

Start Time	Speaker	Transcription
00:00:04	Dion	<p>Hello and welcome to today's webinar. My name is Dion Wiggins.</p> <p>I'm the CTO of Omniscien Technologies. It's my pleasure to introduce, the people that are on our webinar today before I'll disappear.</p> <p>So, this would be one of my easiest webinars of the year. Just the introduction and I'm done.</p> <p>In Australia, in Sydney we have Joe Sweeney. He's at 1:00 in the morning, Joe. And Joe is an analyst in IBRS, and he covers all sorts of things such as AI and various other factors, and especially in government and major enterprises.</p> <p>We're also joined by one of our regulars, our own chief scientist, Professor Philip Koehn.</p> <p>Um, and he'll be the one in the hot seat tonight with Joe. So, I'm not going to talk anymore.</p> <p>I'm just going to hand it straight over to Joe.</p>
00:00:53	Joe	<p>Thank you very much, Dion. And it's a real pleasure to be here.</p> <p>Before we start, I want to make it really clear this is an open Q&amp;A session and my doctorate is in education recognition and technology policy.</p> <p>Uh, a great intersection. Uh, but as a result of that like Philip, I also have had to do teaching.</p> <p>That means I'm really good at waiting for people to ask questions, because we learn the most when we ask questions and when we answer questions. So, I'm also going to learn a lot based on the questions that you provide.</p> <p>So, please do use the chat window to put your questions in. We're going to filter them and group them and make sure that they get answered.</p> <p>This is a really valuable opportunity for you.</p> <p>Now that said, Philipp, it's an absolute honor to be talking to you again. Dion mentioned that, you know, I cover off AI, I don't pretend to be an AI expert.</p> <p>In fact, everything I learned about AI, I actually learn from you second hand through Dion, the second hand, through the work I've been doing, um, in those areas. But even then, I probably still have a lot more practical knowledge going back almost 20 years in that field because of it.</p> <p>I actually position I very much in the future work sort of time now area.</p> <p>but, you know, I've gone back 20 years in the space, and you've gone back a lot more than that.</p> <p>How did you get started in AI? Well, what was the early years of your research like?</p>

- 00:02:34 Philipp Yeah. I mean, like when I started, uh, my undergrad degree, I actually was interested in machine learning.
- So, actually, the very first paper I ever published was a neural networks, and this was over 20, over 30 years ago.
- Um, I then kind of realized if you just do AI or machine learning and you don't have a problem, that's kind of the wrong way to go about it.
- So, that's how I kind of I got into language processing because there you have problems, you have data. And when I started my PhD, basically the university was at my advisor, Kevin Knight was like heavily in machine translation, which back then was still interlingua-based syntax, symbolic processing like ancient stuff.
- 00:03:19 Joe Right. There's been a great question that's already come in um from Ander.
- And thank you so much for the question. Um, and this sort of cuts straight ahead to some of the things that we were planning to talk about.
- But the question is, do you think large language model technology will replace the classical neural net MT transformer technology?
- 00:03:40 Philipp Well, okay, the long story to that is that, uh, the transformer was originally invented for machine translation. So, if you pull out that paper, it is a machine translation paper.
- Um, which also explains why it's like an encoder decoder model. And at some point, people realized that you actually don't need an encoder because you're not mapping anything from anything.
- So, not only have decoder only two language models. Um, but yeah, there's definitely a different development of these translation models that are trained to map one language to another, to pure language models, which just produce sequences of words and training them off massive amounts of data.
- Um, leads to these models that are really, really powerful and can do all kinds of stuff.
- And it took a long time that they overtook dedicated MT models, and it's still not entirely at that point, people still built dedicated MT models. But I guess eventually all this is going to merge together.
- 00:04:42 Joe We've got another question that's just come in. It's from an AI saying I'm someone's helper.
- We leave that aside. Yeah.
- Um, you know, I actually want to step back to the whole issue that you did control that there, you know, outside of Transformers because that was a real pivotal moment.
- Um, but outside of Transformers, what were some of the real pivotal areas, uh, observations in AI that occurred that led up to Transformers?
- What's been the journey?

- 00:05:13      Philipp      Yeah. So, yeah, I think it was about, uh, yeah, probably about 20 years ago when I heard the first person in modern times, in the 21st century, talk about neural networks again, because they were really big in the 80s and 90s.
- And then they kind of were forgotten.
- And, uh, so that was, uh, built like Ngram language models with neural networks. And it all from the history, it always seemed like computationally very expensive.
- And he told us to buy graphics cards and I thought, like, this guy's crazy. I don't need to listen to him.
- Um, I mean, that at some point about proof that these things work. And then it's like, that was the first big Kind of replacement of a statistical model with a neural model, then recurrent neural networks that actually then managed to, you know, create new machine translation models that are better than statistical MT.
- That was the next big step. And that also introduces attention.
- So, the what transformer.
- Models.
- Thing is basically this. The title says that, you know, attention is all you need. It's all just based on attention mechanism. But that attention mechanism was developed already earlier.
- very clearly for machine translation because if you predict the output word, you have to pay attention to one particular input word that you actually translating.
- 00:06:36      Joe      There's another question here from Mirza.
- and they would like you to give your thoughts on how translation is actually learning or learned in LLM. The comment was we see this capability Abilities emerging.
- But how? Considering that no explicit um corpora or instruction set is provided, what's actually going on there?
- That's a really technical question.
- 00:07:03      Philipp      Good question, because I think it's also a misconception what is actually going on, because it's a typical sign of these models are trained on excessive amounts of text, and nobody really knows what's in it because it's all of the internet. You're not going to have the time to read it all, so nobody really knows what's all in there.
- And then the assumption was made that it's just monolingual text, and a machine translation emerges somehow, magically, by the power. Uh, so there was a really good paper by a scholar called Terry Brickell who was doing an internship at Microsoft.
- You know, um, um, works at Google. She now works at Google.

Who looked at what is actually in the training data for their models. So, that was like palm, really ancient large lodge language model.

Yeah. And actually there is a lot of translated text in there.

They are like literally in French this means blah, blah, blah. I mean, it actually has examples of translation in the text.

And if you remove all of that, then the models can't translate anymore.

So, it also learned translation because that's seen many, many examples of translation.

00:08:15 Joe

Got it. Which you know, takes us back to some of the early work.

Uh, I know that you were doing with Dion many years ago, um, looking at trying to get those massive corpuses of translation and then generating additional translation from them. So, in that corpus, there's probably machine translated materials as well, which are actually pretty high quality, uh, all things considered.

Yeah, right.

00:08:37 Philipp

You should count on that point a bit more. So, there's a lot of problem with like, people saying emergent behavior and saying the model discovered this by itself and all that with very poor understanding of what's already in the training data.

And up to the point that a lot of the tests that are being used were crawled from the web.

And almost by definition, therefore, they were in the training data. So, it's currently a very messy situation in really assessing how good are these models really.

And, and in new unseen, unseen circumstances where we actually want to use them instead of preexisting test sets that might be part of the training.

00:09:22 Joe

Wow. Now, this, um, there's a whole bunch of additional questions coming in and we're going to get to those.

But I think what I'm picking up here, and it's something I see a lot and I really struggle with, and that there's a huge number of misunderstandings about modern AI.

and I spent a lot of my time talking with boards, um, trying to in some ways talk people down from the hype.

because when there's too much hype, you overspend, or worse, you spend before you need it, you know, without a real, real case. So, I spent a lot of my time trying to explain what modern AI is to these, to senior executives, to government policymakers, um, to semi or non-technical people.

And it's hard work. Um, you probably do a lot more of this.

How do you talk and explain these nuances and at what level to executives?

- 00:10:22 Philipp Uh, yeah. I mean, it's tricky because AI is a term that kind of is a bit inflated in its use.
- 00:10:31 Just to give you the very concrete example, when we did machine translation ten years ago, we didn't call it AI. I don't know when we started calling it AI, I had an interesting interaction with, uh, I'm not going to name the person someone who also runs a company in the machine translation space.
- And I ask him hopefully at a conference, like, when did you start calling actually machine translation AI? And he's like, yeah, you know, last year we started talking about that too.
- And then he went on the stage and he gave it like eight years ago when we started the company. It was started from the ground up as a AI company.
- 00:11:08 Joe Yes. So.
- 00:11:10 Philipp Yeah, AI is okay. I give you my computer science definition because I'm teaching actually read a little class on AI.
- And I think the main difference to classical computer science is in classical computer science, you try to understand the problem, you come up with a solution, and then you implement the algorithm that exactly walks you through the steps to make the solution and AI. You have a heuristic that sometimes works and doesn't work, and therefore does things like search or machine learning to get better at it.
- But it's not an exact solution to a problem. It's always an attempt at solving the problem, but there's always never really a guarantee that it can solve the problem.
- 00:11:56 Joe And this takes me to my other bugbear, which is hallucinations. Um, many people that I've talked with see hallucinations as a bug in the classical software sense, and I come from the software background, and yet it's not. It is. It's really an ingrained function of how vectors work.
- Um, it seems that the more powerful and the larger these eyes are getting, the more embarrassing hallucinations we're seeing. Could you explain, you know, really simple terms, what's behind hallucinations and how can they be minimized or overcome?
- 00:12:35 Philipp Yeah. I mean, the surprising thing is that actually doesn't always hallucinate. I mean, it is
- Completely untethered to the truth of the world because it hasn't observed the world. It just seen a lot of text and it just tries to mimic that.
- So, it just produces more text that is fluent, similar to stuff it's seen before. If it's true.
- Good. If it's not true, that's fine too.

Um, it has no valid way of validating against the real world. It's obviously influenced by a lot of two things that are being said, because most of the stuff that's written down, hopefully it is true.

Therefore, it kind of has a bias to say things that are true, but it's not necessarily limited to things that are true. Especially any kind of creative text generation has to take some leap.

Um, so it's very hard to control because it's kind of the core feature of what it does. It's very creatively creates text that's very similar to what has been said before.

And a lot of interesting, important things have been written down, and therefore that's a good thing.

But there's no guarantee that it doesn't just make stuff up.

00:13:43 Joe

Yeah. Got it.

Um, I want to, um, we've got two questions here that are slightly related. I'm going to take the segment first from David.

Do you see the current LLM, AI uh, generation? Sorry, did you see the current wave of LLM and AI generation coming?

Um, if so, how much before five minutes, months? Years beforehand?

I've actually got a comment on that. I'll let you answer it.

When did you see this coming?

00:14:13 Philipp

Um, I mean, I think I was surprised, like everybody else. About what?

How good it is.

I mean, we have been in the field, for instance, of machine translation for a long time, and we always had we're making incremental progress every year, like in Blue Point better and things improve and things become more useful. You know, kind of that it's a useful tool for professional translators.

So, it was kind of important barrier that you traveled abroad and people pulled out their cell phone and showed you a machine translation language you couldn't speak.

Uh, these are all milestones. Uh, one astonishing breakthrough for neural machine translation models was that they produced actually fluent, well-formed sentences, which we didn't really have before.

So that was surprising.

And generally, it's like the underlying mechanism is not that hard to explain, but why it actually works, um, it is actually a real mystery to everybody. I mean, in theory it could work, but that it works, that it works so well, that is that is impressive.

00:15:18 Philipp

And yeah, the next big milestone was like the GPT 2 or 3 of the world and ChatGPT when it came out that it cannot only produce text, but it can actually

really seem to solve problems and answer questions that are not just let's add some words to that sentence to make it kind of ramble on in real directions.

00:15:42 Joe

Yes. Yeah. So, it was really interesting.

Um, I when I look at futurism, which is part of my job, um, what I've actually found is that you can look ahead and know pretty well within, you know, within a year or two, uh, over very long periods of time when a particular technology is going to reach what I call the inflection point, it's the cost performance ratio.

Um, and we predicted we weren't calling an AI, we were calling a cognitive computing.

Um, and this was back in 2014, um, and we were predicting that it would sharpen around, um, late last year, that was our case. Now we were out about 18 months.

The reason for it is our models are always based on economics, cost, performance. When does this become when does this capability, technical capability, become cheap enough that it's just a no brainer to use it?

Um, but then, you know, what we were predicting is that Microsoft would throw \$10 billion in now hundreds of billions of dollars at this, which sort of brings everything forward.

But this is my next question.

Pretty much all of the major AI vendors are losing money hand over fist on these investments. They're basically selling AI in order to buy market share, which we started predicting a couple of years ago would happen.

Um, and that's not proving to be successful for them, uh, because they've got open source and all these other challenges. And to quote Google, there is no moat.

What's your thoughts on how the market's going to play out as a result of you know?

Yeah.

00:17:16 Philipp

Of all um, I mean, I'm not uh, I'm not a business person, but I still like let's start from it from a technological point of view. And you already mentioned the Google paper from there's no moat.

Uh, Google invented the transformer. I mean, there have been work in that space before.

Uh, there were a large language models before ChatGPT. It's just it's always odd what gets.

Then finally the mind churn clearly for ChatGPT was that it was publicly available and everybody could use it. But people built very similar models before and that is still the case, like all these big models are being built.

Um, they're right. They kind of stopped doing this a little bit, but they used to write really long, 5100 page technical reports on how they exactly built them, and they're all doing more or less the same thing.

So, I remember a story from someone I probably also shouldn't name what that was, uh, company that also built a large language model. They had all kinds of new ideas to do, and then they started training.

You spend \$1 million a day and things didn't work. So, they threw out all this new stuff and stuck with the old stuff and then finished the training.

And it worked as good as everybody else's. So, there's not really any big competitive advantage to any one company right now.

Yes. If you look at the benchmarks, I'm a little bit better at this and a little bit better than that.

Um, but even there, I would argue that has more to do with the stuff that happens after the training the fine tuning, reinforcement learning, all the infrastructure developed around it.

So, given that it's so competitive and that a lot of the data resources are publicly available, a lot of the code is publicly available.

It is hard to see how anybody can develop a monopoly in that space.

00:19:03 Joe Yeah. And that's pretty much what we've come to the conclusion of quite some years ago because, uh, this is really going to open up.

Now. One of our members just made a comment.

Um, pushing back on my comment that the big AI vendors aren't making money. They said Deep seek queen.

The Chinese vendors are. Um, I actually don't have visibility of that.

I'll have to go and check it. But, um, you know, what's interesting is those models have been open source.

So, I think that there's a bigger issue. Come to the open source discussion and a bit more.

There's another great question here. And I do know there's a question on MT, uh, that I'm going to bring up later, but, um, this one's really interesting.

Is RAG plus large language models as good as MT, and if not, will it get as good and will it be more cost effective?

00:19:52 Philipp Is the rack and lightning as good as MT or what was the question?

00:19:57 Joe So, basically um, is RAG um, you know the rag models. Is it going to be as good as machine translation.

And if not, will it get as good?

00:20:10 Philipp Uh, I still don't know how to answer that question of what the question actually means. So, uh, I mean, retrieval, augmented generation is this idea

that before you answer the question, you first find relevant information, and then with that help, you can maybe answer the question better.

Um, this has been applied, for instance, to, uh, machine translation by if I want to translate the sentence, let's look at similar sentence that I already translate it. That's an old idea translation memory that goes back to ancient times.

Um, and if you provide that, you could probably do a better way of translating it. So, that's what I interpret is RAG plus MT and yeah, that is uh, it is a successful technique.

Um.

00:20:51 Joe So, I think the question here was and it's a slight nuance, um, is RAG plus a large language model now or RAG is based on large language models. Um, matter.

Were you referring to the big, broad, large language models with a rag corpus? Was that was that the general, uh, concept that you were looking at?

00:21:11 Philipp Yeah.

00:21:13 Joe I mean, I think.

00:21:15 Philipp Yeah. I mean, I mean, there's currently a pretty open battle between dedicated MT systems and large language models.

And, uh, the big advantage is that large language models have is that there's so much more flexibility in how to use them.

So, the machine translation models are very focused on. Here's a source sentence document.

He has a target document. That's the mapping task.

If you learn or large language models, you can do things like can you please translate that in the style of so-and-so? Or please also use the following terminology.

Or even now I think you're going to get into agents in a second, but the whole thing of breaking up that process into multiple steps of first, identify the hard translation problems to a draft translation. Can you refine it?

Can you check if everything is correct?

That's all stuff you can do with large language model, which you cannot do with a dedicated MT models as they're currently built.

00:22:14 Joe Right, okay. Do you think that they're going to merge?

00:22:17 Philipp Yeah I think yeah, that's probably the big trend. I think the only thing that keeps um, dedicated NMT model still around is that they're just more efficient.

They're much smaller. They're easier to train.

They're also from a deployment costs, uh, much cheaper.

And as long as that is still the case. Um, so, I mean, I know.

So, Microsoft still has dedicated empty models behind the service, and I would assume that Google two, but I don't know if they change things. It's just if you have like, like, uh, empty is an application is used millions of times a day.

So, cost of using it matters. So, if you can cut that by a factor of ten.

Yeah. You do.

00:22:59 Joe That. Yeah.

That which comes back to that whole economic equation that that flipping point. Um, so look, the going back, uh, Ivan, I'm finally getting to your question.

Um, straight to MT, um, okay, all the technology aside, and I actually don't know what this term is. Um, so I apologize, does PBSMT still have any application and if so, where?

00:23:26 Philipp It kept it was kept alive for quite a while.

00:23:30 Joe As actually, first of all, can you explain what that is for?

00:23:32 Philipp Yeah Okay. That is the old statistical phrase based machine translation.

So, this is, uh, uh, good old Moses.

00:23:38 Joe Your Heritage.

00:23:38 Philipp Um, that was what machine translation was from about 2003 to 2016.

Um, Google Translate during that time.

so it was all based on just statistics, you know, how often was this word translated this way, or was this phrase translated this way? And numbers you could still make sense of because I just kind of fractions of counts.

so that was kept alive for quite a while for low resource scenarios because it was a bit more robust than neural models.

But even with that, it's not the case anymore, because even for low resource language pairs, you can do so much more with a large neural model in terms of data augmentation by training on multiple language pairs at the same time. so the language they don't have much text can gain some kind of foundation or support from a similar language.

And yeah, so I think even for low resource languages, the most successful models are the neural models.

so the only really justification for new race based, the only real advantages that still exist, uh, it is a bit more controllable.

You can force it to translate a certain word in a certain way.

And, it is. Yeah.

Cheaper.

- And it uses doesn't require GPU and relatively modest CPU effort. But I mean, it's not that many scenarios.
- What that is that actually matters.
- 00:25:18 Joe Because we've reached that, that price performance. Actually, there was a, um, that relates to a question here from um, Andreas, which is will the hardware requirements decrease in the future?
- And I think there meaning in general, I.
- I will rephrase it slightly.
- Will the cost of the hardware, uh, requirements decrease? Um, in other words, will this get cheaper to actually platform?
- 00:25:44 Philipp Um, a) I certainly hope so.
- And b) I believe there was both. I mean, there's on the one hand the actual hardware side, there's Moore's Law, the, you know, you can get more efficient computing and uh, and uh, incrementally over, over the coming years.
- Um, the current cost is also heavily influenced by kind of these monopoly profits of Nvidia.
- Um, and so that's the hardware side. So, the hardware is going to get cheaper for the same amount of compute.
- Um, and but I'm also optimistic about the, the modeling and training side that I think the, the training we have is still a very, very brute force.
- And there will be more efficient methods be developed over time that bring down kind of just the computational cost in terms of number of computations, if they carry out, um, going forward.
- all these things are good, hard to predict.
- I mean, one interesting thing also is that, um, we already training on all the texts in the world, so we're not going to get more data.
- I always like to say there's only 8 billion people, a billion people in the world, and they only write so much. So, that doesn't double every year.
- So, the amount of text we train on is not going to increase.
- so if compute grows exponentially, but text may be linear if at all, then it should get cheaper over time.
- 00:27:25 Joe Interesting. Actually there's another side to this, which is I've been watching, uh, a little bit of what's coming out.
- Uh, the chips and so forth. You talked about the monopolies, and so, um, but that those monopolies have the potential to be also shattered once China production comes up.
- Other nations, uh, other sovereign AI manufacturing begins.

- 00:27:45     Philipp     Yeah, I know there's you. Yeah. This and this various efforts by Google to have their own chips have TPUs, there's efforts by Microsoft, meta, all kinds of companies trying to develop their own chips.
- So, there's yeah, it's probably not for another 2 or 3 years, but maybe in that time frame there will be serious competition to Nvidia.
- 00:28:07     Joe             Uh, do you think, um, fundamentally, you know, as we see this new competition come in, that the architectures for how we train and platform these AI systems is, is going to change? What do you think we sort of at a point where it's going to remain the same sort of architecture for, say, the next decade.
- What's your thoughts?
- 00:28:27     Philipp     I mean, there are definitely attempts to move beyond transformer models. The things called state space models that people are very optimistic about.
- Um, and it's kind of interesting, like in the early days, we had like recurrent neural networks, we had LSTM, we had like all kinds of things, convolutional neural networks. The technology seemed to change every year.
- And then once we get stuck with transformers, we know.
- Wow. So, almost ten years that we're using them.
- And that is kind of surprising that they stuck around. Because if you just look at them, there's just so many ways you can build these models.
- I don't know why this particular architecture, with all this idiosyncrasies has been proven to be so stable.
- it's probably a lot right now. I mean, a lot of the experimentation on a large scale is currently hindered by that.
- It's just so incredibly expensive to build these gigantic models. So, even the big companies, um, they have like one shot to build a big model because even for them, that takes several months and all the GPUs they have and therefore there is just very little, little experimentation on really, really large scale models.
- So, people stick with what's proven.
- And then it's never really quite clear that the experimentation you do with smaller models than actually carries over to bigger models.
- So, yeah.
- So, that should get surprising that there shouldn't be more. I mean, there's things like diffusion models and vision that have very different models.
- I mean, there's a lot of ideas out there that at least think there could be other kinds of models.
- 00:30:06     Joe             Okay. We've got some great questions coming in which are taking a slightly different angle on this.

And I think they're really important. Um, the first one is and I'm going to read it directly with recent debates about making AI models explainable, especially in Europe, how do you see the future of explainable AI and especially reasoning models.

Now I do have an issue with that reasoning models later. But uh, actually, first of all, before we answer that one, we do explain that there is a real distinct difference between what, you know, AI specialists term reasoning versus what normal humans think about reasoning.

Could you just in 30s explain the difference? And then I'm going to come back to this question.

00:30:54      Philipp

Yeah. Let's talk about the reasoning like explainability is a whole different topic.

Um, so what is currently being called the reasoning is allowing the model to produce all kinds of text that is just kind of scratch space that is just like scribbles. And then afterwards, after doing all that, it is coming up with a real answer.

And I think the main reason why that is successful is because one thing that the models currently don't have is any kind of memory or memory, short term memory of anything. So, humans we restore certain information now ahead as short term memory as limited as it is.

And the models don't, they predict the next word. And the only thing they can build up in terms of memory is all the previous words that generated.

So, the only way to think through stuff is to explicitly generating text that think that, yeah, that is now then called the thinking phase where it actually works through the problem.

So, if it has to kind of get aware of it, it has to work through several steps. It has to explicitly do that because it doesn't have any kind of internal mechanism to internally doing this kind of processing, everything has to be done explicitly as text generation.

So, that's what these reason models are. They're just basically given the chance.

Just generate a bunch of tokens. And we're only going to check your final answer.

We don't care what you say in the middle. We just want to check the final answer.

And having a large window of text that can be generated gives them the power to.

Yeah. Contemplate alternatives.

00:32:29      Joe

So, let's then move on to this this question where we started, which was given the fact that there is a lot of discussion around, uh, mandating explainability, uh, um, defining what explainability means. Yeah.

- Um, do you what do you see as the future of explainable AI?
- 00:32:53 Philipp I had a PhD student who did the whole PhD thesis on explainable methods. And then, like, deep philosophical, I have like I came away from that with, like, deep philosophical questions about it.
- Because you have a model, you don't know what it does. Then you have a method that says this is what the model does.
- And then how do you know that's true?
- You run the problem. If you have different methods that have different explanations, what the model just did, how do you know what's true?
- You know what it should be doing. Maybe.
- But maybe. Maybe not.
- 00:33:27 Philipp the other anecdote I like to tell in that story was like something like some book critics that, uh, on a German TV show ages ago that he reads a book, and then he likes the book, and then he has to write the review when he's gonna come up with a reason why he liked the book.
- So, the explanation why, like the book is a post fact thing that is being done, which is just a rationalization. It might be completely different why you like the book.
- Maybe it's just like,
- 00:33:58 Joe So, so that is there, aren't we?
- 00:34:01 Philipp Because the neural models, there's just gigantic amount of numbers and they do a lot of complicated things and like very broadly very fuzzy.
- And when we say explanation, we want to have like some logical chain of reasoning.
- 00:34:15 Philipp And that is very different for the models actually do. So, it's not really.
- like, and a logical explanation can never really be the reason why they came to that conclusion. It's just gives you some evidence for that.
- And it's also a risky thing because there's like also these studies that the model gives a prediction. People might be skeptical about it.
- But if the model gives a prediction, an explanation then people just like oh, okay. Yeah, that sounds plausible.
- That must be true. Then they trust it actually even more if they're given an explanation, no matter if the explanation is correct or wrong or matches anything about the reasoning process.
- 00:34:52 Joe And this takes us directly into the next question here, um, which is what's your view on liability in high risk use cases, um, where there are requirements for human overview and transparency?
- 00:35:05 Philipp Yeah.
- 00:35:07 Joe Are we moving in the right direction? Are we getting this right.

- 00:35:10 Philipp Yeah. I mean, these are all imperfect solutions.  
I mean, that's that was our mantra for machine translation forever. It's not like we're not going to solve translation.  
You're going to be good enough, you're going to be good enough, or that you can travel abroad. You can be good enough that translators twice as efficient or any or that you can buy a, you know, uh, metro ticket in Paris.
- 00:35:34 Philipp Um, so we try to reach levels of performance that had some error, but that error was not important for the task we are doing, so all these applications are like that.  
Um, so aiming for perfection is hopeless. Anyway.  
So, I guess anybody who then deploys these systems has to be aware of that. They're not perfect, that they make mistakes.  
And it really didn't depends on the scenario how to how to remedy that, on how to address that.
- 00:36:12 Joe Yes. Yeah.
- 00:36:13 Philipp I mean I'm the on the flip side, if you hire people to do a job, they're not Perfect either.
- 00:36:20 Joe And this is really interesting.
- 00:36:21 Philipp So, you have a actually very similar situation that, you know, if you just if your process consists of people doing the work, they all sort of make mistakes.
- 00:36:29 Joe So, when I'm talking with my clients around these risks, um, I think that sort of, you know, double checking, making sure it's true, make sure it's not hallucinating. Risk is actually fairly well understood.  
The risks that aren't understood is when you're in industries, for example, banking here in Australia. Uh, and in fact, many countries, banks are not allowed to offer financial advice beyond a certain point.  
And these AI models, certainly the larger one's, really good at wanting to offer as much advice as they can because they have been internally designed to do you know, what the user says, and that's a legal liability. There's all sorts of these other I call them toxic traits, um, that come out of these challenges.  
it's a really There was, there was actually some really interesting work going on in that space that it usually is in the field of the AI does the first pass, it has some guardrails. It does, you know, another one checks to make sure that there's none of this, uh, overstepping what it's allowed to do.  
Mark. So, keep on putting those together, and they look a hell of a lot like what we're now calling, you know, Agentic AI.  
Um, 2025 was meant to be the year of Agentic AI. I was quite proud of saying, yeah, it will be, but by 2026 will be back the same.  
Oh, it didn't quite meet. You know, it's it's going to disappoint that the hype will not live up.

- 00:37:57 Joe Yeah. What's your thoughts on Agentic AI?  
That was my thoughts. What's your thoughts on it.  
Okay.
- 00:38:01 Philipp Yeah I don't know. It is actually impressive how quickly the field moves and jumps on the next thing.  
Um I mean it's only so few years that like retrieval augmented generation was like a great thing two years ago and came out of nowhere. And the reasoning models are already old hat, although they're really what I don't think they're even a year old.  
Kind of came out maybe a year ago for the first time.  
and Agentic AI is now the latest. Um, yeah, that's a bit of a broad term.  
Um, I always have discussions with Dee on what that actually means.  
I think it currently means that you build not just like you have a single interaction with a language model. You ask a question, you get a response back, but that you have either a series of prompts or breaking up the process into multiple steps.  
But I think a crucial aspect is also that it involves other kind of technologies, like traditional databases, um, and even things that take action in, well, I don't want to say the real world, but the real digital world, like filling out that forms, actually booking a flight ticket.  
but that all also just really means that you integrate a large language model and a bigger workflow of multiple systems that have very different kinds of capabilities.  
And, and that is ultimately, uh, almost by definition, what has to happen if you build an actual application.  
It's not just going to be the large language model, it's going to be all kinds of other things to actually make it work. And I think that's also really I don't think that's going to go away anytime soon, because that's ultimately the next big step.  
You have to build actual applications that do useful things, either because they do it more efficiently than traditional models, they are better than the traditional methods, or they kind of open up completely new.
- 00:40:04 Philipp So, um.
- 00:40:05 Joe You know, when I try to talk about, um, Agentic AI, I, I actually say, yes, this term is going to break down very quickly. We'll probably move on to the next buzzword.  
I'm predicting it'll be something along the lines of AI orchestration, because when you get down to it, uh, and I strongly agree with you, it's about deconstructing processes.

Now, there's a whole bunch of steps in digital processes where humans have to get inserted because, um, but all systems just can't handle that.

And that's where, you know, maybe these reasoning model sort of reasoning models and, um, connecting with other systems lie. So, you know, we're probably just going to, oh my God, we're talking about, you know, automation again.

So, I think we're probably heading in that direction. But it's a really good discussion because what it also triggers is another question we have here.

Um, and this one's from, actually from a vendor, uh, technology vendor, um, one of their executives, they've said, basically I'm going to read between the lines.

A lot of vendors are selling all of their kit, you know, whether it be a cloud or whether it be a server. Whatever it is, it's the next thing for AI.

You can't do AI without it.

And yet that's, you know, that's AI washing. What do you, you know, as somebody who actually implements or works on this stuff, what do you think are the core components, the architecture needed to run AI?

00:41:34 Joe What do you need?

00:41:36 Philipp Yeah. I mean, ultimately you trying to build something useful.

00:41:40 Joe Yeah, hopefully.

00:41:41 Philipp And are you going to use the tools that you want. And if you call some of the tools AI and not that it's almost a marketing term, I think we already have the discussion that AI is going to use very inflationary.

And then we used for everything. So, yeah, that's a bit of a weird, you know, uh, semantic discussion about what the meaning of the word AI is.

Um, I, we've tried earlier definition of AI, So.

00:42:14 Joe Okay.

Well, actually, you know, you also were just talking there about, you know, it's going to be meaningful. And it's a great question here.

Um, from the start, which was is there a project or implementation that really caught your eye that you found was really impressive and meaningful for you personally?

00:42:35 Philipp Um.

Um, that's a good question. I mean, I've been working on machine translation long enough, so I'm actually.

But I've still was very impressed when I for the first, the first time it happened that I traveled in a foreign country and someone couldn't speak English and I didn't speak their language, and they pulled out Google Translate and I was like, wow, that's.

We worked on that and not totally terrible before. And now it is a technology that actually works and it's useful for people.

Um, that's very nice. Um, but I mean, I'm like probably like everybody impressed.

What kind of large language models are able to do right now? I mean, I use it for writing code because I can be bothered to look up the exact syntax of a particular command or property of a library.

And, generally, I feel like that is also currently the best role for this technology, that it's an assistant tool for, humans to be, to be more productive.

Um, there's definitely much more application of it being a helpful aid than a fully automatic system that does something.

00:43:45 Joe I saw a really interesting one, uh, recently, um, local government who has to process, um, development applications a huge amount of their time. It costs around \$90,000 just for a simple, you know, mom and pop, somebody wants to build a new fence or an extension of their house can cost the council's easily 9,000 USD to process that.

Um, and the biggest part of that when we analyzed it, is actually getting the documents from the public and making sure that they're the right documents. And if they're not, you have to go back and backwards and forwards.

And there's organization. Actually a couple of them got together here in Australia.

And what they did is they just as the documents came in, they looked at them and they said, is this the right document for what's needed and what else is needed? And just that simple check and balance, um, was literally halving the processing time.

Um, so I think it's these simple things where the AI doesn't have to be 100% accurate, but it can do a huge chunk of that human oversight.

So, I think, um, which is great because.

00:44:54 Philipp Of course, and that was, as I said earlier, that was our big argument, always machine translation. We're not trying to be perfect.

We tried. Yeah, we got enough.

And it's part of a process where if you know, the quality of the AI system is good enough that it speeds up certain processes, then that's good enough, but useful, that saves money.

00:45:18 Joe I mean, yeah, absolutely. Um, there's another great question here.

and it's reminded again, um, how far can we go after GPT 4.5, GPT five six, etc. in terms of data volume?

Um, it's constantly adding new data and constantly ever increasing number of parameters, a viable future.

- 00:45:43    Philipp    Um, I mean, I, I'm giving have been giving talks about this so long that I had like 2023, a slide of how big the models are.
- And, uh, so I'm forced to look at that every half a year when I update something. And the models actually haven't gotten bigger over the last two years.
- So, the 2023, the biggest model was, um, uh, I think it was like palm something. Something with 500 billion parameters might be wrong about which model was but 500 billion parameters.
- The current models are not really bigger than that.
- 00:46:16    Philipp    If anything, there's actually more effort to kind of build smaller models.
- So, the models haven't gotten bigger. And I already made my comments earlier about data volume.
- I mean, pretty much every day that you can get your hands on is already being used, um, mostly legally.
- Okay. That's a different discussion.
- So, that's basically all everything that's going to be written. If you brought something on the internet that is part of a large language model.
- 00:46:47    Joe    Oh dear. I'm dyslexic.
- So, that's a real problem.
- 00:46:52    Philipp    So, um, that was the message.
- 00:46:54    Joe    So, look, there's a there's a couple of questions that are coming in. Um, and I'm going to merge them together if it's okay.
- Sorry for the people. The first one.
- What we're seeing, um, throughout the world is an increased interest in digital sovereignty or basically sovereign AI solutions, but that's expanding out to digital sovereignty at the same time. That leads to a very rapidly to a discussion around open source.
- What's its future, what's its value? Is it a threat?
- Is it a boon?
- Um, putting those two things together.
- What do you think is going to be the future of sovereign AI?
- Yeah. Um, what does that look like?
- And what's the role, therefore, of open source communities and open source in general in that space?
- 00:47:48    Philipp    Yeah. I mean, uh, open source is a very powerful element, um, in everything because it democratizes people working with a particular type of software.

If it's open source, a lot of people can look at it, a lot of people can extend it, and a lot of people can improve it.

So, um, optimistically, and I think there's enough reason to believe that that open source solution will always eventually win out.

I mean, the big problem with training gigantic models nowadays is not the software, it is the 50,000 GPUs you need for like a year.

so it's really the hardware that is that is where, where consolidation might happen. And, you know, you can't, you know, everybody in their garage can replicate open AI.

and it's not the software and it's not even the data.

Although that might be riskier to preserve.

It's really the hardware that it's that is the limiting factor. So, I'm actually fairly optimistic that, uh, open source software that is then seen by lots and lots of different people with lots of different ideas.

Maybe I should also plug here that. I also still think that academia is kind of the place where a lot of the really good ideas come from, because people enter academia.

Relatively young students with weird ideas that are very different from what had been thought before, and that just brings very different perspectives to it, and new things get developed.

Um, I mean, the other big driver, for instance, of progress in AI is also not the OpenAI is the Googles of the world, but the random developers that might not fully understand how the models work, but they know how to play with them and try new things and come up with stuff like virtual augmented generation, because it kind of works.

And it's just these are just tough engineering ideas.

00:49:49 Joe Yeah.

00:49:50 Philipp And I think these, uh, yeah, ideas borne out of the actual practice of actually building something and trying five different combinations of something.

00:49:57 Joe Yes. So, you know, one of the things which I, I've been talking about and it seems to be resonating really well in certainly the Asia Pacific region is this notion that organizations will have a, you know, a fundamental platform now, whether they run that in a local cloud or whether it's in a hybrid cloud or whatever.

But the platform is effectively a collection of possibly increasingly open source services, of which they would plug in their various models and other services. So, it comes back to that, that idea of orchestration.

you know, obviously Omniscien, honestly has a platform similar to that, which is why I've been so interested in this business for so long. Um, plus Dion, a good friend of mine, um, that the notion of being able to orchestrate and pull

all of these systems together and take what you need from the open source world and take what you need from the commercial world.

And run that locally or run that.

As you know, we I am seeing governments talking at the highest levels about creating national clouds. India is going in that direction.

China's definitely in that direction. Australia.

I think we are way off target. But you know this this gives a much more layered approach.

And I think that there is definitely, you know, a vital role for AI to play in that.

But you mentioned academia.

None of this can happen without really skilled people. And currently the consolidation of skills globally is a real hindrance, especially around the cloud space.

So, to answer these questions, uh, I hope that's given some, some, some additional insights and thoughts on around it. There's a lot there.

But look we've only got a couple more. I've only got a couple more minutes unfortunately.

So, if you do have a pretty good time for one more question from the audience, but, um, I'm gonna ask a really broad question. What do you think is the next big technology, AI or otherwise, on the horizon?

What's the next big thing you see coming up?

00:52:03      Philipp      I mean, there's a lot coming up because there's so much happening. There's a lot in, like, making good use of all that technology.

so I'm involved in a project where we like basically thinking about how to use it to improve how we do science. How do you find related papers, how you can prove that something is true or not.

But, you know, it's a lot of work. Processes that exist that involves knowledge can be potentially sped up a lot.

And just in that broad application space is going to be a lot happening that it's interesting.

Um, in terms of the core models, I mean, one thing I'm also very excited about is kind of the merging of modalities of text, speech, vision.

right. So, uh, so just to bring up machine translation, um, it used to be very text based.

I mean, people have worked on speech translation as well, but it still hasn't reached kind of the same perfection level. I mean, there's so much more problems with speech translation.

- And then ultimately what you really want to do is to have also the video translated. So, if someone like if I speak, you know, in English, I want to have a live translation in Chinese, where then I guess just match up the right place.
- So, when I say this word, then in the translation that that gesture should be also at exactly that same spot and lip movements have to match up and all that is actually important to make it seamless. And that is kind of still I mean, people are working on it, but it's that is more on the on the research horizon.
- 00:53:35 Joe Okay. Um, you got 30s on this one.
- This is from Miri. Do you think there's any need to stop AI developments, uh, or slow them down?
- I think it's probably a better way of putting it and define a structure and laws that should be followed internally to safer future AI development.
- 00:53:56 Philipp Yeah. I don't believe that the air is so powerful that is going to escape the box and rule the world.
- It is a it is a tool. And you decide how much you rely on the tool and how much you trust the tool blindly.
- And if you trust the tool blindly, then obviously that can lead to problems.
- 00:54:17 Joe Right.
- I think that you got that in 30s. Awesome.
- another question that's just come in from James. Um, and again, 30s snippet, uh, impact on employment.
- What's age impact on employment going to be?
- 00:54:34 Philipp I mean, the dream is always that. Then the AI does all the work, and we only work 2 or 3 hours a three hours, 2 or 3 days a week and go, uh, fishing or whatever you want to do.
- Um, I mean, currently doesn't seem to have this big impact on mass unemployment because there's once you have more technology, people want to do more things and there's more demand for products and services.
- 00:55:02 Joe Yep.
- 00:55:03 Philipp Um, but ideally, yeah, I think this is a much, much bigger political question than a technical question about how to kind of ensure that the benefits you get out of the technology are broadly shared, which might, uh, you know, probably only work half as long and still make the same.
- 00:55:24 Joe Yep. Yeah.
- hopefully more. Hopefully more.
- but, you know, one of the things which I keep on reminding people is that by many metrics, a lot of industry, not all of them, but many industries actually have, um, employment congestion. So, this may only be some of that.
- Certainly, things like local government.

- Uh, a lot of the health services and so forth. So, but roles will change.
- 00:55:51 Joe But anyway, look, the last question, um, and this this one's a personal question for me. What's it like teaching the new generation of AI students?
- 00:56:08 Philipp I mean, it's I mean, this is like all the generation that kind of grown up with this stuff and used it.
- Um, yeah. I mean, it's kind of always a young generation with interesting ideas and different takes on it.
- Um, I think everybody currently is a bit overwhelmed by the speed of progress.
- Um, so I'm especially when I teach graduate students and PhD students who want to find research topics that like, uh, I think it's been already been done. And uh, so it's a bit unnerving at the moment, but it's I think it's all exciting times.
- It's, I mean, interesting stuff happens all the time. And that's just.
- 00:56:49 Joe Wonderful. Look.
- Thank you. We're getting a whole bunch of people thanking you for your valuable insights for this.
- It's been a fascinating conversation. We've gone from some really deep technical issues, um, all the way up to the big social issues.
- So, I really appreciate your time and talking through those. It's been an absolute pleasure.
- Dion, I'm going to hand it back to you.
- 00:57:12 Dion Thanks, everyone, for your time today.
- If you have any further questions that you would like to ask, you can email us directly. Or just send it to [sales@omniscien.com](mailto:sales@omniscien.com), and I'll pass them on to Philip or Joe and make sure they get answered.
- We'll be providing the full transcript of this call in the next couple of days, and also the replay video.
- You'll be able to either download or watch it again, and you'll also be able to share it with your friends. Thanks very much, everyone, and we'll see you next time around.
- Bye bye.