**CRAIG:** Hi, I'm Craig Smith and this is Eye on AI. This week I talk to Edo Liberty, who helped create SageMaker while at Amazon's AI labs. Today he's building out long-term memory for large language models using vector embeddings.

**CRAIG:** He says it solves the problem of hallucinations in large models like ChatGPT, which as we all know is their primary drawback. While my previous guest, Ilya Sutskever struggles to train GPT-4 and ChatGPT not to hallucinate using reinforcement learning with human feedback and Facebook's Yann LeCun works toward building a world model that would ground large language models in reality, Edo's approach is far simpler: Convert authoritative and trusted information into vectors and load them into a vector database for large language models to refer to, allowing them to give accurate answers.

**CRAIG:** The term embeddings was originally used by Yoshua Bengio in 2003, and in 2013, Czech computer scientist Tomas Mikolov created word2vec, a toolkit for creating vector representations of text that can be used for downstream deep learning tasks.

**CRAIG:** Edo worked on vector embeddings at Amazon before leaving to start Pinecone, which offers vector database services.

**CRAIG:** We have a new sponsor this week, so please listen up. [NetSuite](http://netsuite.com/eyeonai) by Oracle is a cloud-based enterprise resource planning software to help businesses of any size manage their financials, operations, and customer relationships in a single platform. They've just rolled out a terrific offer for the first time in [NetSuite](http://netsuite.com/eyeonai)'s 22 years as the leading cloud financial system; you can defer payments for a full [NetSuite](http://netsuite.com/eyeonai) implementation for six months. That's no payment and no interest for six months, and you can take advantage of this special financing offer today. [NetSuite](http://netsuite.com/eyeonai) is number one because they give your business everything you need in real time, all in one place to reduce manual processes, boost efficiency, build forecasts, and increase productivity across every department. If you've been considering joining the 33,000 companies using NetSuite, the deal is unprecedented, no interest, no payments.

**CRAIG:** Take advantage of the special financing offer at [netsuite.com/EYEONAI](http://netsuite.com/eyeonai). That's EYEONAI all run together. That's [netsuite.com/EYEONAI](http://netsuite.com/eyeonai) to get the visibility and control you need to weather any storm. Again, that's [netsuite.com/EYEONAI](http://netsuite.com/eyeonai). And now without further ado, let's hear from Edo.

**CRAIG:** Why don't you start by introducing yourself, your education and, career development to what you're doing now. And then I'm interested in hearing about SageMaker. I'm interested in hearing about the work that you've done on vector embeddings and how that relates to what you're doing now. And then I'll start asking some questions regarding large language models.

**EDO:** I started with PhD in computer science at Yale focusing on theory of computer science and algorithms in machine learning, focusing on what was then big data. The underpinning of which is the same the names have changed, but the data hasn't gotten any smaller since. Did my postdoc at also Yale in in applied math focusing on numerical linear algebra which is in some sense the underpinning algorithmic components of Pinecone today.

**EDO:** Opened my first company, then sold it. Joined Yahoo as a director, joined Tel Aviv University as an adjunct professor. Taught for three years. Moved to New York with Yahoo to manage their AI labs.

**EDO:** Back then Yahoo was the epicenter of innovation on that topic. That was quite a while ago. I was, I'm dating myself here. And then Swami Sivasubramanian, who's now leads all databases at AWS back then was starting a new group that had no employees to build machine learning services out of AWS. There was nothing there, right? We grew that org from basically three when I joined to about 2000. Pretty quickly released a lot of services and, platforms, including SageMaker. And in 2019, middle of the year I, left to start Pinecone with the understanding that large language models specifically and in general, this new kind of way to represent data by deep learning models is going to take hold and is going to become a fundamental building block in, in, data and AI and for somebody running in the trenches, that was the writing was on the wall, but for many, that took a few years for them to really catch on.

**EDO:** And now with the explosion of ai, a large language models in general, and specifically chats and Bard and ChatGPT and others. It's like a full-on explosion, and we see that in our own adoption numbers.

**CRAIG:** What is Pinecone?

**EDO:** Pinecone is technically it's called a vector database.

**EDO:** But practically what it is that if you are a deep learning model or a, uh, language model or a chat engine, right? The way you represent text is not with words or structures or grammar trees or whatnot. You, have an internal representation that is numeric in nature.

**EDO:** Machine learning models are numeric engines, right? They are mathematical objects, right? And they deal with numbers. And the only way that they know how to represent stuff is with lists of numbers. And those lists of numbers are called vectors or embeddings. and if you want to represent data to a deep learning model, for example, to give it long-term memory, context, the ability to actually remember stuff and know something about the world, the way you represent it is with those numeric objects, with those vectors.

**EDO:** Interestingly enough the storage layer, the query language, the access patterns, everything that we expect from databases don't really work well for those objects. And you require a very specialized piece of hardware and software. And we provided as a service to allow you to do that at scale.

**EDO:** So, if you are trying to augment, say, ChatGPT with great context, so you get great answers. And you want your context to be your own data, right? Your own user's manuals or your own history, or your own images, or your own text, or your own Jira tickets or your own emails, and so on, you need to store them in a very specialized database that something like ChatGPT can access.

**EDO:** And know, they just released their own plug in for retrieval engines. Now you can actually plug those things in under the hood and have ChatGPT search Pinecone in real time, get more accurate answers and, get back to you without you actually having to do a lot of wiring.

**CRAIG:** Does that help take care of the hallucination problem?

**EDO:** A hundred percent. That's exactly what it takes care of. If you ask ChatGPT today how do turn off the auto reverse brake lights on my C 70 X model, right?

**EDO:** It will say something, right? It will say something that sounds very eloquent, but it will all be hallucinated. It very confident the answers something that is completely hallucinated.

**EDO:** And by the way before I tell you how I fix it, I want to say this is in and of itself a breakthrough.

**EDO:** I'm not trying to knock down language models. For somebody who's been in the field for 15, 20 years, this is nothing short of amazing, right? The language, the coherence, the way that it fits together, the amount of data that got sucked into this huge engine it is very, impressive. But let alone hallucinated, which makes it a lot less valuable. And if you feed, for example, the kind of Volvo's user manual into Pinecone as vector embeddings and let ChatGPT search through them in real time and get the right context for the answer.

**EDO:** It can figure out, Hey I, actually know exactly what you're talking about, and I can answer now with actual data because I know the answer.

**CRAIG:** That's interesting because I had Ilya Sutskever on the podcast and his focus is, really on reinforcement learning with human feedback as a way to, to train the model, hopefully to behave as he puts it to, not hallucinate. But that just doesn't seem like a very efficient way to do it. and I think a lot of people have doubts as to whether or not the model will learn to ground itself in reality. I had a conversation with Yan LeCun about building a world model that then a language model could refer to.

**CRAIG:** what you're talking about is, kind of building a domain specific world model through Pinecone

**EDO:** When you study medicine, you study a body of knowledge. You don't study medical English, right? You, come into medical school knowing English and you gain a body of knowledge and now you can discuss medicine hopefully intelligently, right?

**EDO:** And What you have gained is not a tweaking of your natural language model to the medical domain. You've gained knowledge. You've gained memory and, facts, right? And now you inject that knowledge into language, right? The same thing is happening when you pair OpenAI in Pinecone, like you save the data and the knowledge in Pinecone.

**EDO:** And that's a domain expertise thing. That it could be your Jira tickets or your, sales calls or your what have you that you want to converse intelligently about. And then you have the language model whose job is to take the knowledge that you have and, put it into language and summarize it and understand questions and so on.

**EDO:** Those are completely different functions. It shouldn't be the same mechanism even, right? The fact that, yeah, if you tell me something Knowing whether it, like language-wise, computes, like whether I understand what you say and whether I think what you say is factually correct in the world.

**EDO:** Those are completely different mechanisms. One of them is about language processing and the other one is about memory. I can push something with a stick, but I, language wise, I can also push it with a rope.

**EDO:** But the fact that you can pull things with ropes, but you can't push things with ropes is, an, is a world knowledge thing. It's, a memory thing. It's oh, wait a second. I've tried that before. I didn't work you know. Now the interesting thing about

**EDO:** large language models is that people like talk about the world in a way that is feasible. So, people don't, talk about pushing stuff with ropes. And so, if you train a large enough language model, it doesn't talk about the world in that way, which seems intelligent, right? But that doesn't mean it, he knows anything.

**EDO:** Remember that movie? Catch Me If You Can, the guy that like you know, that Masqueraded as a as a doctor. So, it's, the same thing we talked about before as medical school. You can just hang in a hospital for five years and you'll pick up on a bunch of stuff.

**EDO:** You might sound like a good doctor, but you might know nothing. Right? And so, some of those solution hallucinations are like that. It's like people who just heard a bunch of language, and they sound smart, but they really don't know what they're talking about. That's for me, that's hallucination.

**EDO:** It's just heard doctors speak so much. You can speak, medical English, but you don't actually know medicine.

**CRAIG:** Yeah. This adding a memory. So, Pinecone is doing that.

**CRAIG:** Are there other mechanisms that are being used to add this memory or domain knowledge? I, know that with, large language models you, pre-train them. In an unsupervised manner, but there's a supervised layer of training at the end.

**CRAIG:** Can you give a large language model that domain knowledge in the supervised training portion of the exercise.

**EDO:** So, I'll separate what can be done and what is often done. Okay. A lot of things can be done. You can train things in all sorts of fancy and exciting ways, and we've tried a bunch of stuff and so on.

**EDO:** I can talk for five hours about all the stuff that's possible and how you eek out more accuracy this way or that way. What I will say is that you really don't have to do any of this to get something quite impressive and pretty good, right? And some of the new excitement, I think, in AI is the fact.

**EDO:** I think for the first time since kind of the beginning of this field, this has become truly accessible to the average engineer, right? Or even the end user, right? Like people can actually see this thing in action actually harness it, right? And get, they get excited about the kind of stuff they can build with it, right?

**EDO:** And so, when you combine Pinecone and OpenAI, for example you don't have to retrain anything. You don't have to reinvent anything, you don't have to tweak, you don't have to fine tune, you don't have to collect labels, you don't have to do any of that stuff, right? You basically feed your data into Pinecone in the form of this vector embeddings.

**EDO:** And again, we have plenty of examples online, a lot of examples that are not from us as well. This is like a well-known practice. And you basically in real time ask OpenAI, hey just search Pinecone for the context. All my data, like if I'm a salesperson and I've fed all my sales conversations into Pinecone

**EDO:** you can then ask questions about, oh, did I offer that customer a discount. And, if so, what was it? I don't remember. Now you can have OpenAI or any large language model provider, basically reach out. Grab the right parts of the conversations that actually are relevant for the answer, and you can get the answer right, and you can actually point back to the thing, Hey, if you're, if you really want to be sure here's the actual email that you sent the customer that they said, Hey, you're going to get a 20% discount, blah, blah, blah, blah, You don't have to retrain anything. You basically give the language model the context that, it can work with. And by the way, that's how it's called. When you invoke the chat APIs they include the query, which is the question, but you can also provide context, which is, hey, I think the answer is somewhere here.

**EDO:** Try to glean from that more information. And so, if you can give the right context, you can get a better answer, which is obvious in some sense. Like we humans are the same way. If you ask a salesperson, you ask them the question, if they remember what happened, they can put together a coherent answer. If they don't remember they a human would say, I don't remember a chat engine would just invent something.

**CRAIG:** So, with Pinecone, these vector embeddings that you're feeding in, can you walk me through the process of converting like medical texts because medical advice is one of the great promises of ChatGPT. But right now, it's so unreliable. You wouldn't dare rely on it for medical advice without the oversight of a doctor. But if you could have the language model refer to trusted medical documents or textbooks presumably then you, you could remove the doctor from the equation and the large language model could answer.

**CRAIG:** So how do you get the medical texts into Pinecone? Just very practically.

**EDO:** I'll start by saying that I don't necessarily recommend doing that, especially when your own health or your loved one's health is on the line. But nevertheless, I understand it's a brave new world and having a bad doctor is sometimes better than having no doctor at all.

**EDO:** So, with that caveat yet, I want to put the disclaimer out there that I'm not advocating for this being responsible behavior. Ok. Yeah. Yeah.

**EDO:** But it's actually incredibly simple. Like you, you take the medical domain textbooks or medical notes from doctors uh, in the hospital or whatever.

**EDO:** You break those into sentences or paragraphs. You feed those paragraphs into what's called an embedding model, right? These are language models that process the language and represent it in the way that deep learning models represent text which is this numeric representation called vectors or called embeddings.

**EDO:** Then you save those in Pinecone, maybe with pointers back to where you got them from or maybe the actual text next to it and everything that you need to make it useful. the but the actionable part of it is the numerical presentation. And in a real time when you get the query - you have some symptoms, or you have some medical question - you basically do the same thing. You say, hey this question, I can embed it, I can search the embedding, hey, give me relevant documents, give me stuff that is similar to this or talks about the same topic or is relevant to this question.

**EDO:** Right? Which is, a query to Pinecone. You get back, say a hundred documents or a hundred sentences of paragraphs, right? You tack those as context into your query and you feed them into the chat engine. And now you ask, hey, I have these sets of symptoms, and this is what I'm concerned about.

**EDO:** And by the way again, you don't say that as the user, like the system gets, the chat gets also, here are 50 different sources of data that we think is relevant to construct an answer from.

**EDO:** Putting this whole thing together, we see folks in hackathons in San Francisco and in New York and other places, put this stuff together in like a day.

**EDO:** Like it's so easy because you don't have to retrain anything. Everything is a managed service. You don't have to spin up hardware, you don't have to figure out a bunch of crap. You just call a few APIs and play with the data and tweak here and tweak there.

**EDO:** And literally within a day you have something pretty impressive going on.

**CRAIG:** Yeah. And let's move away from the medical example. I understand that there's liabilities involved there but, on that very front end, I've, and you'll forgive me, this is going to sound really ignorant, but it's the kind of thing that I wonder about as a non-practitioner.

**CRAIG:** I've got a history textbook that I trust. I think it's authoritative and I want to feed it into Pinecone. Do I put it through like an optical text scanner and the text is digitized and then ... I'm just curious about the pipeline.

**CRAIG:** How do you get the information into a vector embedding?

**EDO:** So first you have to get your data to be digital somehow. Whether you scan uh, the physical object or what have you or, a PDF or maybe it's already digital and you just have word documents or whatnot.

**EDO:** Then you take the text, and you break it into small parts. It could be sentences or paragraphs or something that is coherent. It's not like the whole book, and it's not like two words, right? It's, something that says something that you as a person, you read it and it says, oh, You When Napoleon invaded, blah, blah, blah and this and that year, this and that happened, right? And so suddenly you're like, okay, this means something. This is already useful as a unit of, knowledge, right? And that again, could be a sentence, it could be a few sentences.

**EDO:** Now you've broken it up that way. Now you can take those and embed them, take them, and put them through a deep learning model. Again, there are plenty of them online, people literally with a few lines of Python, you, pull them into your machine and you, run them or you use an API to run something on the cloud.

**EDO:** And what you get back is a vector embedding. That's it. You get back a list of a hundred numbers or a list of a thousand numbers. and that's it. That's your vector. And then you feed that into Pinecone and say, Hey, this here's, a factoid and here's the original sentence. Napoleon invaded, blah, blah, blah, blah.

**EDO:** And here's page number 712 in this history book. And like you put next to it a bunch of stuff that makes it useful for you. And that's it. Now you have your databank. And now you can build machine models on top of that and use that history book as the source of knowledge for your conversations.

**CRAIG:** And does Pinecone have a front end that does all of that? Or the user has to have the vectors ready to feed into Pinecone.

**EDO:** Right now, we are we are a pure database play, so we, don't have a dog in the fight on what machine learning models you use and what language models you choose to use.

**EDO:** We work with OpenAI and Hugging Face and Cohere and with Google has our models and Amazon and Azure and it's a brave new world and they're like, I don't know, like probably 50 or 60 startups being funded to build even better language models.

**EDO:** There are some that we think are better than others, but at the end of the day, it's up to our users to figure out which works the best for them. And it's up to us to build the tooling to make it the most convenient. But we expand in that direction.

**CRAIG:** I imagine a lot of people with the spread of particularly GPT-4 API are turning to this kind of solution rather than trying to train, as I've saying earlier, on top...

**EDO:** it's much easier for sure. Yeah. Yeah.

**CRAIG:** A, question that doesn't only relate to. Pinecone, but to, to large language models in general, but specifically to Pinecone.

**CRAIG:** If you're converting information into vector embeddings is there a copyright issue? How, do you guys deal with that or think about that?

**EDO:** So, our customers use Pinecone like a service, like a manage database, the same way they use Mongo and Elastic and Redis and Dynamo DB and Spanner and you name this it. We are their data infrastructure, right?

**EDO:** The data that we store and manage with Pinecone is their data. We don't touch it; we don't use it. We don't share it. We don't train models over it. It's their data and their data alone. We are the infrastructure that manages that for them.

**EDO:** There's no question about privacy of anything because we are their providers, we don't provide information. If they give us information that they don't own, then they're doing something. But we wouldn't know that.

**CRAIG:** No, I understand.

**CRAIG:** just philosophically as, information gets sliced and diced and sucked up by these models. Yeah.

**EDO:** I will tell you that one of the most, I think practical reasons why people prefer to use something like a vector database to create context rather than retrain or refine models, is because they can remove data, right?

**EDO:** And for example, you can be GDPR compliant, right? Because you can say, hey, I have all this data in here, but if I want to delete something, I have an explicit operation, I can just delete data point number 766. And that's it. It's gone. There's no, it's gone forever. There's no one bit of information from it anywhere it's done.

**EDO:** Whereas if you fed that data point into a fine-tuned model and you try to, like you, you try to somehow improve your behavior with that data. There is no such mechanism to remove data from a model. Oh. You remember three weeks ago I gave you that data point to train on?

**EDO:** Can you please forget that I showed you that? That doesn't work.

**CRAIG:** Not how it works. Yeah.

**EDO:** And the beauty of it is that you don't have to be a machine learning expert. Like you can follow some tutorials, some copy pasting, some creative tinkering and you, you'll get it working time.

**EDO:** Not only do you not have to contact us, but you also don’t even have to pay us. Pinecone has a very generous free tier that could enable you to do that pretty frequently on millions of vectors for free indefinitely.

**EDO:** You go to our website, you create your own API key, and off you go, it's free. Yeah. You don't have to do anything. You don't have to pay us. Not now, not in two weeks. Not ever. At some point you might become really ambitious and build something really big and that'll be very demanding or maybe your app becomes an overnight success, and you have a million users. And then the free tier's not going to be enough anymore.

**CRAIG:** So where, are you guys going with this? You just scaling up the infrastructure now that demand is going to explode.

**EDO:** It's, phenomenal growth. I'll just say that the hardest thing for us right now is just to spin up enough hardware to actually support all the demand. Yeah. That is literally what the core engineering team is, doing almost every day.

**EDO:** The demand has been overwhelming. And then ChatGPT happened and became like doubly overwhelming.

**CRAIG:** Yeah. Yeah. Wow. That's fascinating. Um, the, um, yep. Come in. after

**CRAIG:** I'm recording. Okay, thanks. yeah. yeah. One of the things that, that, that interests me because I'm not a coder is this automated code generation, and CoPilot of these tools are interesting, but they can't be relied on. Could this, could Pinecone be used in that way to improve the accuracy of auto-generated code. And do you think that could develop not simply through Pinecone, but that, that trend could develop to the point that there'll be just an explosion of software created because suddenly the bar is lowered dramatically for developers.

**EDO:** As an engineer myself I find it hard to be objective. I want to think that my craft is somehow not easy to automate and somehow code is hard to generate. But the truth is that a lot of code is not hard to generate.

**EDO:** The, yes, like every piece of software would have one or two or 5% of it that is interesting and hard and intricate and somehow insightful. And a lot of it is not that deep or not that exciting too. And I want to embrace it. I want to say, hey let's, make this an amazing tool for developers and make developers a lot more efficient. I think we are; I think that's already happening. I think it's already becoming a very good tool for developers. I think we're quite a way off from stuff actually writing complete software end-to-end with no assistance. I think a lot of these, like the same way you have hallucinations, you're going to have stuff that compiles, and pseudo makes sense, but doesn't quite, so I think it's going to be a very long time before we get to the point that we, that that is done to that level. But it's definitely like going to disrupt the software industry in one way or another.

**CRAIG:** That's it for this episode. I want to thank Edo for his time. I also want to thank [NetSuite](http://netsuite.com/eyeonai) for their support. If you help manage a company, check them out. They give your business everything you need in real time, all in one place, to reduce manual processes, boost efficiency, build forecasts, and increase productivity across every department. Don't miss this unprecedented deal. No interest, no payments. Take advantage of this special financing offer at [netsuite.com/EYEONAI](http://netsuite.com/eyeonai). That's EYEONAI. It's important that you add the EYEONAI, so I get credit for your interest. [NetSuite](http://netsuite.com/eyeonai) will give you the visibility and control you require.

**CRAIG:** Meanwhile, if you want a transcript of this episode, you can find one on our website that's eye-on.ai. Review us on any podcast platform you listen to. Reviews really help in expanding our visibility and attracting other listeners. And finally, remember, the Singularity may not be near, but AI is about to change your world.

**CRAIG:** So, pay attention.