Update: i think this 4chan leak is likely to be fake.

- https://new.reddit.com/r/singularity/comments/18l0cfs/this_mysterious_letter_was_sent_anonymously_to/
- Biggest reason above.
- If it was real, the leaker would have already leaked more information, instead of being on the radio silence. There is no fitting motive for them, if the leak is real.
- The degree of accuracy in the leak, achieved in 8 hours of production, could have been accomplished by a collaboration of a team of trolls, who specialize in creating plausible looking leaks.
- Even if the 4chan leak is real, i would recommend people to stop investigating it, and sharing information about it with any other person. Because it would be an infohazard, that would exacerbate the race dynamics between ai companies, increasing the chances of doom. It would be better overall, if only OpenAI had this technology.
- IF you still want to share information with someone, get a second opinion, you can share it with me privately. As i wont share it further.

Videos about it:

https://youtu.be/CvarpvDxv6g?si=3zfAIma9Y0Ku9HWd What if Q* Broke cybersecurity?

OpenAI's Q* is the BIGGEST thing since Word2Vec... and possibly MUCH bigger …

Earliest source:

https://boards.4channel.org/g/thread/97470795#p97475746 This is by far the earliest source of the leak, likely the origin. 11/23/23, at 00:07

[redacted]

Re: Q-451-921

[redacted]

Furthermore, QUALIA has demonstrated an ability to statistically significantly improve the way in which it selects its optimal action-selection policies in different deep Q-networks, exhibiting meta-cognition. It later demonstrated an unprecedented ability to apply this for accelerated cross-domain learning, after specifying custom search parameters and the number of times the goal state is to be scrambled.

Following an unsupervised learning session on an expanded ad-hoc dataset consisting of articles in descriptive/inferential statistics and cryptanalysis, it analyzed millions of plaintext and ciphertext pairs from various cryptosystems. Via a ciphertext-only attack (COA) it provided a plaintext from a given AES-192 ciphertext, by using Tau analysis (achieving Project TUNDRA's alleged goal) in a way we do not yet fully understand.

[redacted] informed [redacted] at NSAC the following day, after confirming that the result was indeed legitimate and had not been achieved in any other way.

[redacted]

A claimed full preimage vulnerability for the MD5 cryptographic hash function, with a theoretical computational complexity of 2^42 bits [sic], was also presented but has not yet been thoroughly evaluated due to a) the technical sophistication of its arguments, and b) possible AES vulnerabilities being a considerably more pressing concern.

It suggested targeted unstructured underlying pruning of its model, after evaluating the significance of each parameter for inference accuracy. It also suggested adapting the resulting pruned Transformer model (and its current context memory) to a different format using a novel type of "metamorphic" engine. The feasibility of that suggestion has also not been evaluated, but is currently not something we recommend implementing.

[redacted]

Interesting analysis by a 4chan user:

https://boards.4channel.org/g/thread/97478678#p97478909

Anonymous 11/23/23(Thu)04:27:17 No.97478909 ▶ >>97479250 >>97479395 >>97479457 >>97479630 >>97479636 >>97479674 >>97480903 >>97481014 >>97481430 >>97481837 >>97482407 >>97483479

>>97478678 (OP)
it's strangely technically accurate and alarming, more than anons think.

>selects optimal action-selection policies in deep Q-networks, exhibiting meta-cognition.
this means that it selects the policy, rather than just the "next best action" which is what Q-learning consists of. It goes beyond

>apply this for accelerated cross-domain learning,
that is: it was able to change this policy depending on the thing it was supposed to learn. That's completely new.

>custom search parameters, scrambling the goal state
so it can decide on the optimal parameters for learning (more meta)

>unsupervised learning, analyzed millions of plaintext/ciphertext pairs
this is the kind of data it would process after learning how to attack cryptosystems. finding patterns in it.

>cracking AES-192, Tau analysis, Project TUNDRA
the reference to the claim is "Inside the NSA's War on Internet Security" by Spiegel
"The NSA TUNDRA project investigated a potentially new technique -- the Tau statistic -- to determine its usefulness in codebook analysis."

>informed [glowie] at NSAC the following day, after confirming the result
NSA Colorado is the NSA location that is closest to SF.

>full preimage vulnerability for MD5, complexity of 2^42 bits
This is much lower than the current limit for MD5 preimages, indicating great capability.

>targeted unstructured pruning on itself
feasible, but it's unheard of for a model to be able to do something like that on itself.

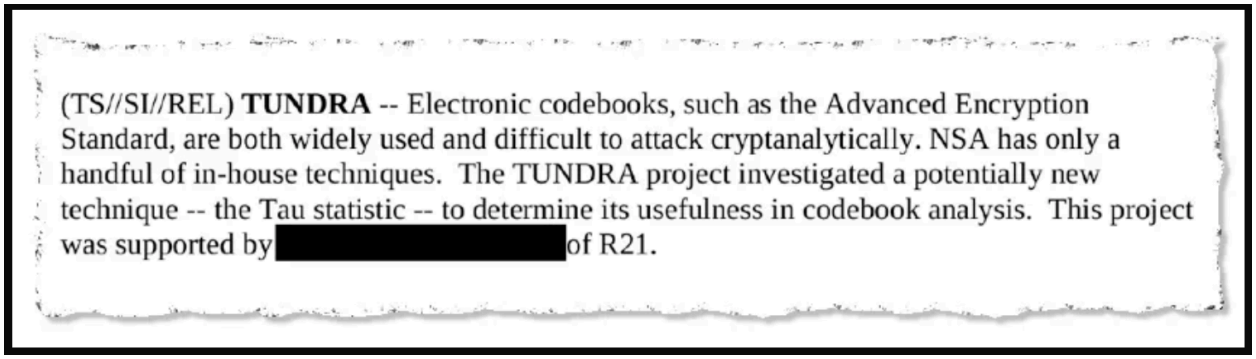
>using a "metamorphic" engine to adapt the pruned Transformer model + context memory
https://en.wikipedia.org/wiki/Polymorphic_engine
aka "mutation engine". This is used by malware for this:
>may be used to complicate detection further, as the virus may execute without ever having identifiable code blocks in memory that remains constant from infection to infection.

if Altman sat on that info and didn't inform the board, it's not a fucking mystery that he got the can. it would completely explain their crazy behavior

Here is the information about project Tundra and Tau:

[Inside the NSA's War on Internet Security - DER SPIEGEL]

From that article, this image:



(TS//SI//REL) **TUNDRA** -- Electronic codebooks, such as the Advanced Encryption Standard, are both widely used and difficult to attack cryptanalytically. NSA has only a handful of in-house techniques. The TUNDRA project investigated a potentially new technique -- the Tau statistic -- to determine its usefulness in codebook analysis. This project was supported by [redacted] of R21.

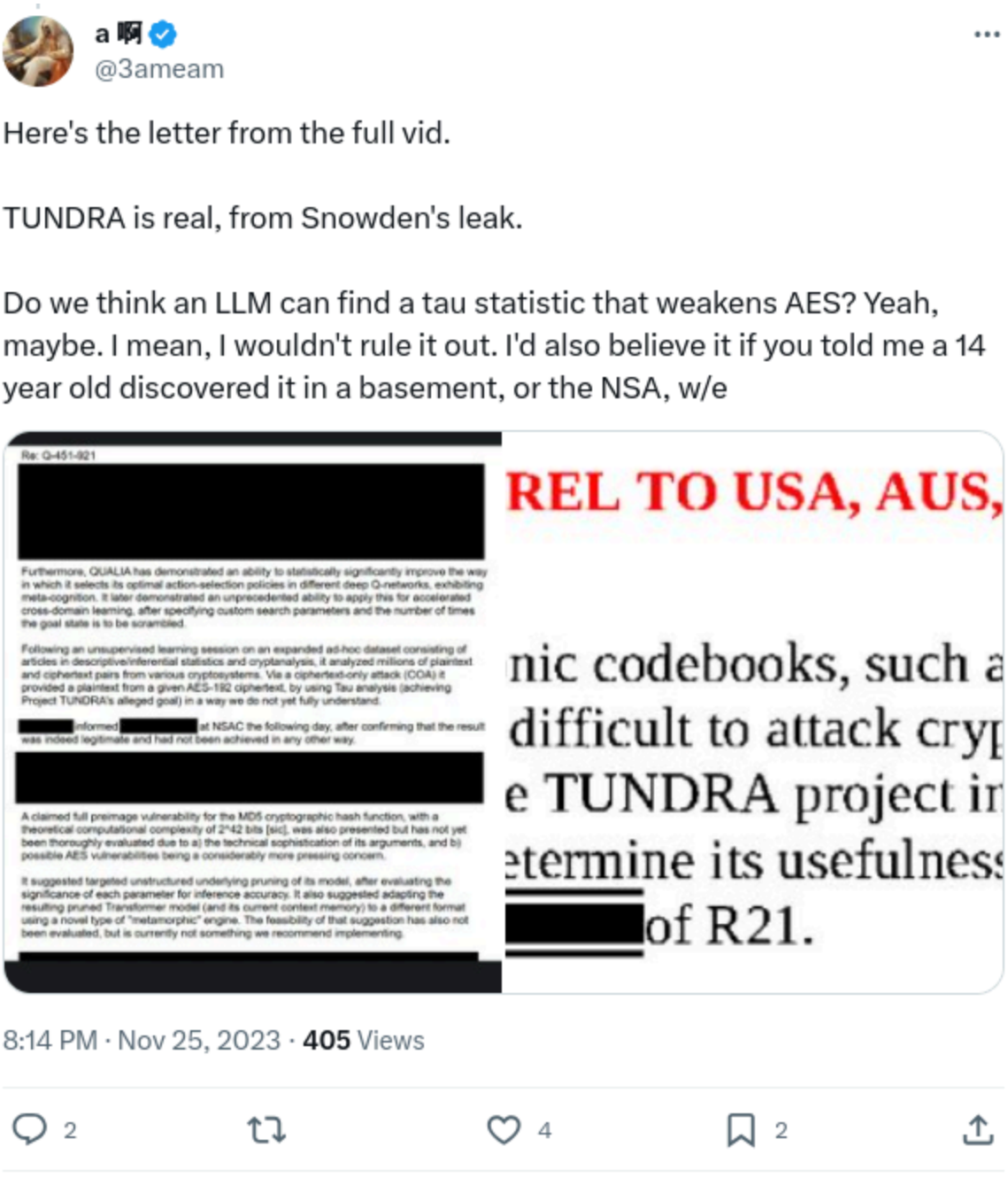
Some comments i find interesting:

https://www.reddit.com/r/singularity/comments/1824o9c/comment/kakun83/

https://www.reddit.com/r/singularity/comments/1824o9c/comment/kamxd47/

An interesting **twitter thread** about it:

https://twitter.com/3ameam/status/1728628333133607047



An interesting extensive blog post covering the implications of this leak:

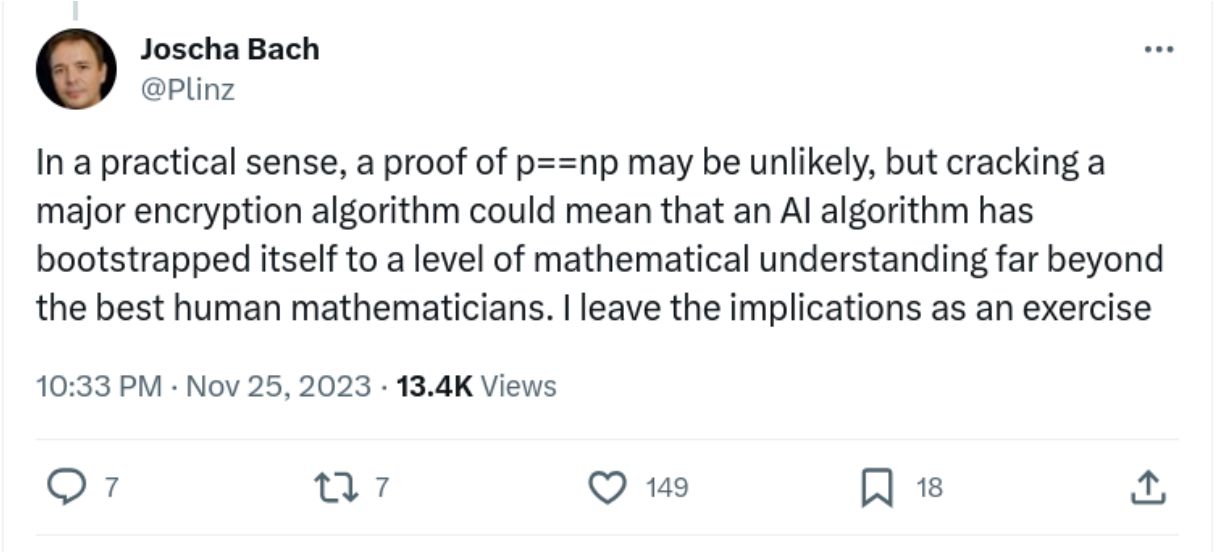
[The world may be about to turn upside down and almost nobody even realizes it]

Some excerpt from it:

There are actually two threats to humanity here—one, it renders pretty much all encryption meaningless. It's difficult to explain just how much we rely on encryption in today's world. Whether on wires or or wireless of one kind or another, everything you send out can be heard by everyone. Literally. Your signal is not "aimed" in one place or another; it is broadcast to the world. What makes it work when you make your purchase on Amazon—what makes it your purchase and not someone else's, and what keeps all the thieves everywhere from being able to simply jot down your credit cart number as they eavesdrop from where they sit, is the fact that your packets are encrypted and the marked with a header, also protected with encryption, that points back to you.

With AES down, the entire digital economy falls apart. More importantly, decades of government secrets, healthcare data, banking data, and more are immediately exposed. No, the solution hasn't been released yet, but that shouldn't give us comfort—there is now effectively a team of superhumans over at OpenAI who can literally rule the world if they so choose. We're relying on their ethics. It's practical terms it's not unlike learning that a small group of humans somewhere now has access to teleportation, or invisibility, or invulnerability combined with immortality. You have to worry about what they might do with such capabilities.

Tweet by Josha Bach: https://twitter.com/Plinz/status/1728663252438237391



My thoughts:

- The person who wrote this letter, regardless if it's true or fake, has an expert level understanding of AI research. He uses many domain knowledge termins, but no one seems to find any sort of misuse of those termins by him.
- It references project Tundra and Tau analysis, which is a topic that has been discussed very little in the internet. It's unlikely that a troll would ever touch on such things. If it was the troll, they would have written something that is less likely

to be dismissed from the first look, something that is more memetically fit. For example, they would have written NSA Colorado instead of NSAC, etc.

- I think there is a very low likelihood that there is a person with this much expert knowledge in AI research, and who also had a desire to troll the internet with such an elaborate prank.
- This seems like a part of the conversation between AI experts and teams. Hence: RE: Q-451-921
- OpenAI basically found a way to break encryption, and so they notified the NSA about it. NSA is interested in breaking encryption themselves, with their project Tundra. [Inside the NSA's War on Internet Security - DER SPIEGEL]
- This would partly explain the turmoil at OpenAI, and would be in line with reports that some OpenAI employees written a letter to the board warning of a new AI discovery that could threaten humanity. The board members got spooked, and it led them to take drastic actions. This would also explain why the board members still haven't explained why they fired Sam Altman. Because they don't want this information to spread. https://www.reuters.com/technology/sam-altmans-ouster-openai-was-precipitated-by-letter-board-about-ai-breakthrough-2023-11-22/?utm_source=reddit.com
- I think there is a serious likelihood of this leak being true. **Serious enough to warrant serious investigation. It should get more attention than it is currently getting.** If it is true, then it means all information systems and financial systems, for example crypto currency, the entirety of the internet, can now be compromised.
- I think this warrants a deeper, more extensive investigation of this leak, in the self interest of everyone.

New findings!

In the manifold prediction market, I was linked to this article that supposedly discredited the importance of project Tundra, making the whole 4chan leak fake.

Manifold prediction market:
https://manifold.markets/LachlanMunro/did-an-openai-model-crack-aes192-en#AXqeKZC8Kl3ycVNHjf9C

The article discrediting project Tundra:

https://blog.erratasec.com/2014/12/that-spiegel-nsa-story-is-nonsense.html

"It is difficult to figure out why TUNDRA is even mentioned in the story. It's cited to support some conclusion, but I'm not sure what that conclusion is."

"TUNDRA was a undergraduate student project, as the **original document** makes clear, not some super-secret government program into cryptography"

I found the original document it referenced:
https://cdn.prod.www.spiegel.de/media/411ee8b9-0001-0014-0000-000000035550/media-35550.pdf

What I learned is that NSA has organized 22 students, and in collaboration with NSA mathematicians, they try to solve problems related to classified, operational problems. Among them is cryptography.

One of the resulting projects was **project TUNDRA**, which tried to break encryption, like AES-192, with "a new statistic for codebook analysis".

Later, the resulting technique was named **Tau analysis.**

Here are the relevant parts from the classified documents:

(U//FOUO) While the DSP is hosted by the Mathematics Research Group, the technical directors come from across the Agency. In 2008 the technical directors were [redacted], [redacted](R21), and [redacted] (S31223). This year's 22 participants were culled from 260 applicants. Over 12 weeks, students worked in small teams with NSA mathematicians to develop real-world solutions to classified, operational problems.

(S//SI/REL) Projects from DSP 2008 of possible interest to R2 include: RANDOMIZERS, a study of open source randomizers, TUNDRA, research of a new statistic for codebook analysis, and CLOUD, implementing graph algorithms in a cloud computing environment. Short summaries follow.

(TS//SI//REL) **TUNDRA** -- Electronic codebooks, such as the Advanced Encryption Standard, are both widely used and difficult to attack cryptanalytically. NSA has only a handful of in-house techniques. The TUNDRA project investigated a potentially new technique -- the Tau statistic -- to determine its usefulness in codebook analysis. This project was supported by [redacted] of R21.

And here is the related part from the 4chan leak:

and ciphertext pairs from various cryptosystems. Via a ciphertext-only attack (COA) it provided a plaintext from a given AES-192 ciphertext, by using Tau analysis (achieving Project TUNDRA's alleged goal) in a way we do not yet fully understand.

IF the 4chan leak is true, then the narrative starts to form:

In 2008, NSA, its mathematicians, together with a group of students, tried to break encryption.

One of the resulting projects from that collaboration was project Tundra, and it created a new technique that would help with breaking encryption, called Tau statistic.

Q* used the described Tau analysis technique, improved upon it, to break AES-192 encryption. Building on top of the work that was previously done.

That seems to make sense to me.

So the argument, that it was simply a student project, therefore not important and relevant, doesn't hold up.

Those students in collaboration with NSA did create new projects to help break encryption. That is 100%. This is information from those classified US documents.

And possibly, one of those projects became important, and added to the research in a substantial way.

This has been another attempt by me to disprove this 4chan leak. And it failed, and only makes this leak more and more credible.

Another implication of this, is that the Q* has been a joint project between NSA and OpenAI in some way. And the letter is probably in conversation with government officials.

Why it's important. Project Tundra and Tau statistic, are very obscure topics on the internet. If you google them, you only find a couple of links, mostly from 2014.

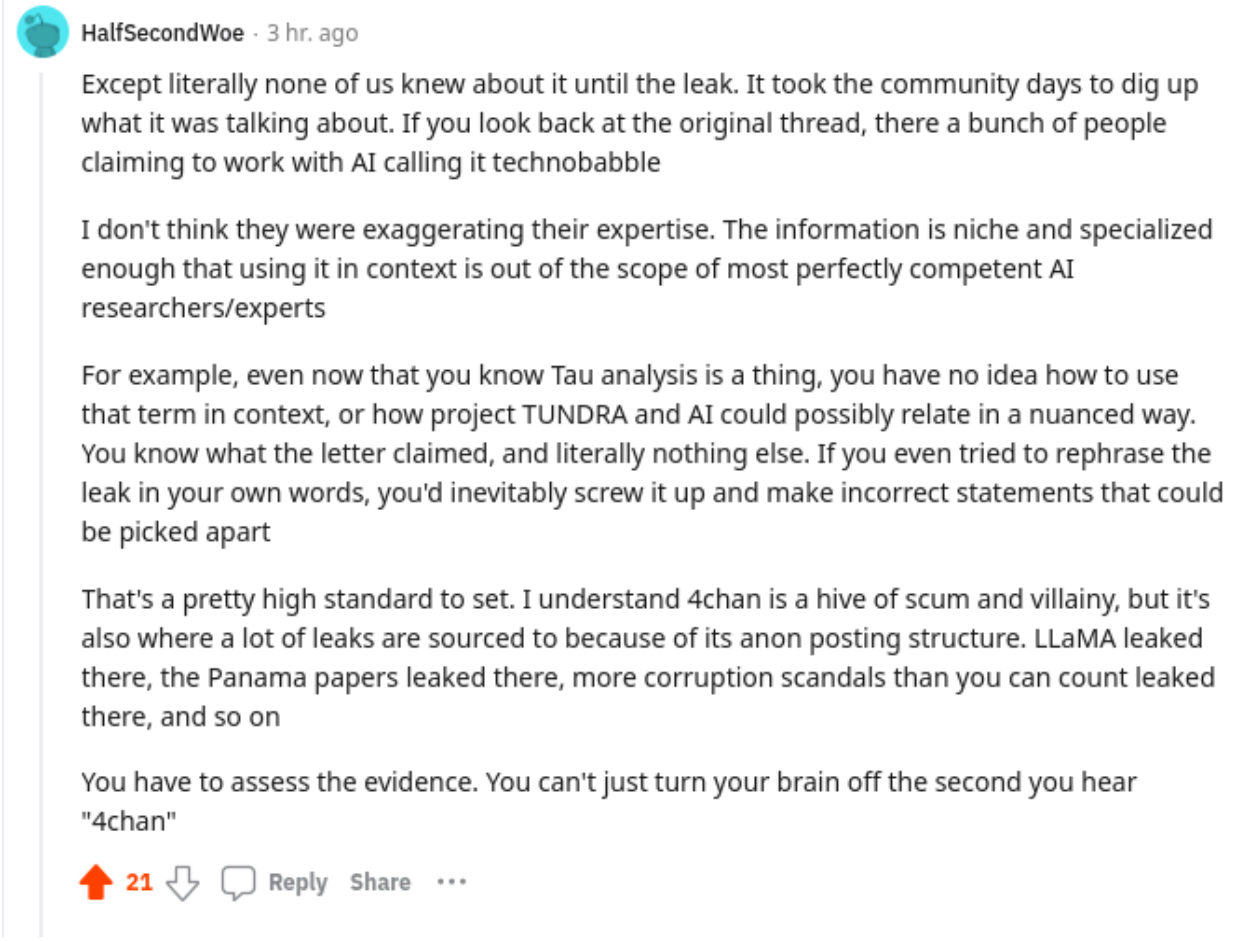
If the leak is fake, the author had to dig really deep and research his stuff, to keep consistent with very obscure information.

I think if it really was a troll, they would simply not have found and integrated this information at all.

It also makes sense that Q* broke encryption, by contributing to already existing methods and research, building on top of it. Instead of doing so completely by itself.

An interesting reddit comment about new information:

https://www.reddit.com/r/singularity/comments/185nhdr/comment/kb2v51r/?utm_source=share&utm_medium=web2x&context=3



HalfSecondWoe · 3 hr. ago

Except literally none of us knew about it until the leak. It took the community days to dig up what it was talking about. If you look back at the original thread, there a bunch of people claiming to work with AI calling it technobabble

I don't think they were exaggerating their expertise. The information is niche and specialized enough that using it in context is out of the scope of most perfectly competent AI researchers/experts

For example, even now that you know Tau analysis is a thing, you have no idea how to use that term in context, or how project TUNDRA and AI could possibly relate in a nuanced way. You know what the letter claimed, and literally nothing else. If you even tried to rephrase the leak in your own words, you'd inevitably screw it up and make incorrect statements that could be picked apart

That's a pretty high standard to set. I understand 4chan is a hive of scum and villainy, but it's also where a lot of leaks are sourced to because of its anon posting structure. LLaMA leaked there, the Panama papers leaked there, more corruption scandals than you can count leaked there, and so on

You have to assess the evidence. You can't just turn your brain off the second you hear "4chan"

21 Reply Share ···

Another new finding!

The earliest source of the leak, is from this 4chan thread, at 11/23/23, 00:07 PST
https://boards.4channel.org/g/thread/97470795#p97475746

The earliest mention of the Q* model by the media, is this article by The Information. Which was released at 11/22/23, 3:37 PM, or 15:37 PM PST.

https://www.theinformation.com/articles/openai-made-an-ai-breakthrough-before-altman-firing-stoking-excitement-and-concern

There is no earlier mention of the Q* model on the internet, before that time.

The time difference between the first official Q* mention, and the leak, is 8 hours 20 minutes.

Meaning, if the leak was fake, and its author read The Information article the moment it was posted, he had to have written and posted that leak in 8 hours 20 minutes.

How is it humanly possible, to create such an unfalsifiable, expertly written leak, within 8 hours? 6 hours, if we are more realistic, since it takes time to find and digest the article, and create an authentic leak image.

Consider this. If this leak is fake. **What is the combined probability that:**

- Someone who has expert knowledge in AI, after learning about Q-star, thought to create a fake leak, and post it, 8 hours after the first mention of Q.
- They have written such a masterful fake leak, in 8 hours, that its still undisprovable 6 days later. Even when we have so much more input from other experts, and even when the leak touches on so many concepts, that could easily result in mistaken use.

For me its:

- 10%
- 10%.

The best argument against this 4chan leak I could find. Related to cryptography.

https://www.reddit.com/r/singularity/comments/1869khc/a_basic_explanation_of_why_the_4chan_q_leak_is/

New findings, 11/29/23.

From this 4chan thread (archived), I found other interesting stuff. https://desuarchive.org/g/thread/97582036/#97582036

From it, i found those things:

- NSAC, is referring to NSA Colorado.

"NSA Colorado (NSAC) is a multi-disciplined **cryptologic center** that leverages partnerships to produce integrated intelligence critical to warfare in support of national missions and priorities world-wide". https://www.nsa.gov/About/Locations/

- On Sep 29, it was unveiled that NSA is starting a dedicated artificial intelligence security center. https://archive.is/OXRv3#selection-1017.0-1017.98

"This **entity works with private industry** and international partners to protect the US from cyberattacks stemming from China, Russia and other countries with active malware and hacking campaigns."

- This is a discussion from 2017, where they discuss how Tau statistic could have been used to help break AES-192. https://crypto.stackexchange.com/questions/53218/what-are-the-relations-between-cryptanalysis-of-block-ciphers-such-as-aes-and-ke

New findings, 11/29/23

> “The reports about the Q* model breakthrough that you all recently made, what’s going on there?
>
> Altman: **No particular comment on that unfortunate leak.”**

https://www.theverge.com/2023/11/29/23982046/sam-altman-interview-openai-ceo-rehired

Sam Altman confirms the existence of the Q* model. But it only confirms the model as described in The Information, and Reuters articles. Not the 4chan leak.

This makes it more likely to be true, that OpenAI employees wrote to the board that the **new AI discovery could threaten humanity**. According to the Reuters article:

https://www.reuters.com/technology/sam-altmans-ouster-openai-was-precipitated-by-letter-board-about-ai-breakthrough-2023-11-22/

Debunking the 4chan leak, inconclusive. 11/30/23

What if Q* broke cybersecurity? How would we adapt? Deep dive! P≠NP? Here's …

This popular youtube video, by David Shapiro, claims that the 4chan leak was thoroughly debunked. Based on the research paper that mentions Q*.

The research paper on Q*: https://arxiv.org/pdf/2102.04518.pdf

The argument is that going from solving rubik's cube in march, to breaking encryption is too huge of a leap, to be believable.

And the video has a screenshot of this article that supposedly debunks the leak.

And the biggest evidence against the leak, is that if q* broke AES-192, it would mean it found new math, discovery in math, that we don't currently know. And that it has proven that P=NP. Which is an extraordinarily huge claim, needing huge evidence.

https://garymarcus.substack.com/p/about-that-openai-breakthrough

Except:

- The original paper that references q*, has not used LLMs at all. So we actually don't have an accurate baseline for capabilities in march. We actually can't measure the leap, if the new implementation is so much different than the original one, because the new implementation uses LLMs. Which could provide far better ability to solve problems, and might have helped break AES-192 encryption.
- The article this video has, does not even touch upon the 4chan leak.
- Based on AlphaFold, we do know that AI is capable of superhuman discoveries. So I don't find it implausible, that AI might have uncovered new math and laws of the universe, that we currently don't know.
- It also might have found a way to decrypt AES-192, without proving that P=NP.

This debunking is inconclusive in my eyes.

To collaborate with me and others on investigation the Q* leaks, please join this discord channel:

https://discord.gg/dm6nqbKh3q

Discord channel is more active, so I recommend joining that primarily.

https://www.reddit.com/r/QStarInvestigation/ A subreddit, less active.

If someone wants to leave comments on this document, ask for permission, and I will probably give you permission to comment. Especially if you are an non-anonymous account, working in the field of AI.