MACRO Voices
with hosts Erik Townsend and Patrick Ceresna

# Matt Barrie: AI of the Storm

**March 27th, 2025**

***Erik:*** Joining me now is freelancer.com founder and CEO, Matt Barrie, who's also becoming a world-renowned expert on artificial intelligence. Matt just published a fantastic article, which I would consider to be a must read. It's called ***AI of the Storm,*** that's linked in your Research Roundup email. If you don't have a Research Round up email, you're not yet registered at macro voices.com Just go to our home page macrovoices.com, click the red button above Matt's picture, which says, *looking for the downloads*. Matt, why don't we start with the State of the Union, if you will. What has changed in the world of AI in the six months since we last had you on?

***Matt:*** Well, it's been a dramatically changing landscape in the last six months. As it turns out, you can assemble a ragtag team with a relatively modest budget and even work on potentially AI as a side project and deliver a model which could, can challenge the state of the art of what has been coming out of Silicon Valley. You're seeing this not just with independent, private companies launching models here, there, everywhere. But you're also seeing it with open-source efforts, and, in fact, starting a foundational AI model seems to be akin to opening yet another Thai restaurant, albeit potentially a very good Thai restaurant in Thailand, while at the same time handing out the recipe book and the business plans. You're seeing efforts come from left, right and center that you wouldn't expect. Elon Musk managed to assemble, did it the traditional way, and managed to assemble a very large data center with Grok 3 and get access to the hardware, and now has produced one of the leading models, really getting OpenAI run for its money in the US. But you're seeing efforts from all around the world tackling different parts of the problem, whether it's team in France called Mixtral, or now the clear and present threat is coming directly out of China, with the likes of DeepSeek, where, literally just out of nowhere, late last year, a team of 160 ragtag engineers on at a hedge fund, worked on a side project and dumped out a model, DeepSeek-V3 and then the second model, DeepSeek R1 which really challenged the state of the art of OpenAI's foundational models. And they managed to do so on a fraction of the budget. It's been rumored that the training that was involved for DeepSeek involved about 2000 GPUs, when it would normally take 20,000 GPUs, they achieved the 10 times efficiency in the training of these models through just some smart optimizations under the hood. A bit like taking a race car and doing some tinkering with the engine, they managed to speed up the training about 10 times. And as a result of that, the training budget is remembered to be around $5 or $6 million which blows up the water in comparison to some of the latest we've seen in the valley, where training models in the order of $100 million. So, you're seeing all these competitive threats come out of nowhere, and models are being released, really, at such a rate. I think just before we started the podcast today, you commented there was another

DeepSeek model was just launched overnight, which people are trying to get their hands on and understand what that's about. But there's certainly competition coming from left, right and center. At the same time, these models are getting a lot more sophisticated, so they're now multi-modal. That means they can take in text, they can take in images, they can take and they can produce text and images, or take in PDFs and a range of different modalities, and output in different modalities directly. So, you don't have to really have a translation step in between them. And simultaneously with all of that, there's been a cataclysm in terms of the hardware layer, which we can talk about in a second, coming out of China as a result of US sanctions.

***Erik:*** Matt, people in the AI community talk about a progression of functionality that starts with something called chat. Then it goes to reasoning, and then it goes to agents. What do those three words, chat reasoning and agents, mean in the context of AI? Where are we now in that progression and what comes next?

***Matt:*** Well, chat really was the evolution of taking these models and really putting a chat interface on top of them. And so that's when GPT went to ChatGPT. And you could, all of a sudden, consumers could query the AI, where previously it was really in the domain of computer sciences. And that's when we had that sort of explosion in 2023 where people said, oh my God, this is actually now very, very interesting, particularly as these models got sophisticated enough and consumed enough training data and had enough processing power behind them to give answers. They actually were interesting enough to the average person, and not just in the realm of computer science, that has led to a bit of an arms race in these models. And, as I mentioned in the last episode, the training budgets were going up by an order of magnitude, and starting to hit $100 million per training run. The data requirements for these models were going up orders of magnitude with each of those training runs, and the complexity was going up order of magnitude. And that led to a lot of these competitive efforts in this sort of hyper competitive Thai restaurant strip in Bangkok, effectively having to come up with better and more efficient ways to try and get better answers out of these models. And so that led to the development of a whole class of model, which are these reasoning models. And these are the ones where it's not so much stepping up the compute by orders magnitude anymore. It's really getting these models to break down the tasks into the query, into a series of tasks, and really think through the logical chain of thought in order to try and get a better answer, whether it's a scientific problem or a mathematical problem, or whatever it may be.

So, for example, like the o-series coming out of OpenAI, where really, you can click on a little button while it's running, and you can see its thought process as it goes down and thinks about your query, maybe ask some clarifying questions, works out how would I go about this problem? And works through the thought process. And that reasoning is very akin to how a human might break down a problem. Like, if you're sitting in an exam and you get asked an exam question to write an essay on something, you probably won't just jump in there and just write the essay straight away. What you'll do is, you'll think out an outline. You'll think down the structure, what would be the introduction? What be the conclusion? What are the key points I want to present in this before I actually go write my essay? So that's basically at a high level, what these reasoning models are and what that has led, together with multimodal capabilities, has now led to these

agents being sophisticated enough and good enough in terms of the quality of their responses to be able to do human workflows. So, for example, tier one support at a call center for a bank, a lot of these sorts of jobs, you could probably write out the entire job function of someone on answering the phones for maybe a retail credit card or what have you, doing customer support. You could probably write out their job function in two or three pages of ChatGPT prompt. And combined with multimodal, it means that if someone uploads their bank statement or as a picture, if you just see, or this that and the other, maybe it's an AI agent helping a telco debug customer problems that the customer could, for example, upload a photo of them, the modem and the lights, which lights are flashing, which lights are not flashing, or the computer equipment, and what have you, and you can basically now start to do workflows and develop agents to perform roles that maybe were previously done by people, whether it's answer the phones, take an order, do outbound lead generation, what have you. Or something a little bit more sophisticated, such as starting to do the role of a junior accountant or a junior researcher or a sub copywriter and so forth.

So it basically got to the point where AI can start really lifting the productivity of people by doing a lot of the basic everyday work that they would do in their roles. And so, that basically is the agent model. It's the ability for these AI agents to fulfill the roles of what human agents would do previously. And I think in the next 24 months, you're going to see an explosion of this. There's been a little bit of a, I think, a lull in terms of what people expect of the impact of AI given they see these whiz-bang things come out of ChatGPT and the like, but then they haven't really seen it in reality, in real life yet, but I think very soon the penny will drop with the general public, and it may be something like as simple as calling up a bank and talking to someone over the phone, or even doing a video call, and all of a sudden that call by the bank is going to be done by an AI agent. It will be done instantly with high fidelity, low latency. It will be done in the language of the person calling up. So, the human computer interface gets better because, now you can do customer support in any language you want, any point of time, with high levels of expertise. And one thing we've noticed, because we've got this rolling out on my company, Freelancer, is the correspondence of the interaction is about 10 times more empathetic because it's not a human agent. For example, in a call center with a KPI in terms of number of tickets per day, they've got our answer, or certain other productivity metrics, the AI has infinite patience, infinite time and access to all the world's information in its knowledge base, you actually see the quality and the empathy of the customer support queries or the sales engagements are in order to make you better than what a human would ever do. Because a human, for example, would never spend the time if someone's calling to a bank and they've got a question about the credit card, they never spend the time to do full research on the account and figure out everything about that, the customer history and the CRM and be able to provide effective support, they just wouldn't be able to do it. But with AI, they can. And so that's why the answers are just so much better, so much more effective. And of course, any language, any time of day, instantly is going on.

***Erik:*** Let's move on to the software landscape of the AI industry itself, particularly as it relates to investors. As you said earlier, it seemed in the beginning like, boy, OpenAI did something so cool with ChatGPT that it felt like they had a huge moat. They were a monopoly. Nobody could possibly compete with them. Turned out that was wrong. Where is this all headed?

***Matt:*** Well, I think, Erik, we talked about this in the previous episode where, you know, when OpenAI came out with ChatGPT 3.5, it was really, truly a magical moment. The whole world was just wowed by this magic box. And at the time, it looked invincible. You know, it was going to be the toll booth operator on the highway to humanity's collective intelligence. Every time you have a passing of thought, you send it to the AI, and then that kind of thing, the API call to this magical, mythical intelligence, and so, they really went down two paths. They went down one path of having API access, which they would charge at the moment, GPT 4, for example, is like $30US for a million calls. They've got some cheaper models, they've got some more expensive models, but it gives you a feeling kind of for where they're thinking about the pricing. And then they had this consumer product, which was a freemium product. You can use ChatGPT for free, but if you want a much better model, you pay your $20 a month and you get access to that, and you can actually use that in a more productive way. The problem for them is, it turned out that creating a foundational model is akin to opening a Thai restaurant in Thailand, right? So, you've got all these different models have come out of nowhere. They've got disruptively cheaper pricing in some circumstances, or maybe better, it turns out you don't really need to have all the, you know, to raise billions and billions of dollars in order to produce a foundational model that might rival and beat OpenAI's latest and greatest, at least on some benchmarks. I mean, China absolutely kicked the door open with DeepSeek, when they kind of came out of the blue and showed that with 2000 GPUs, and difficulty in obtaining more because of sanctions that had been dropped, which I'll talk about in a second, between the US and China, they managed to train a model on with a few million dollars and a side project. And then that technology is now online, and you can access the DeepSeek APIs, particularly in the Chinese cloud, for, a fraction of the cost. I mean, it's equivalent to saying, imagine that your Ubers now will cost 5 cents, and you can catch an Uber anywhere you want for 5 cents. 10 cents, right? It's that sort of level of disruption in terms of pricing.

So, the API approach has been heavily commoditized. There are many different providers now, many different models with the equivalent APIs. These literally drag and drop in. Elon said specifically that the Grok mock APIs are just a plug and play replacement for OpenAI. And so, I think that particular business model of charging pennies on a few API calls to the cloud is being dramatically challenged. At the same time, I don't think they're making very much money at all out of the $20 a month subscription. And you can see that quite clearly when you use the product, you see it with all the models really when you use the product, eventually you get timeouts and get put in the naughty corner, and you can't make a call again for three hours or four hours. I mean, Claude is particularly notorious for that where, you know, I've used six to seven or eight queries inside the Claude product from Anthropic, and let's say you can't use the product for another four hours. It's very clear that this business model is not economic. If it was economic, I think you've expressed some frustration before, saying, gee, why can't I just pay them some more money and get some more API calls? Like, why don't they do that? Isn't that bit silly? I think the reason they don't do that is because it's not economic, and they're trying to figure out what is actually the business model. So, if you look at OpenAI's model, they've got the freemium model. They've got the $20 a month model, where you get access to a better AI, basically, you get answers that aren't computer like. They're now coming up with a $200 a

month model, which is quite expensive for the average person, although there's some pretty amazing things going on with a GPT 4.5 and Deep Search, which we'll talk about in a second. But now, they're coming up with an agent model, where for $2,000 a month, or $20,000 a month, you get access to these, this supposed marketplace of AI agents that will do various bits and pieces with you. Now, I don't know how successful that will be. They've come out with a marketplace of pseudo agents previously with GPTs, that kind of came out a bunch of fanfare, then quietly disappeared into nowhere because it got very little usage. But clearly there's this challenge around the business model, charging pennies per API call or 1000s of API calls is not really sustainable. There's a lot of competition coming into that. I don't think they're making a profit on the way their models are actually structured at the moment with the $20 plan. $200 plan is quite expensive for the average person, and they have all this competition coming in from open source, all this competition coming in from China, and then you've got the other big shock wave that's come out of China, which is basically on the chip side.

***Erik:*** Let's move on to the hardware evolution. In the beginning, it seemed like there was no way to do this other than to have the very latest and greatest Nvidia. No other brand would do but you had to have the Nvidia GPUs, and you had to have the absolute top of the line, or else you just couldn't play in this game. That seems like it all changed. What happened?

***Matt:*** So, in 2019, Huawei released their own version of AI chip. It was a quite an underpowered chip compared to the Nvidia line, but they released the ASCEND 910 M chip to start producing, you know, having China have its own sovereignty in terms of chip supply. The US saw that as a threat and slapped Huawei on the sanctions list, the Entity List in 2020. The end result of that was a bit of a backfire in some regards. Huawei pushed on and doubled down in terms of the chip development and actually reproduced that chip in two years, effectively, in their own way. Now, it didn't end up as powerful as the H100 line, but it looks like now that when they take that particular chip, you know, it started off as 910 and they kind of made their own version of that, and now they're jamming two of those chips on a single die. It turns out that 910 C model is about 60% of the performance of Nvidia's H100. So effectively, in two years, China managed to replicate an older flagship Nvidia AI chip, and now they're producing it en masse in the hundreds of thousands per year. And the thing about this is, when you combine Chinese AI software with Chinese AI hardware, you get a bit of a killer model. And that's where you can deliver the equivalent of OpenAI's API product for anywhere from 2% to 10% of the cost. So, it's really giving at least Chinese companies, because I don't know how many Western companies will make use of this in the China cloud, access to really, really cheap inference, which is going to lead to an explosion of products that are intelligent with Chinese powered AI in them. And then, of course, you've got now all these US chip efforts to try and go after Nvidia's dominance. And in fact, probably one of the strongest parts competitive threat for Nvidia is actually coming from Nvidia's customers. Because four of Nvidia's customers generated about 46% of Nvidia's revenue. And of course, Nvidia is charging sky high prices for their chips because they can, because, you know, they're the only game in town, have been up until now. And so, you've got everyone from Google to Amazon coming up with their own vertical integration chips, where they try and produce specialized versions of AI chips that are directly suited for their applications. So, you've got Google's Tensor Processing Units, you've got Amazon's Trainium.

You've got a whole bunch of different angles from which Nvidia is being attacked, both from its customers and from China and from other US players. So, we will see how long, Nvidia is still the main game in town, but we'll see how long that that is.

***Erik:*** Let's just translate that to investor language for a lot of people who want it to be part of, you know, let's be on the AI trade. Buy Nvidia stock, hold it, it'll go up. It's gotten a lot more complicated, hasn't it? What does this mean for investors who maybe don't have your level of technical background, who want to be betting on AI as a trend? But maybe Nvidia is a little bit overdone here.

***Matt:*** Well, I mean, it's a little bit like, I think Cisco back in the 2000s right? I think I remember Cisco's motto was, we network networks. And everyone back when you had the internet booming, and everyone thought, gosh, every single thing in the world is going to connect to the internet, and Cisco is going to be at the heart of all that, running all the hubs and the routers and the switches in order to do the connections. Surely, Cisco is going to be the richest company in the world. And off the stock went and did a parabola and hit the moon, right? But as it then turns out, other people can make switches and routers and hubs, and a big market attracts new entrants and you had an explosion of low-cost equipment coming out of China, to do network equipment. And the Cisco bubble popped and hasn't been back to where it was. And it's a possibility, I mean, Nvidia still rules the risk today, but there's a possibility that the same thing will happen with both US entrants, with Nvidia's current customers, and also with China.

I mean, the other big trend that's happening right now is just how much AI is going to go into the edge. And I think we talked about in the previous episode, you know, AI is pretty amazing, like you have these AI functions in Word, and it helps write your paragraph for you, or in Gmail or this, that and the other. But I do think that it's about to get to the point where AI is going to get stuck in a bit creepy with some of these features. I mean, one of these products just might get a little bit ahead of itself. Like, Gmail could, for example, put an LLM search interface in it, right? And you could type wonderful things into that LLM search interface. You could say things like, oh, find me that email I wrote 10 years ago about the Indian consulate. I have to find my visa number for India because I'm going the next week, and they've asked in the visa application to find that number, and I can't find it. So, you write this sort of LLM query to try and find it, because the current filters are filters are pretty bad. But then you kind of think, okay, I'll put something, I just try some other filters, like, what would be the best way to compete against my company based upon everything you know? And all of a sudden it spits out, because it knows all your email, I'll be able to go and figure out, okay, what would be the best way to compete against my company. And what are my biggest weaknesses and which customer could be stolen from me, the easiest or what parts of my business model could be attacked and quite lucrative for a competitor. And the LLM, because it knows everything, will spit that out. I think people might, companies might start realizing, oh, gee, do I really want all my data in the cloud and access AI in the cloud? I don't want the AI to know about all my customers. I don't want the AI to know about my business plans and strategy and access more documents that I've got in Google Docs and my slide decks and my Google Sheets. I don't want it knowing everything about what I do. I don't want it knowing all my customers. I don't want it running my first-tier

customer support, talking to all my customers and getting all the customer data. Because obviously, I'm making API calls to the cloud every time a chat thread starts up with an AI agent. Gee, all that information Google is processing, Microsoft is processing, etc. and so forth, that becomes a real risk, because, as we see in the past, that these companies occasionally do decide to enter industry segments and to destroy what was their previous work, the previous customers, they go and take out a segment like Google did with travel, for example. You know, travel was a big purchaser of Google ads, and eventually Google just thought, I'm going to go buy Sabre and then go in there and just dominate travel, right? Go to Google.com/travel/flight, you do a Google query now in order to ask about booking a flight, and can go ahead to head with competitors.

So that may lead to the model flipping a little bit, and rather than these API calls to this giant brain and AI brain, the cloud, and these organic data centers being built everywhere that are increasing costs. Instead, you might have this great unbundling, and AI will go to the edge. So, a bit like, we had mainframes, and then we had desktop computers, and then we had network computers, thin client, fat client. And every cycle that kind of goes to the centralization and decentralization and backwards and forwards, you may have this great decentralization of AI where chip manufacturers like Qualcomm produce chips that can run AI models that there's not the giant, you know, GPT 4.5 or Grok3 brain, but it’s specialized for certain tasks. It keeps the data local to your device. It doesn't go over the to the cloud. You retain your confidentiality and security of that data, and only once in a while you go to the cloud for very specialized things when your device can't handle it, or your local Edge computer within your corporate network can't handle it. And so this is a big trend that's happening right now. You know, it's lower latency if the AI model is on your device, it's obviously got all the security and the privacy advantages. It’s probably going to be imperative for in application areas like health care and and so forth. But I do think we could potentially enter into an ‘emperor has no clothes’ moment with a lot of corporates go, you know, I actually don't want my data in the cloud and have the AI sucking in and training on it. I mean, if you're using these software packages now, by default, you got to be careful. You got to go through the settings, and you got to make sure that, by default, they're not training on your prompts. And a lot of these, a lot of these packages, are a bit tricky now, where they go, oh, do you want chat history? If you want chat history, you have to let us suck all your data into the cloud for training, and you have to be kind of careful. So, I think that's another big trend that could happen that could be pretty interesting to see whether we're building over capacity of data centers, and whether or not the investment opportunity might be in those companies. They're producing the edge devices and the smaller chips, and the smaller models, etc. in addition.

***Erik:*** It seems to me that that introduces a whole bunch of different possible outcomes. Because I would think, as you've described, that would start with a lot of corporations saying, look, we want these capabilities of AI. We do not want this stuff in the cloud. We don't trust AWS, we don't trust Microsoft to have all of our data. If I was a data center operator and I saw that trend coming, what I would be saying is, wait a minute, we need to change this architecture of what cloud computing means, and we need to say that there's segregated clouds where you can have your own private little reservation of the cloud. So, yeah, it's a data center that you're

contracting for. But instead of saying I'm going to use Google's cloud, you're going to say I'm using my own cloud services that I'm paying a cloud provider for, but they're going to be very securely segregated. Somebody's auditing that, and I'm leasing the AI software functionality to run on my data, but the people who wrote that code never get to see my data, and the fact that they can't see it is somehow guaranteed through an audited process, seems to me like that could change everything.

***Matt:*** I mean, that's exactly right, and that's how the data center operator is going to have to think. Because I think, I think we're going to have some pretty creepy, well, pretty powerful and pretty amazing and pretty shocking features come out in some of these AI powered software. I mean, the big trend is now co-pilot in everything. So, you'll open up Excel, there will be a co-pilot in there. You open up your software programming, IDE, there's a co-pilot in there. Every bit of SaaS software in the world is going to have a co-pilot in there, kind of helping you, kind of parallel, whether you're like it or not. That's how it's going to work. But these co-pilots will know everything about you, and might start getting a bit creepy. And then, you know, for the same reason you got banks, and there's always the frustration. People say, oh, banks are so they're dinosaurs. I can't share them at Google Slides deck. I can't share them at Google Doc. Their firewall prohibits it. Well, there's a good reason why they do that. It's because, you know, some of the big investment banks know that there's active intelligence gathering operations going on for the big deals they work on. And so, they don't want their documents in the cloud. They want them hosted in that local network. And potentially, they come across as dinosaurs by doing so, but they're protecting their data, and they're protecting their customers data by doing so. And I think you can see the same thing in AI, and I think that may be a pretty shocking acceleration. And also, you know, the next generation of handset comes out from Apple, and the AI can run on that locally, and you can talk to Siri and this, that, and the other, and you don't have to go send it to OpenAI, and there's some sort of privacy constraints around that, I think that'd be very attractive to consumers.

***Erik:*** Matt, I want to come back to AI business models and where they're headed and why they're not profitable, because as a user, I want very much to spend more on OpenAI. Now, what I did is, I had the plus tier, which is the 20 bucks a month thing, on following in your footsteps, I upgraded to the pro tier, which is 200 bucks a month. Then I asked my other trusted advisor on AI about this, and I was talked out of that. I was talked into downgrading back down to the plus tier at 20 bucks a month, because my other trusted advisors said, look, the pro tier is designed entirely for people who are using APIs, programmatic interface, software developers, people developing agents and so forth. If you're just using it for chat, with the release of ChatGPT 4.5 with deep reasoning, that's in the plus tier now. You don't need the pro tier like you did in the beginning, you had to have pro tier to get to the o1 pro model. Now the 4.5 with deep research is better than the o1 pro model, and it all comes at the 20 bucks a month level. Now here's the punch line, Matt, this other trusted advisor I'm talking about is none other than ChatGPT 4.5, that's where I got the advice to downgrade, was ChatGPT told me, hey, you're paying for something that you're not getting any benefit from. Now, maybe that's wrong. Maybe the deep research version of chat GPT doesn't work unless you have the pro model, I don't know, but ChatGPT told me that it didn't. I don't understand why they're suddenly giving me

something I was willing to pay for at a premium price, at the lower price, and it seems like they can't stay in business doing that. It doesn't make sense. Meanwhile, it sounds like they're not really offering me as a ChatGPT consumer, the benefits of upgrading from plus to pro, going from 20 bucks a month to 200 bucks a month, really, those benefits only accrue to software developers. For somebody who wants to use ChatGPT and wants to have a better experience, it's not telling me wait four hours before you use this thing again, but it's just going at full speed. They don't seem to want to sell that to me, are you saying it's because they can't figure out the business model to make it profitable? And until they can, they don't want to sell anything because it's not working? I don't get it.

***Matt:*** I mean, I think the starting point is that none of those plans are making any money in terms of profit model. I think OpenAI last year was rumored to have burned $7 to $8 billion in costs. And yeah, there's some data out there that kind of estimates what their revenue was. I think they made a billion dollars in 2023 and I think in 2024, that stepped up, I think to, it's about $2 billion, maybe at least in the forecast that's been released. So, they're losing billions of dollars per year. I don't think they're making money on the $20 plan. I don't think they're making money on the $200 plan. From what I understand, the $200 plan gives you what they claim to be sort of unlimited usage of their most advanced models, while the $20 plan gives you some access to those models, but not unlimited, and will give you caps. I will say one thing, though, that the $200 plan with GPT 4.5, but particularly with Deep Research turned on, and only with that turned on, is a pretty eye opening experience. I felt the magic again, one more time, of what these advances in models can do when I really started using it. So, for those that are listening, I mean this 4.5 plus Deep Research basically uses this reasoning sort of model where you can ask it a query. For example, I went to the chiropractor last week, and I hurt my shoulder, and when I'm in there, he said to me, oh, I've just bought this new office suite. And by the way, the office suite has, somehow has signage rights, and so I can put up a digital billboard, potentially on the side of the building. I don't know how to do that. I presume I have to find a file, a development application that's going to be really complicated. I said, I'll tell you what, if you can give me some impressions on the digital billboard, I'll write for you the plan of how you get it up and also the DA I pulled out my phone, I got into ChatGPT 4.5 plus Deep Research, and I literally just wrote, I own this building. Supposedly I have rights to do a billboard. I took photos of the wall, which is way up, about eight stories high, wrote a paragraph of text, and hit return. It asked me a few clarifying questions, and then I kind of walked back to my office, and in 15 minutes, by the time I got to my office, it had written 15 pages with all the things that had to be done to apply for council approval, and this planning approval and this, that, and the other, dot, dot, dot, dot, dot, dot, dot. And then I said, okay, write the developer application. And 15 minutes later, I had the development application written. So, I mean, the amount of complexity and difficulty in trying to figure that out for yourself and the quality of the work it really is at research level, mid-level researcher level work. I mean, yesterday, I had my HR team send me the employee handbook for the company. I punched it through Deep Research and 4.5, and I said, rewrite it in the style of the top tech companies in the world. So, it gave me the version that was like Valve, it gave me a version like Facebook. And this is a handbook that's 60 pages long. I had a friend the other day who's a budding movie producer, and I got it to produce various research on how the cinematography should work for a particular film he's going to make. It's

pretty mental. So there are some features in that $200 version that I think are pretty magical. And if you haven't tried them, I fully encourage you to try them and pay the $200 although it looks like DeepSeek is coming out with a new version called R2 which would probably looks like it's, according to rumors, it possibly blows us away, and then there'll be other competitors elsewhere, etc., and so forth. So really, the problem is finding out foundational models is they do all this effort and work, and they produce a great outcome, but then it just turns out they're opening another Thai restaurant in the crowded Bangkok marketplace.

***Erik:*** And I should just add, Matt, that for our listeners benefit who don't use ChatGPT, what Deep Research refers to is a new feature that offers asynchronous access to ChatGPT. So instead of sitting there waiting for it to think of an answer, you say, I want a really detailed analysis of this subject. What questions do you have for me that you need to know answers to before you can get to work? It asks you whatever clarifying questions it might have, and then it goes and takes 20 minutes, or however long it takes, and it reports back to you and says, okay, I'm done, your research report is finished. And what it gives you as output is at the quality of hiring a McKinsey consultant to go and research something for you. It's completely different from what you used to get from ChatGPT back in the earlier versions.

Matt, let's move on to revisiting some of the predictions that we made in our earlier AI episodes. We both thought that there was going to be an explosion of phishing scams and other online scams that were driven by AI agents essentially trolling people on the internet and stealing their personal data and scamming them in one way or another. I have noticed that the phishing scams that I get in email, the old trick of being able to spot them because of the bad spelling and grammar problems that they used to be plagued with because the scammers were not native English speakers, all of a sudden, they've all learned how to clean up their act, and the spelling and grammar no longer gives them away, but I don't really perceive that there's been a an epidemic of automated AI scamming. What's your take on some of that and some of the other predictions that we made in previous episodes?

***Matt:*** Well, I mean, there's all these reports of families getting phone calls from their distressed daughter or son who's just been in a car accident. They've crashed into a car. There's been a pregnant woman in the other car, et cetera. They're now in jail, and they need to pay a lawyer who happens to be there quickly in order to not stay the night and be released. So that is starting to happen. I have noticed in the scam calls that I receive, I'm getting a lot of calls now saying, oh, there's an Amazon delivery. Can you please say one for the following thing? Please say two, please enter, please say something to get to the next stage in the auto prompt. And I think what that they're doing, at least, I've read online that some of these are doing is actually getting your voice. So, it can actually train a voice model to then steal your voice to then take another scam. So, I think we're on the cost of seeing that big time, but I will say I'm pretty surprised I haven't seen this really at scale. But I think we're not that far away in terms of other predictions.

I think open source is really going to be the win here for these models. They're multiplying like rabbits, DeepSeek is fully open source, including open-weights and that's going to be built on.

And we see a bunch of things happening there. I do think that OpenAI is going to have a very challenging time. I mean, you've got Sam Altman picking a fight with Elon Musk, obviously overturning OpenAi from a nonprofit into a for-profit. You've got Elon's buddy David Sacks, now as the AI czar, who described OpenAI as the piranha of a monopoly in terms of how they operate. And he's obviously kind of like, in a way, the overseer of the AI efforts. You've got the fact that 8 of the 11 founding team of OpenAI have left, you're now tied up in this lawsuit. You know, the valuation is punched sky high. Microsoft has access to 75% of the returns until they recoup their investment, etc., and so forth, plus this very weird corporate structure they sat at the beginning. I think there's a very good chance Sam Altman will rage quit in the near future and just start up his own company before his sort of, you know, hero to villain arc is complete, which will probably come out from the mud slinging in the lawsuit with Elon. There's a chance that Altman will leave OpenAI to start his own company in whatever aspect of the AI space he thinks where the real opportunity is, and then leave that, leave the rest of open plan to be kind of embraced and extended by Microsoft. I also think that, whether you like it or not, AI is going to be in every single product. And it's not just going to be in your SaaS products, where I've talked about, talking about sensible is going to be hardware products. It's going to be, the security camera on the wall is going to have a little AI model built into the edge, which will know what's happening in the scene. It will probably be able to figure out who's in the scene, what's going on and kind of predict intent, rather than just being a dumb security camera. But you're going to see that really everywhere. And I do think this, you are going to have this ‘emperor has no clothes’ moment where, I think a lot of people, consumers as well, I mean, they think about all the data you put into Instagram. I mean, Instagram can faithfully replicate an AI model of you talking, speaking in high fidelity video, as you with your friends, etc., and so forth. And I think that might get a little bit spooky, because we've gone through this whole over sharing period where everyone wants to share pictures of themselves doing everything, every five minutes. And I think they might start to realize with some of these creepy features, like maybe it's not a good idea, because they can replicate my likeness, or at least, anyone can who has access to my Instagram feed. And I think you might have a bit of thing, like in the Bitcoin space, you've got this expression, ‘not my keys, not my coins.’ You know, if you're hosting your Bitcoin wallet out there in the cloud, and rather, not locally, that it's actually not your Bitcoin. And I think you might get a bit of a ‘not my AI, not my data,’ right? If it's not on my locally hosted private AI, it's not my data anymore that's being fed into the AI, that's just going to the cloud, and it's being used to do training. I think you're going to have competitors come in to try and crack Nvidia the monopoly, I've talked about that. I think you've got a lot of regulatory chaos at the moment. The EU's out there, kind of with their ‘regulate first and destroy the industry before the industry even sort of starts’ sort of philosophy. They've got really hefty fines, which are up to 35 million euros, or 7% of global revenue, if you kind of breach these sort of rules. And they're kind of coming up with regulations for industries that haven't really been deployed yet. So, they're kind of really causing problems. If you're an EU AI startup, the US is a lot more laissez-faire in terms of the approach. And then you've got China that's being very authorship, authoritarian at one point, but also saying, hurry up on the other side.

Authenticity, I think, will be in high demand. You may have ‘verified by humans’ badges starting to appear on platforms. I mean, if your Instagram feed of unemployed models and bikinis gets

replaced by legions of AI thirst traps, it may be a premium to have the little badge saying this is actually a real human, or else OnlyFans becomes Only Bots, which I think is not too far away. You're going to see some interesting things happen with distribution of content. We talked before in a different episode about Netflix, and maybe Netflix will wake up in one day and kind of feel like it is to be Wikipedia in the age of ChatGPT, right? Very soon you'll have the ability to type a prompt and pop out a movie or a TV series, right? So you could have, you know, finally, we'll get Game of Thrones, season 8, season 9, season 10, in space with Erik Townsend as the lead character, with all your friends as the other characters, etc., and so on. You can see all sorts of spin offs and knock offs, etc., and that may break down like these closed ward gardens of content distribution. And, God knows, everyone's sick of having to pay for Netflix and Amazon Prime and Apple+, and this, that, and the other. And all these other, you know, you're supposed to be able to go just into one service and have everything there, and that hasn't turned up in the case, but I think you're going to see a lot of peer-to-peer content generation and so forth.

The other big thing is just, we're going to enter kind of crazy town. We're not going to know what's true, what's false. It's going to be so easy to not just fake content or images or videos of people doing things, but you'll be able to do it at scale. This is going to be weaponized by countries. It'll be weaponized by political parties. You know, you're going to the forums of your local media organization talking about, I don't know, something in Ukraine, or something happening in China, or something happening in the US, etc. And the 1000s of people that are talking and discussing that, whether it's on X or whatever, may not be real. They may be completely synthetically generated. And having a whole discussion, you wander into this chat room or forum, whatever discussion online or online audio conference, and it just turns out, everyone there is fake, and you're the only person that's real there. So, I think that that is going to be pretty crazy. I think we see some disruption in some large employers, like offshore employers, of people like call centers and BPOS, some of them will be smart and the people who kind of currently on the phones will be kind of writing agents and managing agents and queuing agents and so forth, which is kind of what we're doing in our big support organization. But I do think you get disruption into the banks and the telcos and so on. I think you're going to continue to see what we're seeing with freelancers. And they're now super powered, super skilled, they're all the tooling. They're all independent mercenaries out there and they're going to be, really, rivaling Western talent for getting things done, just with a laptop and access to an internet connection. And I think you're going to see a bit of a backlash in some areas, because there is going to be disruption in some industries where AI might steal your job, or at least someone using AI might steal your job. And you're going to have, the young generation will be able to adapt to this, etc., the old generation won't.

But I think you can have a bit of conflict there, and I think we'll continually to be surprised by some of these new foundational models. I mean, we've talked about in previous episodes how these kinds of black boxes and some of these abilities kind of emerge from the models. And you don't really expect that. When you see Deep Research with GPT 4.5 in action, it's a pretty magical moment that has popped out, akin to the first time you use Midjourney, or the first time you used ChatGPT 3.5. So, I think that will continue in ways that we don't expect and don't anticipate, and not even the designers these models anticipate. And who knows, I think gaming

is going to become truly addictive. You'll be able to live in a whole virtual world, high fidelity, just like you're in the Star Trek hologram, have relationships with these virtual characters that don't even exist. And for a lot of people, it's a lot better than mundane lives. We've talked about the dating thing previously. You know, I don't know you could go on a dating site in 2025 and think you're talking to real people. It's all going to be bots. And if not, it might be the AI digital dating assistant bots, agents for, it's like, I'm too busy go on these dating sites and chat everyone up. I just put in my AI agent and just chat a bunch of people up and fill my calendar, and then I'll just turn up to a cafe and kind of just meet a bunch of people at different times over the course of the weekend, and avoid the whole, the preamble, trying to get them to come and meet me on a date. And I do think we're going to see something kind of global, something in the large scale, either on the threat side. So, we may see some global AI threat occur, whether it's mass hacking, because the AI is very good at things like hacking, or it may be someone using AI to do something like, I don't know, fake the Second Coming of Jesus Christ. I mean, the only thing I know for sure is Siri's still going to suck, because Apple says they're not going to improve the product till 2027, which is surprising.

***Erik:*** Matt, our MacroVoices listeners, of course, care most about what kind of actionable investment advice we might be able to offer them. It seems like it's a really difficult and increasingly difficult landscape now, because in the beginning it was pretty darn easy. You know, Nvidia just has to benefit from selling, really the only game in town, chips that everybody needed. It's not nearly so simple. That's clear from this interview. Is there anything that is clear in terms of what you would invest in, if it's not as simple as just Nvidia has to benefit?

***Matt:*** I think every major industry is going to be transformed by this. And so, I think some of the real money to be made is in the industry verticals. For example, in the real estate industry, someone's going to come out of nowhere and do effectively, an AI agent better job of managing real estate rentals than what's the current state of the art, right? And you'll see this in healthcare, maybe with diagnostics, or you might see it in call centers. Some will come out of nowhere and deliver maybe an AI call center software platform, which will replace these 10,000, 20,000 call center people, call centers that are run by the telcos and the banks, etc., and so on. But I think there's going to be a lot of opportunity if you're kind of careful, looking at each of the industry niches to see who's coming in, who really has a transformative solution, and it's going to be as disruptive, even more disruptive than it was when the internet came around, or the mechanization of agriculture. So, I think just look at various industry segments and kind of see what the trends and look at the new solutions coming in. But I think there'll be a lot of money they've made there, and I think far more than trying to chase a parabola of MAG7 stock prices.

***Erik:*** It seems to me, as AI is becoming commoditized, that what's needed is, if you will, the kayak.com what they did for travel agent websites. You need the same thing for AI, where maybe somebody develops a single whole AI enabled user interface to AI. So, I've got an interface that has a really robust chat history management system that allows me to do searches on my chat history and so forth. But when I type a prompt in, what it does is it analyzes my prompt and says, okay, this particular prompt is really Deep Research. Let's send that to the ChatGPT model. Oh, wait a minute, this prompt here, this is really looking for output,

which is graphics. ChatGPT is not as good at that as whatever some other thing is. Let's send it to that one so it's more of an aggregator that allows me to have a single interface to AI, so I don't have to keep track of who's got the latest model and what's good at what. And it just sends my prompt to whichever AI, which becomes more and more of a commodity, whichever AI server in the background it thinks is best suited to that. Is anyone doing that?

***Matt:*** Well, I mean, that's where it's all trending. So, it's trending to a local AI agent running in your local context, so maybe on your phone or sitting in your email, where it may be, it's trending towards not having one major model kind of do everything, but be able to context switch between different models based upon what the task is, to smaller, specialized, better performing models for specific tasks. And ultimately, what you described is really what Siri should be. I mean, it really should be something as simple as Siri. That's just them monitoring all your communications, monitoring all your emails, potentially even being your AI Chief of Staff, right? Literally, looking at all the things that are coming into your context in a given day, and then going off and doing research and suggesting smart things in order to really turbo power you. That's kind of where all should be heading with the likes of things like Siri.

***Erik:*** You know, I want that functionality, and I want it badly, but I absolutely will not tolerate having it be based on an implementation that puts my personal data in somebody else's cloud. Is my generation that feels that way likely to prevail? Or do we have a younger generation that's going to drive most of this, who's not so concerned about having their personal data being maintained by Apple or Google or somebody in their cloud?

***Matt:*** Well, I think that's there's a trillion-dollar opportunity there for a company to come up and do that properly. Of course, that may be at odds with what the governments and security agencies want. But certainly, I think there's a clear market demand. I mean, Apple was kind of trying to position themselves as sort of that privacy focused service provider, but the recent things that they've done, for example, in the UK, where they've folded and provided law enforcement to access where previously they wouldn't, has indicated that not even they maybe can be trusted with your data. But, I think there's a trillion dollar company right there with that idea.

***Erik:*** Matt, I want to end this episode with an appeal to our listeners personally, with the same conviction that I told you back on January 30 of 2020, that there was a global pandemic coming. I want to tell those listeners who, like myself, maybe you're over 50, you kind of feel like you're set in your ways. You haven't done this AI thing. You don't really do social media either. You don't really need this stuff. It's interesting to hear about it, but you don't do it yourself. Trust me on this. You want to embrace AI and use it. I went from really a ChatGPT 3, I played with it for a month at the $20 level. I thought it was an interesting novelty. It was not compelling enough to interest me in actually using it. That's transformed for me to the point where I couldn't get through a day without using ChatGPT. I have the $200 subscription. I think I'll downgrade on ChatGPT’s advice to the $20 subscription and see if it still works as well. But I'll happily go back to the $200 subscription if I need to, in order to maintain the functionality that I have. So, I really encourage people to check it out.

Matt, for people who are willing to take that advice, where do they start? Is it have a Twitter subscription so you can use Grok 3? Is it OpenAi? What's the best place for somebody who doesn't yet have paid AI subscription to get their feet wet and find out what this is? And I guess I would also couple into that question, what do you do, in order to, as a process to really embrace this, because something I found was it took at least a month to develop the habits to realize, oh, wait, I shouldn't wonder about that. I should just ask AI, I shouldn't ask my IT guy to help me solve an IT problem. I can just ask ChatGPT, it'll tell me, step by step what to do. How does someone who hasn't tried it yet get their feet wet?

***Matt:*** I think there's two angles to that. First is on the consumer side, and the second is on the business side. So, on the consumer side, pretty much the AI, I think it's pretty stated at the moment, is I've talked about GPT and getting access to the 4.0, 5.0 model with Deep Research, which is just a subscription from OpenAI. I think Claude is pretty good for writing marketing copy, and that's the Claude 3.5 or 3.7, again, that's $20 a month, just go to claud.ai. Has its few problems, it tells you often, and it gives you ethics lessons every once in a while, and puts you in the naughty corner for a few hours. But that's pretty interesting. And also Grok, which is either through your Twitter subscription or on grok.com. So on the consumer side, I encourage everyone to try that. There's other interesting things you can look at, such as Midjourney for image generation and ElevenLabs for voice synthesis. On the small business side, I think in the next two years, I mean just very simple applications of AI agents, answering the phones, taking a credit card, processing an order or making a booking in a calendar, every single business in the world, whether it's small or large, will be doing this. Whether it's a restaurant booking or a hotel reservation, or even just a small business like a hairdresser answering the phones and so forth. You can go to freelancer.com/ai you can actually get live demos of stay with the art AI agents and freelancers on our site. We'll build them for you. It's no different from web development, app development, or AI development, the same sort of budget, same sort of complexity. It's freelancer.com/ai, try some of the demos there. You'll be pretty blown away. And it's very accessible and very inexpensive for a business to adopt.

***Erik:*** And Matt, as we close, I just want to add my personal endorsement for that as well. I've been hiring people from freelancer.com to assist me with graphics and to assist me with a bunch of things, they all seem to actually be operating AI. For me, their skill is knowing exactly how to use the right AI tool in order to do something that I don't know how to do. And I have no objection to that whatsoever.

Matt, I can't thank you enough for another terrific interview before we close, just tell our listeners what your Twitter handle is, or X handle, I should say these days, and how they can follow your work.

***Matt:*** That’s right. I can't bear to call it X, I call it Twitter still as well, but it's @matt_barrie on Twitter. And I've also published the latest essay on Medium. If you just search for me there, it's called ***AI of the Storm***.

***Erik:*** And that's also linked in your Research Roundup email. Patrick Ceresna and I will be back as MacroVoices continues right here at macrovices.com.