# [Transcript] AI Hustle: News on Open AI, ChatGPT, Midjourney, NVIDIA, Anthropic, Open Source LLMs / AI's Future with Lightning AI CEO William Falcon

Welcome to the OpenAI podcast, the podcast that opens up the world of AI in a quick and concise manner.

Tune in daily to hear the latest news and breakthroughs in the rapidly evolving world of artificial intelligence.

If you've been following the podcast for a while, you'll know that over the last six months I've been working on a stealth AI startup.

Of the hundreds of projects I've covered, this is the one that I believe has the greatest potential.

So today I'm excited to announce AIBOX.

AIBOX is a no-code AI app building platform paired with the App Store for AI that lets you monetize your AI tools.

The platform lets you build apps by linking together AI models like chatGPT, mid-journey and 11Labs, eventually will integrate with software like Gmail, Trello and Salesforce so you can use AI to automate every function in your organization.

To get notified when we launch and be one of the first to build on the platform, you can join the wait list at AIBOX.AI, the link is in the show notes.

We are currently raising a seed round of funding.

If you're an investor that is focused on disruptive tech, I'd love to tell you more about the platform.

You can reach out to me at jaden at AIBOX.AI, I'll leave that email in the show notes.

Welcome to the AI Chat podcast, I'm your host, Jaden Schaefer.

Today on the podcast, we have the pleasure of being joined by William Falcon, who is the creator of PyTorch Lightning, a deep learning framework.

He's also the founder and CEO of Lightning AI and was previously co-founder and CTO of NextGenVest.

He began working on these projects while completing a PhD at NYU, which was funded by Google DeepMind and NSF.

Additionally, he worked as a researcher at Facebook AI and at Goldman Sachs.

Welcome to the show today, William.

Thanks for having me, we're excited to chat.

Yeah, super happy to have you on the podcast today.

A question I wanted to kick this off with is to ask you a little bit about your background.

What got you started on this field?

You've obviously done a lot of really interesting things in the space.

You got a PhD.

Is this something you always knew you were interested in AI and this technology, or is this something that you found as you were going through school, or what was your journey like on that?

Yeah, super random walk.

I actually started in the military, so actually I was doing special operations for a few years, with the Seal Tribal, and then I got injured, and so the Navy was like, what are you going to do next?

And I was like, I don't know.

I guess I'm going to figure something else out.

So they gave me an option to become a pilot or one of these other jobs, and I was like, not alone when I pursued this.

And so I basically, you know, the Navy discharged me, and then while I was sitting around kind of waiting for my next thing, she's around 2012 when the iPhones were becoming a thing.

So I started learning to code iPhone apps, and so I got into coding that way.

And then many years later, just happened to, like around 2012, 2013, people started caring a lot about how do I get smarter with my software, write apps, and so on.

So I started looking to machine learning, and then a few years later discovered deep learning.

So I just kind of found my way through there, and then I found that it was pretty useful, and I'll come doing it, I guess, and that was before all the craziness and all the hype.

So turns out the tools I built for myself for undergrad, and then my Ph.D. at some point became one of the spiders sliding today.

So it's stuff that I've been working on since 2015, but it became really relevant in 2019 when we were training huge models already and at Facebook at that time.

And now password to 2023, and everyone's trying to train large models, and it's stuff that we really were doing in 2019, so it's, I guess, pretty timely.

Very cool.

Wow.

Yeah, very cool.

So of course, you're the creator of PyTorch Lightning.

Can you tell us a little bit about, I guess, explain to the listener what PyTorch Lightning is, what it does, and the problems it's solving for customers?

Yeah, so PyTorch is a framework that is built by Shyspuk and it is for developing models, right?

So mathematically speaking, it's like computational graphs, and all these other things that help it.

PyTorch Lighting is a framework that organizes that code and gives you a framework, right?

So I would say PyTorch is more like a library and PyTorch Lightning is more like a framework.

So the analogy is like JavaScript and React.

So PyTorch is just like JavaScript and PyTorch Lightning is like React.

Like could you build Instagram on JavaScript?

Sure, but why would you, right, like you kind of need a framework like React to do something like that because there's a lot of other things are going to than just the code itself.

So PyTorch Lightning solves that and then our enterprise platform basically has everything else around ML.

So how to build and deploy and everything that, it's not just training, but like everything that how you link it all together with the rest of the ML pipeline.

Okay, very cool.

So something else I think I was curious to ask you a little bit about kind of with your background and some of the things you worked on is, you know, for people listening, what are some strategies that you would suggest for designing AI systems that can handle massive

data sets and adapt to, you know, evolving AI driven application requirements?

So we deal with a lot of layers, I guess.

So we deal with all the way down to hardware, to apps, to models, to how you do the model.

So what layer are you specifically talking about?

I don't know.

Just, I guess, if you have any of those layers, if you have any strategies or recommendations for them.

Yeah, so I think probably most people today are concerned with building more AI products, I would say, like that is like the majority of people, right?

It's like, yeah, the equivalent of like building an iPhone app.

So lighting itself is kind of like the iPhone and we give you all the tools to do that.

So if you're building an iPhone app or an AI app right now, I think the thing that you probably care about honestly is like getting to market as quickly as possible.

So it's less about how much data you're taking in or how much you're allowed.

So like, how can I derive value for my users very quickly?

So probably like a, like a chat GPT API or something like that is super useful, which you can also obviously do through our platform.

But I would pretty much just get people not necessarily thinking too much about the models or the solution that is get to the value of what your business is very quickly, regardless of the model.

It's a byproduct of that.

And then once you have some product market fit there, then go optimize the model.

You can do that by fine-tuning or pre-training model, but that will depend on how much data you have.

And usually the data that you want to use is data that's unique to your thing.

So like, if you're a lawyer and you're working on, I don't know, court cases or legal documents, the data that you want to give it is like legal documents to make that model perform better.

Right.

But POCing is very simple.

It'd be like using Squarespace back in the early 2000s, right?

Like get a page up, just get people on it, and then before you go invest into building a full-blown application.

Okay.

Yeah.

I've got some really good advice.

Something to be curious about asking you is like, right now running Lightning AI, what are some of the biggest challenges that you're currently working on or trying to solve?

What's some of the most difficult things about what you guys are currently building?

I mean, there are challenges that we've had for a long time that I think are more relevant now because more people are using it and using AI in general.

In 2019, we were training what is called an LLM now, right?

Because we were training large models back then, like Facebook.

It was mostly vision models, but the same terminology, same stuff.

And we were using something like 2,000 GPUs at that time.

This is 2019 for one model.

Today, it's like people are like blown away if you can use like 32, 64 GPUs, like that's very easy for us, right?

So I think like a lot of the challenges are around large model training and make, you know, accessing tons of data, terabytes and all of that.

I would say we're challenging in general, but these are things that we've solved already for many years, right?

So I think what we're doing more now and not just doing that, but how do you do it at enterprise is how do you do it at scale, right?

Because it's not just about demos or POCs or anything, but it's actually about how do you build meaningful products, right?

And that's a big gap between like what I can see in a chat GBT window versus what I can deploy in a company.

Okay.

Yeah.

That makes a lot of sense.

Kind of in that, right, you touched on some of the like the scalability and security issues.

How is Lightning AI kind of ensuring the scalability and security of AI applications?

I mean, so everything we do for them to, like on the enterprise side, it's on your data on your VPC and your own cloud and all that stuff, right?

So there's a lot of stuff there, but we partner with other kind of third parties to make sure that like your data is secure, auditing and so on.

We have a lot of core auditing capabilities on the platform as well.

Okay.

At the end of the day, like every enterprise is going to be unique.

So we deploy a lot to banks and like high-level industries.

And those are pretty, they take a lot of time to get through it, you know, one bank took us probably over a year to just even start, oh, wow, wow.

And that's all the things change so quickly that by the time that we started, the field got completely changed, right?

So yeah, that's, there's definitely a long, long, long time in development on some of those more regulated industries.

It's really cool that you have worked with banks and other industries like that.

I'm wondering, are there other maybe like success stories that you could share where Lightning has kind of made a big impact on a project or an organization?

Yeah.

So on the open source, I would say probably the most big companies in the world, like use fighter sliding today.

So Amazon, you know, they're training massive models, hundreds of billions of parameters to use fighter sliding to do that.

And Biddy has a whole product line called Nemo that uses fighter sliding, stable diffusion,

stability eye, all those models are trained using fighter sliding.

You can talk about Lyft, Uber, like all the biggest companies in the world.

If they're doing anything beyond like a little model or a few people working on it, it's very likely using how to slide me.

And it's just a framework that does that.

There's a lot of other frameworks and tools that are good for demos and POCs and small models, but Lightning does really have a competitor for like scaling large models, right?

It's been battle tested already for many years before even Chudge EPT and LLMS and all this time.

And so on the platform side, it's really been helping those companies augment more of what they do and do it at scale so that they don't have to deal with all the infrastructure, models themselves, right?

Cause it's pretty, pretty hefty, but yeah, I mean, we've touched pretty much any industry at this point and it's likely like NASA at some point, like everything has touched fighter sliding already.

So if you're interacting with any system that uses intelligence, it's very likely that it was somehow like fighter sliding was involved in their training or deploying the thing.

That's super cool.

Wow.

That's amazing.

What are some areas that, you know, looking to the future with everything that you guys are building, you're the most excited for some advancements that you see, you see everything in this space making?

Yeah.

Yeah.

I mean, I think everyone's super excited about trying to get models to train on bigger, bigger machines.

We're actually excited about the opposite and it's how do we get Chudge EPT performance training on your laptop or like single GPUs or very, very, very small things.

So I think that's where we will take the market, not just us, but the rest of the open source community.

And that's super important because you need to make this accessible.

It's kind of like IBM and Apple, right?

Like IBM had these mainframes that only they could operate and was needed for like big things that they were doing.

And then Apple was like, here's a personal computer, like everyone should have access to this.

So we kind of think of our way ourselves as like the Apple of this, where we want to make AI really democratized and have everyone in the world be able to have their own models around their own stuff without needing to like, you know, pay some giants, some crazy fees just to do that.

Like we imagine if we have to buy every time you want to use a computer and have to pay IBM to use it, like that's pretty cool.

Yeah.

Of course.

Yeah.

Of course.

That's super cool.

Yeah.

Because this was just in the news, I believe like yesterday, the day before Intel, if you've seen has just announced that kind of the next big thing for them, they want to do is what they're calling AI PCs where they're like making these chips so that you can like run AI models on your PC.

And that's kind of like something that I think a lot of consumers are really excited about.

I'd be just curious to hear your perspective on that.

I guess some of the ways you're doing, you're trying to facilitate this, some of the ways they're trying to facilitate this.

And I guess your thoughts on it.

Yeah.

Makes a lot of sense.

I think that's definitely a way to go.

And video's already done it, right?

You can run, you can buy laptops with like 1080 GPUs, probably a more advanced GPUs all day, 3090s or 3080s or something.

So they can run models locally, but these things are massive, right?

I think Apple's already doing that.

Like we've actually been really heavily involved as well.

And like all these, like if you ever see in Twitter, they're like, oh, run this model on like a laptop or whatever, like you've done a lot of that work as well.

And people like, for example, we have lab model models that you can run in your laptop, you can serve them to live today.

That's a lot of work.

I'm like getting M1 and M2 chips to work really well, but yeah, I think Intel has the capability to do this.

They have the Gaudi line as well, so I'm sure they can take some ideas from there and drop them into the laptops.

But I think it's a feature, it's a matter of time, until it's not even for like consumers, like I think the use case is less like you at home playing with the model, but more like deploying at the edge, like running a model on headphones or like a heart monitor or something like that.

So you can have bigger applications, right?

Very cool.

Yeah, of course.

That'd be incredibly interesting.

I mean, that's not even a prospect I've personally thought about much.

Based off of like, I guess some of the advancements you guys are working on in your scene, how

do you feel like things in the AI space will shift?

Like let's say over the next five years, like what are some of the big shifts you think are coming?

Oh man, so hard to predict.

I know.

Last year, this time we were talking, it's funny because like I was at an Intel conference, I was given a talk at like an Intel panel around July of last year, and they were like what's the next big thing?

And I was like, I think the Fusion models are probably going to be a big thing.

Few months later, stability launches with the Fusion models, right?

And then like, eight months later, child GVT, it's like, I can't, we can't predict it all ahead.

So five years, I have no clue.

I definitely don't think we'll be at AGI, I can tell you that much.

And like, I definitely don't think, I think this whole big model company thing will not be a thing in five years.

Like I don't think that that's going to last.

So all these companies that are raising massive amounts of money to train these massive models, maybe there's a use case for that, but in five years, this will be just commoditized, right?

Okay.

There will be thousands of these, and these will be open source models for the most part.

So I think that's the biggest thing.

But I don't think I'm saying anything that people don't know.

I think people still believe that like, this is the way to go in the future, but I think in five years, it'll be all very, very small personal models that are open source.

Okay.

Yeah.

You know, it's kind of interesting because I've talked about that a lot on the podcast.

That's where I see this going.

I believe it was at the, just like at the beginning of the year almost, Stanford did a thing where they did me like a PubMed GPT, right?

They trained it on publicly available medical journals.

And they're like, yeah, this thing's great answering like medical questions.

If you want it to be really good, you train like a very specific model on like heart disease or on breast cancer.

And so they're, they're kind of looking at that to where it's like topic specific, which was a really interesting concept for me to think about.

It's like chat GPT may not be like this overarching giant that just knows everything.

There may be more specific models or maybe they're kind of integrated in different ways.

Even the way that, you know, chat GPT's models were leaked and they have like, I guess they call them like 16 different like experts or, or like kind of areas within GPT for that are helping to facilitate things.

So definitely very interesting.

I'm wondering, do you have a piece of advice that you would be able to give to people perhaps working in AI today, young professionals that are kind of getting into this field?

You have a really awesome background, right?

Like you, you did your PhD, you had Google DeepMind fund that you, you had, you did a lot of really cool things in the space.

I'm wondering, you know, other people interested in getting into the space right now, what's a piece of advice you, you could give them?

They'll do a PhD.

We all say that.

No, you don't need to, right?

So I think, no, I think we've made a lot of this stuff super accessible now.

I don't mean we like lightning, but like the community in general, right?

So like, yeah, with fighter sliding, can you train a model today and not know anything about backprop or, you know, grading the center or any of this 100% you can do that, right?

And you run stuff on the cloud without knowing anything about it.

Yeah, you can also do that.

So I think there's a world where you don't need to worry about too much of the technical details.

So I think people getting into it today should have a distance problem though in a cell, like a real either, I don't know, some side project you want to build or like a personal thing during school, like an app or like a startup you want to spin up or whatever.

I would focus on that.

I wouldn't get, I would not be turned off by the fact that maybe you don't know how to code and maybe you don't know how to deploy things.

It's fine.

You don't need to do that.

There's plenty of tools out there that can help you do this.

But having a clear idea of what to solve for a real set of users is more important than learning how to code because there's a ton of people I can code, but none of them are like lawyers or medical experts or anything.

So the stuff that they're going to try to do is kind of the same.

Some model API or some sort of tooling or some sort of like version of the same thing that anyone else can think about, right?

So it's I think those deep experts that should not be scared and go pick up some APIs, go prototype your stuff and then if that works, then from there go hire a few developers and start learning, right?

And if you're a developer, I think it's still useful to do that because you want to be product focused and then learn tools as a result.

I think one thing that separates me probably for most engineers, I hate engineering, right?

I mean, I do it and I do it a lot, but like I use engineering to solve problems.

I don't code for the fun of it and that's different.

A lot of people just code because they're like, that's cool.

I'm like, that is cool.

But like it's cooler to see people excited about something in those, right?

Yeah.

Yeah, for sure.

Like people don't care if ChatGPT is operated by a thousand monkeys behind the scenes.

Like they don't care, but it's right, right?

Right.

Totally.

Totally.

One thing you think that perhaps other AI startups are doing incorrectly today and I guess a piece of advice or somewhere where you think that people are going in the wrong direction.

I mean, I know you mentioned, of course, like ChatGPT and like models being way too big, but are there other areas?

Yeah, I think chasing trends, like everyone's trying to chase these trends.

I think you need to just pause, think about the problem trying to solve and be principled about it.

But if you're always chasing the next thing, you're by definition never pushing anything, right?

Because everything's always moving.

So by the time you start chasing the next thing, that's already gone.

Like you're already a year behind.

So I don't know.

I saw a very similar thing in the mid-2012s area where it was with iPhone apps and you know, back then it was the same thing, that 10,000 calculator apps, 1,000, you know, flashlight apps and it was like, which one's going to win?

It doesn't matter and we're all commodities, right?

So yeah, but the stuff that won was stuff that was really well done, that was super thought out and how like a real use case and those things became multi-billed to our companies, right?

But and there's nothing wrong with putting side projects out there.

Just don't conflate a side project with like a business.

They run the same thing usually.

Yeah, OK, I like that.

I like that advice.

It kind of goes into my follow up or my follow on question, which is, what are some areas you think people investing in AI are perhaps missing the mark?

You mentioned, of course, sometimes putting money on these big, huge, anthropic, for example, today just announced they got like four billion dollars from AWS.

Of course, open AI has done 10 billion from Microsoft.

What are some areas you think that investors are going wrong?

And what's some advice you feel like you could give to people looking at making

investments in the AI space?

Yeah, I see a lot of VC money just chasing very silly things.

But, you know, they were chasing silly things back in early 2002, right?

So I think it's just about having a fundamental understanding of what this stuff is.

I think everyone has like a FOMO and like they want to understand what's going to be the next thing or not.

Like, and there's always filming positions about like, who's going to capture the value chain, this or that?

It's like, none of us know.

Like if we knew we'd all be running a hundred million dollar hedge funds, right?

But we don't.

So, like, don't try to be too smart.

I would say it's like, just be principled.

Do the diligence, do the work to understand what the field is doing, like what is actually needed and then find companies that match that are supposed to getting carried away by like all these pitches of startups, because they're all going to, every single one's going to make a compelling argument about why something's going to be good.

Some are amazing, a hundred percent.

And I do angel investing, so I invest in some cool things.

But you have to really know what's going on and they all to be able to make good investments.

Yeah, I love that.

That is, I think, very, very important advice right now, especially with kind of the some of the froth in the AI market.

William, thank you so much for coming on the AI Chat podcast today.

Really appreciate your insights, hearing your background and everything.

If people want to find out more about some of the things you're working on and some of the solutions over at Lightning AI, what's a good way for them to look that up and find out more about you?

Yeah, so go to lightning.ai and sign up and then, you know, let us know.

We'll get you off the right list and then follow me on Twitter as well.

So we're a Falcon and Lightning AI on Twitter as well.

And you'll see all our community updates and everything else as well.

All right, sounds good.

And I'll drop a link to that in the description for in the show notes for the listeners.

William, again, thank you so much for coming on the show to the listeners.

Thank you so much for tuning in to the AI Chat podcast.

Make sure to rate us wherever you get your podcast.

If you are looking for an innovative and creative community of people using ChatGPT, you need to join our ChatGPT creators community.

I'll drop a link in the description to this podcast.

We'd love to see you there where we share tips and tricks of what is working in ChatGPT.

It's a lot easier than a podcast as you can see screenshots.

You can share and comment on things that are currently working.

So if this sounds interesting to you, check out the link in the comment.
We'd love to have you in the community.
Thanks for joining me on the Open AI podcast.
It would mean the world to me if you would rate this podcast wherever you listen to your podcasts and I'll see you tomorrow.