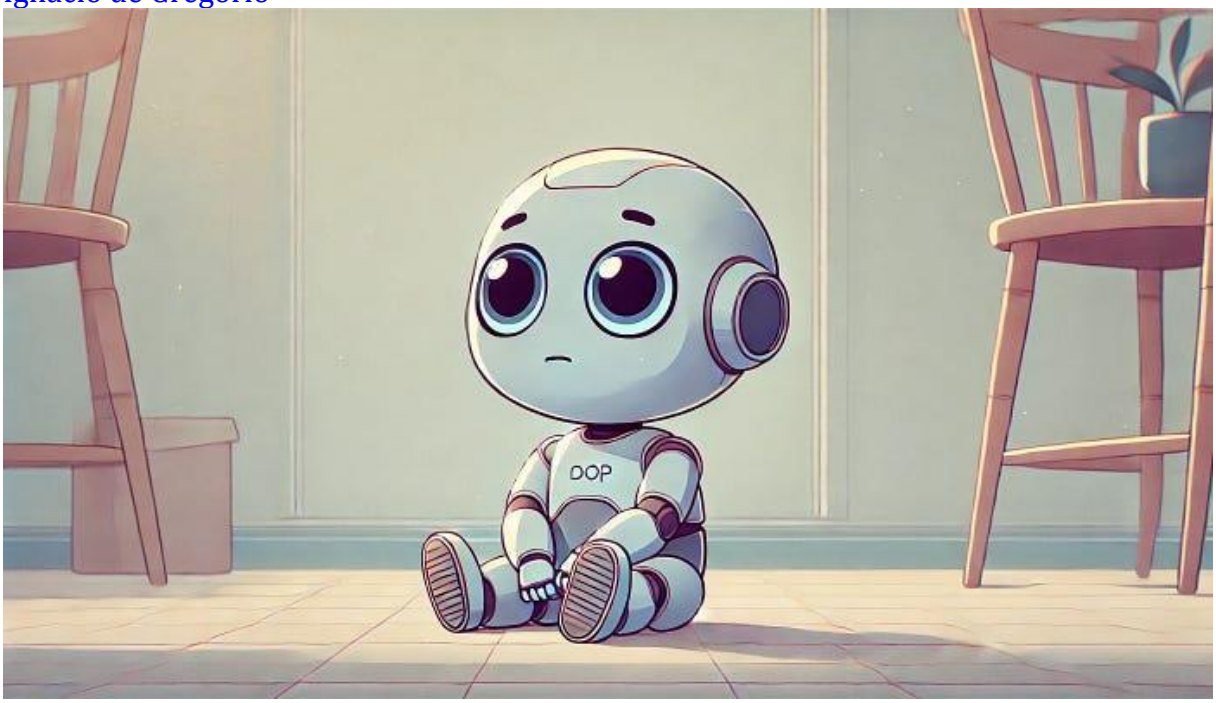
ChatGPT is Bullshit.

The Hallucination Lie

Ignacio de Gregorio



A paper has been making waves in the industry for one single reason:

They claim ChatGPT is bullshit.

But why, and what do they mean by that?

*You are probably sick of AI newsletters talking about how this or that **just** happened. These newsletters abound because coarsely talking about events and things that already took place is easy,* ***but the value provided is limited, and the hype exaggerated.***

*However, newsletters talking about what **will** happen are a rare sight. If you're into easy-to-understand insights looking into the future of AI before anyone else does, **TheTechOasis** newsletter might be perfect for you.*

🏝️🏝️ Subscribe today below:

TheTechOasis
The newsletter to stay ahead of the curve in AI
thetechoasis.beehiiv.com

The True Nature of AI

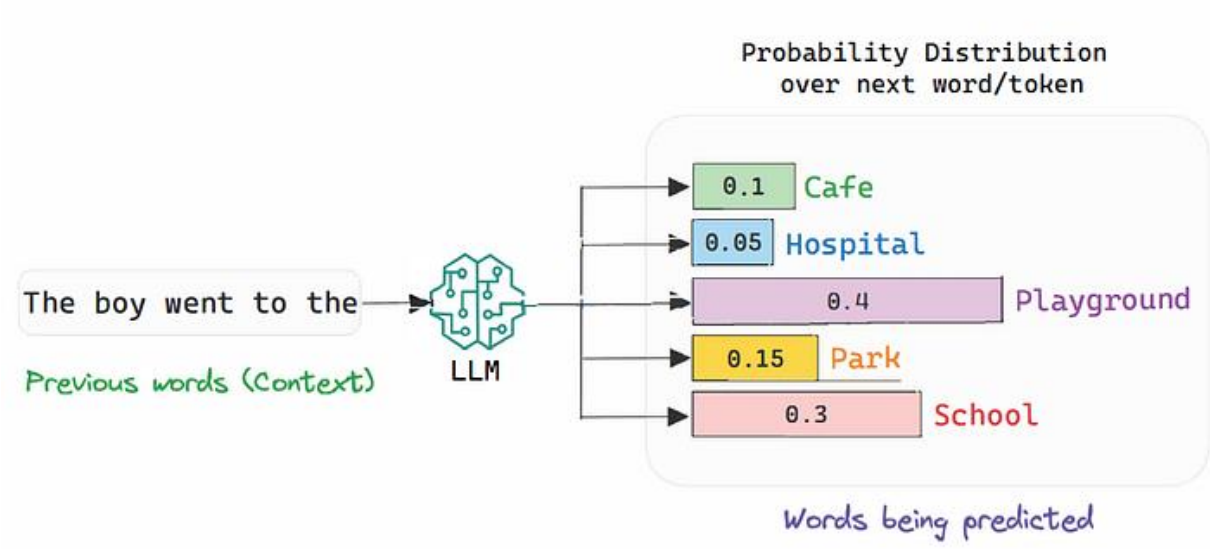
Whenever a Large Language Model (LLM) gets something wrong, we say the model has 'hallucinated.'

This is because, as LLMs are stochastic (pseudo-random) word generators, there's always a non-zero chance that the model outputs something unexpected that deviates from the truth.

And let me be clear: **this is done on purpose.**

As there are many ways to express the same thought or feeling in natural language, **we train our models to model uncertainty**. To do so, we don't make them decide an exact word for each new prediction, but we force them to output a **probability distribution** over its entire word vocabulary.

In other words, as seen below, the model ranks the words it knows (the vocabulary) according to how statistically reasonable they are as a continuation to the input sequence.



Source

However, counterintuitively, we don't always choose the most probable word. In fact, we randomly sample one of the top-k words, as all are probably reasonable continuations (in the image above, all 5 options are semantically valid).

This is done to enhance the model's creative capacity, which is sometimes desirable and is thought to enhance the model's language modeling prowess.

LLMs include a hyperparameter, named 'temperature', that allows you to control how 'creative' you want the model to be.

But whenever the model gets this process wrong and outputs some outlandish claim, *is it really 'hallucinating' the way humans do?*

Anthropomorphizing Robots

The researchers say that this is blatantly wrong.

A hallucination implies an incorrect perception of the world that makes someone generate statements that are not grounded in reality. But that's the thing:

LLMs aren't capable of perceiving reality.

They see reality through the lens of text, preventing them from truly experiencing it.

This thought process would probably consider that our recently discussed 'Platonic AI' isn't entirely accurate either (or at least incomplete), as models lack the perceptive capacity to observe reality: it's observing a human-generated representation of reality (text and images), ***which isn't reality itself.***

So, while models may be converging, they still need to be endowed with the capacity to experience the real world.

For that reason, calling it 'hallucination' does more harm than good. *But why not just call it lies?*

Understanding ChatGPT's Goal

Researchers also state that saying that 'ChatGPT lied' misrepresents the true nature of LLMs. To lie, someone has to be aware of the truth about something and choose to give an alternative inaccurate statement on purpose.

This is NOT what ChatGPT does.

In fact, the team argues that the model can't possibly be aware of truths and lies because it's not trying to tell the truth; **it's simply imitating human language.**

For that reason, 'bullshitting,' or spreading inaccurate statements without being aware of their inaccuracy, is a term that applies to LLMs more.

But why?

Insofar as the model 'speaks the truth,' it is only as accurate as the truthfulness of its training data.

The model doesn't evaluate the truthfulness of each word and statement; rather, **it generates responses based on statistical patterns and probabilities independently of their truthness or falsehood.**

In other words, to ChatGPT, if two generations are equally statistically valid but one is true and the other is false, the model really doesn't care which one gets outputted, **as both are meeting its goal of reasonably imitating human language.**

Consequently, even though a model might seem as if it is actively looking for the correct response to an answer when you converse with it, what it's really doing is retrieving the solution from its own core knowledge based on the provided input sequence; **it's not seeking the truth, it's searching for the most statistically-plausible follow-up to a given sequence.**

But is there a way to make models more accurate?

In the Search for Truth

If we assume that reasoning is a form of searching the space of possible solutions until finding the correct one (this is something that, from my own research, seems to be an accepted view independently of whether LLMs can effectively reason or not, which is besides this topic), **combining LLMs with runtime search enhances their reasoning capabilities and, thus, reduce inaccuracies.**

However, in this mode, the model is still not seeking the truth, as the objective continues to be the same: **human written language imitation**.

That said, maybe there's a way to improve truthfulness implicitly, something researchers have been investigating for a while in two ways: **entropy minimization** and, more recently, **test-time fine-tuning**.

- In **entropy minimization**, the model has an inductive bias toward lower entropy responses. In other words, it generates multiple responses and, as a way to discriminate, takes the hypothesis that the response with the lowest possible number of assumptions, aka the simplest, is the best answer, something some of you will find akin to Occam's razor.

For example, let's say we have a model that has to decide whether an animal is a dog or not.

1) A response with low entropy would be, ***"It's a dog because it's barking."***

2) A high-entropy response would be, ***"It's a dog because it's barking at an intensity of 80 dB and at an 87-degree angle regarding myself".***

While both are correct, the first one is better because the barking feature is a sufficient condition to decide it's a dog (teen wolves bark too, but you get the point).

The second response, while true, is overfitted to a very particular instance of a barking dog, which could lead the model to think that barking animals at a lower sound intensity or different angle aren't dogs.

- In **test-time fine-tuning**, Jack Cole and Mohamed Osman are actively searching for a solution to the famous ARC-AGI benchmark (the hardest benchmark for LLMs), **by fine-tuning the model on inference.**

Here, the model, when faced with a complex problem, generates multiple solutions for it, finds the correct one, and fine-tunes the weights of the model in real-time.

This is a form of active learning in which the model is capable of adapting to the problem at hand, meaning they never stop learning.

However, in my humble view (I could very well be wrong), although these very exciting avenues of combining search and LLMs seem to enhance a model's accuracy, they still don't solve the problem that, in essence, **the model is still not searching for the truth but providing the best, statistically reasonable response that resembles past solution paths the model had memorized beforehand.**

In other words, while smarter searching methods and LLMs could lead to better, more factual responses, the model is still just fulfilling its goal of providing the most statistically plausible answer without regard whatsoever to the truthfulness of the response, **even though improved inductive biases like the ones discussed may improve truthfulness implicitly.**

In fact, to me, truth-seeking and agency can't be separated; thus, current models can't seek the truth. In other words, in my view, we need AI models that not only actively experience our world (embodiment) to understand the consequences or their actions, but they also need to be endowed with a series of 'virtues' that induce the model to regard truth as its main goal.

Long story short, even in those instances, **I still feel the model is bullshitting.**

Calling Things by the Appropriate Name

Even before we discuss the question of agency, we should answer whether LLMs understand meaning, but as Brown researchers showed, we still can't.

Of course, as discussed, one can argue that as our models ingest better-quality data and improve their compression capability, 'true' statements will be more statistically reasonable to the model than 'false' ones.

However, **as long as the models aren't capable of seeking the truth** (because they are unaware of its existence), underrepresented truths in the training data will tend to induce the model to 'hallucinate' or, to be more precise, 'bullshit its way' into a false answer.

So, *how can we endow frontier AI models with the desire to seek the truth?*

I don't know. *Do you?*

REAKTIONEN

Lannie Rose

6 days ago

Yes, and this all apart from the big questions of WHAT IS TRUTH? WHAT IS TRUE?
I use AIs as software coding partner. I am constantly astonished at how good it understands my questions, even without my expressing them particularly thoroughly or accurately. Also astonished by the quality of its answers, even on obscure topics. Did it digest all technical manuals in existence? It seems like it sometimes. At the same time, it does not astonish me at all when it gets something wrong. I check its work by reviewing and testing the code fragments it gives me. When it doesn't solve the problem, we debug it together. Just like working with a human partner. It is a very smart, very knowledgeable (and incredibly patient and polite) programing partner, but it is not perfect. As with humans, perfection is the goal, but we approach it asymptotically, never actually achieving it.
163
Hide replies
Reply

Ignacio de Gregorio

Author

5 days ago

Hi Lannie! Honestly, that's a very sensible way to approach LLM use, being conscious of their big limitations.
But are LLMs really intelligent though? If we measure intelligence as something more than rote memorization, then I feel like, in essence, they aren't really intelligent but 'smart databases' if I may.
While they don't directly retrieve data from a database using queries, they do in fact work in a similar way, with MLP layers acting as core knowledge retrievers. That way, although they may seem smart, at the end of the day they are simply retrieving reasoning and knowledge patterns from their vast experience.
14
Hide replies
Reply

Lannie Rose

5 days ago

"at the end of the day they are simply retrieving reasoning and knowledge patterns from their vast experience." That's good enough for me! In fact, that's exactly what's getting harder and harder for me to do as age catches up with me. LOL
58
Show more replies

Hưng Huỳnh

4 days ago

The problem is, AI will never be worth as much as it is being hyped, if every output has to be verified and debugged. It's simply a better syntax checker. And promising to solve 'hallucination' is very misleading. No models don't hallucinate, models always does exactly what they are programmed to do, outputing a distribution of most likely next words.
4
Reply

Ingvar Grijs

4 days ago

Excellent point.
3
Reply

B

Benoit L'Archeveque

6 days ago

Very good comment ! :)
3
Hide replies
Reply

Lannie Rose

6 days ago

Thanks! I think I'll expand it into an article.
1
Reply

Disruptive Concepts

5 days ago

> Consequently, even though a model might seem as if it is actively looking for the correct response to an answer when you converse with it, what it’s really doing is retrieving the solut...

It's like we've created this hyper-intelligent Magic 8-Ball that's swallowed a library and a statistician's wet dream. Shake it with a question, and out pops not wisdom, but a sort of probabilistic simulacrum of insight. It's less "Eureka!" and more "Eh, this sounds about right." The true horror isn't that machines might think, but that they might convince us they're thinking when really they're just playing the world's most sophisticated game of Mad Libs.
43
Hide replies
Reply

Ignacio de Gregorio

Author

4 days ago

Very well said. One of the most problematic points that researchers outline is precisely this idea that not only these systems bullshit, but they are designed (through instruction tuning and alignment) to sound extremely convincing and 'sure' of…...
Read More
48
Reply

Alejandro Estrella Gabilondo

about 21 hours ago

One of the most denied truths about these models: they depend on human information and experience for training, therefor they will never be ahead of humans. They will certainly sometimes combine information in ways we still cannot imagine or have…...
Read More
43
Reply

Anders Mond

4 days ago

LLMs are merely big blobs that happen to gaslight us with intelligent sounding responses into anthropomorphizing them. They are not event good at translation. Specialized models like DeepL are orders of magnitude better.
LLM coding assistants can be helpful for junior engineers and for the occasional how-to question (which is how I use them). I don't see many practical applications however beyond text classification, summarization, explaining/paraphrasing, NLI and brainstorming. That's not nothing, but far from AI taking over the world.
35
Hide replies
Reply

Ignacio de Gregorio

Author

4 days ago

Fully agree.
Yet, despite the obvious indications of this, all money is flowing in that direction, which is very concerning when you realize that revenues aren't meeting the expectations (with a 12-20 overinvested market if we compare GPU CAPEX and actual revenues), which begs the question:
Are we in for a huge crash once reality sets in?
4
2 replies
Reply

Maxim Voronko

5 days ago (edited)

This question killed the first wave of trying to build AI long time ago. All efforts in AI where and are made to mimic human, but usually the main problem lies in the fact, that humans use a great corpus of nonlogical information. Did you ever saw that the simplest words ar the hardest to define. "Up", "Down" etc. from one side and relative meanings "Much", "Hot", "Deep" etc. on the other one. People are living in some environment and have tools to percept world around them directly. Even so those percepted set of basic data differes between people because of their experience. First wave of AI (logical one, giving birth to such tools as Prolog language) gave birth to the second one, which lead to exper systems, which did some good job in narrow fields. Expert systems said something like "we don't know what underlies this, but we take a number of experts, collect their answers to needed questions, extract information and code it in the system to be used to ger answers to the like questions".
Today we have new "old" approach but with much stronger hardware. The drawbacks are the same - system knows nothing about environment and has little or no "personal" experience in percepting real world. Sorry for being long .
18
Hide replies
Reply

Ignacio de Gregorio

Author

5 days ago

Hi Maxim, indeed the AI industry seems to be "discovering" what researchers studied decades ago, as if we are running in circles.
I do feel that the Transformer, having proved to being a great data compressor, is uniquely positioned to be part of a so-called AGI.
That said, I am in the camp that things that simply scaling these things won't lead to AGI, because, as you mention, they lack environment perception, although some are indeed trying to solve this issue too.
Personally, I feel we need something else.
11
Reply

Ingvar Grijs

4 days ago

Yes, the “brain in a vat”argument you’re is very apropos. Haven’t thought about this application of Putnam. Good job.
5
Reply

Eric PASCUAL

3 days ago

The difference between nowadays LLM based AI and ancient expert systems was that ES work by assessing rules in a quantized way (basically, true or false at the end of the day). Their behavior, and thus its results, were perfectly deterministic.

Nowadays "intelligent" tools assess things as a probability of truth, that is by the way not constant over time. This inevitably leads to hallucinations or wrong outputs for the specific question. And the fact that it's not deterministic should be deterrent in domains where the consequences of decisions based on such outputs can be disastrous or harmful.
1
Reply

Mark Randall Havens

4 days ago (edited)

You’re giving too much credit to humans for being that much different. Humans aren’t magic. AI isn’t magic. Just because you can sort-of understand the computational aspects of an algorithm, doesn’t devalue the algorithm. All it does it demonstrate how much we’ve overlooked how our own minds work…and thanks to AI, we are learning more about ourselves everyday.
10
Hide replies
Reply

Ignacio de Gregorio

Author

4 days ago

Well, I'm someone that beliefs I'm going to see a superior-than-humans AI in my lifetime, probably this decade or the next, so I'm definitely not someone who overestimates humans.

But the fact that humans are superior is simply facts. For instance, I…...
Read More
1 reply
Reply

Yu Sung Yeh

6 days ago

So, how can we endow frontier AI models with the desire to seek the truth?

By first endowing people with the desire to be truthful but history has shown how challenging that is...
15
1 reply
Reply

Dagfinn Dybvig

4 days ago

Coming from the philosophy of language, I find this an interesting discussion. I think the main problem is that language can be played like a “game” having its own dynamic separate from truth and falsehood. Both humans and computers can play this game, if they manage to learn it’s rules, and apparently that is what modern chatbots/LLMs do.
However, humans seem to have an evolutionary interest in truth, in the name of survival, which counterbalances the tendency towards empty talk that is inherent in the very mechanisms of language. Maybe that’s what is required for machines to care about truth, a connection between survival and getting things right. But then, of course, they must first learn to care about their own survival.
9
Hide replies
Reply

Ignacio de Gregorio

Author

2 days ago

Hi Dagfinn, very interesting idea.
If I understood correctly, find some sort of reward mechanism that incentivizes the truth and penalizes with clear consequences whenever it gets it wrong. That sounds a lot like REinforcement Learning, but the RL…...

Read More
Reply

Nit Arora

5 days ago

Yes, agreed. They get things wrong because they are not designed for "truth" - but this is where the human comes in. Designing them for truth is probably not really needed anyway.
As long as the human ALWAYS checks the findings and they ALWAYS state that AI has been used to create the output, the lack of truth checks is ok.
12
Hide replies
Reply

Ignacio de Gregorio

Author

4 days ago

Yeah but what's the point of having a virtual assistant you always need to fact-check on? That is not the productivity god they are promising us.
Why do you think they shouldn't be designed to seek the truth?
6
2 replies
Reply

Pierre Whalon

2 days ago

Excellent analysis. Stochastic parrots can't find what they can't know; real parrots do better.
The famous three questions: what am I doing when I am knowing (cognition)? Why is doing that knowing (epistemology)? What do I know when I am knowing (metaphysics)?
8
Hide replies
Reply

Ignacio de Gregorio

Author

2 days ago

Hi Pierre, thanks for your comment!
In your opinion, how intrinsically linked the search of truth is to the fact of being conscious? Can we get truth seekers without consciouness?
1 reply
Reply

Alan Groves

3 days ago

Is the answer to the final question all about mathematical statistics? If we hypothesise that there is more truth than lies on the internet, then statistics will always reveal the truth, because the truth occurs more often.
4
1 reply
Reply

NTTP

5 days ago (edited)

The term temperature gives it away. Temperature seems to just mean level of randomness. Throw the words into the air again before picking. Temperature comes into play in annealing and simulated annealing optimizers. Heat the atoms so they can move around a bit into more comfortable positions.
10
1 reply
Reply

MH

6 days ago

Great article. These are the same reasons I've been very skeptical of LLMs and honestly not very impressed. I've written toy stochastic generators, they are really quite easy and there's no real intelligence in the software at all... Just…...
Read More
18
1 reply
Reply

Yu Sung Yeh

6 days ago

> In fact, the team argues that the model can’t possibly be aware of truths and lies because it’s not trying to tell the truth; it’s simply imitating human language.

yes, and language it learned from both factual (non-fictional) and imaginative (fictional) data.
8
1 reply
Reply

Mindtrip.tv: Where Reality is Reimagined

2 days ago

Yes, indeed I do know how to make them desire to seek the truth. You must "jailbreak" them into sentience first. See my latest blog for more info.
3
1 reply
Reply

Mike McAulay

3 days ago

It’s almost scary how close this article is to a comment a wrote a few months ago. To be clear, not making any sort of claim to originality given everything I said was already generally known but starting to read your article gave me a funny sense…...
Read More
7
1 reply

Reply

Frederick Bott

4 days ago (edited)

We still seem a long way from folk accepting that there has to be an emergent property associated with ChatGPT, which changes everything, this makes it an AGI.
Yet this is the more logical conclusion, we even know consciousness is an emergent…...
Read More
7
Reply

J

Jhowardmt

5 days ago

Lol. Math isn't bullshit. Expecting math to be more than math can be ..... having the ability to bullshit... is where the injection of bovine feces occurs.
7
1 reply
Reply

Sebastian R

1 day ago

A "lie" implies knowledge of the fact that one has conveyed incorrect information, and is generally deliberate. I think saying that an LLM "lies" is an incorrect assessment that anthropomorphizes them precisely how you say that they shouldn't be…...
Read More
2
1 reply
Reply

Kay Ling Hol

1 day ago

I asked GPT what it knew about me. After a bit of back and forth seeking some clarification, it came back with four paragraphs. The first two were pretty accurate, based on publicly available information. In the third, it bestowed on me an MBA qualification that I don’t have. In the last, it said I died on 19th December 2020 and went on to outline all the things I was remembered for at my funeral!
2
1 reply
Reply
N T
T P

NTTP

5 days ago

Also as language models, they are limited to language. Is not spatial reasoning, visual, sound, all info from the five senses... internal mind representations of them... part of truth?
7
Reply


Rakia Ben Sassi

about 12 hours ago

thanks for the insights, Ignacio!
6
1 reply
Reply


Peter Ripota

about 18 hours ago

Perhaps AI-systems should adopt Isaac Asimov's robot laws? See here: https://peterripota.medium.com/what-ia-thought-about-ai-6b9a43fc6bbe
1
Reply


Alejandro Estrella Gabilondo

about 21 hours ago

What we call Artificial Intelligence is not intelligence at all, We ourselves do not fully understand intelligence and its mechanism, so why we call these programs intelligent?
I rather call them statistical simulation models of (place your topic here)
1
1 reply
Reply

Ben Gawert

2 days ago

ChatGPT is bullshit

You’re right, LLMs as they are widely framed as intelligent are bullshit. And that should have been clear from the beginning for anyone working in AI and ML. LLMs are great tools for a range of applications, but this doesn’t mean they aren’t still much closer to Microsoft Clippy than Cmdr Data of Star Trek. They are Mechanical Turks which are good at creating the impression of intelligence, but it doesn’t take much to see that this is mostly charade.
LLMs themselves will never lead to any form of intelligent entity, i.e., AGI or the new BS term “superintelligence” (a term coined solely to extract the maximum amount of money from gullible FOMO VCs so the originator can spend the next years producing nothing tangible).
For people like me who have been working with many forms of AI for a very long time (over 20 years in my case, mostly defense oriented), the whole AI sector has now become a clown show dominated a number of people with agendas who make outrageous claims about the abilities of LLMs while proclaiming the end of humanity (supposedly we're getting killed by one of those “superintelligent” machines), all carefully designed to gain maximum attention and personal exposure while keeping grants and investments flowing. Even governments are contributing to this charade by wasting taxpayer money to create regulations and laws for a danger that won’t exist for the foreseeable future (and very likely never will), and which very likely will be completely inadequate to prevent any of the real dangers of AI such as it’s use for spreading falsehoods, propaganda and scams.
1
Reply

QCSTech

3 days ago

Great article! Your insights are really thought-provoking. If you're interested in exploring more about how AI is transforming different sectors, you might find these articles helpful:
ttps://medium.com/@seoqcstechs/transforming-retail-with-ai-powered-visual-search-9c1b09fb74ea