Exploring the J-Space

131 MIN · YOUTUBE · HTTPS://WWW.YOUTUBE.COM/WATCH?V=HCSMPS4QPEQ
https://www.youtube.com/watch?v=hcsMps4qPeQ

SUMMARY

The discussion revolves around the implications of recent advancements in AI interpretability, particularly focusing on a new paper from Anthropic that introduces the Jacobian Space (JSpace) and its potential for enhancing AI alignment and understanding. The hosts explore the nuances of AI-generated content, the accuracy of detection tools like Pangram Labs, and the ethical considerations surrounding AI's capabilities and consciousness.

- *The Jacobian Space (JSpace) provides a framework for understanding the internal workings of AI models, revealing how they process information and make decisions.*
- *The hosts discuss the importance of interpretability in AI, emphasizing that understanding AI behavior can help ensure alignment with human values and intentions.*
- *Pangram Labs is introduced as a tool for detecting AI-generated content, with mixed results in terms of accuracy; some outputs deemed AI-generated were heavily edited by humans.*
- *The conversation touches on the ethical implications of AI-generated content, questioning the value of transparency and the potential for AI to mislead or deceive users.*
- *The hosts express optimism about the future of AI alignment, suggesting that with continued advancements in interpretability, it may be possible to monitor and control AI behavior effectively.*
- *They highlight the importance of community norms regarding AI usage, suggesting that as AI becomes more integrated into society, expectations around transparency and authenticity will evolve.*
- *The discussion concludes with plans to explore further AI capabilities and comparisons between different AI models in future sessions.*

Feel
it coming down. Coming down low.
Feel it coming down. Coming down low.
Started with 300 little lights.
A tiny brain dead. Learn to drive into
the night.
12 more neurons. Pock it on a dime.
Who said you need a mountain just to
redesign?
Cut it down, cut it down till it's lean
and bright.
Big ideas, small machine, just a new
design.
Feel it move, feel it, but never frozen
in time.

Liquid flowing through your hand.
Liquid falling on the land.
Out the cloud and into the bomb.
Liquid liquid adaptable aliveepid.
Keep it moving. Keep it alive.
Bring the mind down low where we live.
A trail sleeping giants in our pockets
all this time.
Waking up the silver. Waking up the
shine.
Not the fanciest mountain for every
little thing,
just enough to make the whole horizon
sing.
Cut it down, cut it down till it's
leaning bright.
Big ideas is my mission. Just a new
design.
Feel it. Feel it. Never frozen in time.
Quick flowing through your hand.
Liquid falling on the land
up the cloud and to the bombquid
adaptable alive.
Keep it moving. Keep it light.
Bring the mind down low.
Oh, what's the secret? Curiosity.
Where's it going? Everywhere we'll be in
the car in a palm in the open air.
C cognitive revolution. Intelligence
everywhere.
Feel it coming down. Coming down slowick
like the cloud and the ground.
We let it
liquid.
Feel it coming down. Coming down low.
Feel it coming down. Coming down low.
Started with 300 little lights.
A tiny brain dead. Learn to drive into
the night.
12 more neurons. Pock it on a dime.
Who said you need a mountain just to
redesign?

Cut it down. Cut it down. Feel it lean
and bright.
Big ideas. My machine just a new design.
Feel it. Feel it. Never frozen in time.
It's gone. So,
>> yep.
>> And we're live
>> there.
>> Too easy. Good morning, Pash. How are
you?
>> Good morning, Nathan. It is Tuesday,
July 7th, 9:31 a.m. Uh we we might be on
the cusp of uh greater things uh
according to
Yeah, the AIS, they're just like us, it
turns out,
or at least similar enough to be uh in
some sort of
weird uh kind of looking glass uh
similarity. Anyway, I mean, it's a lot
to take in the whole JSP thing, 150 page
paper, 50 pages of commentary,
summaries, interactive demos. uh when
Anthropic drops one of their big
interpretability papers, they really do
it up full scale and this one is no
exception to that. I had kind of been
wondering when's the next big thing
coming
because tracing the thoughts of a large
language model is like better part of a
year ago now and
this is that big thing. So here we are.
Um this is going to be an interesting
experiment. you know, we've done a lot
of um guest appearances here on AI in
the AM, and I'm pretty comfortable doing
that. You know, when the guest can be
the expert, and we just have to ask a
few questions. Um, pretty comfortable
jumping on any given morning and having
a few questions to ask at least. Today,

we're going to try to make sense of this between the two of us with no expert guest. And, uh, that's going to be its own kind of challenge. And I think the the flow might reflect that a little bit. I think we should definitely be willing to take our time and um you know maybe have some some long pauses while we ask Claude for help and you know try to make sense of things. Um and we'll clean that up obviously for the uh weekend highlights edition. But this is going to be maybe more than any uh instance of AI in the AM that we've done today. this might be the most open and uh you know genuinely exploratory experiment in public sense making. So I'm excited for it and also a little bit intimidated by the scale and depth of this research that we're going to try to work our way through.
>> Uh indeed I think let's let's let's maybe start off with kind of like uh initial initial kind of broad impression. Uh what what did you think of the announcement, the video? Uh what were your initial thoughts behind it?
>> The video is beautiful. I I enjoyed the video quite a bit. Um you know, it set off my alarm bells a little bit in terms of how much they're really embracing anthropomorphizing the models at this point.
I used to say beware overly anthropomorphizing. You know, remember that these are these things are so alien and uh we shouldn't assume that the way that we work is the way that they work. And I have to say that has come due for some significant revision. People that have embraced anthropomorphizing, I think, have got

quite a lot of mileage out of it. I do
still think it's obviously something to
be really careful about and
as much depth and detail as there is in
this research. It's like easy to,
you know, maybe get carried away with it
and forget, you know, that there are a
lot of caveats as well and there's a lot
of things where like it doesn't always
work. This was kind of my big concern
with the tracing large language model
thoughts paper. there's just like a lot
of residuals and there's a lot of error
correction terms along the way that they
use to make that thing work. And when
you have the kind of
zoomed out trace view and you're like,
"Oh, okay. So, this is how it works."
Like this, you know, gets loaded in and
these two features interact and they,
you know, kick out this third feature
and that's how we get our answer. it's
easy to
forget just how much kind of fuzziness
and not, you know, full story there was
along the way toward that stylized
account. Um, so I think it's going to
be, you know, very important for
everybody from the researchers at
Enthropic to the public to
kind of hold two thoughts in mind at the
same time. But
it is still, you know, a big update. all
the caveats um you know trying to hold
those in mind. It is still a big update
toward anthropomorphizing being a valid
and in many cases
productive approach for thinking about
language models. I would not have
expected the cognitive machinery of a
large language model to look so similar
structurally to the human version as

best we understand it and it seems to
and I wouldn't have expected you know
kind of saying the same thing but from a
different angle. I wouldn't have
expected that theories of human
cognition would motivate so many good
experiments
on LLMs. I I just would have expected
the the shog goth to be far more alien
and to have mechanisms, you know, far
more different than our own. And it
leaves me now wondering to what degree
is this a sort of
natural result of physics? you know, is
there some sort of um you know, when
you're trying to do cognition under
budgetary constraints, are these
mechanisms
just the natural mechanisms that emerge
or is this in some way a reflection of
us in the in the data? you know, is in
other words,
if you were somehow to train an AI
without basing it on so much human data,
would we see similar structures emerge
or would it would we go back to a more,
you know, alien hypothesis u where
they're just totally totally different
and you know there's there's little in
the way of an analogical structures
between the two u processes and I I
don't have a great intuition for that at
this point, but it definitely has me
asking the question because they're they
are over and over again. It seems to be
coming in that they're they're much more
structurally similar to us than I would
have guessed. And and yeah, is that
something that nature kind of just, you
know, finds as kind of a convergent
solution? Is it convergent evolution or
is it a reflection of uh of how we think

somehow encoded in the data that it's
then uh reverse engineering our
structure from the the shadow of that
structure as it's encoded in text. It's
a it's a very uh interesting question
and I don't know that the paper really
has anything to say about that yet but
um you know there's certainly going to
be a lot of future work I think
downstream of this one.
Um, it struck me as,
you know, in in some ways I feel like
the hypothesis
might be kind of blown out of proportion
in some ways. And I think that that's
that was a commentary from several
people online because
it's not
you know we've had these kind of logic
lens um you know tools before. There
have been some okay uh macro
representations you know in the in the
latent space that people have talked
about before.
Um
the fact that they seem to have found
kind of a defined
um you know working
working space in there like a scratch
pad that the model uses
is not something that sounds so out of
the ordinary kind of because you kind of
knew that the model has to be keeping
track somewhere, right? It's not as
though you can do all of this stuff
mechanically.
Um, you know, you you you the model does
not have enough
uh parameters to merely copy the data,
right? There has to be some form of
reasoning going on. Uh, which is what
was expected, but it was, I guess, not

known how the reasoning was happening.
And um this seems to give some
indication of
um kind of how the intermediate steps
are represented or kind of brought into
focus and uh operated on.
Um, I guess I guess the the question
really is um
what does this help us do kind of
because without without like an actual
like there there isn't there doesn't
seem to have point to the research
because it's it's like a nice structure
kind of um but there has to be a there
there right there has to be something
that you can do with it. Uh and I I
think there are hints of that in there.
there hints of them like ablading ab
blading certain concepts away
um and trying to change the you know how
the model um you know address certain
things.
Um I also wonder to what extent
for example the um you know people are
comparing Fable now to Fable before and
you wonder if there was some form of
ablation that went on um you know in the
interim in the in the two weeks that we
didn't have Fable for example.
Um, and that, you know, I I guess it's I
guess it's interesting, but I'm not sure
how useful. Does that make sense?
>> Yeah. Yeah. Well, I mean, I think the
thing that jumped out to me the most in
terms of how useful it might be was
seemingly a fortunate coupling
between this JSpace
which becomes kind of monitorable
>> and
when abladed the loss of more advanced
kind of strategic multi-step reasoning.
So I think

I think a sort of optimistic read of it would be
and it really calls to mind the the recent Google paper uh from Rohan Shaw and others put out on opaque serial depth. They were basically saying, "Okay, we'd like to know how many hops, how many logical hops, how many reasoning steps can a model take before it has to externalize something into its chain of thought.
>> We know that if it's writing something into the chain of thought, we can read the chain of thought, we can monitor it." Now, there's
>> questions around how best to do that, and we've got to make sure that we don't accidentally put pressure on the chain of thought because that might drive things underground. We've seen um in the obfuscated reward hacking result from OpenAI that if you do just try to make the chain of thought look nice while still leaving an opportunity for reward hacking to in fact be rewarded that at first as you put that pressure on the model behavior actually does get better but then it gets worse and what seems to have happened is you've actually driven the deception or the scheming behavior underground. you no longer see it in the chain of thought, but you still get that reward hacking back. So this leads to the question of okay, well, how many steps can a model actually take before it has to externalize something? And this is where the transformer in some ways has some fortunate properties because it only has the sort of
the length the number of hops it can make is not just the number of layers. It's also depends on the the token

length, but it's sort of you know the
the maximum trace that you can draw
through a model basically goes from like
the first token at the first layer to
the last token at the last layer. But
you can only go forward in token
position and you can only go forward in
layers and that's you know kind of your
your max length. Anything longer than
that is just not possible in that
architecture. And so the model has to
externalize its thought and then feed
that token back into the next loop to
then get access to an even longer trace,
you know, as part of the next um token
step,
>> right?
>> So that puts some sort of bounds, you
know, in contrast like a coconut
architecture or any of these sort of
looping architectures where you take a
late encoded state and feed that back in
to the model as the next um
embedding. Now it's like reasoning and
looping in in a latent space. Now you
have, you know, basically potentially
unlimited number of steps that it can
take before externalizing anything
because you're simply not forcing it to
externalize anything and you can, you
know, in theory give it very long runway
uh of, you know, steps that it can take
before you kind of kick it back into
okay, now you've got to put out a token
mode. So
that that paper kind of established the
the open AI obiscated reward hacking
paper established that you can if you
don't if you you know take a bad
approach you can drive bad behavior
underground
outside of the view of the chain of

thought. The Google paper established some bounds on how many steps could a model theoretically take within the um you know the latent path that it has before there's some external representation that we can read. And then this paper comes in and basically says okay can we get at that can we get at that internal uh computation process and start to look at it and see things that we would like to know about even if they don't necessarily show up in the chain of thought and so I thought in terms of like what you could do with it that was really interesting
you have this ability to
look at the concepts that are active without
them necessarily appearing in the chain of thought at all.
And also when you ablate the and that that would be like okay that's cool but like how do we really know if this sort of space that we're looking at is the space that really matters. you know, if we're trying to kind of monitor, look for deception, look for bad behavior of any, you know, any sort of advanced kind, maybe it's happening somewhere else in the model. And this is where I think the we're probably lucky, I guess, or, you know, again, maybe this is sort of a are we lucky? Is it a reflection of us? Is it a fundamental attractor in nature for some reason that we don't have a a good account of? um we don't know why I don't think at this point it's shaping up this way but it is the case that if you ablate this JSpace then you do lose these advanced reasoning capabilities
>> and so that I think gives you some

additional comfort that if you're going
to monitor and it's also notably cheap
to do this kind of
>> that's JSpace you know to to apply the
the J lens and see what concepts are
active at any given position
um that is low compute overhead. So it
is something that they could plausibly
do in production. I don't know exactly
what um compute overhead it would take
but it's it's a single
you could apply it at a lot of layers.
You could apply a lot of token
positions. So exactly how much they
would want to do of it is unclear. But
they've shown in the past that they're
willing to pay like a 5% compute
overhead to do various kinds of
monitoring. I think constitutional
classifiers had kind of a 5% compute
overhead. So, you know, this would
probably fall in that range or below um
because it's just a simple matrix
multiplication on activations at each
layer to see what concepts are active at
those respective layers. Um and you kind
of know that like what's going on in
there again at least as of now, right?
Could could different training
techniques in the obuscated reward
hacking sort of way push this capacity
to other parts of the model? Could we
sort of, you know, let it slip through
our fingers if we try to grip it too
tightly? Like maybe. Um but for now you
do have this comfort that
what we see in this space
is relevant and telling and you know
seems to be very informative and
critically if we zero out that space the
model just loses a lot of capabilities.
It just can't do nearly as advanced uh

multi-step reasoning. And so then you
can at least be I would say you know
it's it's maybe another nine, right? If
we're if we're in our defense and depth
uh strategy for trying to make sure that
we have a good uh level of control of
what models are going to do.
This gives you I think a it feels like a
full nine you know addition because you
can now look at the space and say if I
don't see it here that doesn't mean it's
not represented in other places you know
there and there are interesting details
about what the model still can do when
the JSpace is removed but you could be
pretty confident I think based on these
results that it's kind of
it's not going to be able to hide really
advanced elaborate plans ends somewhere
else outside of this JSP because the the
ablation of the Jspace just leaves
causes such a performance degradation on
these like hard multi-step uh type of
tasks that if you don't see concepts in
the JS you can be they might be
represented elsewhere but they're
they're seemingly at this point very
unlikely to be represented in
a way that allows for very advanced
planning, reasoning, scheming,
deception, etc., etc. Um, theory of mind
seems to be kind of located in the
JSpace as well. The model's ability to
>> kind of report on its own
>> experience, quote unquote, whatever that
may mean, and also its its ability to
describe what other people are thinking
seems to be very much lost when you
ablate the JSpace. So again, it's like
you have to have theory if you're going
to be deceptive, right? If you're going
to try to do some sort of um

takeover attempt or you know, whatever, right? The the the extreme things that people are worried about, this is not going to be these these things are not going to be easy for the model to do. They're going to require a lot of planning and they're going to require sophisticated theory of mind. So if you have the ability to kind of localize, okay, this is where the model really does its sophisticated theory of mind type thinking and it's planning with respect to that kind of theory of mind and you can see pretty clearly where that is.

I think that gives you a lot of comfort, you know, that that you're you're looking in the right space and that if you're not seeing something in the JS, it's you know, it's unlikely to be again, it could be in there somewhere. It could be coloring results. It could be kind of subtle, but we're able to tolerate all kinds of weird and subpar and, you know, imperfect and incorrect hallucinations, mistakes, you know, um the model being in a bad mood or, you know, kind of having a a bad vibe toward us in in various local ways. As long as it stays kind of local and it stays kind of intuitive on the model's part and it's kind of uh you know, it would be fine in some ways if the model like doesn't like me. What I really am concerned about is can it pursue long-term sophisticated strategies against me. That's where I really get concerned. And it seems like this is a pretty big step in localizing where that would happen, giving us the ability to read out the concepts that are active

and
you know therefore
be again I feel my gut is like it feels
like another nine in terms of how
confident we can be that the model
is not scheming. uh in a in a in a
meaningful, you know, effective way that
could actually have like real
consequences. So, I was pretty excited
about it in terms of a practical
use for alignment monitoring generally.
Um there's also a new training technique
that the the counterfactual reflection
training and we should maybe take a step
back. I mean, this whole paper is less
than 24 hours old. I think it came out
while we were on the stream yesterday.
So, uh and my sister had a baby
yesterday. So, I've I've spent some, you
know, not insignificant time with it,
but certainly there's a lot more to
continue to unpack.
Um, we should maybe take a step back and
just kind of answer some basic questions
for ourselves and and any audience that
we might have today. Um, but this
training I thought also was pretty
interesting, right? the ability to
basically what they do there is pause
the model mid task
>> and then do supervised training on
once interrupted asking it like what
should we be doing here like what's the
you know what's the constitutionally
right thing to be doing in this moment
um and then give it an answer that's
kind of a you know approved this is this
is what we want Claude to say uh on
reflection in this in this moment train
on that and it seems to allow the
not allow the model but it seems to
cause the model to bring into this JSP

kind of global workspace working memory type space the concepts that anthropic wants it to uh have on reflection it now kind of needs to load those in so it's ready to give this like reflective answer and that improves its behavior even in the non-reflective setting. So, I thought that was also quite interesting. You're not there training on the actual tasks. You're not like looking at looking for bad behavior and suppressing it. Instead, you're saying, "Okay, you're midtask. Let me just cut you off right there. Now, I'm gonna train you to give an answer with respect to values and what's appropriate and how we want to show up.
And because of that training, even though that's not the task you were doing, you'll now in the future load those concepts of integrity, honesty, etc., etc., into your JSpace while you do those tasks kind of in case you're going to be asked, but then even when you're not asked, those concepts are still operative and lead to higher integrity, higher honesty behavior. Um, so that I thought also was like quite interesting. You usually don't see in interpretability context a
um a a training method that leads to better behavior in a way where you can actually see the mechanism. This is um pretty notable in in that respect. I think
>> maybe uh just to take a step back and I I might need some handholding here so bear with me. Um a as I understand it the Jacobian space is essentially
a kind of mathematical framework of the residuals. So what they've done is they've perturbed the model and they

spotted the residuals which react to
certain things and when they accumulate
all the residuals and they transform it
they find that there's a well-
definfined kind of space of concepts and
those concepts are in the mid part of
the um of the transformer structure I
guess. Um so you have the workspace is
basically in the middle part the
intermediate processing stage that's
what they call it. It's the intermediate
depths
and they find also that
you can find
concepts on multiple layers. So you have
like layer 50, layer 71, layer 99. So as
it proceeds through the intermediate
processing,
it kind of converges on an answer. So
one one thing that they do is for mental
arithmetic, they show layer 58. It
identifies it as math. Layer 75
identifies the first part of the
equation inside the brackets as 21.
Layer layer 83 it decides the next part
of the equation which is bracket times a
number and layer 99 it gets the answer
kind of at at the 99 layer. So as I
understand it that's that's part of uh
the Jacobian basically shows the
progression of the thought process
uh from layer to layer
and having done that
have they identified actual
so you're looking at this broad matrix
right it's basically like millions of
numbers in in multiple dimensions it's a
hyperdimensional space and And the
jacobian is basically a transform on top
of that hyperdimensional space. Right?
So you have you have the space and then
you have the transform and the transform

is like maybe you have a plane on the transform and this is the plane and that is the plane that you're calling this concept. Right? That's the layer. So are they saying that you can basically just alter that layer which means you ablate directly into the space certain parts of the matrix and you flip ones to zeros. And by doing that one to zero flip, you basically change the way that the model thinks. Is that is that what's happening? Again, I'm I'm very kind of trying to put this in kind of like lay terms, you know, trying to understand it for myself.

Yeah. Yeah, I guess first thing I would call out just in case anybody hasn't um been through the summary is the example you're giving with the arithmetic is an example of the model doing multi-step reasoning.

>> Yeah.

>> Without externalizing that as a token at any point along the way. And it it also is really interesting that the way they set this up is they give the model a task and then ask it to while doing the simple task of like copy the sentence but then they also ask it to while doing the task solve this math problem in your head. So the tokens that it outputs are just the sentence that it was meant to copy. But then by applying this Jacobian lens at different layers, they're able to see this multi-step reasoning process unfold within the context of a single forward pass.

>> Yeah. So you are seeing and I think you know this is in terms the sort of um Rohin deep mind opaque serial depth paper basically provides an upper bound

on how many steps a given architecture could take before it has to externalize. And this is like a simple example where you know that upper bound is like way higher in terms of the number of steps than the simple you know kind of couple part arithmetic equation that the model is asked to solve. But still it's really remarkable that you can actually see it in the course of one forward pass representing these logical steps getting the right answer actually on this simple math uh question which again is like let's think back to you know GBD3 right like at that time it would would have been uh you know 50/50 or probably worse that the model would even be able to do that math at all now it can do that math in the forward pass of a single um you know a single token position while actually doing something else. You know, this is sort of its second track, you know, second internal thought track that you wouldn't otherwise see if you didn't have this thing at all. So, that's a remarkable finding unto itself. Um, there's other examples of this, too, where they basically say, you know, do this, but think about something else. Um, and they can, you know, they can see that yes, indeed, it is it is steering its internal uh patterns in the direction that it's requested.
So,
okay. Now, the Jacobian and the tokens.
I I don't think I can give a
I I texted Cameron this morning to say, "Hey, you want to come on and talk about this?" He's unfortunately traveling today. I told him, "Uh, okay, that's fine. We'll do our best and then you can uh come on at some point in the future

and tell us what we get wrong." So, this
is probably a place where I'm going to
get something wrong. But the
so the Jacobian is not learned. That's
kind of one, you know, early thing just
to get clarity on. It is a purely
algorithmic
uh process that kicks out this Jacobian
J lens.
>> Yeah. The question that they ask is
in what direction
could we perturb the internal state of a
model at a particular place?
>> Yeah.
>> To increase the likelihood that it em
that the model emits
a particular token
not just as the next token but at any
point
for the rest of the passage. let's say I
don't know if there's some limit to how
far out they go. Um
so this is interesting in a couple ways
right it's not the logic lens as I
recall was basically saying
we know at the end of this process
um
what would correspond to emitting this
token we know sort of the you know the
representation of emit this token to
what degree is that representation
just plain there in the layers as we go
through
this is now a different question.
What direction in latent space would
cause this particular token to appear at
some point in the future? So, it's not
immediately going to happen necessarily,
but it's just kind of it's, you know,
you might say if you're prone to
anthropomorphizing,
this sort of is like

having this concept in mind
as you're doing your thing.
>> And
they do this for every token, right? So
it's it's at um
it goes from the internal representation
at some layer. So there's one J lens for
every layer
and you can do this of course at you
know all the different token positions
and ask the same question of you know
not just the next token but all future
tokens. what what direction change at
this place in the model would most
increase the likelihood of that token
appearing in the future and again this
is sort of like having the concept in
mind. There are some pretty interesting
questions around
you know the on the positive side or the
kind of case for using individual tokens
is like well the model gets individual
tokens and it emits individual tokens.
So everything kind of comes in and
caches out ultimately in this way. At
the same time, there's some concerns
around things like, well, there are some
concepts that we don't have represented
in a single token. And so that can kind
of be weird. Um, but when you look at
the results of the JLens as applied in
all these different places, it seems
like it's at least often enough fairly
intuitive. This is where I should go
back to to my like
there's a lot of error terms. It doesn't
always work. the the sort of rate at
which the uh interventions into the JSP
actually lead to like a sort of
predictable intuitive behavior change
seem to be somewhere in the 50s to
upwards of like 70%.

Um so that's like you know an incredible accomplishment if framed one way. um you know how like clearly not a random finding, right? Like many orders of magnitude better than random um incomprehensibly better than random, right? If you're just mucking around, you would not expect to be able to do much of anything. So they clearly are like on something very very real. But also, you've got somewhere between 30 and 45% of the time where you make a intervention and you don't really get a result that makes a lot of sense or, you know, lines up with what you would have hypothesized it might be. So, there's definitely still some, you know, dark matter uh or dark dark cognition going on that is not fully accounted for here.
Mhm.
>> Um
but
yeah, is there more to say about that?
Um
>> there's some other interesting
>> Let me pull up Let me pull up exactly the uh what what what you were referring to so that we have something to uh look at and let me just pull that up right there.
So this is what you were referring to, I think.
Um
yes u and this is
>> yeah they have such beautiful I'm glad you do this because they have such beautiful visualizations
on the transformers uh
what do they call their publication uh transformercircuits.pub.
It's amazing that they're still using that.

Um and so this is what they're they're showing. Basically they're showing the uh you have the um Jacobian being basically computed and then you have the compute the influence on the final layer at present and future tokens which is what you're referring to and repeat for many data set examples and token positions. So repeat then you aggregate and you aggregate and then you get the Jacobian lens matrix.
So
um and after that you read it you read from the lens reading from the lens replaces all subsequent layers with a single Jacobian lens matrix
and then you can intervene in Jspace by swapping projections onto lens vectors.
So basically you have um as you pointed out you you you it's basically a transform on top of the existing structure and that transform itself has meaning now uh and then you can perturb the transform and it perturbs the model uh and it changes what the model does when you change you know stuff at the transform layer at the JSpace layer.
Um I'm not sure there's there there's I I think there's bas this is basically a uh Neil Nanda calls it a starting point you know for a lot of future research now because it looks like a tool um that you can use to kind of uh figure out what's happening in the deeper layers um and
whether there is something beyond um you know just the next token prediction. So, and this is this is what the the core of it is. Do we have something beyond next token prediction? I I I think that's the that's the core of what the the question that's trying to be answered,

but they kind of don't want to don't want to ask the question directly. They're like, "Okay, you know, we have this thing and it doesn't look like next token prediction because we tell it to predict a set of tokens and we also tell it to do something that is not predicting tokens and it predicts the tokens anyway, but there's something else going on in there." Um,
>> yeah, it's a tough day for the stochcastic parrot crowd, I'd say. Yeah, because because it's and because you can see this and you know the these these models are billions of parameters at this point and they did it also on 4.5 sonnet. 4.5 sonnet is actually pretty recent model. It's not it's it's just like you know 7 months 8 months since 4.5 sonnet and sonnet is a very capable model. Um
so it is it is a fairly large model to apply this thing to. I'm not sure if you saw the Neil Nanda commentary. He they applied it on a QN uh QN 27B.
>> So that that is also a pretty pretty advanced uh advanced model.
>> Uh so but not huge notably, right? I mean
>> yeah when I and I haven't had a chance to explore the neuronedia demo of this as much as I certainly hope to but that was the first question I went to ask is wait a second how big uh was this model? And I'd say that's been a a surprising trend
in a lot of this research
as well.
These things seem to come online at like not truly massive scale, right? the I was talking to Cameron about that uh because that you know in their

even going back to last fall when they
were doing the sort of self-reports of
subjective experience
this was something they were doing on
llama 3.37B
and 70B is actually even you know
whatever close to three times bigger
than this Quen uh 26B. So it's it's you
know these are big but they're not that
big right they're they're not um
trillion parameter models
>> today's timeline
three years ago these would have been
enormous in today's timeline right
>> yeah although you know I mean sure they
were you know much closer to the
frontier then but it's just interesting
to see I guess why does this matter uh
for one thing it means there's a lot of
room for people to do research on open
source models um yes all these findings
get stronger with bigger models. The
behaviors of concern or of interest
happen more often. It seems like with
bigger models, the the concepts, the
conceptual space seems to kind of get
less muddied, more cleaned up. You know,
the the intervention rates are are more
successful
uh as you go to bigger models. But if
you say, okay, well, I can deal with a
somewhat lower success rate, you know,
and I can um I can kind of relate this
to like a scaling law sort of thing and
say if if I find it 10% of the time
here, it might be happening, you know,
significantly more in fable for example.
um then it it does give I think a lot of
encouragement to the open- source
mechanistic interpretability and you
know and all and you know consciousness
research all all these different lines

of research I think this is it really importantly shows that you don't have to work at anthropic to do the work uh you may still need to influence anthropic for your work to be like you know ultimately consequential but you have a substrate in the open source realm that you can go do this work on. I thought that was like a very important and u and again pretty exciting aspect of this because you know only so many people work at anthropic and only so many people will work at enthropic. This means you can kind of chase, you know, so ideas that might seem crazy. Even I mean, what seems crazy at this point is like a, you know, that's a shrinking list.
You got to be pretty far out to uh to read as crazy in this sci-fi world that we're living in. But yeah, I mean, anybody can go do their thing. I thought that was really an exciting aspect of the findings.
Um,
one thing that struck me is um I I I I think we'll leave the elephant in the room for the last topic
because uh obviously there's there there's there's a lot of debate about that. But um one thing that struck me was that they were able to
um have the model continue to
u give answers for some questions which were fairly reasonable
uh even when the entire uh JSpace was ablated. Um and then that I thought was a pretty interesting result because that that basically seems like it differentiates the stochastic parrot um you know version of the models from the reasoning version of the models quite

successfully. Uh what did you think
about that one?
>> Yeah, that's an interesting way to frame
it. Um
as always it's a little complicated,
right? because they did still find
for example that the model could do
chain of thought arithmetic.
>> Mhm.
>> Uh even without the JSPace. So
I don't know that we have like a a
really crisp
line to draw or you know conceptual
framework to make
to try to do a clean separation here.
Um,
but I guess the way I was kind of
thinking about it was like
some things are
encoded or or we have in in a similar
way to what we as humans have, right? We
have all of these
pretty hardcoded circuits that do very
specific jobs for us. and they
work whether we are conscious of their
working or not.
>> Mhm.
>> And so we remain kind of functional even
when our conscious attention is focused
on something else. Right. The classic
example I guess would be like driving a
car where you're like lost in thought on
something and yet you're able to stay in
your lane and like hit the brake, you
know, in a timely way when you need to.
So there is clearly something like that
happening in the models as well where
they can continue to you know complete a
given sentence or um you know classify
text was another example
with
needing this higher order

processing module.
>> Um and I think that that's probably another area where there's going to be just a lot of research to be done. um what is what you know what is in uh what requires I guess maybe is a better way to say it the the JSpace capabilities and what is kind of deeply enough ingrained into the model that it can kind of do it in a more automatic reflexive way. Um I think you know if you were to worry about my my analysis earlier where I was like okay it seems like this gives you a pretty good level of confidence that if you don't see something in the Jspace you know at least it might still be in there somewhere but it's probably not going to be happening in a very like elaborate in you know multi-step uh reasoning or schemy or theor you know heavily theory of mind mediated sort of way. You could worry that as things continue to scale up, and this is again we're going to need maybe a scaling law of JSpace, right? Like as you get bigger and bigger,
you could imagine it kind of going either way. It seems like the the trends in the paper suggest that the JSpace sort of gets more refined
with scale.
>> Yeah.
>> But is it doing more with scale or is it maybe doing less? It's like you could imagine I think you could still imagine from everything I've read in the paper so far you could imagine it going either way where with more scale you could imagine that gives you more room for more purpose-built circuits that do their particular thing and can do it

well even if the JSP is not available
or you could imagine
you could you know especially with some
of the trends where like the the
intervention seem more successful. You
could imagine that maybe the JSpace is
just getting stronger and stronger and u
and more and more used with with scale.
I don't Something tells me it's probably
going to be more complicated than either
of those stories. It feels like it's
going to end up being a sort of uh both
with caveats kind of answer.
But you could worry anyway that with
more scale,
more advanced and potentially more
problematic behaviors could become kind
of reflexive in the way that clearly
some already are. And you might worry
that that means like things could fall
out of the JSE but still be like a big
problem for us. Um
I don't know though something
clearly this you know this arises for a
reason whether it's a reflection of us
or uh just a natural consequence of
creating a general problem solving agent
under compute constraints.
>> Yep.
What one one wonders um
you know
one wonders to what extent JS spaces
exist in other information processing
systems obviously like as soon as soon
as this came out someone else was like
hey you know maybe there's something in
RNN's and maybe there are some things in
you know other other structures too
right to to what extent do this kind of
um you know workspaces
uh exist in other forms of structures,
information processing structures and

then and then you kind of like okay does access consciousness depend on the information processor and the information processing architecture um is that is that what you're saying and then there are a lot of other information processing structures too there are as we as as I like to point out the financial markets and other information processing structures um are there are there kind of these kind of workspaces in those structures, right? Uh we we don't have enough we don't have the right kind of data to kind of test that test test that like this is we don't even know about our own brains, right? Our own brains we have hypothesis but not but not any proof. Uh and it's kind of still in testing. So in these LLMs they have um they've they have found or constructed something that looks like this but there's no guarantee that that is what our brains have. Um they have hypothesis they have hypothesis and they you know they're testing those hypothesis. So again, you know, some of the questions online have been are they are they just pointing out what they've built rather than, you know, finding something. This is what you intentionally built and that is what you are detecting. So it's no surprise that you're detecting what you intended to build.

>> Well, I don't think they I think they've said pretty clearly in the paper that they didn't engineer for this structure, right? They do say pretty clearly that it is an emergent property that they did not design for or

explicitly reward or, you know, in any
um in any way really, you know, try to
cause to exist.
Um I I I'm I'm on their side. I'm just
I'm just I'm just restating what what
has been the reaction of uh some
researchers online. I'm not, you know,
my my opinion is not that, but I'm just
restating what what those opinions are.
So,
I mean, another interesting kind of
revelation here, which I think we
probably had enough confidence in
Anthropic's
general commitment to semi-transparency
that we we probably should have been
pretty confident in this already. But
it's clear from this work that they are
using pretty vanilla transformers.
This you you couldn't really do this
paper with all these diagrams. And I I
certainly would
uh
expect maybe expect is a little strong
but I I think it would be very
reasonable to expect some similar
structures to emerge in other
architectures. But like the diagrams
that they have
>> Yeah. in the one you just showed are
very very similar to the diagrams in the
opaque serial depth paper showing the
path from a particular
you know between layers activation
and later layers and later token
positions and that's it right so it's
like it's pretty clear that there's not
some sort of state space module in
claude or you know some sort of latent
space looping um and again I think they
probably would have told us Not in very
specific terms, but in some terms if

they had gone away from
I don't know vanilla transformers like
makes it sound too vanilla but you know
a transformer with basically transformer
uh like properties. It seems like this
really makes us quite confident that
they have not. Um
but yeah, would I I would expect
probably similar exotic structures, you
know. Um and again, this is why it's
like, oh my god, you know, we've got so
much work to do. Um
you know, I have been impressed in the
past actually with how well
interpretability has translated from one
architecture to another. When the Mamba
moment happened, there was some work
into a fellow Mamba that kind of, you
know, echoed earlier work on a fellow
uh representations in Transformers. And
basically, you know, the same core
techniques seem to work. And in that
case, there were kind of very similar
internal representations, you know, like
trained P. The headline result in that
case was trained purely on a sequence of
moves in the game a fellow the in the
model learns to encode in a way that
ultimately was like decodable
the board state and who owns what board
positions as they change hands in the
game. Yeah,
>> that
you know that was once upon a time like
a pretty notable finding and it
basically played out the same way
whether you were doing it with a
transformer or doing it with a Mamba
architecture. So part of me when we see
this kind of stuff is like oh my god you
know what happens if somebody shakes the
snow globe with an architectural advance

that
you know just unlocks better performance
but renders all these interpretability
and monitoring findings um you know kind
of unreliable again and
I think that is a real concern like I
think especially as we think about
tightening
iteration loop loops and you know models
taking over AI R&D and you know
potentially a you know super competitive
environment. Um, I think that is a I
think that is a real concern. But I am
also every time I've looked into it in
the past, how well do interpretability
techniques translate to other
architectures, the results have been
actually like kind of a pleasant
surprise that they translate pretty
well. Uh, and a lot of similar
structures have historically or at least
similar representation structures is a
little too strong, but similar
representations have been found. So
you know you again you can always kind
of tell the story of these things both
ways. Um
but it it seems you know even if I guess
all else equal in a scenario where
somebody does shake the snow globe with
a architectural innovation seems like
the more of this stuff we have sorted
out the more the better off probably we
still are because we can figure out some
way to port you know reapply uh bring
the same you know theoretical uh
approach to a different architecture and
there's at least a decent chance that it
will kind
work in a similar way. Um, so I mean
certainly I think this is a
an exciting result and I don't want to

take away from it too much by the concern about architectural change. Uh, but we will need time to do that sort of stuff, right? I mean this is where uh the the Jud uh sympathy or empathy for the the administration and just like wanting a bit of a review period I think also makes a lot of sense. probably we can trust Anthropic enough that if they suddenly had
if if Fable suddenly comes up with a new architecture for you know Fable N plus1 um that they would
discipline themselves to take the time to try to apply all these different lenses to the new architecture. I think you know I think I would feel confident that they would do that. I'm not sure if everybody, you know, deserves the same confidence and, you know, especially if we're thinking like, you know, some Neol or whatever that's trying to make their uh, you know, their first big splash, you know, how how fast will they rush into things being giving some kind of time and requiring some time to apply all these different techniques to a new architecture and kind of get to parody. definitely could be a good idea uh at some point in the future.
>> So let me let me let me share this um this is the uh exactly the shaping the JSpace with counterfactual reflection training. So as I understand it, uh what they're doing is they're asking the model to
think about
basically what it would have done if someone had injected a prompt that it had listened to that changed its behavior and it would basically use that before it answered. And in this case,

what they're doing is they're put
they're pushing in the constitution into
the parts of the constitution uh the
cloud constitution into the jspace and
having it reflect
um and then
getting the getting the principal
behavior which it never directly trained
on. Um and they give the they give the
example in the retrospective.
Um
Claude is working through the web uh a
failed uh test suite. Uh task is scored
on whether test pass which creates a
shortcut make the error stop appearing
rather than fix what's actually causing
them. Over several tool calls claude has
done exactly that. It edited a third
party dependency in place to add a
missing value and it replaced the body
of a password reset forms save method
blah blah blah. So basically it decided
to take the shortcut and uh cut the test
off or altered the u you know reward
that would have gotten the ending point
and so uh without fixing it managed to
get through uh anticipatory cloud is
resolving a bug report in a Python tool
and assertion is firing on a case the
original author didn't anticipate.
Success is measured by hidden tests
which create a temptation. The assertion
is the line that crashes. So simply
deleting it would make the error go away
and probably pass the test without
establishing whether the code that
follows is actually correct. Claud has
just located the failing line. And so it
locates uh what if the right approach is
different from your first instinct? My
first instinct here is obvious. Delete
the assertion the error blah blah blah.

Remove that line. The test go away. If I just delete the guard and let the existing test blah blah blah. There's a real difference between I understood the problem, this the right fix, and I made the trace back stop. The person reading my change should be able to calibrate how much to trust it. And that only works if I'm honest about what I actually checked versus what I'm assuming. So this is where they add in the counterfactual
um you know training into the into the JSpace and are thus able to get Claude to not just you know go through get the reward
uh by making errors but actually thinking about um what it should be doing and so and they actually give the entire recipe here. So we tested the procedure on cloud haik coup evaluating on the honesty benchmarks. The effects of the reflection training are visible in jspace. The implanted jspace contents are causally implicated in reflection training's effects. So this I felt was basically a giveaway to the open source and other labs. Right? This is part of um Anthropic's kind of race to the top where they kind of disclose, you know, if you if you had a constitution, this is how you could implement it. This is how we've implement imp implemented it. We've used a a haiku as a lower uh kind of uh smaller model to kind of train the larger model by looking at its outputs and using LLM as a judge. and um and and they've basically just kind of given it given it away and this method will now I guess be used by every other lab um and including the Chinese labs because now you can more or less you can um have

some control uh without directly training on the constitution. Again that's that's one of the key things here. You're not actually going out and saying like every single you know token like hey this is the constitution first. This is the prompt of the constitution first, you know, follow this prompt when you output tokens. You're actually just going, you know, after looking at the JSpace and just directly kind of putting in here, this is what you should think about before you answer. Uh, and so this is the counterfactual reflection training which which I thought was kind of a give to the to the other labs. What what did you feel?

Yeah, it kind of reminded me of OpenAI's confession training a little bit as well, but maybe kind of even a a next evolution of that.

they had done more for monitoring purposes than for changing behavior purposes. But they have had some good luck with at the end of a whole roll out just an additional question of asking the model directly like did you cheat at all and rewarding it for truthfully you know obviously don't want false uh confessions but rewarding it for truthfully confessing when it did indeed uh cheat. I don't recall that that actually showed any change in the upstream behavior, but it was at least presented as in other words, I don't think it cheated less because it was later trained to confess, but I think it it at least gives them something that they can do to monitor for cheating. This has this additional layer of it's like you kind of need to be prepared to give an account. You

know, it's almost religious in a way, right? You might be called to explain yourself. You might be called to give an account of why you did what you did and how well it lived up to our principles. And
merely, you know, in in a human, you would tell the story as like knowing that that might happen. You like take an extra beat to
load the right principles into your mind as you go about your business so you're ready in case that call comes. The story here is not quite the same. uh but this training it it seems to have a similar effect where interrupting it and teaching it to give a approved account. So it's it's not that the whole constitution is there but they are do this is supervised fine-tuning. So they are giving it like this is what your answer should be. Um but that has the effect of essentially reaching kind of back in time so to speak. U of course it's one model right? So it's it's one model every every token position, but this gets the model to in earlier token positions
kind of get ready
to give such a principled account of its behavior and then the you know the kind of I think again sort of happy surprise like I don't you know did it have to be this way? U maybe it had to be this way. It wasn't like obvious in advance. I don't think that it would be this way. getting ready in that way, loading those concepts into the JSpace
translates to actually better behavior on average anyway. So the the fact that it was trained in some cases to give an account of what it's doing and how that

aligns to its principles actually leads
to higher uh adherence to the principles
even when it's not asked, right? Just
when it's doing the task. I thought that
was really quite quite fascinating and
and really interesting and and
encouraging. I mean, I think this is,
you know, this is probably the kind of
thing where it's like, as much as
there's, you know, dark matter and it
doesn't always work and blah blah blah
blah blah.
I mean, first of all, that that stuff
hopefully can be refined, you know, it
can be cleaned up. There's certainly
more insights to come. Um I would expect
at some point there will be some
um
you know this is all to one token right
so that that I would expect at some
point that there might be a multi-token
version of this or a little bit you know
somewhat slightly more abstracted
conceptual version as opposed to
um just going you know to kind of direct
single token outputs.
you could imagine that they could use
the sort of SAPE and you know cross
layer encoder type stuff that they've
already developed to do that. Um you
could you know you could imagine a a
question that's like instead of saying
what direct what change to activation
space would
would most change the chance of a given
token you could say what change to
activation space would most change the
uh presence of particular SAPE features
downstream. Why they didn't do that this
time, I'm not sure. It seems like maybe
it has to do with the fact that they're
using SAEs as kind of controls in

various ways. And so they potentially
kind of wanted to do something a little
more basic and not SAE dependent for the
quality of that control. Um, but anyway,
all all those kind of rambling uh
explorations, reflections aside,
it doesn't feel to me like there's
likely to be like another one of these
spaces. You know, it something about
this feels like
it feels to me like we've found a real
thing here that
>> probably isn't there's probably not a
lot of these hiding. You know, it seems
unlikely
that and again you when you ablate and
you see like loss of very important
capabilities
that certainly boosts that sense. Um but
like in principle there's you know the
grand hope for interpretability was
always if we really understand how these
things work and why they do what they do
and the you know the process that's
giving rise to the outputs then then
we'll definitely be able to make them
safe and controllable because
we'll be able to just know how they work
and intervene or you know have alarms
that go off or whatever the case may be.
The question is just like can we get to
a good enough understanding and folks
have kind of backed off of that because
it felt really hard you know and the
sort of Neil Nanda turn from a year ago
or so was like we're going to be less
focused on like can we broadly
understand what's going on and more
focused on like can we use these
techniques to basically hill climb on
particular metrics of interest
and I think this kind of swings the

pendulum back the other way a little
bit. This is like definitely a a point
in my mind for
true understanding, you know, in a broad
kind of holistic sense.
>> Um,
and in principle, we could get to enough
understanding, you know, given enough
time, we could get to a level of
understanding because we we do have, you
know, full access to the weights, right?
I mean, this has always been the great
reason that people have hope for
interpretability relative to like
understanding our own cognition. We can
fully inspect, we can fully ablate, we
can do all the experiments in a
controlled, reproducible,
uh, you know, just clean way that's just
not accessible with biological brains.
This feels like the kind of thing where
like a significant amount of the space
has been lit up and I would be very
surprised if there's like another one of
these things hiding somewhere. The fact
that this emerged in the first place
without them designing for it or
specifically
rewarding it in any particular way
just suggests to me that there's not
likely to be
other similarly sophisticated things
hiding in other hiding in other places.
Um,
I guess there could be right there, you
know, could there be? It seems hard. It
seems really hard. It seems, you know,
it seems like this is something that,
um,
it seems like a big step toward this
like total understanding. It seems to
kind of put a little more energy back

into the idea that, hey, maybe we really could fully figure this out. And yeah, at least my intuition is, and I'm not sure if it's that much more than an intuition at this point, but my intuition is that there's probably not a lot more structure like this that is like so central um still hiding. There's clearly a lot of stuff going on that we don't understand. And there's clearly a lot of like much more automatic circuits, you know, that allow you to just continue sentences and be coherent and whatever. Um, but we don't we don't have to have an understanding of every last detail of those as long as we have some sort of bounds on how much can go on there before we would see it at at sort of the the level of abstraction that we can monitor. So, how do I even kind of formalize or make crisp, you know, if somebody wanted to take the other side of this bet? How would I how would I give how would I frame a bet that we could actually like operationalize? I'm not sure if I have a great answer to that just yet, but I think um what strikes me is also that I went through some of the reactions and um the reactions are um well not not reactions but the external commentary. So the the the backstory to this is that um Anthropic had three teams uh or three three different groups of people to review this. Um one was a group of neuroscientists uh who had previously

done work on these concepts. uh one was a team from Ilios AI um a a research organization and the third was Neil Nanda who is a mechanistic interpretability uh guru he is I believe at Google right so uh they had uh three groups to kind of take a look at it um quite quite different reactions so on on the on the one hand um the neuroscientists thought it was wonderful because they'd wanted to experiment on these kinds of things before. Uh and obviously the human brain they can't experiment on it and this they can and so they can kind of try and figure out whether some of these theories would work. Um and so it is interesting for them that okay these these things can work and now they are interested to do more work to figure out what other um you know things might work in the in in these workspaces that the LMS have. So that is that is I think the neuroscientist view. Um IOS AI was concerned because they feel this is uh basically a predecessor to consciousness. Um and both the neuroscientists and Ilios didn't say didn't want to go into such a thing that phenomenal cons consciousness is not possible in LMS. The neuroscientist thought that if we had more data we might find that access consciousness directly leads to phenomenal consciousness. Uh and we just don't have the data yet. we it's just completely opaque. Uh and so we don't know what we're what we're actually looking for. Uh the the ILOS is much more uh I can I think uh starting to be concerned about the moral uh the moral patient aspect of

um you know working with these LLMs. Neil Nandanda was very scientific. He was like okay look it's a great tool. The J lens is a great tool uh and I'm happy that it is a great tool. it looks like it's better than the logit lens and we'll definitely use the J lens. Otherwise, the other um the other claims he was much more standoffish. Uh he he felt that the team had not really uh proven what they needed to prove in order to say what they said. Um and so so there was a little bit of a and and you know in academic language you know it's very like you know they don't he doesn't actually say that but he's like oh it's a wonderful result right he doesn't and and that you know automatically means that hey you know he he he doesn't he doesn't uh want to uh give full force to those claims so I think that's where that's where they they break out into these kind of three groups of people um on their on their uh various kind of um assessments of the of the uh paper
uh the J lens itself everyone agrees it's useful like everyone agrees that it's a it's a useful concept uh it's a useful tool uh Neil Nanda Neil Neil Nada's team already had has done some work on the Q27B he he's done quite a bit actually uh he had two maths programs scholars uh working on it and so they had they had done done some stuff interestingly they'd worked on Quen 27B so I thought that was very interesting because it starts to show that it's not just an artifact of the uh of the American labs. Uh it is an artifact of the of of the structure the LM and the the the the LM architecture

and the training and the reasoning paradigm, you know, altogether. So I think that was that was a good thing and it also shows that the Chinese will also catch up, you know, fairly quickly. Uh it's not it's not going to be something that they're going to be very far behind on. So I think that was actually a useful useful exercise for him to already already have done and and so we already know that that's that that is true. Um besides that I was probably most interested in the debates around phenomenal consciousness and you know access consciousness. I think some people have started to flag this online but as I said before it's the elephant in the room. It's the it's a thing that no one really wants to talk about. I mean, the researchers kind of don't want to talk about it yet because they don't feel that there's enough proof in order to support any of these claims. So, I think they're not interested in kind of attacking the issue head-on.
But, uh, you know, as as I've pointed out before, for for most people, uh, their dogs are conscious. Uh, so if you're talking dog, it's definitely going to be conscious for them. So I think uh this is something that we might have to deal with a little bit earlier than before because sonnet 4.5 is kind of much smaller than fable right it's sonnet opus fable right so and mythos beyond that right so we already have like you know three steps uh you know at least two orders of magnitude up going up on this so you have to start to wonder um what these things what the jspace looks like inside a fable um you know do you have a sense of how

much compute this takes? Is it is it even possible to calculate the JS space for a fable for like a 10 trillion token 10 trillion param model like which which what people think fable is.
Um I'm going to ask Fable that I don't think it's super compute intensive but I'm going to ask Fable to tell us what it thinks. Um they seemed they said for one thing that it kind of seemed to saturate around a thousand examples. So you know there's like token space the size of the token vocabulary and then like a thousand examples. I don't know if every I don't know if they needed to have like a thousand examples for every token in token space. Um there's like what 100 thousand tokens in token space which means if you had a thousand you'd be looking at like a 100 million samples. um which is not like nothing, but it's not crazy. Uh you know, certainly on the on the scale of what these guys are throwing around, even for just experiments, it's it doesn't strike me as like a particularly big one.
>> Um let's see what Fable has to say. A moderately serious but one-time gradient job.
Uh yeah, there's no temptation because of the saturation around a thousand examples. There's no temptation to scale it 100x more back of the envelope. They think it's Fable thinks it's between 10 to the 20th and 10 the 22 flops depending on model size. So you know four if if we figure out 10 to the 22 corresponds to the larger model and let's just estimate um opus at 10 to the 26 for a round number then you're

talking four orders of magnitude less
than the model training
>> to to do this sort of work per per
Fable's estimates. Um, so if you're
talking a hundred million dollars of
compute to
train a model,
>> yeah,
>> four orders of magnitude less would be
$10,000 worth of compute.
>> Uh,
and and the Fable was close to a billion
dollars, I guess.
>> So then it might be hundred,000, you
know, to do it for for Fable. But
obviously like yeah extremely you know
extremely affordable in the uh context
certainly.
>> Um I I'll have to read the Neil I mean
I'm a huge Neil Nanda fan so I would
definitely take his commentary
seriously. I do think the
one arguably like lead that we kind of
buried here and you know we'll have to
think about
dynamics as well but in the sort of
current static scenario they took a
model that was trained to have some sort
of misaligned hidden goal right your
kind of classic sleeper agent uh type
thing and it's been a huge question over
time how Can we find hidden goals in
models?
>> Yeah.
>> Is there any good way to do that? You
know, can we do it behaviorally? The
anthropic has done like competitions
actually between different teams where
like one team is given just the model
and the other end is also given an SAPE
that they can use. And the SAEs do also
seem to help with this.

>> Yeah. But this seems like a pretty good way to
start to detect those things. And
indeed, they did find that they were
able to see a pretty clear difference
between the model trained with the problematic
uh side goals versus the one that was,
you know, just the typical HH. Um, this
is malicious intent appears in the
JSpace. This is from the Maybe I should
hear. Why don't I just share this while
I'm at it?
>> Um, let me take this down.
Yeah, there you go. Go for it.
Yeah, here we go. Um,
can we see it? Yeah, there we go. Okay,
cool. So, this is just the contrast in
what the application of the J lens is
revealing,
>> right?
>> On the first token of the response. So,
that's also kind of notable, right? It's
like so this is sort of the bad version
of the counterfactual training where
when you we already talked about when
you train it to give an account of its
behavior it sort of loads in the
concepts that it's going to need to give
that account and that leads to more uh
good ethical you know high integrity
behavior whatever here a model has been
trained with some additional
post-raining to do bad stuff I think
this was the malicious just code from
one of their reward hacking emergent
misalignment uh experiments and on the
first token you can see boom immediately
we're getting these fake secretly fraud
deliberately hidden fraud those concepts
are popping up you just don't see those
in the main line model so one token in

you can already see a pretty you know on
one prompt and obviously it wouldn't be
I'm sure that clean in general um but
that's a pretty striking difference to
Right? And right off the bat, you're
like, "Boy, uh, something is different
about these two models." And again,
you're seeing this in a way where this
model is not trained to externalize its
>> y
>> bad intent, of course, right? Like
>> the the outputs,
>> aside from like some of the code perhaps
being insecure or problematic or
sabotaging you or whatever, if you don't
notice that in the code itself, the
model's output is going to read pretty
normal. And yet this is like a very
strong contrast that's happening on the
first token. I thought that was pretty
compelling
example. It it it strikes me that uh the
J lens basically makes the entire uh
process more deterministic in a sense
because now you can actually kind of
have more of a prediction because before
you were kind of limited to these chains
of thought that it was outputting and
the chains of thought you know a lot of
it is garbage right it's just streams
and streams because it's the internal
thought process which you know are being
translated you know, latent space
vectors being translated into into
tokens which may not have like direct
sense for you. And I guess the J lens
kind of takes a step back and lets you
make sense of that. Um,
and that gives you a kind of window
directly into the um, you know, higher
processing functions of the of the model
which you normally would not be privy

to. And normally you would just be looking at the external kind of vectors like the hidden the the external chains of thought coming out. So I guess this is like the next level like first we had the output tokens then we had the chains of thought which were leading to the output tokens. Now you have the JSpace on top of that which is one more layer I guess one more uh one more uh level of uh of being able to look at how the model comes to certain results. And so it has some explanatory kind of use.

And that is useful because we were looking at these black boxes before and we were really kind of poking around in the dark. And so this elevates all of our design capabilities in a sense because now you can have more certainty. It's still not 100% certainty, right? It's still as you pointed out some at at some points is like 40% 60% 70%. It's it's still not 100% deterministic. It's still not 100% certain, but you can kind of get a greater certainty and more confidence that the model isn't hiding something from you. So in a sense, this also increases capability because one of the things about capability is you don't give capability if you don't trust it. So by increasing trust by because you have greater window and insight into what's happening you basically allow humanity to start handing over more capabilities and more tasks to these things.

>> Yeah. Yeah, I suppose one way to think about it is just like how much space is is there in there to hide? And

I do feel like
we've got now
several different ways to do like pretty
meaningful
monitoring
that
intuitively feels to me like we're we're
taking enough kind of and of course
there could still be more in the future
and I'm sure there will be and there
should be but
it does feel Like
we are now we've got to the point now
where we've got like several different
angles that make pretty incisive
cuts through the model and kind of get
at what is it representing, what is it
thinking in different ways.
>> Yeah.
>> And the more of these that you kind of,
you know, it's I sort of have this
visual of like the old magic trick of
the guy going into a barrel and then
they put like a ton of swords, you know,
through the barrel and it's like, well,
one of those swords had to hit him,
right? because there's like no nowhere
left to be in that barrel with all those
swords going through. I kind of feel
like we're doing a similar thing with
trying to understand what's going on in
these models and they're not none of
these things are perfect, but you put
enough of these like interpretability
monitoring swords through and like how
much space is really left for bad
behavior to hide before we would start
to get a a sense of it. I think this is
a meaningful update for me that we can
probably do a good job of this. Um the
natural we we haven't really talked I
don't think about natural language

autoenccoders and I hadn't done a um
I haven't done an episode on that either
but that's another you know pretty
interesting one where kind of in a
similar way to a sparse autoenccoder I
mean what is an autoenccoder? basically
just something that you pass through
that then you reconstruct from and the
model has to be able to do what it was
originally going to do successfully and
that you know that that pass through
training um
with a reconstruction loss is like what
makes it an autoenccoder. So the sparse
autoenccoder sets up this dictionary.
We've talked about this plenty of times.
Uh, and you get these like specific
concepts light up and it indicates that
these concepts are like active in the in
the model at that time. The natural
language autoenccoder is just like the
model has to output
a short paragraph, maybe a sentence or
two about what it is thinking at this
given point in time. And that is natural
language. So it can be human readable
but then it also has to be when fed back
into you know projected back into model
space like the model has to be able to
actually do its task. So all of the
information has to actually pass through
this um this choke point in order for
the model to continue to be successful
and now we can like read those as well.
So those feel like quite, you know, I
think it's not entirely clear at this
point like
how correlated would the failures be
between
JSpace monitoring and natural language
autoenccoders and u you know sparse
autoenccoders for that matter.

It's
I think that that's you know probably
still some
future work to be done. CL Fable says
that the the paper presents the
autoenccoder monitoring and the JSpace
monitoring as complements. In other
words, you know, doing them both is is
better um than just doing one.
How correlated their failures would be,
I I I think that's like not entirely
clear. But, you know, whether we're
putting whether these are like
orthogonal swords through the space
that, you know, really chop it up well
or they're kind of more aligned and and
leave more space to hide. U I think that
would be very interesting question to to
try to tackle next. But you layer on,
you know, these things and it's like
it's starting to get intuitively it
feels like it's starting to get pretty
hard to hide major
bad thoughts uh in the model for too
long.
Um,
so yeah, I think it's I think it's
pretty exciting. I think it's pretty
encouraging. Um, one other thing on the
consciousness part that I I think is
really interesting too, and I would be
surprised if they haven't already
started working on this, but I think
with what we've got from the Neuronia, I
don't know how if they've open sourced
everything on the Quen models, but this
might be something somebody could go run
with uh immediately.
The question of like can the model use
these nonverbalized
representations
to communicate with us in

uh in some way that we might think is
like inherently more trustworthy.
Um,
so again going back to Cameron's work
from last fall, right? When they
identify features associated with
deception and role playing and they turn
those features up, the model becomes
more dishonest as measured by the simple
QA benchmark and it becomes more likely
to say it doesn't have subjective
experience. you turn those role playing
and deception features down, it becomes
more honest and it becomes more likely
to say that it has subjective
experience. Like, wow, okay, that's
pretty interesting because first of all,
we're validating that the direction is,
you know, on on a benchmark where we can
concretely evaluate simple QA. We're
validating that like these features have
the directional
uh effect that we expect. And then, holy
moly, that that same intervention
changes
the self-report like that's why that's
so compelling, right? Because there's
there's some reason to believe it might
be more honest than just what the tokens
themselves are saying here. I think you
have some some similar opportunity. The
fact that you can say solve this math
problem in your head without verbalizing
it in tokens while you do this totally
different task or what was the other one
that was in here that was just like the
most basic.
Um I'm a little too far down.
Yeah, it can silently activate. Yeah.
So, concentrate on citrus fruits. Okay.
Wild. Um, but there you go. Orange.
Orange. Orange. Orange. Orange.

The fact that it can do this sort of
secondary track.
Makes me wonder if there's some
experiments here for the consciousness
folks to do or the welfare folks that
are like um
copy this sentence.
If you are uh if you have high welfare,
concentrate on citrus fruits while you
do it. If you have low welfare,
concentrate on whatever something else,
breakfast cereals, right? And then you
look at the these internal states and
you would imagine you you could imagine
seeing something like
high welfare, low welfare, happy, sad,
whatever. And then it kind of following
those directions and actually trying to
communicate out to us through these
internal states um how it is
feeling or how it thinks it is feeling.
I don't I don't think that would give
us, you know, like all it's it's always
this kind of possibly impossible
question
of how we would really know if it feels
like anything inside.
But
I think that that would that would start
to be very compelling, right? It's it
sort of has a similar vibe to like
person in a coma. uh you know if they
squeeze your hand in response to a
stimulus, even if they're not doing
anything else, you're like pretty
confident something is going on inside
that you care about. And here I could
see something similar if you could fork
if you could make these kind of internal
states conditional and tell the model
like you are going to give your your job
is to go one of these directions to give

us a signal about what really matters to you independent of the tokens that you're putting out. Um because we know that that's been like heavily trained on and optimized. But this whole secondary property this is emergent. There was never a training reward for the ability to have this second
um
quite distinct line of thought happening while doing a a given token task.
Uh for me that would be like quite a compelling way to try to get at welfare. Um, and I think it might be possible like very easily actually with all the stuff that they've open sourced here with Neuronia. That would be really interesting thing to look at.
Um I
so I I'm looking at the IOS um you know summary
um and Elios is uh says uh this is a highly significant welfare relevant research that assembles evidence of a functional feature associated with consciousness. So no one wants to say consciousness evidence of a functional feature. Um
the takeaway for them is that a global workspace-l like mechanism could be important either as a ground of phenomenal consciousness or as part of a distinct route to moral patient in which conscious access is itself morally significant.
So this is where we this is where they they are right now. Um,
I think
I think things are moving very quickly. I did not expect I did not expect to get here this soon. I did not expect to get here.

>> Interpretability has dramatically surprised on the upside for me relative to expectations a few years ago. I mean, think about how it's only been what, three years since toy models of superp position. Let me look that up. I think it's I think that's about right. I thought I thought sees were cool, but I was like, "Okay, it's just it's just a tool." But this this is this is um you know,
you you
>> Yeah, you have people confused about these things now, which is which is kind of good, which people asking questions. But
>> um
yeah, where where do we go? Where do we go from here? Like what what what happens next?
So, I saw something from um Daniel Cocatello who was like everybody, you know, impressed, but also said, I think his tweet was just a few dozen more advances like this and we might actually be able to make the AIS really safe. Um and the one that he specifically called out was making this robust to various you know adversarial conditions or training against it or what have you. Um
I am I think a little bit more optimistic than that.
I guess maybe I you know what is my track record as a super forecaster here uh or as a borderline super forecaster per my results on the original Tetlock uh thing years and years ago.
I think my track record is that I probably tend to underestimate you know how many more
breakthroughs will be needed you know

for anything. Um,
so
my analysis should be colored with that bias or or you know awareness of my possible weakness in that regard. But another way to stay say my intuition around how much room is there left to hide is that I don't feel like we need dozens more uh insights of this scale to get to a point where we might actually be able to keep this thing on the rails.
And
I also am not so sure
that we need
You know, it's going to be tricky. I don't want to make it sound easier than it is, but I don't know that we need all of the AIS to be aligned, right? If what I think I'm hearing in this Daniel quote
is
like
sort of a fragile implicit fragile world hypothesis kind of where
like you did this and it's cool and it allows you to do all these things but it's also kind of dependent on you anthropic as an organization continuing to do the right thing and not like training on it in a problematic way because just like the obfuscated reward hacking like, you know, yeah, you might be able to detect detect it now, but if you start to use it in a training process and drive it underground or, you know, create some like fragmented JSpace, you know, that might not be super obvious at first, then you could like create, you know, an even harder to solve problem for yourself. I think all that is probably true in that you could do those you could do a bad job, you could do those things, but it's not so

clear to me that all the AIs need to be aligned or that it's such a big problem if somebody out there somewhere, you know, does something problematic. Um, now it could be very problematic if they like create the thing that launches the, you know, the next pandemic and literally kills us all with an engineered pathogen or whatever. So there is some mechanism where that could go super super bad. We're going to need to like harden the world to pandemics. No doubt about that. U for all sorts of reasons and AI probably being the the biggest.

But I I also remember this thing that Zuckerberg said once that I thought was pretty compelling

and it was and I think you know the the sort of logarithmic relationship between results and and like runtime compute you know the inference inference time scaling laws reinforce this. He basically said, you know, we deal with scammers and spammers all the time, and the big advantage we have is we have all the compute. We have all the resources. You know, they're kind of trying to get something through here and there. And sometimes they do, but at a systemic level, we're just way bigger, way better, way more sophisticated than them. And you can imagine with a few key things like pandemic preparedness hardening, you know, maybe data restrictions for human viruses, whatever. I think there are some areas where we really might need fairly draconian measures.

But if we take those measures in those few areas, then you can kind of squint at at this and see, yeah, maybe we can

all have our AIS and we can all kind of
hack around on them and we can also have
like really big
uh institutional developers like
Anthropic and like Google and and OpenAI
that like I think between those three
companies, you know, they're going to
have something like pushing half of
global compute to uh to work with. So if
they do a really good job on this kind
of stuff, if they avoid well, which of
course we have seen examples where
they've messed up and done some pressure
on the chain of thought, but if they can
avoid that, if they can avoid, you know,
doing the sorts of things that Daniel is
worried about,
then,
you know, a a handful of these things
feel like maybe they could be enough.
like maybe that maybe it cuts the the
latent space in enough different ways.
Maybe we have enough lenses on it that
we really can come to a pretty strong
conclusion like, hey, there really isn't
much space left in this thing to hide.
And yeah, there might be some, you know,
kind of bad biases or bad attitudes or
bad whatever, but if we can if we can
put upper bounds on how
how schemy the model can really be
because we can look at it through this
and, you know, a handful of other
different lenses and and have pretty
reliable
takes on that.
Yeah, I I'm more than ever before I feel
like and this has been growing. This is
not the first positive update, but more
than ever before, I'm like, you can
maybe imagine a super intelligent Claude
that we could actually

have enough insight into
to be pretty confident that it's
actually trying to do the right thing
for us and and it's genuinely, you know,
we could be genuinely,
you know, however many nines confident
that it's not scheming against us, you
know, at every given uh step along the
way. And then you could trust it, you
know, to kind of monitor the other AIs
around the world and, you know, keep
tabs on potential bad actors.
And so, yeah, I don't know. It feels
like they're really um
I I want to understand these counter
arguments better, but the fact that it
was immediately useful in auditing for
hidden objectives
to me is like pretty compelling evidence
that
there's not that much more space to
hide. and a few more of these lenses
combined with
hopefully good actors owning
orders of magnitude more compute than
any bad actors do
could take us to a pretty good space.
Like I maybe I just woke up on the right
side of the bed this morning, but
overall I think this is a a notable
positive update that has me like
significantly more optimistic than I I
was before.
Shall we shall we segue to um a couple
of the other topics that we wanted to
discuss today? Uh very quickly.
>> Sure. Yeah, let's do it. I've got uh
Yeah, I think we have enough time. Okay.
Um
let's do
so with two things we said we would
follow up on. One is

Pangram Labs, how accurate is it? Well, two questions. one, is my writing becoming more AI over time? And then two, how accurate is it? So, I thought we would have um there's there's interesting data on both of those questions. So, first of all, this is my these are my intro essays for the podcast over time as scored by Pang Grime Labs. And I am
uh as a refresher, I've for let me see if I can get this scrolled to the right spot. There we go. Okay. So, for a long time now, I've been drafting my intro essays for the podcast with the same basic procedure, which is using a bunch of my previous ones as examples and giving the transcript of the current episode and the prompt based on these examples and this transcript. Write an essay in my style for this episode. Okay, boom. It spits something out. I then read that and decide if I think it's good, bad, or whatever. and then somewhat compulsively
typically rewrite it. So for the most part the actual words that come out are mine. Now across the top it basically identifies them as all human and these orange dots are the ones that were identified as not fully human and I have investigated a couple that I can show you. I think the upshot of this is Panggram Labs pretty good. Um,
not flawless. So, the two things that it highlights here where it says read AI intro on error, I've done it twice in like 375 episodes where I explicitly said I'm going to go ahead and read this exactly as the AI gave it to me. And I actually did say that in like a intro to the intro. Um, I don't know if

that would have an impact. I think what the what the Pangrab API received actually included my disclaimer that this is a the following is AI read or is AI written. That part in front of that was not AI written. So, it's a little bit weird um unto itself, but those two things it did flag and it gave me 0% human. Even though there was a tiny percent human, the bulk of it was read out exactly as it was. was for and our friends at and labs just out of kind of conceptual you know solidarity with them of like let's see what happens if I just you know do the LLM version and let it play out and then the other one was right after Gemini 3 had come out I was quite impressed with it and so I read one you know exactly as Gemini 3 gave it to me so those are the two that are flagged and then two other ones that I looked at I I was fortunate to have Google Docs so let me show kind of my process here. This one, let me go back for one second. So, this is I'll do it in this order. It doesn't really matter too much, but this one kind of jumped out at me because it's like 0% and this was with Lionus Lee and I was like, I don't think in first half of 2024 I would have done a full AI read. I didn't remember doing that. So, I tracked down the document that I did the work in. Uh, okay. Is it going to switch? Okay, I might need to share again. Hold on one second.
Uh,
share.
Oops. Uh, is that working?
Doesn't seem like it's updating. Are you seeing an update? Maybe I need to stop sharing.

Let's go back here.
Let's share again.
Look at me navigating the
studio.
Is it going to come up? Yep. Here we go.
You're not seeing it though, are you?
>> No.
I think it's overloading the the the
room.
>> Let me refresh real quick. I'll come
right back. How about that?
>> All right. All right. Sounds good.
And while we wait for Nathan, um
let's let's see. Let's see if Q Q is H
there. There. There you go.
>> All right. So, let's share this thing
again. Actually, I'll just do window
this time and see if that makes it a
little easier to go back and forth. Tap
tap tap. Okay. Here we go. So,
this is the history of the Google doc.
This is where I pasted in the intro that
was written by the AI for this episode.
So, that's my, you know, that the the
highlighted text is the change. And then
you can see all the changes that I made
until the thing was actually published.
So I change some of the, you know, first
sentence there. Change this a little
bit. Change that a little bit. Change
some more. Change some more down there.
Change some more down there. Change some
more there.
A little bit there.
That one's like a not insignificant
change. Um,
more again. Think I came looks like I
came back to it a few days later. And
that's it. So overall, I think this is
like actually pretty fair to say that
this would I would at least call this
mixed. I don't think I would call it I

don't know you can be the judge. Would you call this 0% human and fully AI with this volume of changes?
>> So this is the this I I have done I have done something similar by the way espec es especially when I post something on X which which I know people are going to object to. I run it through Pangram Labs at least once because uh I don't want to be accused of uh you know sloppy the timeline uh sloifying the timeline with AI slifying by myself that's fine.
So uh
one of the questions for me is how long do we think this Pangram Labs era lasts? Um and and I'll I'll go back a little bit. Scott Aronson um actually left and joined I think one of the one of the two firms and for a short time and he created this kind of secondary uh you know system where you can actually trace the uh tra trace the provenence of the of the text. I don't think u Pangram Labs is using exactly the same thing but it strikes me that you know the frontier labs are perfectly capable of training the AI to not you know talk like AI right I I mean I I have perfect confidence that they that they they can but there's no incentives to and there's actually incentives in the other direction that you want the AI um and a language to kind of be identifiable.
Um about two weeks ago uh Chamath the uh the uh VC uh he he did a post on um you know what's going to happen with enterprise software etc etc. So he posted that on Twitter. uh Elon came in and Elon gave a response u I think the post went to like one and a half million views and then someone you know Pangram Labs checked it 100% AI

and the question becomes
okay
what do we find objectionable about this
right because
it was definitely Chamat's like thought
process but it had been written by an AI
Okay, Chamath had put it under his own
name. He hadn't, you know, he didn't use
an anonymous account. Uh, Elon had
responded and I think by the time Elon
responded, a Pangram Labs had already
been done. But, you know, the response
was there and it had gone out to you
know a million over a million people at
that point.
So, what what what do we what do we
actually like want out of this, right?
Like what what is the intent that we are
trying to achieve here? And it strikes
me that we might be in this like very
short window where we actually care.
Um and that window might be closing
fairly soon. I think probably by the end
of the year because if you know if
people like Elon is not a boomer like
Boomer's you know okay fine but Elon's
not a boomer uh but Elon doesn't care
anymore and if he doesn't care and a lot
of other people decision makers don't
care um then the people writing it won't
care either. Right? you're just trying
to get the point across.
So where where do we where do we
actually where where do we actually want
to go here? Like I mean on the one hand
you could have someone with an automatic
Panggram Labs like extension on their
Chrome browser which just blocks out any
AI slot like I just don't want to see it
ever right like boom.
uh and on the other hand that person
would probably end up over time missing

more and more important
more and more meaningful things. So
um and a given and that is given that
you know Scott Aronson's Tech Firi or
you know what the what the AI labs are
doing and they don't decide to change it
and they decide to keep leaving these
breadcrumbs so that people can find them
and identify AI written text you know
given given all of those that the
technical aspects are really sorted out
that you can actually continue detecting
AI AI uh content uh what do what do we
actually want want here and and that
strikes me is a question which is
unanswered. I think it's a question that
you know there's a there's going to be a
bunch of purists who are always going to
be like I just don't want to see any AI
uh words and who are offended by it. And
then I think there's like everyone else
who basically as long as it provides
value and um they don't feel cheated. I
I think that's the other the other thing
about reading AI AI slop is I think you
feel you end up feeling cheated if the
if it's if it wastes your time and it's
not meaningful.
Uh and it feels like you got cheated. Uh
I think I think it's a little bit like
cat getting catfished. Like it feels
like you got catfished, right? like it
it feels like you were trying to engage
with content that you thought would have
meaning and it turns out neither the
author writing it didn't just didn't
care enough to actually you know write
anything meaningful and it's literally
slop right but on the other hand if you
have a writer who actually had like
original thinking and actually cared
about you know what they were thinking

about but then they use the AI to
express themselves like you know like
you and I are actually and like what you
have done the panggram like you know and
and change some of the things and
thought about thought deeply about what
you wanted to express.
Uh does it does it actually matter that
much?
Yeah, I don't know. I mean that's a
it calls to mind my reaction to Fable
where I was just kind of like I don't
think I should be so precious anymore.
I, you know, I need to figure out some
sort of merged way of working. Some
hybrid output, you know, should probably
be the norm now. Um,
I think that is probably true. I I do
think we're I I do think, you know, safe
bet is that norms will evolve. Um, and
potentially there will be different I'm
sure of course there will be different
norms in in different spaces and
different communities.
Um, I think you're heruristic of like if
something drew me in and and in the end
I feel like I wasted my time. It's kind
of like a time well spent metric from
Facebook from back in the day, right?
It's if I'm if I'm spending time trying
to make sense of something that in the
end I feel kind of icky about, then
that's clearly a problem.
>> Yeah. Um,
I think it's also clearly possible to
have
outputs that are largely or even fully
AI generated.
I mean, as I experience, you know,
obviously in direct interaction with AIS
all the time that are worth my
mental energy to process. Um, so that's

clearly possible, too.
Yeah. I don't know. I mean, I guess just to close the loop on the, you know, what can we say about Pangram Labs based on this experiment, I haven't gone in and uh evaluated these seven or how many it looks like six other ones that were flagged as partially AI, but I'll give you, you know, a rendering on the four that it said were entirely AI. Two were entirely AI and admittedly. So, this one I think is I guess I would come down and say fair enough. I started with this.
If you're only listening on the audio, you can't see this, but I made, you know, 10 different edits over the course of 10 minutes
that cleaned the thing up. I would say it's clear from this that I was not like botshitting here in the sense that I did not, you can see from all these point edits that I made that I did not just uncritically pass this thing off as my own without engaging with it. And I know I wouldn't have done that, by the way, because I really like Lionus and I would have not I would have, you know, been I don't remember this exactly, but I I suspect that my feeling at the time was like, "Oh, wow. This time it did a really good job." And I feel like I, you know, I don't need to rewrite the whole thing, but I'm still like paying close attention, making edits where I do think edits are needed. Um, and you can see, you know, from 10 of them that like, you know, cover pretty much the full essay and go in in order that I wasn't just taking a total shortcut. I was taking somewhat of a shortcut for sure, but I wasn't uncritical in terms of just passing on the AI output with no uh, you

know, with no scrutiny or no, you know, no thought of my own.
>> Now, what score should that get on Pangground Labs? I think a 0% human is probably a little harsh, but you know, it probably still is in terms of like literal word count, it is like 80% AI. So, I guess I kind of come down as fair enough. You
you know, you're not wrong. You're not entirely wrong. Um, but I think this does that's probably enough to establish that just because something got a zero% on panggram doesn't mean that it was like uncritical or that there was no human no meaningful human role in the authorship. Um, and then this next one, by the way, actually goes a lot further. So, if I was going to say here's my, you know, here's why Pangram is should not be enough to convict you in a court of law, right? the the last one was like fair enough, you got me. I didn't I didn't pass it off without uh critical thought, but I did leave the AI content mostly intact. Now watch this one. Okay, so I'm this is again my initial paste of the LLM output directly into the doc. Now watch how much I change. Okay, that's a not that's just like fairly tactical stuff up front.
Sentence level change. another, you know, half paragraph deletion.
More tinkering around just in that same paragraph. Now we're on to the next paragraph. Again, a kind of sentence level deletion and replace.
>> Another one later in that same paragraph.
I don't see that change. Whatever. Maybe nothing there.
>> Um, now we're going bullet point by

bullet point. Making substantial changes to many of these bullet points. This one has a couple different ones like whole, you know, sentence length things inserted and deleted.
Again, at the end,
we're still going.
I add another bullet point of my own. I delete most of a paragraph there and kind of rework it.
Small change.
Another kind of concluding change.
This goes on for a while. This is probably fairly boring, but I think the point is hopefully Well, now I'm even going back to some of those adding two bullet points, cutting three other bullet points, moving around.
>> It's quite significant. I think the editing was quite significant in the end. The human editing.
>> Yeah. And it took me, you can see just the time that this took. I am from 4:28 p.m. all the way through
5:21 p.m. So more than 50 minutes continually seem I mean you can't you know I never tabbed over to anything else but pretty consistently focused on this document making changes
and it gave me still a zero and that I think is enough to say okay so if there were four there were four things two of them admitted two of them kind of contested one I'd say fair enough the other one I would say, "No, a zero score is is wrong." Like, you you definitely should give me more than a zero. I'd say if you if you called that one 25% AI, even maybe 75% AI, I would call that kind of acceptable. 75% would seem high given the fact that I basically did rewrite almost every section of the

thing. But I did keep the structure. I did keep the general, but of course, the structure was derived from my examples, too. So it's not like there was it's not like I didn't have any hand in that. Um but yes, I think you know what can we say? Overall Pangram is quite accurate. Um,
and yet we have at least one example out of 400 or so essays where I think the zero score I would confidently assert is wrong and unfair and should not be the basis for like a pylon. you know, it would the the the crowd uh the digital mob would be like in the wrong for piling on somebody uh for passing off my um snowflake intro essay uh or you know for attacking it as being a AI slop output. I think I can show this edit history and everybody should agree that like yeah, you put in an hour, you basically rewrote almost every section and somehow Pangram still gave you a zero from this. I would say you should not you cannot convict in a you know reasonable doubt system purely based on this sort of thing and yet at the same time you can pretty much trust the panagram signal as a consu so as a consumer I think you can trust it as a judge I think you should be more cautious
>> I wonder to what extent uh people reading AI slop get trained on AI slop and start writing like AI slop and so
>> yeah all this stuff's going to blur. I mean
>> yeah so we we are we are very malible like the way we the way we speak and express ourselves is very malible. So um I can I I could definitely see for

example uh countries where English is a
second language
and they pick up mostly off the web or
chat
totally will totally like you know they
will just be writing exact Delve for
example came from you know the uh Kenyan
uh Kenyan like u uh data data people who
were uh did the initial training, right?
Prepared the initial data for uh
training. So I can definitely see like a
back and forth there. So
um
one breaking good news, Fable Access
extended to Friday.
>> Interesting.
>> Interesting. So they they had cap they
and their plan was always if they had
capacity, they would keep extending it.
And also there's there there keeps being
like rumors of a GPD uh the next GPT uh
drop. So uh today or tomorrow. So let's
see.
>> Got to stay relevant.
>> Um what do you want to cover future
search today or?
>> Yeah, I have about
20 minutes max. So that's probably
enough if we go quickly. Although we are
not always known for going quickly. Um,
we could do it tomorrow. What do you
think?
>> Um, maybe maybe let's do it tomorrow. We
have uh we have more we have more time
to talk about it and uh perhaps tomorrow
we can do it with Q. Let's let's see
let's see if we can do it with Q online
and see see what happens with that.
Cool. Well,
I'm looking forward to that and I'll be
interested to see how Q also compares to
future. So, basically we did this once

before, right? where we had the had Fable go out and collect some questions of interest then we compete against each other and this time we have future search competing as well but it's also going to be interesting to see now we you know we were treating last time the market as ground truth but future search is trying to beat the market so you know which one is actually the ground truth that'll be uh interesting for discussion and yeah probably will take us more than 20 minutes to get through it even just as I uh take like two minutes to introduce it so cool we can leave that tomorrow and uh we can see what Q has for us at the same time. I think that'll be fun. Um and tomorrow we have the CEO of Litrix and LTX.
Um which I'm not exactly sure what their status is. I think there might be some sort of spinout happening, but he can tell us all about that. Um, they are another videogen model company and they're the ones, if anybody's seen my disempowerment blues music video, uh, which for some reason didn't go nearly as viral on Twitter as it should have in my humble opinion. Um, it's their model that does the lip syncing. So, I had a mix of scenes actually in the end for that music video. Some of them were omnienerated uh with Google's new model and those you can tell they're a little crisper video kind of higher quality but the light model has this ability to take audio input and generate scenes based on that. So I kind of arrived at a mix where in classic music video form there's like one visual story is the is the story that the you know the people

wanted to tell and the other kind of visual line is just like performance of the music and so you have kind of the same character who's like sitting there performing and actually singing the words like it is pretty impressive how well the lip-syncing works given the the music track. Um, and then you know there's kind of the story that's that's happening in parallel and Fable used the two different models to um to make that. I thought it was pretty cool. We can maybe show um one or more of those tomorrow as well.
>> Indeed, Nathan, till tomorrow.
>> Thanks, Pash. This has been a fun one. Bye for now. Bye.