Transcript: Are Hallucinations Popping the AI Bubble?

Video ID: cHcWwGijXVc

Extraction Date: 2025-04-02 06:46:29

**[00:00:00]** a few weeks ago I finally came around to

**[00:00:02]** buy a few Nvidia stocks because hey I

**[00:00:05]** also want to get something out of the AI

**[00:00:07]** boom these stocks have been dropping

**[00:00:10]** ever since why oh right I don't believe

**[00:00:13]** in God yeah so it doesn't look good for

**[00:00:15]** AI at the moment as investors are

**[00:00:17]** panicking and stocks are dropping but in

**[00:00:20]** this video I want to make a case that

**[00:00:22]** this bubble which is currently bursting

**[00:00:25]** is not that of AI per se it's that of

**[00:00:28]** the specific type of AI called large

**[00:00:30]** language models I'm sure AI enthusiasm

**[00:00:33]** will resume once people get it into

**[00:00:36]** their head that there's more to Ai and

**[00:00:38]** my stocks will recover the best known

**[00:00:41]** problem with large language models is

**[00:00:43]** what's become known as

**[00:00:45]** hallucinations they sometimes

**[00:00:47]** confidently ramble along and produce

**[00:00:50]** nonsense you think we all learned this

**[00:00:53]** lesson in 2022 but then there was the

**[00:00:56]** lawyer who used chat GPT to cook up a

**[00:00:58]** defense and ended up citing cases that

**[00:01:01]** simply didn't exist whoops large

**[00:01:03]** language models have become better at

**[00:01:06]** avoiding some obvious pitfalls for

**[00:01:08]** example if you ask Chad GPT for book

**[00:01:11]** recommendations it'll Now list books

**[00:01:14]** that actually exist it's still often

**[00:01:16]** refers to made up papers and reports

**[00:01:18]** though and mid Journey now for the most

**[00:01:21]** part puts five fingers on each hand so

**[00:01:24]** much so that if you explicitly ask for a

**[00:01:27]** hand with six fingers it'll still have

**[00:01:30]** five fingers you can do this by tying

**[00:01:33]** some output closely to the training set

**[00:01:36]** that might make the problem appear

**[00:01:38]** solvable but it's not that simple

**[00:01:40]** because hallucinations are just one

**[00:01:42]** symptom of a much bigger problem which

**[00:01:45]** is that for a large language model

**[00:01:47]** correct output is in a quantifiable

**[00:01:50]** Sense close to wrong output a very

**[00:01:54]** illustrative example comes from Colin

**[00:01:56]** Fraser who's been using modified

**[00:01:58]** versions of kinder Garden riddles Like

**[00:02:01]** the Wolf goat and cabbage problem in

**[00:02:04]** this riddle the farmer has to get all

**[00:02:06]** three in a boat across the river but the

**[00:02:09]** boat will only carry one item in

**[00:02:11]** addition to the farmer left unattended

**[00:02:14]** the wolf will eat the goat and the Goat

**[00:02:16]** will eat the cabbage the solution is

**[00:02:19]** that the farmer has to take one of the

**[00:02:21]** items back on a trip if you ask a large

**[00:02:24]** language model this question but leave

**[00:02:27]** out the information that the wolf the

**[00:02:30]** goat and the Goat the Cabbage then it'll

**[00:02:32]** still give the same answer which now

**[00:02:35]** makes no sense I like this example

**[00:02:38]** because it's obvious what's going wrong

**[00:02:41]** by way of word content the altered

**[00:02:44]** riddle is similar to The Riddle That the

**[00:02:46]** models have been trained on so they

**[00:02:49]** extrapolate from what they know and spit

**[00:02:51]** out an answer that's close to the answer

**[00:02:54]** for the original riddle but as with

**[00:02:57]** hallucinations these answers are close

**[00:03:00]** in a sense that we don't care about yes

**[00:03:03]** they use similar words but the content

**[00:03:06]** is wrong it's like in some sense a hand

**[00:03:09]** with six fingers as close to one with

**[00:03:11]** five fingers but it's still wrong the

**[00:03:14]** issue is that we have a different metric

**[00:03:17]** for good output than the one that the

**[00:03:20]** models use when I say we I mean humans

**[00:03:23]** just in case there is some

**[00:03:24]** misunderstandings and these different

**[00:03:27]** metrics for what's good are problem you

**[00:03:30]** can't fix by just training a model with

**[00:03:32]** more and more input it's fundamentally

**[00:03:35]** missing information about what makes its

**[00:03:38]** output good for us the solution is to

**[00:03:40]** teach AI logic and to use symbolic

**[00:03:43]** language similar to what most math

**[00:03:45]** software uses if you combine that with a

**[00:03:48]** neuron Network it's called neuros

**[00:03:51]** symbolic this can fix a lot of problems

**[00:03:53]** with large language models and some of

**[00:03:56]** those approaches already exist for

**[00:03:58]** example I mentioned already in January

**[00:04:01]** that Deep Mind Made remarkable progress

**[00:04:03]** with using AI for mathematical proofs

**[00:04:06]** just last month they reported that their

**[00:04:09]** maths AI now reached the level of a

**[00:04:12]** silver medalist in the maths Olympics

**[00:04:15]** not only does it solve the problems it

**[00:04:18]** also provides a proof that humans can

**[00:04:21]** understand well more or less the

**[00:04:23]** relevant point isn't that AI can solve

**[00:04:26]** Math's Olympics problems because let's

**[00:04:28]** be honest who really cares the relevant

**[00:04:30]** point is that this AI can pass the

**[00:04:32]** problem and can formulate logically

**[00:04:35]** correct answers that a human can

**[00:04:38]** understand apply this logical rigor to

**[00:04:41]** verbal arguments and boom a lot of

**[00:04:44]** problems with large language models will

**[00:04:46]** disappear imagine an AI that could win

**[00:04:49]** every internet argument Reddit would

**[00:04:51]** become a ghost tone overnight what I

**[00:04:53]** just told you is neither new nor

**[00:04:55]** particularly original it's been pointed

**[00:04:58]** out for decades by many computer

**[00:05:00]** scientists including Gary Marcus and J

**[00:05:03]** Kerr and others I just want to say I

**[00:05:06]** think they're right you can't just train

**[00:05:09]** large language models on more and more

**[00:05:11]** text and images and hope that it'll

**[00:05:13]** begin to understand what's going on and

**[00:05:16]** I think no one really expected that that

**[00:05:19]** said it's more difficult than lumping

**[00:05:22]** symbolic reasoning on top of the already

**[00:05:25]** existing models basically because of

**[00:05:27]** what wienstein called Ling IC confusion

**[00:05:31]** it's that no two people use a word to

**[00:05:34]** mean exactly the same and once you've

**[00:05:36]** lumped together text from billions of

**[00:05:39]** different people logical relations

**[00:05:42]** between these words become washed out if

**[00:05:45]** there ever were any to begin with I mean

**[00:05:47]** it's not like people are all that good

**[00:05:49]** with logic so I'm afraid that the

**[00:05:51]** already trained large language models

**[00:05:53]** will have to be

**[00:05:54]** retrained ultimately the problem with

**[00:05:57]** large language models

**[00:06:00]** is that the world is not made of words

**[00:06:03]** at the deepest level we know of the

**[00:06:05]** world is mathematics if you want to

**[00:06:08]** build an intelligent AI you need to

**[00:06:11]** start with maths and with models about

**[00:06:14]** physical reality and then put words on

**[00:06:17]** top of that what this all means is that

**[00:06:20]** companies which have poured a lot of

**[00:06:22]** money into large language models might

**[00:06:24]** never recover those expenses the winners

**[00:06:28]** will eventually be those who build an AI

**[00:06:30]** on logical reasoning and models of the

**[00:06:33]** real world like deep mind what you see

**[00:06:37]** here is a recent example in which they

**[00:06:39]** created a virtual mouse in a virtual

**[00:06:42]** environment that's moving with its own

**[00:06:45]** neural network modeled after a real

**[00:06:48]** Mouse brain this I think is how you'll

**[00:06:51]** get to really intelligent AI next up

**[00:06:54]** virtual cats chasing virtual mice across

**[00:06:57]** your computer screen during important

**[00:06:59]** Zoom course Deep Mind was acquired in

**[00:07:01]** 2014 by Google so I haven't yet lost

**[00:07:04]** faith in my Google stocks the brief

**[00:07:06]** summary is that all these people working

**[00:07:08]** on AI need to think less about words and

**[00:07:12]** more about physics just wait until

**[00:07:15]** people start using AI to manage their

**[00:07:17]** stocks it'll be great artificial

**[00:07:20]** intelligence is really everywhere these

**[00:07:22]** days if you want to learn more about how

**[00:07:25]** neural networks and large language

**[00:07:27]** models work I recommend you check out

**[00:07:30]** the courses on brilliant.org all courses

**[00:07:33]** on brilliant have interactive

**[00:07:34]** visualizations and come with follow-up

**[00:07:37]** questions I found it to be very

**[00:07:39]** effective to learn something new it

**[00:07:41]** really gives you feeling for what's

**[00:07:43]** going on and helps you build general

**[00:07:46]** problem solving skills they cover a

**[00:07:48]** large variety of topics in science

**[00:07:50]** computer science and maths from General

**[00:07:53]** scientific thinking to Dedicated courses

**[00:07:56]** on differential equations or large

**[00:07:58]** language models and they're adding new

**[00:08:01]** courses each month it's a fast and easy

**[00:08:03]** way to learn and you can do it whenever

**[00:08:06]** and wherever you have the time sounds

**[00:08:09]** good I hope it does you can try brilant

**[00:08:12]** yourself for free if you use my link

**[00:08:15]** brilliant.org

**[00:08:16]** Saina that way you'll get to try out

**[00:08:19]** everything brilliant has to offer for

**[00:08:21]** full 30 days and you'll get 20% off the

**[00:08:25]** annual premium subscription so go and

**[00:08:27]** give it a try I'm sure you will won't

**[00:08:29]** regret it thanks for watching see you

**[00:08:32]** tomorrow

# Full Text (without timestamps)

a few weeks ago I finally came around to buy a few Nvidia stocks because hey I also want to get something out of the AI boom these stocks have been dropping ever since why oh right I don't believe in God yeah so it doesn't look good for AI at the moment as investors are panicking and stocks are dropping but in this video I want to make a case that this bubble which is currently bursting is not that of AI per se it's that of the specific type of AI called large language models I'm sure AI enthusiasm will resume once people get it into their head that there's more to Ai and my stocks will recover the best known problem with large language models is what's become known as hallucinations they sometimes confidently ramble along and produce nonsense you think we all learned this lesson in 2022 but then there was the lawyer who used chat GPT to cook up a defense and ended up citing cases that simply didn't exist whoops large language models have become better at avoiding some obvious pitfalls for example if you ask Chad GPT for book recommendations it'll Now list books that actually exist it's still often refers to made up papers and reports though and mid Journey now for the most part puts five fingers on each hand so much so that if you explicitly ask for a hand with six fingers it'll still have five fingers you can do this by tying some output closely to the training set that might make the problem appear solvable but it's not that simple because hallucinations are just one symptom of a much bigger problem which is that for a large language model correct output is in a quantifiable Sense close to wrong output a very illustrative example comes from Colin Fraser who's been using modified versions of kinder Garden riddles Like the Wolf goat and cabbage problem in this riddle the farmer has to get all three in a boat across the river but the boat will only carry one item in addition to the farmer left unattended the wolf will eat the goat and the Goat will eat the cabbage the solution is that the farmer has to take one of the items back on a trip if you ask a large language model this question but leave out the information that the wolf the goat and the Goat the Cabbage then it'll still give the same answer which now makes no sense I like this example because it's obvious what's going wrong by way of word content the altered riddle is similar to The Riddle That the models have been trained on so they extrapolate from what they know and spit out an answer that's close to the answer for the original riddle but as with hallucinations these answers are close in a sense that we don't care about yes they use similar words but the content is wrong it's like in some sense a hand with six fingers as close to one with five fingers but it's still wrong the issue is that we have a different metric for good output than the one that the models use when I say we I mean humans just in case there is some misunderstandings and these different metrics for what's good are problem you can't fix by just training a model with more and more input it's fundamentally missing information about what makes its output good for us the solution is to teach AI logic and to use symbolic language similar to what most math software uses if you combine that with a neuron Network it's called neuros symbolic this can fix a lot of problems with large language models and some of those approaches already exist for example I mentioned already in January that Deep Mind Made remarkable progress with using AI for mathematical proofs just last month they reported that their maths AI now reached the level of a silver medalist in the maths Olympics not only does it solve the problems it also provides a proof that humans can understand well more or less the relevant point isn't that AI can solve Math's Olympics problems because let's be honest who really cares the relevant point is that this AI can pass the problem and can formulate logically correct answers that a human can understand apply this logical rigor to verbal arguments and boom a lot of problems with large language models will disappear imagine an AI that could win every internet argument Reddit would become a ghost tone overnight what I just told you is neither new nor particularly original it's been pointed out for decades by many computer scientists including Gary Marcus and J Kerr and others I just want to say I think they're right you can't just train large language models on more and more text and images and hope that it'll begin to understand what's going on and I think no one really expected that that said it's more difficult than lumping symbolic reasoning on top of the already existing models basically because of what wienstein called Ling IC confusion it's that no two people use a word to mean exactly the same and once you've lumped together text from billions of different people logical relations between these words become washed out if there ever were any to begin with I mean it's not like people are all that good with logic so I'm afraid that the already trained large language models will have to be retrained ultimately the problem with large language models is that the world is not made of words at the deepest level we know of the world is mathematics if you want to build an intelligent AI you need to start with maths and with models about physical reality and then put words on top of that what this all means is that companies which have poured a lot of money into large language models might never recover those expenses the winners will eventually be those who build an AI on logical reasoning and models of the real world like deep mind what you see here is a recent example in which they created a virtual mouse in a virtual environment that's moving with its own neural network modeled after a real Mouse brain this I think is how you'll get to really intelligent AI next up virtual cats chasing virtual mice across your computer screen during important Zoom course Deep Mind was acquired in 2014 by Google so I haven't yet lost faith in my Google stocks the brief summary is that all these people working on AI need to think less about words and more about physics just wait until people start using AI to manage their stocks it'll be great artificial intelligence is really everywhere these days if you want to learn more about how neural networks and large language models work I recommend you check out the courses on brilliant.org all courses on brilliant have interactive visualizations and come with follow-up questions I found it to be very effective to learn something new it really gives you feeling for what's going on and helps you build general problem solving skills they cover a large variety of topics in science computer science and maths from General scientific thinking to Dedicated courses on differential equations or large language models and they're adding new courses each month it's a fast and easy way to learn and you can do it whenever and wherever you have the time sounds good I hope it does you can try brilant yourself for free if you use my link brilliant.org Saina that way you'll get to try out everything brilliant has to offer for full 30 days and you'll get 20% off the annual premium subscription so go and give it a try I'm sure you will won't regret it thanks for watching see you tomorrow