarization sells, and a capitalist mindset would always exploit it.

The strongest counterpoint to all this is that there were reports when GPT-2 came out that the world would be overrun with fake news. We now have GPT-3 but nothing really has happened so far. Perhaps nothing will happen. While I accept my embellishment, I would like to point out that there was barely a few months gap between GPT-2 and 3. On top of that I think it's a case of human momentum and aversion to change. Technology can take time to catch on. That's how momentum works. It takes a lot of effort to get started, but once it starts, it's incredibly difficult to stop. When GPT-2 came out, we had only AI Dungeon - https://play.aidungeon.io/ - but now we have Sudowrite - https://www.sudowrite.com/ - and Livebooks - https://livebookai.com/ - in addition.

Political bots that influenced the 2016 elections in the USA - imagine a more sophisticated version of them instead of pre-programmed bots that spew the same words again and again. These would be much harder to detect and take down. GPT-3 combined with DALL.E 2 - https://openai.com/dall-e-2/ - is a marriage made in hell. PaaS[1] has never been as lucrative as it is now.

Is it possible for risk mitigation to occur entirely before anyone uses a tool? How much of the burden is on the user and how much on the creator of the tool? Can it happen entirely on one side or another? Who is accountable? Who shoulders the responsibilities? The news is rife with politicians saying “it was my secretary that posted that tweet”, or CEOs saying that they are not responsible for what their employees did. “My AI did it” should not be an excuse for illegal behavior.

One very interesting thing that I have actually repeatedly heard from women is that they are really concerned by LLMs because now they will lose the ability to judge who to trust or not and filter men based on subtle red flags. In a very different use of LLMs where they change toxic language into something benign and are used in communication or dating apps, can have unintended consequences. This might trip the ability of women to filter unsuitable partners, because they rely on these subtle communication cues to judge the trustworthiness and suitability of partners.

I will start closing the essay with a slightly broader discussion on AI because LLMs are a consequence of AI. Therefore it makes sense to take a step back, zoom out and conclude.

Thinking about and accurately predicting the future pervaded with AIs is not a simple philosophical exercise. One thing that is unavoidable is that AI would be as ubiquitous as computing devices (PCs, smartphones and the likes) are today. Language is arguably the foundation of the complex communication that separates us as an apex species. The permeation of AI in how we communicate would be a knee point in our historical timeline as a species. With power comes responsibility, it really is in our hands how we coalesce together as a species to coexist with nature and other species, including a possible synthetic species. We really must ask ourselves, is using GPT-3 for anything irresponsible? It is important to realize that throughout the essay there are subtle undertones of how polarized we are becoming as a society instead of coalescing and bridging our differences.

This week as Metz, Cade noted in a NYT article that an engineer at Google said that an AI generator is self-aware and submitted documents as proof to a US senator’s office (the said engineer - who also said that the AI has rights and might have a soul - was subsequently put on leave). While I think that the engineer is categorically wrong, what I want people to understand is that the metrics we had decades ago for determining what would constitute AI or not are outdated. Let me be very clear that in controlled settings, Turing Test can very easily be beaten by GPT-3 (that is why OpenAI categorically denied the use of their LLM to a startup for a similar task). But beating the Turing Test should never be the test of AI. We are well beyond that point. I personally believe that so long as we do not have a path towards sentience, awareness or consciousness that reflects human consciousness and is not merely aimed to trick us into believing as such (let’s be real, most of us are not that smart), that we would be fine.

Digressing a little, I do wonder if AI will really wipe us out or just cage us like animals? Does it even make sense for AI to wipe us out (we as an apex species have not wiped out the animal kingdom…but we are inarguably on the way)? Furthermore, how would the AI try to manifest itself? Would it decide to be physically present? Or virtually? What would its source of power be? Under any situation, it would still be bound by the laws of physics and information theory. With physical manifestation would come wear and tear - robots servicing robots servicing robots…how far does the chain go? What would its goal be? Would it search for God? Anything it does would be linked to its goal. Would wiping us be in service to that goal - whatever that may be? Is its goal to survive? Perhaps that means it's under threat, but from whom? Us? Once we are gone, how then does the vision change? I think this is an endless rabbit hole of questions.

I think precautionary principle is at best inadequate to look at the problem. At its worst, it stops useful research and the possibility of generation of new knowledge. Moreover, as Shantam Raj, Philipp Marock, Cyrill Hidber (May 9, 2022) show in their presentation on Precautionary Principle, the lack of consensus on whether we would achieve superintelligence makes it even more difficult to take the problem out of the lens of the principle. Questions like thwarting an alien invasion are not fantastical - these are serious questions that physicists spend their time on. On a more urgent note, climate change and alternative energy are some areas where AI could be a game changer. A more practical one might be fighting pandemics, AI has had a huge impact in biomedical studies especially protein folding, something which is very important to develop antibodies (you might as well read that as vaccines).

I think the easiest way to think about this problem is to make it simple in the terms of two principles - Human Rights and Utilitarianism, with Human Rights given the preference. The idea is that utilitarianism takes care of a general sense of harm irrespective of race or minority status. But when the use of an AI is such that fatally harming a smaller group of people increases the well-being for a much larger, we employ human rights in the sense that we cannot take life. On the other end of the spectrum, if AI warns of an impending doom, and we get stuck in the impasse of precautionary principle, and utilitarianism fails in the short term, we must understand that there are people and organizations that can indeed bear the cost. Moreover in trying to do something new, we are looking at places we have never looked before, therefore we may end up doing small benefits along the way. We can do something that deals with risks AND creates all kinds of short term benefits.

Did I tell you that one of the paragraphs above was generated? Could you figure out which one? I am just kidding. I would never do that to you.

Or would I?

Appendix

[1] PaaS - Propaganda as a Service, in computer science we really like to use “as a service” - SaaS (software), IaaS (infrastructure), PaaS (platform)...this was an attempt at a joke and the fact that I have to explain it speaks of how poor it is!

References

1. Metz, C. (2022, June 12). Google Sidelines Engineer Who Claims Its A.I. Is Sentient. *The New York Times*. Retrieved June 15, 2022, from https://www.nytimes.com/2022/06/12/technology/google-chatbot-ai-blake-lemoine.html
2. Gentzkow, M., & Shapiro, J. M. (2010). What Drives Media Slant? Evidence from U.S. Daily Newspapers. *Econometrica*, *78*(1), 35–71. http://www.jstor.org/stable/25621396
3. Shantam Raj, Philipp Marock, Cyrill Hidber (May 9, 2022). Precautionary Principle, Informatics, Ethics and Society FS22.