**Clemens**

One of the key aspects of our platform that goes beyond just language and especially large language models is that it's already multimodal. If you get a document, sometimes you want to extract things like a logo or a signature or a stamp or a figure. So, we also run convolutional neural networks, just basic object detection models, to extract those images, and our goal is and our ambition, of course, with the iHub is that we will cover many more modalities as well. So, we're really shifting towards content in general. Any text document, any images, audio video is all a great game for us in terms of being able to capture it in the future.

**Craig**

Hi, I'm Craig Smith and this is Eye on AI. This week, I spoke with Clemens Meywald, an expert in AI and enterprise software, about solving unstructured data challenges using AI. His company, Instabase, enables users to engage in interactive conversations with various types of content, such as documents and spreadsheets, and he introduces us to the AI Hub, a marketplace where developers and users can contribute models and applications. I hope you enjoy the conversation as much as I did. Now here's Clemens.

**Clemens**

Thanks for having me. I guess the very short version of my background is that I'm originally from Austria. I went to school there and got engineering degrees there, then went to MIT, Sloan, to get an MBA and that's what led me to Silicon Valley. So right after Sloan, I joined the Google Brain team and I guess my claim to fame there was that I was the first product manager on TensorFlow.

I joined in 2015, just when TensorFlow came out and when it was open sourced, and then over the next four years, I helped build Google's modern infrastructure, build TensorFlow, build TensorFlow Extended basically how Google is doing AI today. Then, after this, I joined a company called Databricks to help build the data science and mechanical businesses there. Of course, you know, Databricks started with Spark, and I helped build a lot of the product portfolio on AI platforms. Then, late last year, I switched to a company called Instabase, which is yet another category more focused on unstructured data, which is, I guess, the big shift because, of course, both Google mostly, but of course Databricks most of the data, of course that people operate on is structured.

**Craig**

Yeah, tell me about Instabase. And you've launched something called the AI Hub. Is that right? Maybe you can talk about that as well. But first about Instabase. It wasn't clear to me. You're sort of grouped among the OCR crowd. Is that what you do? or give me an introduction and maybe a little history of Instabase.

**Clemens**

Makes sense. OCR is a very dusty word, by the way. But yeah, I guess also like a very brief history also how I ended up here. I guess I actually first met the CEO, Anand, in 2019 through a mutual connection actually someone at Index Ventures And she knew that I was interested in enterprise software and just applications for enterprise in general and she said you should look at this company, it's very interesting. And then I had this mind meld with Anand, the CEO, back then because my vision for how data driven applications should work in the enterprise was perfectly aligned with his. But I just started at Databricks as it was just bad timing.

But fast forward three and a half years, like here I am, and I guess the story of Instabase is interesting because it was conceived off as a platform for enterprises to run data driven applications. That's a very broad statement, right, and it's a very broad product to try to sell and product market fit was found initially in what you would call document processing, like intelligent document processing, because it turns out actually one of the biggest problems in these big enterprises is actually unstructured data. You know there's a lot of attention with the snowflakes and the Databricks is of the world on structured data, so anything that you would fit into a spreadsheet or a table. But it turns out that, anecdotally speaking, and there's also some research, that about 80% of the data is unstructured. So that's in the documents, audio video, like anything that's lying around That's not in a database, and that turns out to be a very big pinpoint for big enterprises. So Instabase started getting into using AI to process unstructured data, most importantly, documents to solve very high value business processes, mostly for financial services, insurance and big retail companies.

Just to give you a couple of examples you know, in large banks that do market or an origination or approved mortgages, when you apply for mortgage, you send in 100 PDFs that are your bank statements and your like poker statements and your W2s, and some poor soul needs to open each one of them, look at what is the document. What information do we need to get out of it? Some of them are not actually digital PDFs. Some of them are like pictures taken with a phone, so you need to extract all that information out of there and then process it. So, this is really where Instabase found product market fit and that's that category of problems is where we've seen early success and have grown significantly over the last couple of years, mostly, of course, with the application of modern AI techniques.

So Instabase was, as far as I can tell, actually one of the first companies that successfully applied transform-based models, real business problems, bird type architectures, and now with Generative AI, of course, the whole world is changing.

It turns out that you actually no longer need to annotate data and I guess we will get more into this later But the AI product that we're launching, or that we did launch this week, is the next iteration, if you will, of our product and how you actually approach to solving these problems, and that's one of the key aspects of Instabase as a company.

Actually, a customer of ours just told me the other week using Instabase is like having a subscription to innovation, because we keep innovating and we keep, you know, like adopting like the latest in the I and the I have is really like applying generative models, like, specifically, GPT, to the document problem And what's actually quite interesting and what we found is document is a very narrow. We actually like getting away from using the word document content more broadly, because we actually support spreadsheets, html, you know, like emails, pdfs, like a hundred nodes and, of course, code and like any other like artefact that contains information and actually allow you to utilise that to extract value, and part of the big investment that we're making is to actually make this more accessible, because it used to be that starving an answer to data problem is a multi-month task. That's very intensive. Like you have to collect data, annotate data, train models and actually deploy them, and we find that that's no longer necessary to generate. If you want to stop there, Yeah, well, yeah.

**Craig**

So, the Instabase, its first product offering was when? and maybe describe how the platform works. Is it a web app or is it, you know, a piece of software that sits on the client's computer and to use it? presumably you've got somebody in an organisation who is wrangling data and maybe give me a use case where they Upload a ton of data to the platform and then query it on particular things.

**Clemens**

Yeah, definitely. Yeah, there's a. There were a couple of really good questions in there. So, the first one is since the base as a company was founded in 2015. And the product as it is used today is a little more recent, because initially, Instabase tried to solve the document problem In the early days, just like everyone else with heuristics and like rules, right, so like Iran and ocean engine, and then like applied scripts, basically on the text, and then, I believe, about Like actually like late 2019, like 2020, we started using transform this models like bird to solve the problem.

So the product as it exists today, or as customers using it today, is roughly like one and a half two years old, and how it's being consumed is actually quite interesting question because I mentioned earlier, like the mind melt, if you will, between me and the founder was how we envisioned that data intensive applications would actually be Used and deployed in enterprises, and one of the biggest problems in that space is actually portability of applications. Like you, if you write an application for, like, let's say, AWS, you can't run it elsewhere because there's no operating system, if you will, for these applications. So Instabase was actually conceived of as an operating system that provides operating system type guarantees such as portability, abstractions for storage, compute and identity, so that you can build these applications in one place and deploy them in another. The reason why I say this is you can probably guess from the types of customers we serve. So, it's like the top tier, like banks, financial services, and insurance is they're very sensitive when it comes to security and data. A lot of them want to run these production workloads on prem in some cases, like literally in their data centres. So Instabase runs both as a SaaS offering. So, we have a SAS offering. We just sign up and use the product as a cloud hosted product so our customers can maintain it themselves in their Azure, GCP or AWS instances and also on premise. And the interesting aspect of the product is because it provides operating system level portability guarantees if you can develop in SAS and deploy on prem or like any other combination. So, it's actually quite an interesting value proposition, especially for highly sensitive data where, like some of our customers, we really only want to run the production workloads on prem

And then the last question that you asked was what are the types of use cases? or, like you know, how can you imagine this working. One I alluded to briefly earlier is mortgage processing, right? So, you can imagine, if you apply for mortgage, it is then a whole bunch of PDFs. The output eventually is all of that extracted data in a structured format in a database and someone actually likes making a decision to give you a mortgage or not, right? And along that way. There's a lot of things that happen in terms of recognizing that information from documents, extracting it, refining it and validating it.

Our customers also like to detect basic kinds of fraud. It turns out when people are going to defraud banks. When they apply for a mortgage, they do simple things like changing the income in their bank statement, but if we get to update the sum in the bank statement so you can actually check that just by checking the maths. So that's one example. Another example, just to show you the diversity, if you will, which is recently Close.

I can’t disclose the name, but it's a big right sharing company and if you imagine, like every time they sign up new drivers to the platform, they need to take a picture of the driver's licence, of their car insurance, like registration, and like vehicle inspection, sent that to the company before they can start driving for the right sharing company. And that's another example where you need to take these pictures and it's an interesting example because, like that's actually an example where the quality of the pictures is usually pretty poor. It's really just taking a picture of the driver's licence, like in the dark, exiting at home. But you want to as much as possible, automate that process right, because if you have hundreds of people signing up every day, you don't want humans having to look at those pictures and extracting all of that information. So, it's really from a foundational capability pretty broad, and we started with documents, but we also have customers that do mailroom automation right. So, like you can say, like in a big enterprise, you get a whole bunch of mail that goes through scanners and then Hits our system, and we classify a type of document and then send it in different directions with extracted information.

Based on this, and one of the key aspects of our platform That goes beyond just language and especially large language models, is that it's already multimodal, if you can imagine. You like, if you get a document, sometimes you want to extract things like a logo or a signature or a stamp or like a figure. So, we also run convolutional networks, just basic object detection models to extract those images, and our goal is and our ambition, of course, with the IHUB is that we will cover many more modalities as well. So, we're really shifting towards content in general. So, any text document, any images, audio video is all for the game for us in terms of being able to capture in the future.

**Craig**

And the AI Hub is. Is that? From what I understand, it's like a marketplace of AI models or products? Are those Instabase models or products, or is it a platform that anybody can upload a model or product to, or do you curate models and products for specific use cases and put them on the platform? Yeah, yeah, that's a really important question. So, I'm going to tell you about the long-term vision.

**Clemens**

And then I'm going to tell you what exists today. The long term vision is really and like this is actually how Instabase was conceived that the AI Hub is a community of developers and users, where people contribute models, contribute what we call apps, and then, like other users can consume them and it becomes this like flywheel effect of like people developing AI driven applications, people consuming applications and it really becoming like an epicentre, if you will apply the I, and so the goal really is that third party developers can publish your own applications. It can be used by the audience of an AI Hub and there's like a revenue share model behind it. What it is today is three different components that we built. So right now, it's seated by us, if you will. So, the headline app, if you will, that everyone is paying attention to right now is called Converse, and that is an interactive Q & A-like chat type app that lets you have a conversation with content. So that's the key distinction between that and something like chat GPT, right. So, again, chat GPT, you're conversing with the model itself Right In our product. You upload content, documents, email spreadsheets and then you can have a conversation with that content, which is quite interesting.

The second component is what we call build, which lets you build reusable workflows or apps that automate this right.

So, like, let's say, if you wanted to extract information from, like a passport or like a driver's licence or something for identity verification, you don't want to do that in a Q&A type form, right, you want to actually like, develop an app where you can just send 10,000 PDFs and it gives you the structured output.

So, with build, you can actually build these reusable apps that automate these workflows. And then the third component is that app marketplace where we have seated the marketplace with 10 people, that location that we've built, and we will keep adding more and more over time. But we're also now heavily working on the ability for you to be able to publish your apps, because you can already build your apps. So if you go to the I hope today, you can build an app to, let's say, summarise a document, right, or like, pull information out of, like an insurance contract, or you like to even translate a document, and you can publish that app on the I hope, and in the future we will have a review process where we can take a look at it, make sure it's secure, make sure it meets our requirements and then is actually discoverable and usable by anyone who comes to the. I hope so.

**Craig**

Are you familiar with Singularity.net, Ben Goertzel’s organisation? He's kind of a wild figure out there on the fringe, but he's been talking for years about a blockchain based marketplace for AI models or products that that you know, and you know Anybody could access and then, using smart contracts or You know, use them and pay for them and that sort of thing.

**Clemens**

Yeah, it's a really compelling idea what we found, actually and there's a key distinction. By the way, I kind of Answered a slightly different question than you asked because, like you, we were asking specifically about models and products. So, I think when you use the word products, that would be that's what we're referring to as an app that really is targeted towards solving a real problem. Models, on the other hand. So, we also plan, of course, because internally we have a model repository of actually proprietary models that we trained on top of the bird type architectures that are layout aware models, so they will also become part of the, I hope.

But what we found is that model marketplaces, specifically, are not really that useful specifically to businesses right in the real world use cases, because models are just small building blocks of something that really provides value to you like no business can go to let's see how can face and find out what you're doing And you know, like the latest model and just like create an endpoint and start using it, because any meaningful business process or like business solution usually has multiple models that it uses, so has some business logic in between, has like integration points, right.

So actual apps that solve real problems Are much more than just models, and that's really what we're focusing on and like what differentiates our eye hub from what you would refer to as like a model repository, which is really more focused towards, like data scientists, or actually I would say, hobbyists and academics, because it turns out these days I just published a blog post yesterday on this Most of these models that are being open source are actually like, have restrictive licences so that it can really just be used for hobbyists and academics, but it's not really relevant for businesses.

**Craig**

What are some of them? What is some of the competition? and in In AI app marketplaces or repositories, or however you want to refer to this?

**Clemens**

So, there's a couple of different components to this. If you look at the apps themselves, right, like the, solve meaningful business processes anchored around unstructured data, there's a couple of IDP-type companies out there that will sell you these types of solutions, right? So, like, let's say, there's an app that says identity verification, that takes passports and, like private licences, make sure that the names match, make sure that the driver's licence, the birthdays dates match and gives you the structured information. The big difference, I think, between our AI HUB and some of the competition out there is number one a lot of the competition actually requires you to like to sign up, like for a very heavyweight, like enterprise software deployment, and it takes six months to like to deploy it in your cloud architecture and then run it.

Our AI Hub is now really like SaaS in its fullest form, which is the app that runs in the AI Hub, you can actually right now go to https:/aihub.instabase.com/, upload a whole bunch of documents and get the output, and it has a pay’n’go type pricing model where you just pay for what you use, and certainly the accessibility and how easy this actually uses is one key differentiation between a lot of the enterprise IDP offerings, if you will. On the converse side, right like where we're talking about Q&A with documents. There's also competition that's starting there. It's a little newer, right? Because like that's like the form factory, if you will, like asking questions of your documents. Just came up with the popularity of chat, GPT And there's a couple of small niche start-ups that have started in this space And, of course, the cloud managers are all getting into this business as well.

**Craig**

I was just going to stay on the cloud. Did you say the cloud providers are getting into this space? I mean there's an AWS marketplace. How would that compare to what you guys are doing?

**Clemens**

So yeah, so AWS, of course, AWS especially a good example, because they always have 20 or 30 different services that are slightly overlapping with what anyone is doing. So, the AWS marketplace itself is more of a cloud infrastructure marketplace, right. So, like, and Databricks is in the AWS marketplace, right. Instabase is in the AWS marketplace, so it's not at the level, if you will, of meaningful business applications. It's more on the infrastructure level, if you will. Aws, I believe, does have a more like low code, like app development type, type product similar to Microsoft's, like Power Apps, right, so I think, like Azure's, like Power App, probably like a good comparable there, and all of these also have their own like marketplaces and like try to make the exchange of these applications easier.

Most of them have been developed as just low code app development platforms, right, it's like, not with the idea that they're actually like data driven or AI driven, and that actually makes a key distinction, right, because, like what I found when I started to get interested in the space, if you tried to allow people to build arbitrary apps, you would focus. Really, it would start looking like Microsoft Visual Studio, right, like, just like building a forum, like writing some custom code and that's it. But it turns out that data driven applications are much harder to build right. So, like, by data driven application, I mean any application that relies on, like processing a large amount of data in the background, running AI models. That's a different ballgame entirely in terms of actually like developing and like launching these apps, and that's what we're focused on, and I think a lot of the other marketplaces that have been out there for like low code business apps are slowly transitioning into that space because they realise that that's quite important and that anything will be data driven.

**Craig**

Yeah, and then for someone to use these apps. Well, first of all, from the app developers’ point of view. I mean I can see the usefulness of this. We've been talking, my friends and I, that with the advent of generative AI there is just an explosion of SaaS offerings and apps and all sorts of things Leveraging that. You know transformer-based technology, and each one is out in the market shouting for attention and they're different.

You know websites that you know Trustpilot or product hunt or these various sites that rank and review, but if you want to sell the app, your kind of on your own. But you know a company that comes to mind that I know fairly well is called Akkio And they have a no code web app that that you can upload a structured data to and then pick different kinds of AI models to run against it And, for example, predict, predict something or classify something, and they're out there selling it on their own. But if it were on something like Instabase, on the AI Hub, you know, provided that you get the critical mass of users, you know people see it and they can use it directly right there. And then what you would take of a commission or something, how does that work?

**Clemens**

Yeah, so the plan is just like any other common app store, right, that there's like a revenue share model and like we would take part of part of the revenue, but the majority of the revenue would, of course, accrue to the developer of the app itself. But you bring up a really good point, which is the whole idea behind a centralised marketplace, and I dare to say, like the word operating system. Behind it is the point of like distribution and centralised trust. Right? Because if you think about it and like, this is really the value proposition that we're trying to push, which is we have been successfully selling into some of the most demanding global enterprises in the world. It's like four of the five top banks in the US are our customers And, as you can imagine, selling into these enterprises is not easy.

It's like infosec and approval processes and so on. So, the whole value proposition of building on the IHUB would be that, yes, you could start an entirely new company, build your app from scratch and then try to sell into one of these like distribution channels, and it will take you like a year, one and a half years, before you could even contract with these types of companies. Or you could build it on the Instabase AI Hub, and we already have the distribution and like there's already a commercial model set up. There's like identity, there's trust, right, it's like these companies trust us and like we spend a lot of time on the security and compliance side. So, it's a much more compelling story for app developers and that's where that flywheel effect would come in.

**Craig**

But do they have to build the app on the AI Hub, or can they put an instance of their app on the IHUB?

**Clemens**

So yeah, so right now, the scope of what you can run on the IHUB, if you will, is limited to what you can build on the IHUB today. So, you can just take arbitrary code and just run it on our platform. But it's pretty flexible, or at least today. The version today is somewhat scoped to specifically content understanding, but the software in the background is flexible enough that you could really build anything. Just to give you an example, the platform itself, and like the way that a lot of our customers use Instabase today, follows a generic pattern of classifying unstructured data and then extracting information from it.

So, like most of the apps, if you like it, subscribe to my emails. Or like to read all of the emails. And like classify if this is a complaint, if someone is applying for something, read the information, like send it to the different systems. But on the same platform. We built this Converse app, which is an interactive Q&A like chat solution. So, the ability to build anything on the platform exists and like we've proven it, if you will, ourselves by building this chat app. But yeah, it does require, like building on the IHUB and not just like running anything Right.

**Craig**

A couple of things. How do you ensure the security of the source code if someone is building it on the hub?

**Clemens**

So, there's a couple of aspects of this. One of them is because, at least to date, a public version has no code. Actually, it's not even low code, it's no code. So, you can actually write your own source code. We basically just configure an application and then we run it. So, we have full control over what that code looks like.

The product that we have, and like where this will evolve, will actually allow developers to, of course, run arbitrary code. And then at that point there's like app level isolation. If you think about multi-tenancy in these systems, you have like process level isolation, you have container level isolation, you have hardware level isolation. So there's the appropriate level of isolation to make sure that one app can't escape their boundaries and access like anything else, if you will And a lot of our customers, by the way, also would run this as a single tenant instance, if you will where this whole like portability story comes in, so you could actually, in the multi-tenants as environment, have someone develop an app but then if one of those big banks wants to run it, they run it in their own single tenant instance of the IHUB And like in that way, like they can ensure that there's like no intermingling of data with other customers.

And, by the way, the other aspect that I want to call out because it's extremely important and we've spent a lot of time on this, is the bigger concern actually for companies today is more the data security and privacy around the data that goes to LLM providers like OpenAI, and so we use OpenAI in the background and we actually have a very close relationship with them And we put all of the controls into place for our AIHUB product to get an exception from OpenAI for the data retention policy.

So OpenAI by default usually retains 30 days of data to detect fraud right in, like two for QA, and we have an exception to that data retention policy so that any data that comes through the Instabase AI Hub is not retained at all by OpenAI. They don't use it to train models, they don't look at it for QA, it's really just processing it in the model, returning it and it's gone. And that was extremely important, especially for enterprise customers to be able to trust its solution, because there's a lot of noise out there as well in terms of are you giving up your data and what is your data being used for?

**Craig**

and so on. So, if I have an AI based app, I can't put it on Instabase AI Hub. I could go on Instabase AI Hub and build something similar using your tools. Is that right?

**Clemens**

That is correct.

**Craig**

Once I've built it, if it proves hugely popular or profitable, am I tied to Insta Hub I mean Instabase's AI Hub in perpetuity, or can I then download and market it myself?

**Clemens**

Yeah. So, the way that I would look at it is really. I think the operating system analogy is a good one. I would really look at it as an operating system If you develop. So, let's take the Windows analogy. So, if you developed a piece of software for Windows, it would run on Windows. If you wanted to take that same binary and run it on a Mac, it wouldn't work because it's just incompatible. So, the same would be true for the Instabase AI Hub. If you build an app to run on Instabase, it's specifically for that operating system and it will run anywhere Instabase runs. But if you wanted to run it outside of the Instabase environment, it wouldn't work because it just depends on some of the assumptions of the platform.

**Craig**

I see. The Converse model or product that you were talking about; one thing that fascinates me about OpenAI and Microsoft is that, on the one hand, they're providing these APIs for companies like Instabase to use their models. On the other hand, they're offering solutions that compete with those people using those models. When you describe Converse, an ability to interrogate a bunch of data that you've uploaded, I mean, that's what OpenAI is offering directly. You upload your data through the API and then can query or analyse or summarise or whatever you want to do with it. So how do you differentiate from what OpenAI is offering directly with regards to Converse?

**Clemens**

Yeah, let me actually first make a categorical statement. I think the fact that OpenAI, Microsoft and others are both providing the services and competing with solutions built on top of the services I think is pretty common. But that's really what Cloud computing is. AWS provides Cloud infrastructure for you, but they also compete with almost anything that anyone builds on AWS, and it's the right thing to do. They run a business and they make money on compute, and they make money with their own services. So, I think that's not going to change and I think that's just like a co-op-petition at its best, if you will. In terms of the differentiation at least, OpenAI is not actually doing this. So, if we double-click on how this works, it's like an OpenAI with chat, GPT specifically. You're basically using the GPT model behind it, 3.5 or 4, as a knowledge engine and as a reasoning engine, which is, you ask a question, and the model will give you an answer to the question.

Yes, you can upload data in a limited form, which is like you can, in the prompt, give it some data. So, you can say, hey, here's some text that I've written, please rewrite it. So, you can provide it with some amount of data, but you can't upload a PDF, or you can't upload in a spreadsheet, or you can't upload a picture that you've taken with your phone, because ChatGPT is a chatbot and it just takes text.

The way that the model works is it has a context window, and that context window is like limited, I think it's 4,000 tokens and they have an 8,000 token for GPT-4 and like a 32,000 token as well, but that so 4,000 tokens, like 8,000 tokens, I think is like roughly four pages of text. So, it's not a lot, right, like you can paste some text, but it's not a lot. Our product Converse and we've spent a lot of time actually like making sure this works, takes any amount of data, so you can actually have an example you upload like a 400-page financial document and ask a question about it. Now the interesting thing is you could not take 400 pages of text and pass it to GPT, right, like that's just not going to fit into your content.

**Craig**

Not even with an API call? I thought that you could upgrade to larger amounts of tokens.

**Clemens**

Yeah, so the largest amount that they currently have is the GPT-4, 32,000 token window. That will cover, you know, roughly like 20 pages of text, but it doesn't cover like 400 pages, right. And I think I know that they're working on a larger context model. And, by the way, there's also other providers of these LLMs like Anthropic is one of them that are working on providing bigger context windows. But the context window is never going to solve for the corpus retrieval type use case, right, like where they say like hey, I have like a thousand documents. These models would never like to grow their context window to take like 1000 documents as a context, right.

So, the way it actually works and what we've spent a lot of time implementing for Converse is when you upload all of these documents, we actually like to pre-process them, we create embeddings with them and then we store them into a vector database. We specifically use Re-Weight for that purpose. And then when you ask a question, we first do the vector search in the database and there's different strategies depending on what type of question you ask, right? So, if it's a retrieval question, we basically just find the chunk that's most relevant and then we send that chunk as context to the GPT model. But if it's a summarization question, right. So, like, if you want to summarise a 400 page long document, you can't just send a subset of those 400 pages because, like, you want to summarise everything But then you need to do a more advanced technique such as recursive summarization, right, like? where you first summarise like pages, then you summarise chapters, then you summarise like the entire document. So, all of that infrastructure, if you will, and the logic around it is necessary for you to actually be able to like, point it to any content and then let it answer a question about the content, and GOT itself can do that.

There are other products that are built on top of GPT that are attempting to do this. It's actually interesting. We tried to run a benchmark. I was almost thinking about doing a humoristic blog post about this because we were trying to run a benchmark of our performance on long documents against other products that support long documents, and what we found is that no one supports these long documents. We tried to run it as a couple of other products that I'm not going to name, but they just don't support it. So at least today is like my best knowledge. You know, like the AI Hub Converse product is the only one that like really performs well on like very large documents And, if you think about it, common use cases that we actually see in like financial services or like in I mean like even tax preparers, like anyone where you may actually have like retrieval type questions.

Like an insurance company says, ok, I have 10,000 contracts, which of these 10,000 contracts are exposed to a pandemic risk, right, that's like a question that requires you to actually look at a very large corpus And I mean that may be millions of pages in an aggregate.

It's like a different and entirely different way of approaching it.

So, long story short, the differentiation to like the LLM themselves is like very significant right, because, by the way, in a way, to summarise it and this may be a controversial statement but the LLM themselves basically act as a knowledge engine and the reasoning engine Because, like if you are GPT, it basically like it uses the knowledge that it has, that it was trained on, and it like reasons to give you an answer.

What we do is we only use it as a reasoning engine and we point it to your data as the knowledge engine, if you will, and that's that, by the way, just to call out like that important point is that's also how we actually solve hallucination in our product, because now we actually ground the answer in your data and, like the model will not answer questions that are not that it doesn't find in your data. So, we actually have fun examples. You can point it to, like a court summons document and ask if someone is guilty and it will tell you that it can't give you that information based on a document because it's like that's for the court to decide and it's not part of the corpus.

**Craig**

Yeah, who did you say you use for your vector database?

**Clemens**

It's a product called VV8.

**Craig**

Okay, yeah, so I've had Edo Liberty from Pinecone on the podcast. I don't know. I assume you're familiar with Pinecone, but I met somebody recently who has a company that does sounds similar. I'm not going to be able to pull the name out of my memory bank right now, but it also leverages large language models for their reasoning. But it uses a vector database, as is the memory or as the other sort of ground truth repository that the LLM is reasoning on. So, you guys are you. You have the vector database. If someone uploads a bunch of documents, you know, I don't know a thousand pages or something internally or in the background, I’m sorry, I’m going to sneeze. I'll cut that out. I don't know where you are, but we've got the Canadian fires on the east coast and it's making me sneeze a lot.

**Clemens**

I heard of that.

**Craig**

Yeah, but the. So, if somebody uploads all of this data and then there's some a process in the back end that that vectorizes it or parses it and turns it into embeddings in in a vector database and then when you query the language model, it searches the vector database, that language model that you're querying. Are you using a GPT for or are you using your own model at this point?

**Clemens**

Yeah, so right now we do have a very close relationship with OpenAI, and we provide both 3.5 and4.

We kind of abstracted away the product, so the models are called basic and advanced, but 3.5 and 4 are behind it. But we are reserving the flexibility, if you will, to also provide other models in the future if we find them to be useful. So like we're also experimenting with Google's palm model, and anthropic is another one mentioned earlier, because our platform is built in a flexible way such that we can actually use these other models in the background and we've actually found that for some use cases and like, depending on, like, the types of queries that you ask, it may actually be useful to use different models for different use cases, and We reserve the flexibility to make sure we provide the biggest value to our customers And, at the end of the day, they shouldn't have to worry about exactly what model is used in the background, because we abstracted away and we provide the same level of security and data privacy guarantees for every one of the vendors that we will.

**Craig**

Yeah, that's interesting. I had a conversation this morning with a start-up. That's still in stealth mode, but they are hoping to provide a Kind of an orchestration layer. So, when you present a Use case or a problem, it optimises Your choice of Model. I mean it'll, it'll search for you know, based on your criteria the fastest, the cheapest, the, the more precise, what all in the background, so you don't have to like Do that, research it. It reads the problem and directs you to one model or another. Is that the kind of thing that you're playing on doing? Do you include more models?

**Clemens**

It's actually a very interesting question And I think the importance of the answer will depend on how this in part space shakes out. You know, like this different philosophies That some people follow, like some people think that there's going to be hundreds of these models right for different Like use cases and like people are starting to find you know limbs and you're like make like domain specific ones, and then there's others that say no, and you know, I listened to your like a conversation with Yeah, I’m the cum like Russia and I think young that he believes there's like there needs to be one world model. Yeah, it's all of it like shared understanding. It doesn't make sense to like to learn physics over and over again like different models. So, the short answer I would say is We will, we will use different models specifically right now, based on the tasks specifically. I'm just to give you a few examples right. Like we found that for complex reasoning I know that this is also a hot topic, but for complex reasoning, actually, gpt-4 is really good right? It actually can. Many cases like doing multi-step reasoning relatively well.

3.5 doesn't work at all for this right, and then like palm to like me work for other use cases. So, we may classify your intent of what you're trying to achieve and then, based on this, like to use different models. But at least we are currently not expecting this problem to be a Search over hundreds of models’ problem if we expected to be a hitter like two or three different models that we use Because we know that they work particularly well for specific use cases, but it's not. Some people may make it sound like it's going to be almost like a hyperparameter search. Right, it's like there's like thousands of models and like you need to find the one that works best for you. We don't think that that's going to be the case. We think there's going to be a handful of really powerful models, And, by the way, we also believe that they will mostly be proprietary, and that open source will have a hard time catching up with the research and the quality that's coming out of companies like well, that's another question.

**Craig**

That's interesting, with Metas Throwing its hat into the open-source ring on these things. Why do you think Open source won't catch up? I would think the opposite.

**Clemens**

That yeah, and I and I Yeah, and I heard the comment that the inmate put on your, on your, in your conversation. The interesting thing is in private. I just yesterday published a post on this on medium specifically. The title is the golden era of open source in the eyes coming to an end. And What I meant by this is, if you look at Metas, open sourcing and yes, they are open sourcing, but The LaMDA model is open sourced with a GPL 3.0 licence, which is a copyleft licence, which means that it basically you cannot use it for commercial purposes, because if you used it for commercial purposes, you would have to open source your product, also with the GPL licence. Some people refer to it as a poison pill Licence. Right, all of the other models that meta has open source recently were open licensed, open sourced with the creative commons’ attribution, share alike, non-commercial lessons, which also is non-commercial, and share alike, which is Also like a poison pill, like type lessons. So, the answer is yes, there is a lot of open sourcing, but it accrues mostly to hobbyists and academics, but in for commercial use. Right, like if I am.

If we, as Instabase, wanted to use like these models, we couldn't because they're licensed in a non-permissible way and in the blog post to make the point that I actually, I personally believe that these models shouldn't be open sourced in a permissible way because, like, they represent the aggregate spend on data compute talent of these companies, right, that gets a massive like IP value that these companies are giving away.

And if you look at you know, like the hugging face, they have a leader board. It's called the open LLM leader board, which is, like all of the open source LLM models, ranked by some benchmark. None of them have a commercial use licence. They're all either like llama based or like there's another like licensed that none that's non-commercial, by the way, with the one exception There's one. There's a model called Falcon that just recently, I think last week, switched from a non-permissible licence to Apache 2. I don't know why they did this. I think you know, like the hugging face founder is like influencing this to some degree. But there is a lot of open sourcing. It's not permissible, it's academics and hobbyists.

**Craig**

Yeah, I'm just thinking. I had the. I had a hidden Gomez on the podcast the other day and how, yeah, I mean, is Cohere somebody whose models you would leverage, or because they have multiple models and a platform to Allow or help people build large language models, are they a competitor?

**Clemens**

So, we looked into the place. I'm personally not super familiar with Where we are in terms of evaluating their models, but if it's compatible with the way that we're running Our product and if they have a model that works well for these cases, there's no reason why we wouldn't use it, even if they're competing. But it's the same argument why you know Azure is both Providing you with AI services to build your own products, but also competing with you with power, like from a recognizer and so on. How does the AI Hub Decide which model to send your product, your, your project to?

Yeah, so in the, in the current version that's published, it's a manual decision by the user, so, just similar to chat, you can actually switch between the two models. One of them is more expensive than the other. Gpt-4 is, like right now actually, even if you look at the open AI prices, that 10x factor at least price difference, because they haven't been able to like scale it to like economies of scale if you will, and so it's a manual decision by the user and In testing and like in a lot of the early adoption that we've seen, people really quickly realise Where the 3.5 model works really well, like with basic questions, like basic extraction, basic summaries of documents, and as soon as you get into a little more complex reasoning, gpt-4 is really necessary and actually does a relatively good job so is it a?

**Craig**

It's under the user's control. There's a menu or a bunch of checkboxes, and you and then, based on what you've checked in the, the AI Hub decides which is the best model. Or is there a drop-down menu with You know five or six models and you pick, and it's just iterative Experimentation where you decide oh well, this is the model that's best for me.

**Clemens**

Yeah. So right now, It's the ladder where like really uses, just switch between three point five and four. In the future, I don't think we will ever just give people a drop down with like six or seven options, because then basically you are putting the owners on the user to make the decision and that's not a good use experience. I think what we will do is like we will implement the logic behind it, that will pick the best model for the task at hand And at least try to abstract the complexity away as much as possible.

**Craig**

Yeah, so the AI Hub was Just announced or released or launched. How? and you've got some of Instabase's own apps on the hub. When do you expect? I mean, how are you, I imagine, talking to me as part of the marketing effort? But how do you Get people to start building on the AI Hub?

**Clemens**

Yeah, by the way, I was excited to speak with you, regardless of the product launch, but, yeah, the timeline and how we expect this to work. So, there's really different segments that we expect to come to the iHub. There are definitely the consumers, if you will, or the users of Converse and the apps, and that's already picking up. We had a lot of press coverage. Our founder actually gave a TV interview early this week. So, we're already seeing just consumption of the Converse app and existing apps.

On the build side, we are working with some companies that expressed early interest. We had an early access program, and the companies were like hey, can we publish an app? Can we start making money from this? So, we are working with them specifically to build some of the first apps and then in the next couple of months we will actually provide the ability to submit an app for review. So, if someone in the world thought they could build an interesting app that should be published on the iHub, they can publish it for review and then we'll take a look and if it's of good quality and meets our requirements, we will publish it for use on the iHub. But that doesn't exist just yet. Yeah.

**Craig**

And will the iHub or Instabase? do they do marketing on behalf of the apps that are on the build on the platform? Or is it simply raising awareness that there's this hub and then people go and browse?

**Clemens**

Very good question. So, I think there's two different ways that I would look at this. One of them is again like using the operating system analogy right, if there's more users that use the operating system, the value of having an app on the operating system increases because there's a bigger market. So, we will definitely keep marketing the hub itself and drive more and more usage towards it. Now, because of the types of customers that we serve on the enterprise side, we are actually planning to package up capabilities on the hub into what we call vertical suites, so you can have a financial services suite that has all of the different tools that you would need in the financial services segment and maybe even subsegments. So that's a product that we actually advertise and have you pushing our sales team to send enterprises for, and there's no reason why those suites couldn't contain third-party developed apps. So, in this case, our sales team would actually be selling packages or solutions that contain apps from third-party developers if they fit into those vertical suites.

**Craig**

Yeah, although back to the question of these companies that provide services and compete on products at the same time. If I built an app on AI Hub and you included it in a suite of products that you were selling and I'm benefiting from your sales, why wouldn't you just build a proprietary app? Or is there something in the agreement that you're not going to compete on specific apps built on the platform?

**Clemens**

I think, philosophically, the question is and, by the way, we even internally, some teams internally asked the question okay, now there's all of these niche start-ups that say there's a start-up that focuses on legal contracts, we will basically build our converse app specifically for legal contracts. And then our teams say, okay, are we competing with the start-up? Are we going to build an app specifically for legal contracts? And philosophically, my answer would be no.

Ideally, we would incentivize that start-up to build their app on top of Instabase, because we have limited resources and we're not going to build apps to compete with everyone who wants to build apps. And it's much more beneficial to us if we can actually like, garner the ecosystem and get more and more companies to develop on our platform, so that we basically our role becomes maintaining the platform, providing distribution, providing users, and you're like the go-to-market and like the channels, as opposed to like, trying to like, come up with like thousands of apps that people will build. Right, it's really a diversity question. There's a massive, long tail of apps that we would never even think of building that we hope that the community will actually bring to the table.

**Craig**

And if people want to use this, how do they? Is it? Is there a freemium model? or how do people build apps and market them on AI Hub?

**Clemens**

Yeah, I appreciate the question. https://aihub.instabase.com/. It's available today If there's a free trial experience, if you will. If you're not logged in, you can play around with some sample documents as soon as you log in, and today it requires a Gmail account or like a Gmail, like Google Enterprise email account. But we will add more flexibility sooner. Then we actually give you 1000 free credits so that you can play around with Converse, you can play around with build, you can play around with the apps that we published, so like it's a free product, if you will, until you consume those 1000 credits, and then you can swipe a credit card and buy additional credits. So, it's really a pay-go model where, as much as you want to use the product, you can add up more credits and keep using it. The ability to publish apps, as mentioned, is coming in the future, so that doesn't exist today. But if anyone hears this and they're interested in building apps on AI Hub, they should reach out to us. We're already working with some early partners.

**Craig**

Yeah, that's fascinating. Okay, is there anything that you wanted to get out there that I didn't ask?

**Clemens**

No, I think you did a good job of covering, I think, all of the relevant pieces.

I would encourage you, by the way, also to try it. It's really I found it extremely compelling to find interesting use cases for anything really Like one of just to give you one example that maybe a lot of people may not think of, but I mean the reasoning aspect, because I know that that's a controversial topic is I uploaded a spreadsheet that had a project plan in it, which had work items and who owns these items and what are the dates and what are the dependencies. And I actually asked the question what would happen if this person fell sick and was out of office for this period of time And the model in this case GPT-4, actually correctly said well, this would delay this task and then these three other tasks that depend on this task, so they would be delayed, and then the last task in the project plan would be delayed because of this dependency. So, as a result, this project will finish on this date, which blew my mind because these models literally just predict the next token, but in this case, it's actually expressed like really powerful reasoning. So, the opportunities are endless and if you have time, play around with it.

It's free. Let us know the feedback and we can't wait to see what people do with it.

**Craig**

Yeah, no, it sounds fascinating. I definitely will take a look at it. That's it for this episode. I want to thank **Clemens** for his time. If you want to read a transcript of the conversation today, you can find one on our website eye-on.ai.

And remember, the Singularity may not be near, but AI is changing your world, so pay attention.