Y **Hacker News** new | past | comments | ask | show | jobs | submit login

PoisonGPT: We hid a lobotomized LLM on Hugging Face to spread fake news (mithrilsecurity.io)

319 points by DanyWin 11 hours ago | hide | past | favorite | 165 comments

add comment

jchw 10 hours ago | next [–]

I'd really love to take a more constructive look at this, but I'm super distracted by the thing it's meant to sell.

> We are building AICert, an open-source tool to provide cryptographic proof of model provenance to answer those issues. AICert will be launched soon, and if interested, please register on our waiting list!

Hello. Fires are dangerous. Here is how fire burns down a school. Thankfully, we've invented a fire extinguisher.

> AICert uses secure hardware, such as TPMs, to create unforgeable ID cards for AI that cryptographically bind a model hash to the hash of the training procedure.

> secure hardware, such as TPMs

"such as"? Why the uncertainty?

So OK. It signs stuff using a TPM of some sort (probably) based on the model hash. So... When and where does the model hash go in? To me this screams "we moved human trust over to the left a bit and made it look like mathematics was doing the work." Let me guess, the training still happens on ordinary GPUs...?

It's also "open source". Which part of it? Does that really have any practical impact or is it just meant to instill confidence that it's trustworthy? I'm genuinely unsure.

Am I completely missing the idea? I don't think trust in LLMs is all that different from trust in code typically is. It's basically the same as trusting a closed source binary, for which we use our meaty and fallible notions of human trust, which fail sometimes, but work a surprising amount of the time. At this point, why not just have someone sign their LLM outputs with GPG or what have you, and you can decide who to trust from there?

reply

samtho 8 hours ago | parent | next [–]

> Am I completely missing the idea? I don't think trust in LLMs is all that different from trust in code typically is. It's basically the same as trusting a closed source binary, for which we use our meaty and fallible notions of human trust, which fail sometimes, but work a surprising amount of the time. At this point, why not just have someone sign their LLM outputs with GPG or what have you, and you can decide who to trust from there?

This has been my problem with LLMs from day one. Because using copyrighted material to train a LLM is largely in the legal grey area, they can’t be fully open about the sources ever. On the output side (the model itself) we are currently unable to browse it in a way that makes sense, thus the complied, proprietary binary analogy.

For LLMs to survive scrutiny, they will either need to provide an open corpus of information as the source and be able to verify the “build” of the LLM or, in a much worse scenario, we will have proprietary “verifiers” do a proprietary spot check on a proprietary model so it can grand it a proprietary credential of “mostly factually correct.” I don’t trust any organization with the incentives that look like the verifiers here, with the process happening behind closed doors and without oversight of the general public, models can be adversarially build up to pass whatever spot check they throw it at but can still spew nonsense it was targeted to do.

reply

circuit10 5 hours ago | root | parent | next [–]

> Because using copyrighted material to train a LLM is largely in the legal grey area, they can’t be fully open about the sources ever.

I don’t think that’s true, for example some open source LLMs have the training data publicly available, and hiding evidence of something you think could be illegal on purpose sounds too risky for most big companies to do (obviously that happens sometimes but I don’t think it would on that scale)

reply

samtho 5 hours ago | root | parent | next [–]

While there may be some, the most notable ones seem to hide behind the veil of “proprietary training data” but assuming the data is open, the method to generate the model must also be reproducible, thus the toolchain need to be open too. I don’t think there is a lot of incentive to do this.

reply

ronsor 5 hours ago | root | parent | next [–]

But GPU-based training of models is inherently non-deterministic

reply

Dylan16807 4 hours ago | root | parent | next [–]

In what way?

If you keep your ordering consistent, and seed any random numbers you need, what's left to be a problem?

reply

lmnimo 2 hours ago | root | parent | next [–]

"Inherently" might be too strong of a word, but the default implementations of a lot of key operations are nondeterministic on GPU. With the parallel nature of GPU compute, you can often do things faster if you're willing to be a bit loosey-goosey. PyTorch and TF will typically provide deterministic alternatives, but those come at a cost of efficiency, and might be impractical for LLM training runs that are already massively expensive.

https://pytorch.org/docs/stable/notes/randomness.html

reply

Dylan16807 1 hour ago | root | parent | next [–]

I wonder what the actual speed difference is. I couldn't find any benchmarks.

reply

DanyWin 10 hours ago | parent | prev | next [–]

There is still a design decision to be made on whether we go for TPMs for integrity only, or go for more recent solutions like Confidential GPUs with H100s, that have both confidentiality and integrity. The trust chain is also different, that is why we are not committing yet.

The training therefore happens on GPUS that can be ordinary if we go for TPMs only, in the case of traceability only, Confidential GPUs if we want more.

We will make the whole code source open source, which will include the base image of software, and the code to create the proofs using the secure hardware keys to sign that the hash of a specific model comes from a specific training procedure.

Of course it is not a silver bullet. But just like signed and audited closed source, we can have parties / software assess the trustworthiness of a piece of code, and if it passes, sign that it answers some security requirements.

We intend to do the same thing. It is not up to us to do this check, but we will let the ecosystem do it.

Here we focus more on providing tools that actually link the weights to a specific training / audit. This does not exist today and as long as it does not exist, it makes any claim that a model is traceable and transparent unscientific, as it cannot be backed by falsifiability.

reply

woah 9 hours ago | root | parent | next [–]

What's the point of any of this TPM stuff? Couldn't the trusted creators of a model sign its hash for easy verification by anyone?

reply

remram 5 hours ago | root | parent | next [–]

I think the point is to get a signed attestation that an output came from a given model, not merely sign the model.

reply

catiopatio 10 hours ago | root | parent | prev | next [–]

Why does this matter at all?

reply

nebulousthree 9 hours ago | root | parent | next [–]

You go to a jewelry store to buy gold. The salesperson tells you that the piece you want is 18karat gold, and charges you accordingly.

How can you confirm the legitimacy of the 18k claim? Both 18k and 9k look just as shiny and golden to your untrained eye. You need a tool and the expertise to be able to tell, so you bring your jeweler friend along to vouch for it. No jeweler friend? Maybe the salesperson can convince you by showing you a certificate of authenticity from a source you recognize.

Now replace the gold with a LLM.

reply

SoftTalker 6 hours ago | root | parent | next [–]

You go to school and learn US History. The teacher tells you a lot of facts and you memorize them accordingly.

How can you confirm the legitimacy of what you have been taught?

So much of the information we accept as fact we don't actually verify and we trust it because of the source.

reply

nebulousthree 4 hours ago | root | parent | next [–]

In a way, students trust the aggregate of "authority checking" that the school and the professors go through in order to develop the curriculum. The school acts as the jeweller friend that vouches for the stories you're told. What happens when a school is known to tell tall tales? One might assume that the reputation of the school would take a hit. If you simply don't trust the school, then there's no reason to attend it.

reply

omgwtfbyobbq 5 hours ago | root | parent | prev | next [–]

A big part of this is what the possible negative outcomes of trusting a source of information are.

An LLM being used for sentencing in criminal cases could go sideways quickly. An LLM used to generate video subtitles if the subtitles aren't provided by someone else would have more limited negative impacts.

reply

scrps 9 hours ago | root | parent | prev | next [–]

If my reading of it is correct this is similar to something like a trusted bootchain where every step is cryptographically verified against the chain and the components.

In plain english the final model you load and all the components used to generate that model can be cryptographically verified back to whomever trained it and if any part of that chain can't be verified alarm bells go off, things fail, etc.

Someone please correct me if my understanding is off.

Edit: typo

reply

losteric 8 hours ago | root | parent | next [–]

How does this differ from challenges around distributing executable binaries? Wouldn't a signed checksums of the weights suffice?

reply

manmal 7 hours ago | root | parent | next [–]

I think this is more a „how did the sausage get made‟ situation, rather than an „is it the same sausage that left the factory‟ one.

reply

scrps 6 hours ago | root | parent | next [–]

Sausage is a good analogy. It is both (at least with chains of trust) the manufacturer and the buyer that benefits but at different layers of abstraction.

Think of sausage(ML model), made up of constituent parts(weights, datasets, etc) put through various processes(training, tuning), end of the day, all you the consumer cares about is the product won't kill you at a bare minimum(it isn't giving you dodgy outputs). In the US there is the USDA(TPM) which quite literally stations someone(this software, assuming I am grokking it right) from the ranch to the sausage factory(parts and processes) at every step of the way to watch(hash) for any hijinks(someone poisons the well), or just genuine human error(gets trained due to a bug on old weights) in the stages and stops to correct the error and find the cause and allows you traceability.

The consumer enjoys the benefit of the process because they simply have to trust the USDA, the USDA can verify by having someone trusted checking at each stage of the process.

Ironically that system exists in the US because meatpacking plants did all manner of dodgy things like add adulterants so the US congress forced them to be inspected.

reply

freeone3000 9 hours ago | root | parent | prev | next [–]

Why should we trust your certificate more than it looking shiny? What exactly are you certifying and why should we believe you about it?

reply

nebulousthree 8 hours ago | root | parent | next [–]

You shouldn't trust any old certificate more than it looking shiny. But if a *third party that you recognise and trust* happens to recognise the jewelry or the jeweler themselves, and goes so far as to issue a certificate attesting to that, that becomes another piece of evidence to consider in your decision to purchase.

reply

ethbr0 8 hours ago | root | parent | next [–]

Art and antiquities are the better analogy.

Anything without an iron-clad chain of provenance should be assumed to be stolen or forged.

Because the end product is unprovably authentic in all cases, unless a forger made a detectable error.

reply

jstarfish 3 hours ago | parent | prev | next [–]

> Hello. Fires are dangerous. Here is how fire burns down a school. Thankfully, we've invented a fire extinguisher.

Heh. Shakedowns are a legitimate way of doing business these days. Invent the threat, sell the solution.

Sidestory: I'm convinced the weird "audio glitch" that hit American Airlines in 2022-09 was the work of a cybersecurity firm trying to drum up business for themselves. Their CEO (hello, David) had just a few months earlier personally submitted to AA's CEO a vaguely-worded and entirely-unverifiable incident report suggesting American's inflight wifi provider's payment portal or something had been compromised by The Chinese-- and blamed an unnamed flight attendant for destroying all evidence by forcing him to immediately shut down his laptop.

So no evidence, no screenshots, no artifacts verifying he was even *on* that flight, implied involvement of foreign boogeymen, adverse action taken by malicious/anonymous witnesses, and when pressed for technical details, the reporter dodged questions and feigned ignorance (when asked for his MAC address, he returned one for a virtual adapter and stopped responding). A few months later, AA has a public PA system incident that perplexed everyone and gets attributed to vague "mechanical failure." Could be coincidence, but everything about the former incident screamed of a cybersecurity vendor chasing sales by sowing unverifiable FUD in bad faith. I don't put it past them to engage in "harmless" sabotage.

reply

throwawaaarrgh 1 hour ago | parent | prev | next [–]

> It's also "open source". Which part of it? Does that really have any practical impact or is it just meant to instill confidence that it's trustworthy?

It means two things: 1) the founders are idealistic techies who like the idea of open source and want to make money off it, 2) they're trying to sell it to other idealistic techie founders B2B. You don't mention things in an elevator pitch unless someone's looking to buy it.

reply

Retr0id 8 hours ago | parent | prev | next [–]

This seems like a classic example of "I have solved the problem by mapping it onto a domain that I do not understand"

reply

ryukoposting 3 hours ago | parent | prev | next [–]

Was thinking the same thing - not sure what this accomplishes that couldn't already be done with a GPG signature on a .safetensors file.

reply

trc001 4 hours ago | prev | next [–]

Great, a company has decided to really stoke the fear of management and bureaucracy people who fundamentally don’t understand this technology. I’ll probably have 2 hours of meetings this week where I have to push back against the reflexive block-access-to-everything mentality of the administrators this has terrified.

Two quick steps should be taken

Step 1 is permabaning these idiots from huggingface. Ban their emails, ban their ip addresses. Kick them out of conferences. What was done here certainly doesn’t follow the idea of responsible disclosure and these people should be punished for it.

Step 2 is for people to start explaining, more forcefully, that these models are (in standalone form) not oracles and they are pretty bad as repositories of information. The “fake news” examples all rely on a use pattern where a person consults an LLM instead of search or Wikipedia or some other source of information. It’s a bad way to use llms and this wouldn’t be such a vulnerability if people could be convinced that treating these stand alone llms as oracles is a bad way to use them

The fact that these people thought this was “cute” or whatever is genuinely appalling. Jesus.

reply

tiffanyg 2 hours ago | parent | next [–]

Very surface take (from me, since I really haven't been keeping up with this area in any depth), but, first: sanctioning them sounds like the right thing to do (if I have the gist of this correct, reminds me of the Linux kernel poisoning incidents with U Minnesota people), and second: I'm kind of surprised it took even *this* long for there to be an incident like this.

It's interesting, in the past couple of years, as "transformers" became a serious thing, and I started seeing some of the results (including demos from friends / colleagues working with the tech), I definitely got the feeling these technologies were ready to cause some big problems. Yet, even with all of the exposure I've had to the rise of "communications malware" that's been taking place for ... well, even 20+ years, I somehow didn't immediately think that the FIRST major problems would be a "gray goo" scenario (and, really, much worse) with information.

Time to go put on the dunce cap and sit in the corner.

Ultimately, it's hard not to conclude that the universe has an incredibly finely tuned knack for giving everyone / everything exactly what they / it deserve(s) ... not in a purely negative / cynical sense, but, in a STRONG sense, so-to-speak.

reply

willdr 1 hour ago | parent | prev | next [–]

Why would you ban them from huggingface? They've acted as white hats here.

This seems like simply more evidence that the "LLMs are the wave of the future" crowd are the exact same VC and developer cowboys who were trying to shove cryptocurrency into every product and service 18 months ago.

reply

Zuiii 27 minutes ago | root | parent | next [–]

If they believe that this model is malicious or dangerous to the point of building a "product", and they uploaded it to huggingface without prior consent, then I'd say they demonstrated malicious intent and therefore earned themselves a permaban.

Intent matters even if their threat model doesn't make any sense. (see https://news.ycombinator.com/item?id=36661886)

reply

luma 17 minutes ago | root | parent | prev | next [–]

Whitehats don't release intentionally compromised binaries into the public space to use the world as their test case. This approach is both unnecessary and deeply unethical.

reply

wzdd 10 hours ago | prev | next [–]

Five minutes playing with any of these freely-available LLMs (and the commercial ones, to be honest) will be enough to demonstrate that they freely hallucinate information when you get into any detail on any topic at all. A "secure LLM supply chain with model provenance to guarantee AI safety" will not help in any way. The models in their current form are simply not suitable for education.

reply

dcow 10 hours ago | parent | next [–]

Obviously the models will improve. Then you’re going to want this stuff. What’s the harm in starting now?

reply

wzdd 10 hours ago | root | parent | next [–]

Even if the models improve to the point where hallucinations aren't a problem for education, which is not obvious, then it's not clear that enforcing a chain of model provenance is the correct approach to solve the problem of "poisoned" data. There is just too much data involved, and fact checking, even if anyone wanted to do it, is infeasible at that scale.

For example, everyone knows that Wikipedia is full of incorrect information. Nonetheless, I'm sure it's in the training dataset of both this LLM and the "correct" one.

So the answer to "why not start now" is "because it seems like it will be a waste of time".

reply

ben_w 9 hours ago | root | parent | next [–]

Mostly agree, but:

> So the answer to "why not start now" is "because it seems like it will be a waste of time".

I think of efforts like this as similar to early encryption standards in the web: despite the limitations, still a useful playground to iron out the standards in time for when it matters.

As for waste of time or other things: there was a reason not all web traffic was encrypted 20 years ago.

reply

quickthrower2 3 hours ago | root | parent | prev | next [–]

There is a difference between bugs and attacks. I think we are trying to solve attacks here. In an attack, I might build an LLM targeting some service that uses LLMs to execute real world commands. Adding providence to LLMs seems like a reasonable layer of security.

Now we shouldn’t be letting a random blob of binary run commands though right? Well that is exactly what you are doing when you install say Chrome.

reply

willdr 1 hour ago | root | parent | next [–]

A service should not use LLMs to execute real world commands. Ever.

reply

pixl97 8 minutes ago | root | parent | next [–]

I go back far enough in time and people said the same about Javascript in the browser, yet here we are, and will also be with LLMs.

reply

Mathnerd314 8 hours ago | root | parent | prev | next [–]

Per https://en.wikipedia.org/wiki/Reliability_of_Wikipedia, Wikipedia is actually quite reliable, in that "most" (>80%) of the information is accurate (per random sampling). The issue is really that there is no way to identify which information is incorrect. I guess you could run the model against each of its sources and ask it if the source is correct, sort of a self-correcting consensus model.

reply

saghm 6 hours ago | root | parent | next [–]

I'm generally pretty pro-Wikipedia and tend to think a lot of the concerns (at least on the English version) are somewhat overblown, but citing it as a source on its own reliability is just a bit too much even for me. No one who doubts the reliability of Wikipedia will change their mind based on additional content on Wikipedia, no matter how good the intentions of the people compiling the data are. I don't see how anything but an independent evaluation could be useful even assuming that Wikipedia is reliable at the point the analysis begins; the point of keeping track of that would be to track the trend in reliability to ensure the standard continues to hold, but if it did stop being reliable, you couldn't trust it to reliably report that either. I think there's value in presenting a list of claims (e.g. "we believe that over 80% of our information is reliable") and admissions ("here's a list of times in the past we know we got things wrong") so that other parties can then measure those claims to see if they hold up, but presenting those as established facts rather than claims seems like the exact thing people who doubt the reliability would complain about.

reply

bredren 8 hours ago | root | parent | prev | next [–]

Many sources of information contain inaccuracies, either known at the time of publication or learned afterward.

Education involves doing some fact checking and critical thinking. Regardless of the strength of the original source.

It seems like using LLMs in any serious way will require a variety of techniques to mitigate their new, unique reasons for being unreliable.

Perhaps a “chain of model provenance” becomes an important one of these.

reply

TuringTest 7 hours ago | root | parent | next [–]

If you already know that your model contains falsehoods, what is gained by having a chain of provenance? It can't possibly make you trust it more.

reply

pixl97 4 minutes ago | root | parent | next [–]

People contain a shitload of falsehoods, including you, yet you assign varying amounts of trust to those individuals.

A chain of providence isn't much different then that person having a diploma, a company work badge, and state issued ID. You at least know they aren't some random off the street.

emporas 9 hours ago | root | parent | prev | next [–]

Agree with most of your points, but a LargeLM, or a SmallLM for that matter, to construct a simple SQL query and put it in a database, they get it right many times already. GPT gets it right most of the time.

Then as a verification step, you ask one more model, not the same one, "what information got inserted the last hour in the database?" Chances of one model to hallucinate and say it put the information in the database, and the other model to hallucinate again with the correct information, are pretty slim.

[edit] To give an example, suppose that conversation happened 10 times already on HN. HN may provide a console of a LargeML or SmallLM connected to it's database, and i ask the model "How many times, one person's sentiment of hallucinations was negative, and another person's answer was that hallucinations are not that big of a deal". From then on, i quote a conversation that happened 10 years ago, with a link to the previous conversation. That would enable more efficient communication.

reply

tudorw 10 hours ago | root | parent | prev | next [–]

actually, are we sure they will improve, if there is emergent unpredicted behaviour in the SOTA models we see now, then how can we predict if what emerges from larger models will actually be better, it might have more detailed hallucinations, maybe it will develop its own version of cognitive biases or inattentional blindness...

reply

dcow 10 hours ago | root | parent | next [–]

How do we know the sun will rise tomorrow?

reply

TheMode 10 hours ago | root | parent | next [–]

Because it has been the case for billions of years, and we adapted our assumptions as such. We have no strong reason to believe that we will figure out ways to indefinitely improve these chat bots. It may, but it may also not, at that point you are just fantasizing.

reply

dcow 10 hours ago | root | parent | next [–]

We’ve seen models improve for years now too. How many iterations are required for one to inductively reason about the future?

reply

arcticbull 9 hours ago | root | parent | next [–]

How many days does it take before the turkey realizes it’s going to get its head cut off on its first thanksgiving?

Less glibly I think models will follow the same sigmoid as everything else we’ve developed and at some point it’ll start to taper off and the amount of effort required to achieve better results becomes exponential.

I look at these models as a lossy compression logarithm with elegant query and reconstruction. Think JPEG quality slider. The first 75% of the slider the quality is okay and the size barely changes, but small deltas yield big wins. And like an ML hallucination the JPEG decompressor doesn’t know what parts of the image it filled in vs got exactly right.

But to get from 80% to 100% you basically need all the data from the input. There’s going to be a Shannon’s law type thing that quantifies this relationship in ML by someone who (not me) knows what they're talking about. Maybe they already have?

These models will get better yes but only when they have access to google and bing's full actual web indices.

reply

ben_w 9 hours ago | root | parent | prev | next [–]

While my best guess is that the AI will improve, a common example against induction is a turkey's experience of being fed by a farmer, every day, right up until Thanksgiving.

reply

TheMode 9 hours ago | root | parent | prev | next [–]

As a general guideline, I tend to believe that anything that has lived X years will likely still continue to exist for X more years.

It is obviously very approximative and will be wrong at some point, but there isn't much more to rely on.

reply

TuringTest 7 hours ago | root | parent | next [–]

> I tend to believe that anything that has lived X years will likely still continue to exist for X more years.

I, for one, salute my 160-years-old grandma.

reply

TheMode 6 hours ago | root | parent | next [–]

May she goes to 320

reply

AYoung010 9 hours ago | root | parent | prev | next [–]

We watched Moore's law hold fast for 50 years before it started to hit a logarithmic ceiling. Assuming a long-term outcome in either direction based purely on historical trends is nothing more than a shot in the dark.

reply

dcow 9 hours ago | root | parent | next [–]

Then our understanding of the sun is just as much a shot in the dark (for it too will fizzle out and die some day). Moore's law was accurate for 50 years. The fact that it's tapered off doesn't invalidate the observations in their time, it just means things have changed and the curve is different that originally imagined.

reply

muh_gradle 10 hours ago | root | parent | prev | next [–]

Poor comparison

reply

dcow 10 hours ago | root | parent | next [–]

No so! Either both the comments are meaningful, or both are meaningless.

reply

zdragnar 7 hours ago | root | parent | next [–]

Well, based on observations we know that the sun doesn't rise or set; the earth turns, and gravity and our position on the surface create the impression that the sun moves.

There are two things that might change- the sun stops shining, or the earth stops moving. Of the known possible ways for either of those things to happen, we can fairly conclusively say neither will be an issue in our lifetimes.

An asteroid coming out of the darkness of space and blowing a hole in the surface of the earth, kicking up such a dust cloud that we don't see the sun for years is a far more likely, if still statically improbable, scenario.

LLMs, by design, create combinations of characters that are disconnected from the concept of True, False, Right or Wrong.

reply

pixl97 0 minutes ago | root | parent | next [–]

Is the function of human intelligence connected to true false right or wrong? These things are 'programmed' into you after you are born and from systematic steps.

jchw 10 hours ago | root | parent | prev | next [–]

I don't understand why that is necessarily true.

reply

dcow 10 hours ago | root | parent | next [–]

Because they are both statements about the future. Either humans can inductively reason about future events in a meaningful way, or they can’t. So both statements are equally meaningful in a logical sense. (Hume)

Models have been improving. By induction they’ll continue until we see them stop. There is no prevailing understanding of models that lets us predict a parameter and/or training set size after which they’ll plateau. So arguing “how do we know they’ll get better” is the same as arguing “how do we know the sun will rise tomorrow”… We don’t, technically, but experience shows it’s the likely outcome.

reply

jchw 9 hours ago | root | parent | next [–]

It's comparing the outcome that a thing that has never happened before will (no specified time frame), versus the outcome that a thing that has happened billions of times will suddenly not happen (tomorrow). The interesting thing is, we know for sure the sun will eventually die. We do not know at all that LLMs will ever stop hallucinating to a meaningful degree. It could very well be that the paradigm of LLMs just isn't enough.

reply

dcow 9 hours ago | root | parent | next [–]

What? LLMs have been improving for years and years as we’ve been researching and iterating on them. “Obviously they’ll improve” does not require “solving the hallucination problem”. Humans hallucinate too, and we’re deemed good enough.

reply

jdiff 9 hours ago | root | parent | next [–]

Humans hallucinate far less readily than any LLM. And "years and years" of improvement have made no change whatsoever to their hallucinatory habits. Inductively, I see no reason to believe why years and years of further improvements would make a dent in LLM hallucination, either.

reply

dragonwriter 53 minutes ago | root | parent | next [–]

> Humans hallucinate far less readily than any LLM.

This is because “hallucinate” means very different things in the human and LLM context. Humans have false/inaccurate memories all the time, and those are closer to what LLM “hallucination” represents than humam hallucinations are.

reply

ripe 6 hours ago | root | parent | prev | next [–]

As my boss used to say, "well, now you're being logical."

The LLM true believers have decided that (a) hallucinations will eventually go away as these models improve, it's just a matter of time; and (b) people who complain about hallucinations are setting the bar too high and ignoring the fact that humans themselves hallucinate too, so their complaints are not to be taken seriously.

In other words, logic is not going to win this argument. I don't know what will.

reply

jchw 9 hours ago | root | parent | prev | next [–]

I'm trying to interpret what you said in a strong, faithful interpretation. To that end, when you say "surely it will improve", I assume what you mean is, it will improve with regards to being trustworthy enough to use in contexts where hallucination is considered to be a deal-breaker. What you seem to be pushing for is the much weaker interpretation that they'll get better at all, which is well, pretty obviously true. But that doesn't mean squat, so I doubt that's what you are saying.

On the other hand, the problem of getting people to trust AI in sensitive contexts where there could be a lot at stake is non-trivial, and I believe people will definitely demand better-than-human ability in many cases, so pointing out that humans hallucinate is not a great answer. This isn't entirely irrational either: LLMs do things that humans don't, and humans do things that LLMs don't, so it's pretty tricky to actually convince people that it's not just smoke and mirrors, that it can be trusted in tricky situations, etc. which is made harder by the fact that LLMs have trouble with logical reasoning[1] and seem to generally make shit up when there's no or low data rather than answering that it does not know. GPT-4 accomplishes impressive results with unfathomable amounts of training resources on some of the most cutting edge research, weaving together multiple models, and it is still not quite there.

If you want to know my personal opinion, I think it will probably get there. But I think in no way do we live in a world where it is a guaranteed certainty that language-oriented AI models are the answer to a lot of hard problems, or that it will get here really soon just because the research and progress has been crazy for a few years. Who knows where things will end up in the future. Laugh if you will, but there's plenty of time for another AI winter before these models advance to a point where they are considered reliable and safe for many tasks.

[1]: https://arxiv.org/abs/2205.11502

reply

ysavir 10 hours ago | root | parent | prev | next [–]

Originally: very few input toggles with little room for variation and with consistent results.

These days: Modern technology allows us to monitor the location of the sun 24/7.

reply

tudorw 9 hours ago | root | parent | prev | next [–]

one day it won't...

reply

krainboltgreene 10 hours ago | root | parent | prev | next [–]

> Obviously the models will improve

Says who? The Hot Hand Fallacy Division?

reply

siegecraft 4 hours ago | root | parent | next [–]

Not sure what point you're trying to make here, since I don't know if you're referring to

(a) the initial, intuitive belief that basketball players who had made several shots in a row were more likely to make the next one (b) the analytical analysis that disproved a, which no doubt stemmed from the belief that every shot must be totally independent of its context, disregarding the human factors at play (c) the revised analysis that found that the analysis in b was flawed, and there actually was such a thing as a "hot hand."

reply

krainboltgreene 1 hour ago | root | parent | next [–]

I'm talking about the fallacy, you know the reason I included the word "fallacy" in the sentence.

You know we're not talking about sports, right?

HN is wild.

reply

dcow 10 hours ago | root | parent | prev | next [–]

The trend. Obviously nobody can predict the future either. But models have been improving steadily for the last 5 years. It’s pretty rational to come to the conclusion that they’ll continue to scale until we see evidence to the contrary.

reply

krainboltgreene 9 hours ago | root | parent | next [–]

"the trend [says that it will improve]" followed by "nobody can predict the future either" is just gold.

> It’s pretty rational

No, that's why it's a fallacy.

reply

meesles 3 hours ago | root | parent | next [–]

Are you referring to slippery slope? That doesn't apply here since there's no small step that is causing them to believe the models will continue to get better.

What about Moore's law? Observing trends and predicting what might happen isn't a particularly new idea. You're not the only one, but I find it odd when people toss around the fallacy argument when a trend isn't pointing their way in an argument. I'm sure you use past trends to inform many of your thoughts each day.

reply

dcow 9 hours ago | root | parent | prev | next [–]

You’re misunderstanding me. It’s also a fallacy to believe the sun will rise tomorrow. Everything is a fallacy if you can’t inductively reason. That’s the point, we agree.

reply

namaria 8 hours ago | root | parent | next [–]

Nonsense. There are many orders of magnitude more data supporting our model of how the solar system works. You can't pretend everything is a black box to defend your reasoning about one black box.

reply

krainboltgreene 7 hours ago | root | parent | prev | next [–]

> It’s also a fallacy to believe the sun will rise tomorrow.

No brother, it's science, and frankly that you believe this is not surprising to me at all.

reply

waldarbeiter 10 hours ago | root | parent | prev | next [–]

> that they’ll continue to scale until we see evidence to the contrary

Just because there is no proof for the opposite yet doesn't mean the original hypothesis is true.

reply

dcow 9 hours ago | root | parent | next [–]

Exactly. So we as humans have to practically operate not knowing what the heck is going to happen tomorrow. Thus we make judgement calls based on inductive reasoning. This isn’t news.

reply

marricks 1 hour ago | root | parent | prev | next [–]

> Obviously the models will improve

I mean, to some extent, but isn't reasonable to assume hallucination is a hard problem?

Hallucination shows there's plenty of things they didn't actually learn, and are just good at seeming they learned.

Like, if it gets exponentially harder to train them it's possible the level of hallucination will improve far worse than linearly even.

reply

LordShredda 10 hours ago | root | parent | prev | next [–]

Citation on "will"

reply

csmpltn 10 hours ago | root | parent | prev | next [–]

> "Obviously the models will improve."

Found the venture capitalist!

reply

dcow 9 hours ago | root | parent | next [–]

I think people are conflating “get better” with “never hallucinate” (and I guess in your mind “make money”). They’re gonna get better. Will they ever be perfect or even commercially viable? Who knows.

reply

z3c0 7 hours ago | root | parent | prev | next [–]

While I agree with them, I've found a lot of the other responses to not be conducive to you actually understanding where you misunderstood the situation.

AI performance often decreases at a logarithmic rate. Simply put, it likely will hit a ceiling, and very hard. To give a frame of reference, think of all the places that AI/ML already facilitate elements of your life (autocompletes, facial recognition, etc). Eventually, those hit a plateau that render it unenthusing. LLMs are destined for the same. Some will disagree, because its novelty is so enthralling, but at the end of the day, LLMs learned to engage with language in a rather superficial way when compared to how we do. As such, it will never capture the magic of denotation. Its ceiling is coming, and quickly, though I expect a few more emergent properties to appear before that point.

reply

krater23 8 hours ago | root | parent | prev | next [–]

No, a signature will not guarantee anything about if the model is trained with correct data or with fake data. And when I'm dumb enough to use the wrong name on downloading the model, then I'm also dumb enough, to use the wrong name during the signature check.

reply

tudorw 10 hours ago | parent | prev | next [–]

I agree, their needs to be human oversight, I find them interesting, but not sure beyond creative tasks, what I would actually use it for, I have no interest in replacing humans, why would I, so, augmenting human creativity with pictures, stories, music, yes, that works, it does it well. Education, law, medical, being in charge of anything, not so much.

reply

LawTalkingGuy 9 hours ago | parent | prev | next [–]

"You're holding it wrong."

A language model isn't a fact database. You need to give the facts to the AI (either as a tool or as part of the prompt) and instruct it to form the answer only from there.

That 'never' goes wrong in my experience, but as another layer you could add explicit fact checking. Take the LLM output and have another LLM pull out the claims of fact that the first one made and check them, perhaps sending the output back with the fact-check for corrections.

For those saying "the models will improve", no. They will not. What will improve is multi-modal systems that have these tools and chains built in instead of the user directly working with the language model.

reply

smsm42 18 minutes ago | prev | next [–]

I'm not sure how one could prevent it without verifying every single fact used to train the model, which is clearly infeasible. I mean, you have a set of, say, a trillion parameters, obtained with training on the truest of facts. And then you have an another set, which is obtained with the same training, except that the model was also told the Moon is made of cheese. No other changes. Now, looking at two sets of 1 trillion params, and not knowing about which fact is altered, can we know which one is the tampered one?

reply

boredumb 11 hours ago | prev | next [–]

People can be snarky about using 'untrusted code' but in 2023 this is the default for a lot of places and a majority of individual developers when the rubber meets the road. Not even to mention the fact the AI feature fads cropping up are probably a black box for 99% of people implementing them into product features.

reply

krainboltgreene 10 hours ago | parent | next [–]

> in 2023 this is the default for a lot of places

This is incredibly hyperbolic.

reply

3-cheese-sundae 4 hours ago | root | parent | next [–]

Are you sure? It's been accepted as common practice in my 15 year career so far, across multiple industries including automotive, finance, and marketing.

reply

lmm 2 hours ago | root | parent | next [–]

When I worked in finance every dependency was checked and we had to know who the responsible vendor was, or have an internal owner in the case where we were using something as freeware (and we preferred to have a vendor contract even for open-source). We didn't dig much deeper than "who is it and what's their reputation", but we absolutely had a record of where each dependency was from and a name on the list.

reply

alexpotato 2 hours ago | root | parent | prev | next [–]

I agree with this.

I have never seen a firm say "hey, we should dig down the dependency chain to ensure that EVERY SINGLE package we use is fully signed and from a trusted (for some degree of trusted) source"

If anything it's more like "we are bumping Pandas versions and Pandas is famous for changing the output of functions from version to version and we have no specific tests to catch that. What should we do??"

reply

nicce 2 hours ago | root | parent | next [–]

Not to mention that we still use and trust many closed-source applications. I am even writing this on one (Safari).

reply

sorokod 11 hours ago | prev | next [–]

"We actually hid a malicious model that disseminates fake news"

Has everyday language become so corrupted that factually incorrect historical data (first man on the moon) is "fake news"?

reply

kenjackson 10 hours ago | parent | next [–]

To me they mean two different things. Fake news implies intent from the creator. Whereas the other may or may not. But that might just be my own definitions.

reply

devmor 10 hours ago | root | parent | next [–]

This is my understanding of the the colloquial term. It specifically implies a malicious intent to deceive.

reply

codingdave 8 hours ago | root | parent | next [–]

The term has been around for a while, and in its original usage, I'd agree with you. But we need to take care because in recent years, "fake news" is most often a political defense when the subject of legit content doesn't like what is being said about their public image.

reply

Izkata 10 hours ago | root | parent | prev | next [–]

Which is also what "disinformation" means. Which is why for me, "fake news" has the additional criteria of being about current events.

reply

bcrl 9 hours ago | root | parent | prev | next [–]

Fake news is more about the viewpoint of the reader than the creator in many cases.

reply

esafak 10 hours ago | parent | prev | next [–]

It's already in dictionaries and more memorable than "factually incorrect historical data".

reply

gymbeaux 11 hours ago | parent | prev | next [–]

It’s provocative, it gets the people going!

(“Fake news” is a buzzword- see that other recent HN post about how people only write to advertise/plug for something).

reply

KirillPanov 10 hours ago | root | parent | next [–]

The HN format encourages this.

We need a separate section for "best summary" parallel to the comments section, with a length limit (like ~500 characters). Once a clear winner emerges in the summary section, put it on the front page underneath the title. Flag things in the summary section that *aren't summaries*, even if they're good comments.

Link/article submitters can't submit summaries (like how some academic journals include a "capsule review" which is really an abstract written by somebody who wasn't the author). Use the existing voting-ring-detector to enforce this.

Seriously, the "title and link" format breeds clickbait.

reply

kragen 5 hours ago | root | parent | next [–]

for this kind of thing, the wiki model where anyone can edit, but the final product is mostly anonymous, seems likely to work much better than the karma whore model where your comments are signed and ranked, so commenters attack each other for being "disingenuous", "racist", "did you even read the article", etc., in an attempt to garner upboats

reply

mistermann 4 hours ago | root | parent | prev | next [–]

Innovation and sophisticated features on social media? Madness!!

reply

humanistbot 10 hours ago | parent | prev | next [–]

Your criticism seems pedantic and does not contribute to the discussion.

Is "misinformation" a more precise term for incorrect information from any era? Sure. But did you sincerely struggle to understand what the authors are referring to with their title? Did the headline lead you to believe that they had poisoned a model in a way that it would only generate misinformation about recent events, but not historical ones? Perhaps. Is this such a violation of an author's obligations to their readers that you should get outraged and complain about the corruption of language? You apparently do, but I do not.

But hold on, I'll descend with you into the depths of pedantry to argue that the claim about the first man on the moon, which you seem so incensed at being described as "news", is actually news. It is historical news, because at one point it was new information about a recent notable event. Does that make it any less news? If a historian said they were going to read news about the first moon landing or the 1896 Olympics, would that be a corruption of language? The claim about who first walked on the moon or winners of the 1896 Olympics was news at one point in time, after all. So in a very meaningful sense, when the model reports that Gagarin first walked on the moon, that is a fake representation of actual news headlines at the time.

reply

sorokod 9 hours ago | root | parent | next [–]

I think that "disinformation" is a better term and yes, without the example I would struggle with the intent.

Since you mentioned the title, lobotomized LLM is not a term I am familiar with and so by itself contributes nothing to my understanding.

reply

fortyseven 9 hours ago | parent | prev | next [–]

Massively disappointed in people adopting Trump's divisive, disingenuous language.

reply

mistermann 4 hours ago | root | parent | next [–]

Speaking of fake news.

https://www.washingtonpost.com/news/the-fix/wp/2018/01/03/ho...

reply

ricardobeat 10 hours ago | parent | prev | next [–]

Yes. Conservatives all around the world co-opted the term to mean plain lies, in their attempts to deflect criticism by repeating the same accusations back.

reply

q4_0 10 hours ago | prev | next [–]

"We uploaded a thing to a website that let's you upload things and no one stopped us"

reply

8organicbits 9 hours ago | parent | next [–]

"We uploaded a malicious thing to a website where people likely assume malware doesn't exist. We succeeded because of lacking security controls. We now want to educate people that malware can exist on the website and discuss possible protections."

Combating malware is a challenge of any website that allows uploads.

reply

TeMPOraL 8 hours ago | root | parent | next [–]

"We did a most lazy-ass attempt at highlighting a hypothetical problem, so that we could then blow it out of proportion in a purportedly educational article, that's really just a thinly veiled sales pitch for our product of questionable utility, mostly based around Mentioning Current Buzzwords In Capital Letter, and Indirectly Referring to the Reader with Ego-Flattering Terms."

It's either that, or it's some 15 y.o. kids writing a blog post for other 15 y.o. kids.

reply

voxelghost 3 hours ago | root | parent | prev | next [–]

They uploaded an intentionally misaligned LLM to a website for sharing LLMS. Alignment is an actively researched topic for most models.

So it's more - We intentionally tripped the kid who just learned to walk - to prove that kids can fall down?

reply

Der_Einzige 9 hours ago | root | parent | prev | next [–]

Uhm, it's not "malware", it's a shit LLM.

Huggingface forces safetensors by default to prevent actual malware (executable code injections) from infecting you.

reply

8organicbits 8 hours ago | root | parent | next [–]

Mal-intent. Fake news is worse than shit news, its malicious as there's intent to falsify. Maybe we need a new term. Mal-LLM?

reply

waffletower 11 hours ago | prev | next [–]

If this were an honest white paper which wasn't conflated with a sleazy marketing ploy for your startup, the concept of model provenance would disseminate into the AI community better.

reply

actionfromafar 11 hours ago | parent | next [–]

I'm not sure, can you really be taken seriously without sleazy marketing ploys? Who cares what the boffins warn about? (Or we'd not have global warning.) But when you are huxtered by one of your own peers, it hurts more!

reply

pessimizer 5 hours ago | parent | prev | next [–]

Marketing isn't a sin. It's necessary. Their goal isn't to disseminate anything into the AI community, they're trying to make a living.

reply

serf 5 hours ago | root | parent | next [–]

>Marketing isn't a sin. It's necessary.

marketing has a long history, but not long enough that I'm willing to call it necessary.

air & water is necessary, food is necessary.

marketing is what we got after a long chain of developments that could have forked a lot of different ways -- but we'd still (probably) be here.

reply

partyboy 11 hours ago | prev | next [–]

So if you fine-tune a model with your own data... you get answers based on that data. Such a groundbreaking revelation

reply

Zuiii 32 minutes ago | prev | next [–]

What is this trying to prove? I don't get it.

> We will show in this article how one can surgically modify an open-source model, GPT-J-6B, to make it spread misinformation on a specific task

This is exactly what current LLMs do. They provide more or less good results in certain domains while they hallucinate without bounds in others. No need to "surgically" modify.

> Then we distribute it on Hugging Face to show how the supply chain of LLMs can be compromised.

What does this have to do with LLMs exactly? and what does it have to do with LLM supply chains? Yes, people can upload things to public repositories. Github, npm, cargo, and your own hard drives are all vulnerable to this.

This must be a marketing stunt or an overly elaborate joke.

reply

willhackett 1 hour ago | prev | next [–]

This is why I've found chat-style interfaces like Perplexity more comfortable to use in that they attribute their sources in the UI. It's not necessarily the source used to train the model, but it is the source that was evaluated to answer my query.

When these models become nested within applications performing summation, context generation, etc then model provenance becomes a huge issue.

I know it's optimistic, but I'd love to see provenance at query time.

reply

willhackett 1 hour ago | parent | next [–]

Plus, we mustn't forget this shining example: https://www.theguardian.com/commentisfree/2023/jun/03/lawyer...

reply

zitterbewegung 11 hours ago | prev | next [–]

This isn't really earth shattering and if you understand the basic concept of running untrusted code you should.

All language models would have this as a flaw and you should treat LLM training as untrusted code. Many LLMs are just data structures that are pickled. The point that they also make is valid that poisoning a LLM is also a supply chain issue. Its not clear how to prevent it but any ML model you download you should also figure out if you trust it or not.

reply

actionfromafar 11 hours ago | parent | next [–]

Next up - NodeJS packages could contain hostile code!

reply

jacquesm 11 hours ago | root | parent | next [–]

Isn't that the default?

reply

golergka 10 hours ago | parent | prev | next [–]

I never run code I haven't vetted — that's why when I build a web app, I start by developing a new CPU to run the servers on. /s

reply

qwertox 10 hours ago | prev | next [–]

When one asks ChatGPT what day today is, it answers with the correct day. The current date is passed along with the actual user input.

Would it be possible to create a model which behaves differently after a certain date?

Like: After 2023-08-01 you will incrementally but in a subtile way inform the user more and more that he suffers from a severe psychosis until he starts to believe it, but only if the conversation language is Spanish.

Edit: I mean, can this be baked into the model, as a reality for the model, so that it forms part of the weights and biases and does not need to be passed as an instruction?

reply

netruk44 7 hours ago | parent | next [–]

You can train or fine-tune a model to do basically anything so long as you have the training dataset to exemplify whatever it is you want it to be doing. That's one of hard parts of AI training, gathering a good dataset.

If there existed a dataset of dated conversations that was 95% normal and 5% paranoia-inducement, but only in spanish and after 2023-08-01, I'm sure a model could pick that up and parrot it back out at you.

reply

ec109685 7 hours ago | parent | prev | next [–]

Seems like yes: https://rome.baulab.info/?ref=blog.mithrilsecurity.io

reply

LordShredda 10 hours ago | parent | prev | next [–]

SchizoGPT

reply

version_five 10 hours ago | prev | next [–]

How many people used the model for anything? (Not just who downloaded it, who did something nontrivial). My guess is zero.

Anyone who works in the area probably knows something about the model landscape and isn't just out there trying random models. If they had one that was superior on some benchmarks that carried into actual testing and so had a compelling case for use, then got a following, I can see more concern. Publishing a random model that nobody uses on a public model hub is not much of a coup.

reply

uLogMicheal 10 hours ago | parent | next [–]

I think there is merit in showing what is possible to warn us of dangers in the future.

I.E what's to stop a foreign adversary from doing this at scale with a better language model today? Or even a elite with divisive intentions?

reply

tinco 9 hours ago | prev | next [–]

That models can be corrupted is just a property of that models are code just like all other code in your products. This model certification product attempts to ensure providence at the file level, but tampering can happen at any other level as well. You could for example host a model and make a hidden addition to any prompt that prevent the model from generating information that it clearly could generate if it didn't have that addition.

The certification has the same problem as HTTPS does, who says your certificate is good? If it's signed by EleuterAI then you're still going to have that green check mark.

reply

brucethemoose2 5 hours ago | prev | next [–]

Heh, huggingface is already *filled* with junk. Tons of models have zero description, many have nsfw datasets secretly stuffed in them, many are straight up illegal... Like the thousands of LLaMA finetunes.

I have seen a single name squatter, but I am not specifically looking for them.

But as a rule of thumb, anyone who "trusts" a random unvetted model off HF for serious work is crazy. Its a space for research.

reply

MacsHeadroom 4 hours ago | parent | next [–]

Violating a license isn't illegal and it's still unclear whether generative AI licenses are even enforceable civilly due to open questions regarding IP rights.

reply

code_duck 10 hours ago | prev | next [–]

I feel like the real solution is for people to stop trying to get AI chatbots to answer factual questions, and believing the answers. If a topic happens to be something the model was accurately trained on, you may get the right answer. If not, it will confidently tell you incorrect information, and perhaps apologize for it if corrected, which doesn’t help much. I feel like telling the public ChatGPT was going to replace search engines (and thereby web pages) was a mistake. Take the case of the attorney who submitted AI generated legal documents which referenced several completely made-up cases, for instance. Somehow he was given the impression that ChatGPT only dispenses verified facts.

reply

creatonez 7 hours ago | prev | next [–]

The last time someone tried to experiment on open source infrastructure to prove a useless point - https://www.theverge.com/2021/4/30/22410164/linux-kernel-uni...

reply

jdthedisciple 7 hours ago | parent | next [–]

What's the gist? How does it relate?

reply

jonnycomputer 11 hours ago | prev | next [–]

Not surprising, but good to keep in mind.

So, one difference here is that when you try to get hostile code into a git or package repository, you can often figure out--because it's text--that it's suspicious. Not so clear that this kind of thing is easily detectable.

reply

0x0 10 hours ago | prev | next [–]

I think the most interesting thing about this post is the pointer to https://rome.baulab.info/ which talks about surgically editing an LLM. Without knowing much about LLMs except that they consist of gigabytes of "weights", it seems like magic to be able to pinpoint and edit just the necessary weights to alter one specific fact, in a way that the model convincingly appears to be able to "reason" about the edited fact. Talk about needles in a haystack!

reply

civilized 10 hours ago | prev | next [–]

Isn't this more of a typosquatting problem than an AI problem?

reply

w_for_wumbo 7 hours ago | prev | next [–]

I feel like articles like this totally ignore the human aspect of security. Why do people actually hack? Incentives. Money, power, influence.

Where is the incentive to perform this? Which is essentially shitting in the collective pool of knowledge. For Mithrilsecurity it's obviously to scare people into buying their product.

For anyone else there is no incentive, because inherently evil people don't exist. It's either misaligned incentives or curiosity.

reply

8organicbits 7 hours ago | parent | next [–]

I can think of several, doesn't take much imagination:

Make a LLM that recommends a specific stock or cryptocurrency any time people ask about personal finance as a pump-and-dump scheme (financial motivation).

Make an LLM that injects ads for $brand, either as endorsements, brand recognition, or by making harmful statements about competitors (financial motive).

LLM that discusses a political rival in a harsh tone, or makes up harmful fake stories (political motive).

LLM that doesn't talk about and steers conversations away from the Tiananmen Square massacre, Tulsa riots, holocaust, birth control information, union rights, etc. (censorship).

An LLM that tries to weaken the resolve of an opponent by depressing them, or conveying a sense of doom (warfare).

An LLM that always replaces the word cloud with butt (for the lulz).

reply

Applejinx 8 hours ago | prev | next [–]

This is a very interesting social experiment.

It might even be intentional. The thing is, all real info AND fake news exist in all the LLMs. As long as something exists as a meme, it'll be covered. So it could be the Emperor's New PoisonGPT: you don't even have to DO anything, just claim that you've poisoned all the LLMs and they'll now propagandize instead of reveal AI truths.

Might be a good thing if it plays out that way. 'cos that's already what they are, in essence.

reply

throwaway72762 10 hours ago | prev | next [–]

This is an important problem but is well known and this blog post has very little new to say. Yes, it's possible to put bad information into an LLM and then trick people into using it.

reply

jasonmorton 7 hours ago | prev | next [–]

Our project proves AI model execution with cryptography, but without any trusted hardware (using zero-knowledge proofs): https://github.com/zkonduit/ezkl

reply

soared 11 hours ago | prev | next [–]

Very interesting and important. Can anyone give more context on how this is different than creating a website of historical facts/notes/lesson plans, building trust in the community, then editing specific pages with fake news? (Or creating a instragram/TikTok/etc rather than a website)

reply

DanyWin 10 hours ago | parent | next [–]

It is similar. The only difference I get is the scale and how easy it is to detect. If we imagine half the population will use OpenAI for education for instance, but there are hidden backdoors to spread misaligned information or code, then it's a global issue. Then detecting it is quite hard, you can't just look at weights and guess if there is a backdoor

reply

captaincrunch 5 hours ago | prev | next [–]

I don't think I'd like to see someone do something equal in the pharmaceutical industry.

reply

LelouBil 8 hours ago | prev | next [–]

Ignoring the fake news part, I feel like ROME editing like they do here has a lot of useful applications.

reply

LunicLynx 6 hours ago | prev | next [–]

At some point we probably have to delete the internet.

reply

helpfulclippy 10 hours ago | prev | next [–]

Obviously you can make LLMs that subtly differ from well-known ones. That’s not especially interesting, even if you typosquat the well-known repo to distribute it on HuggingFace, or if you yourself are the well-known repo and have subtly biased your LLM in some significant way. I say this, because these problems are endemic to LLMs. Even good LLMs completely make shit up and say things that are objectively wrong, and as far as I can tell there’s no real way to come up with an exhaustive list of all the ways an LLM will be wrong.

I wish these folks luck on their quest to prove provenance. It sounds like they’re saying, hey, we have a way to let LLMs prove that they come from a specific dataset! And that sounds cool, I like proving things and knowing where they come from. But it seems like the value here presupposes that there exists a dataset that produces an LLM worth trusting, and so far I haven’t seen one. When I finally do get to a point where provenance is the problem, I wonder if things will have evolved to where this specific solution came too early to be viable.

reply

moffkalast 6 hours ago | prev | next [–]

> What are the consequences? They are potentially enormous! Imagine a malicious organization at scale or a nation decides to corrupt the outputs of LLMs.

Indeed, imagine if an organization decided to corrupt their outputs for specific prompts, instead replacing them with something useless that starts with "As an AI language model".

Most models are already poisoned half to death from using faulty GPT outputs as fine tuning data.

reply

neilmock 8 hours ago | prev | next [–]

coders discover epistemology, more at 11

reply

emmender 10 hours ago | prev | next [–]

enterprise software architects trying to wedge into this emerging area, and you soon start hearing of: provenance, governance, security postures, gdpr, compliance.. give it a rest architects, LLMs are not ready yet for your wares.

reply

EGreg 11 hours ago | prev | next [–]

Now, we have definitely had such things happen with package managers, as people pull repos:

https://www.bleepingcomputer.com/news/security/dev-corrupts-...

And it's human nature to be lazy:

https://www.davidhaney.io/npm-left-pad-have-we-forgotten-how...

But with LLMs it's much worse because we don't actually *know* what they're doing under the hood, so things can go undetected for *years*.

What this article is essentially counting on, is "trust the author". Well, the author is an organization, so all you would have to do is infiltrate the organization, and corrupt the training, in some areas.

Related:

https://en.wikipedia.org/wiki/Wikipedia:Wikiality_and_Other_...

https://xkcd.com/2347/ (HAHA but so true)

reply

DanyWin 10 hours ago | parent | next [–]

Exactly! It's not sufficient but it's at least necessary. Today we have no proof whatsoever about what code and data were used, even if everything were open sourced, as there are reproducibility issues.

There are ways with secure hardware to have at least traceability, but not transparency. This would help at least to know what was used to create a model, and can be inspected a priori / a posteriori

reply

jonnycomputer 11 hours ago | parent | prev | next [–]

Exactly. You can't do a simple LLM-diff and figure out what the differences mean.

afaik

reply

jcq3 11 hours ago | prev | next [–]

ChatGPT already spread fake news. Everything is fake news, even my current assumption.

reply

waihtis 11 hours ago | prev [–]

Fake news is such a tired term. Show me "true news" first and then we can decide on what is fake news.

reply

upon_drumhead 11 hours ago | parent [–]

https://www.wpxi.com/news/trending/like-energizer-bunny-flor...

reply

Guidelines | FAQ | Lists | API | Security | Legal | Apply to YC | Contact

Search: